

**UNIVERSIDAD COMPLUTENSE DE MADRID**  
**FACULTAD DE MEDICINA**



**TESIS DOCTORAL**

**Investigación de la distribución de los alelos HLA en poblaciones sanas y enfermas mediante la aplicación de nuevas metodologías de secuenciación**

**Examination of the HLA allele distributions in healthy and diseased populations by the application of novel sequencing methodologies**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

**Gonzalo Montero Martín**

DIRECTORES

**Marcelo Aníbal Fernández Viña**  
**Jorge Mauricio Martínez Laso**

Madrid

# UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE MEDICINA

Departamento de Inmunología, Oftalmología y ORL



**TESIS DOCTORAL**

**Investigación de la distribución de los alelos HLA en poblaciones sanas y enfermas mediante la aplicación de nuevas metodologías de secuenciación**

(Examination of the HLA allele distributions in healthy and diseased populations  
by the application of novel sequencing methodologies)

MEMORIA PARA OPTAR AL GRADO DE DOCTOR  
PRESENTADA POR

**Gonzalo Montero Martin**

DIRECTORES

**Marcelo Aníbal Fernández Viña**

**Jorge Mauricio Martínez Laso**

**Madrid, 2020**

**UNIVERSIDAD COMPLUTENSE DE MADRID**

**FACULTAD DE MEDICINA**

**Departamento de Inmunología, Oftalmología y ORL**

**Programa de Doctorado en Investigación Biomédica R.D. 99/2011**



**Investigación de la distribución de los alelos HLA en poblaciones sanas y enfermas mediante la aplicación de nuevas metodologías de secuenciación**

(Examination of the HLA allele distributions in healthy and diseased populations  
by the application of novel sequencing methodologies)

**Gonzalo Montero Martin**

**Madrid, 2020**







# **TESIS DOCTORAL**

## **TÍTULO:**

**Investigación de la distribución de los alelos HLA en poblaciones sanas y enfermas mediante la aplicación de nuevas metodologías de secuenciación**

## **AUTOR:**

**GONZALO MONTERO MARTIN**

## **CO-DIRECTORES:**

**MARCELO A. FERNANDEZ VIÑA Y JORGE M. MARTINEZ LASO**

## **LUGAR DE REALIZACIÓN:**

**- Histocompatibility and Immunogenetics HLA Laboratory, Stanford Blood Center, Stanford University School of Medicine, Department of Pathology. Stanford, California (Estados Unidos).**

**- Unidad de Inmunogenética-Área de Inmunología, Centro Nacional de Microbiología-Instituto de Salud Carlos III. Majadahonda, Madrid (España).**



**Los Co-directores,**

**Don Marcelo Aníbal Fernández Viña**

**y**

**Don Jorge Mauricio Martínez Laso**

**CERTIFICAN:**

**Que Don Gonzalo Montero Martin, ha realizado  
bajo su dirección el trabajo de Tesis Doctoral que lleva por  
título:**

***“Investigación de la distribución de los alelos HLA en poblaciones  
sanas y enfermas mediante la aplicación de nuevas  
metodologías de secuenciación”***

**Revisado el presente trabajo, consideramos que tiene la  
debida calidad para su defensa y calificación**



*A mi esposa, a nuestros bebés y a nuestras familias.*

*“Obstacles don’t have to stop you.*

*If you run into a wall, don’t turn around and give up.*

*Figure out how to climb it, go through it, or work around it.”*

***Michael Jordan***

*“When the going gets tough, the tough gets going”*

***Knute Rockne***



# AGRADECIMIENTOS

En primer lugar quiero agradecer enormemente al Dr. Marcelo Fernández Viña, como co-director y principal mentor del presente trabajo de tesis doctoral, toda su ayuda, infinita generosidad desinteresada, ánimo y enseñanzas en todo este periodo de trabajo de tesis así como en otros proyectos de investigación en el campo de la inmunogenética para el estudio de HLA en poblaciones humanas y asociación de enfermedades. Y por haber sido un ejemplo para mí no sólo en el aspecto profesional sino también como persona, esposo y padre de familia. Donde la humildad, el trabajo en equipo y la pasión por lo que uno hace son las mejores bases para crecer e ir siempre hacia delante con positividad y esperanza, como dice el título del libro “*HLA beyond tears*”.

Del mismo modo, a mi otro co-director de tesis, el Dr. Jorge Martínez Laso, con el que inicié este periodo de tesis doctoral, le agradezco mucho toda su dedicación y experiencia a la hora de introducirme en esta fascinante área de la histocompatibilidad e inmunogenética y en el tipaje molecular de genes HLA.

Al Dr. Edgar Fernández Malavé, por su ayuda y guía como tutor académico de esta tesis doctoral.

A todos los miembros/as participantes en este proyecto del Grupo Español de Trabajo en Histocompatibilidad e Inmunología del Trasplante (GETHIT) de la Sociedad Española de Inmunología (SEI) por su gran compromiso, entusiasmo y espíritu de colaboración con el envío de datos de genotipaje HLA molecular a resolución intermedia y la obtención de las muestras de ADN de individuos sanos procedentes de las distintas regiones estudiadas de España. Y de modo especial al Dr. José Luis Caro y resto de su equipo, del Banc de Sang e Teixits (Barcelona), por su contribución al envío de datos de genotipaje HLA molecular a resolución intermedia de la cohorte de individuos sanos procedentes de Cataluña. Así, de igual modo, a todos los voluntarios sanos que donaron muestras de sangre con fines para la investigación.

A los Doctores Albert Saiz, Pablo Villoslada y resto de su equipo, del Departamento de Neurología del Hospital Clinic (Barcelona), por la obtención y envío de muestras de ADN de pacientes con esclerosis múltiple. Igualmente al Dr. Jorge Oksenberg (Department of Neurology, Weill Institute for Neurosciences, University of California San Francisco) por la coordinación en la colección de muestras. Así como a todos los pacientes que donaron muestras de sangre con fines para la investigación.

A todos los compañeros de laboratorio e investigación que he tenido a lo largo de este periodo, cuya contribución a mi formación técnica y teórica ha sido muy importante para la consecución de este trabajo.

A todo el staff (supervisores, profesores, investigadores, tutores, personal de laboratorio y personal administrativo) del Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory, del Stanford School of Medicine-Department of Pathology, del Área de Inmunología del Centro Nacional de Microbiología-ISCIII y de los Departamentos del Programa de Doctorado en Investigación Biomédica de la Facultad de Medicina de la Universidad Complutense de Madrid.

Al Stanford Blood Center (SBC) y al National Institute of Health (NIH), cuya financiación ha permitido llevar a cabo este trabajo de tesis así como otros proyectos de investigación relacionados.





## **DEDICATORIAS**

A pesar de las dificultades y desafíos que se me han presentado durante este período de tesis doctoral, estoy muy agradecido por todo el aprendizaje, enriquecimiento y oportunidades que he tenido en todos los ámbitos de la vida.

Este periodo también me ha concedido la oportunidad no solamente de conocerme mejor y crecer como persona e investigador pero también de valorar aún más a muchas personas importantes en mi vida.

Así, me gustaría agradecerse enormemente y dedicárselo a todas nuestras familias (a la española y a la estadounidense), y muy en especial a mi esposa, por su apoyo incondicional y hacerme siempre mejor persona, esposo, padre y trabajador.

Al mismo tiempo, quiero también dedicárselo a todas/os y cada una/o de las/os profesoras/es, tutoras/es, mentoras/es, entrenadoras/es y compañeras/os que he tenido a lo largo de mi vida académica y que con una generosidad infinita me han dado siempre la oportunidad de descubrir, aprender y crecer junto a ellas/os.



***TABLE OF CONTENTS***



---

|   |            |
|---|------------|
| <b>SUMMARY</b> .....  | <b>2</b>   |
| <b>RESUMEN</b> .....  | <b>18</b>  |
| <b>INTRODUCTION</b> .....   | <b>36</b>  |
| <b>I. HUMAN LEUKOCYTE ANTIGEN (HLA) SYSTEM</b> .....                          | <b>38</b>  |
| <b>1. DEFINITION, DISCOVERY AND GENERAL ASPECTS</b> .....                     | <b>38</b>  |
| <b>2. GENOMIC ORGANIZATION OF THE HUMAN MHC OR HLA SYSTEM</b> .....           | <b>45</b>  |
| 2.1 Human Major Histocompatibility Complex (MHC) or HLA Genomic Map .....     | 45         |
| 2.2 Human xMHC Gene Clusters .....  | 49         |
| <b>3. HLA CLASS I REGION</b> .....  | <b>52</b>  |
| <b>3.1 HLA Class I Genes</b> .....  | <b>52</b>  |
| 3.1.1. Classical HLA Class I Genes (HLA Class-Ia) .....                       | 52         |
| 3.1.2. Non-Classical HLA Class I Genes (HLA Class-Ib) .....                   | 53         |
| 3.1.3. HLA Class I-like Genes .....   | 53         |
| 3.2 Genetic Organization of HLA Class I Genes .....                           | 54         |
| 3.3 Structure of HLA Class I Molecules .....                                  | 56         |
| 3.4 Biological Function of HLA Class I Molecules .....                        | 59         |
| <b>4. HLA CLASS II REGION</b> .....   | <b>62</b>  |
| <b>4.1 HLA Class II Genes</b> .....   | <b>62</b>  |
| 4.1.1. Classical HLA Class II Genes (HLA-IIa) .....                           | 62         |
| 4.1.2. Non-Classical HLA Class II Genes (HLA-IIb) .....                       | 65         |
| 4.2 Genetic Organization of HLA Class II Genes .....                          | 65         |
| 4.3 Structure of HLA Class II Molecules .....                                 | 68         |
| 4.4 Biological Function of HLA Class II Molecules .....                       | 70         |
| <b>5. IMPORTANT ASPECTS OF THE HLA SYSTEM</b> .....                           | <b>73</b>  |
| 5.1 Polymorphism .....  | 74         |
| 5.2 Nomenclature .....  | 80         |
| 5.3 Linkage Disequilibrium .....  | 84         |
| 5.4 Diversity and Evolution .....   | 85         |
| <b>6. ROLE AND RELEVANCE OF HLA IN MEDICINE AND POPULATION GENETICS</b> ..... | <b>89</b>  |
| 6.1 Clinical Transplantation .....  | 90         |
| 6.2 Disease Associations .....  | 93         |
| 6.3 Pharmacogenetics .....  | 98         |
| 6.4 Population Genetics and Anthropology .....                                | 100        |
| <b>II. MOLECULAR HLA GENOTYPING METHODOLOGIES</b> .....                       | <b>105</b> |

|  |            |
|--|------------|
| <b>7. MOLECULAR HLA GENOTYPING METHODOLOGIES.....</b>  | <b>105</b> |
| <b>8. TRADITIONAL MOLECULAR HLA GENOTYPING METHODOLOGIES: ADVANTAGES, LIMITATIONS AND UNRESOLVED AMBIGUITIES .....</b>                     | <b>108</b> |
| 8.1 Sequence-Specific Oligonucleotide (SSO) Probes .....   | 109        |
| 8.2 Sequence-Specific Primers (SSP).....   | 110        |
| 8.3 Real-Time PCR (RT-PCR).....  | 111        |
| 8.4 Sanger Sequence-Based Typing (SBT).....  | 111        |
| 8.5 Limitations and Unresolved Ambiguities of Traditional HLA Typing Methods .....   | 113        |
| <b>9. NEXT GENERATION SEQUENCING (NGS)-BASED HLA GENOTYPING STRATEGIES: IMPACT AND RELEVANCE.....</b>                                      | <b>121</b> |
| 9.1 Three Generations of Sequencing Technologies and Its Application to High-Resolution HLA Typing .....                                   | 121        |
| 9.2 Main Characteristics of NGS-based HLA Genotyping Workflow Approaches.....  | 142        |
| 9.3 Impact and Relevance of NGS Technologies on HLA Research and Clinical Applications ..  | 202        |
| <b>III. PREVIOUS STUDIES ON HLA DIVERSITY IN SPANISH POPULATION.....</b>   | <b>268</b> |
| <b>10. SPANISH DEMOGRAPHIC HISTORY, GENETIC LANDSCAPE OF THE IBERIAN PENINSULA AND OVERVIEW OF HLA SPANISH POPULATION STUDIES .....</b>    | <b>268</b> |
| <b>IV. HLA ASSOCIATION STUDIES IN MULTIPLE SCLEROSIS (MS).....</b>   | <b>292</b> |
| <b>11. OVERVIEW OF MS GENETICS AND HLA-MS ASSOCIATION STUDIES .....</b>  | <b>292</b> |
| <b><i>OBJECTIVES</i> .....</b>   | <b>322</b> |
| <b><i>MATERIALS AND METHODS</i> .....</b>  | <b>328</b> |
| <b>1. STUDY POPULATION, DESIGN AND DATA COLLECTION .....</b>   | <b>330</b> |
| 1.1 17 <sup>th</sup> IHIW-Spanish Unrelated Healthy Control Reference Group .....  | 331        |
| 1.2 Spanish Multiple Sclerosis Cohort .....  | 334        |
| 1.3 Northeast Spain (Catalan) Healthy Control Reference Group.....   | 335        |
| <b>2. HLA CLASS I AND II NGS GENOTYPING BY A LONG-RANGE SHOT-GUN BASED SEQUENCING STRATEGY USING A SHORT-READ SEQUENCING PLATFORM.....</b> | <b>338</b> |
| 2.1 Specimen Collection and Preparation.....   | 343        |
| 2.2 Long-range PCR of HLA Genes .....  | 344        |
| 2.3 Quantification, Balancing and Pooling of PCR products.....   | 348        |
| 2.4 Construction of DNA Sequencing Library .....   | 350        |
| 2.4.1 Primary DNA Library Preparation .....  | 351        |
| 2.4.2 Index-Adapter Ligation.....  | 352        |
| 2.4.3 Consolidation of Adapter Ligated Products .....  | 353        |
| 2.4.4 DNA Library Size Selection .....   | 354        |
| 2.4.5 Amplification of Size-Selected DNA Library .....   | 355        |

|   |            |
|---|------------|
| 2.5 Preparation of Final DNA Library for Sequencing .....   | 356        |
| 2.6 HLA Allele Calling and Genotype Assignment Bioinformatics Analysis of Sequencing Reads .....  | 358        |
| 2.7 Ambiguity Groups Criteria and Standardization Assignments .....   | 368        |
| <b>3. STATISTICAL ANALYSES .....</b>  | <b>392</b> |
| 3.1 Hardy-Weinberg Equilibrium Proportions (HWEP) Test.....   | 393        |
| 3.2 HLA Allele Frequencies Calculation.....   | 394        |
| 3.3 Pairwise Linkage Disequilibrium Estimation.....   | 395        |
| 3.4 Ewens-Waterson Homozygosity (EWH) Test.....   | 397        |
| 3.5 Extended HLA Haplotype Frequencies Inference via Expectation-Maximization (EM) Algorithm.....   | 399        |
| 3.6 Genetic Distances and Dendograms .....  | 402        |
| 3.7 Case-Control Analyses for HLA-disease Association Study .....   | 402        |
| <b>RESULTS .....</b>  | <b>408</b> |
| <b>I. NGS-BASED HLA STUDY IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP) .....</b>   | <b>410</b> |
| 1. EVALUATION OF CONCORDANCE OF HLA GENOTYPING RESULTS IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....   | 410        |
| 2. EVALUATION OF DEVIATIONS FROM EXPECTED HARDY-WEINBERG EQUILIBRIUM PROPORTIONS (HWEP) IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....  | 412        |
| 3. HLA ALLELE FREQUENCIES IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP) .....   | 413        |
| 4. IDENTIFICATION OF TWO NOVEL HLA ALLELES IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....   | 421        |
| 5. EWENS-WATERSON HOMOZYGOSITY (EWH) TEST OF NEUTRALITY IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....  | 428        |
| 6. 2-LOCUS HAPLOTYPE LINKAGE DISEQUILIBRIUM (LD) ANALYSIS IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....  | 430        |
| 7. GLOBAL MEASURES OF PAIRWISE LINKAGE DISEQUILIBRIUM (LD) FOR <i>HLA-A</i> , <i>-B</i> , <i>-C</i> , <i>-DPA1</i> , <i>-DPB1</i> , <i>-DQA1</i> , <i>-DQB1</i> , <i>-DRB1</i> AND <i>-DRB3/4/5</i> LOCI IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)..... | 443        |
| 8. ESTIMATION OF EXTENDED HLA HAPLOTYPE FREQUENCIES IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....  | 445        |
| 9. EVALUATION OF 3-/4-FIELD EXTENDED HAPLOTYPE DIVERSITY GIVEN BY INCLUSION OF <i>HLA-DPA1</i> AND <i>-DPB1</i> LOCI, IN 17 <sup>TH</sup> -IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP).....  | 449        |



---

|  |            |
|--|------------|
| <b>10. HLA ALLELE FREQUENCY DISTRIBUTIONS AND RELATEDNESS WITHIN SPANISH REGIONAL GROUPS FROM 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)</b> ..... | <b>452</b> |
| <b>II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION</b> .....  | <b>460</b> |
| <b>11. HLA ALLELE LEVEL ANALYSES ON FIRST CASE-CONTROL STUDY</b> .....   | <b>461</b> |
| <b>12. HLA HAPLOTYPE LEVEL ANALYSES ON FIRST CASE-CONTROL STUDY</b> .....  | <b>474</b> |
| <b>13. HLA ALLELE LEVEL ANALYSES ON SECOND CASE-CONTROL STUDY</b> .....  | <b>485</b> |
| <b>14. HLA HAPLOTYPE LEVEL ANALYSES ON SECOND CASE-CONTROL STUDY</b> .....   | <b>487</b> |
| <b><i>DISCUSSION</i></b> .....   | <b>494</b> |
| <b>I. NGS-BASED HLA STUDY IN 17TH-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)</b> .....   | <b>496</b> |
| <b>1. HLA ALLELE LEVEL ANALYSES</b> .....  | <b>497</b> |
| <b>2. HLA HAPLOTYPE LEVEL ANALYSES</b> .....   | <b>542</b> |
| <b>3. FINAL REMARKS</b> .....  | <b>610</b> |
| <b>II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION</b> .....  | <b>615</b> |
| <b>4. HLA ALLELE LEVEL ANALYSES</b> .....  | <b>617</b> |
| <b>5. HLA HAPLOTYPE LEVEL ANALYSES</b> .....   | <b>627</b> |
| <b>6. FINAL REMARKS</b> .....  | <b>631</b> |
| <b><i>CONCLUSIONS</i></b> .....  | <b>634</b> |
| <b>I. NGS-BASED HLA STUDY IN 17TH-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)</b> .....   | <b>636</b> |
| <b>II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION</b> .....  | <b>637</b> |
| <b><i>BIBLIOGRAPHY</i></b> .....   | <b>640</b> |
| <b><i>APPENDIXES</i></b> .....   | <b>708</b> |
| <b>APPENDIX 1:</b> .....   | <b>710</b> |
| <b>APPENDIX 2:</b> .....   | <b>729</b> |





***SUMMARY***



## Introduction

Increasing our knowledge of the HLA system, including both the complete sequence description and the assessment of its diversity at the worldwide human population-level, is of great importance for elucidating the molecular functional mechanisms of the immune system and its regulation in health and disease. Furthermore, assessment of HLA allelic and haplotypic diversity of each human population is essential in the clinical histocompatibility and transplantation setting as well as in the pharmacogenetics, immunotherapy and anthropology fields. Nevertheless, the inherent vast polymorphism and high complexity presented by the HLA system have been an important challenge for its unambiguous and in-depth (high-resolution) characterization by previously available legacy molecular HLA genotyping methods (e.g. SSP, SSO and even SBT). Recent application of novel next-generation sequencing (NGS) technology for high-resolution molecular HLA genotyping has enabled to obtain, at a high-throughput mode and larger scale, full-length and/or extended sequences and genotypes of all major HLA genes, thus overcoming most of these previous limitations.

## Objectives

I) Characterization of HLA allele and haplotype diversity of all major classical HLA genes (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*) by application of NGS of a first representative cohort of the Spanish population that could also serve as a healthy control reference group. Respective statistical analyses were performed for this immunogenetic population data.

II) Characterization of HLA allele and haplotype diversity of all major classical HLA genes (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*) by application of NGS of a respective cohort of multiple sclerosis (MS) patients in the Spanish population (recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain). A first case-control

study was carried out to examine HLA-disease associations with MS in these Spanish population cohorts as well as to attempt a fine-mapping of these allele and haplotype associations by full gene resolution level via NGS. In addition, a second analysis exercise (i.e. test case) of this case-control study was carried out using an alternative healthy control group dataset, exclusively from the Spanish northeastern region of Catalonia in this second case, to evaluate possible differences in the findings of HLA-disease association with MS due to plausible regional HLA genetic variation within mainland Spain (i.e. as a statistical way to try controlling for any possible existing population stratification).

## Materials and Methods

HLA-disease association was examined between HLA class I and class II alleles and extended haplotypes with MS in Spanish population cohorts comprising 238 MS patients (recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain) and 282 healthy unrelated ethnically matched individuals (coming from different regions across Spain) as controls (HC) using high-resolution NGS for the 11 major classical HLA loci (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*) at the 3- to 4-field resolution. All collected and received de-identified genomic DNA samples were genotyped using the MIA FORA NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA), following manufacturer's semi-automated protocol. Research version 3.0 of the MIA FORA™ NGS FLEX HLA genotyping software was used for allele calling (version 3.25.0 (July 2016) IPD-IMGT/HLA database was available at the time of the study). Statistical analyses were performed for this immunogenetic population data, firstly using Pypop v.0.7.0 software to carry out Hardy-Weinberg Equilibrium (HWE) test, Ewens-Watson homozygosity (EWH) statistic, determination of allele frequencies and 2-locus haplotypes and linkage disequilibrium (LD) estimates. Hapl-o-Mat v.1.1 software was used for estimation of extended haplotypes frequencies via an expectation-maximization (EM)

algorithm. Calculation of genetic distances (Nei genetic distances ( $D_A$ )) and construction of respective dendograms were performed using POPTREEW (web version of POPTREE software). Statistical analyses for the case-control studies were performed using R language for statistical computing with the BIGDAWG v.2.1 R package. Additionally, a second available HLA dataset of another healthy control group (comprising HLA genotyping data only available for *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci, only resolved at the 2-field allele resolution level and according to version 3.35.0 (January 2019) IPD-IMGT/HLA database in this case) from 196 de-identified unrelated ethnically matched individuals exclusively from the region of Catalonia (in Northeast Spain)) was also used to verify the findings of HLA-disease associations from this initial case-control study (aforementioned) and to evaluate the effect of plausible regional HLA genetic variation within mainland Spain.

## Results

I) Firstly, in relation to the Spanish population healthy cohort (as part of the 17<sup>th</sup>-IHIW) studied here:

1) At the HLA allele level, the following main findings (defined at a very high-resolution) are noteworthy:

1a) At the 3-/4-field allele resolution level, no overall deviations from expected Hardy-Weinberg Equilibrium Proportions (HWEP) are observed in any of the *HLA* loci analyzed with the exception of a minor but significant departure at the *HLA-DPA1* locus ( $p$ -value = 0.0104). NGS HLA genotyping data of the present thesis work is completely concordant with available lower resolution typing data from Spanish local participating laboratories.

1b) Respectively, 36 *HLA-A*, 53 *HLA-B*, 40 *HLA-C*, 14 *HLA-DPA1*, 29 *HLA-DPB1*, 23 *HLA-DQA1*, 24 *HLA-DQB1*, 37 *HLA-DRB1*, 5 *HLA-DRB3*, 5 *HLA-DRB4* and 3 *HLA-DRB5* distinct alleles (k) were identified. Relative to HLA class I region, NGS-based HLA



genotyping data at the 3-/4-field reveal a significant diversity at the nucleotide level for *HLA-A* and *-C* loci in contrast to *HLA-B* locus. Similarly in HLA class II region, higher level of heterozygosity is found at the 4-field level for HLA class II loci encoding alpha subunits (*HLA-DPA1* and *HLA-DQA1*) in comparison to HLA class II loci encoding beta subunits (*HLA-DPBI*, *HLA-DQBI* or *HLA-DRBI*). 4-field allele resolution level by NGS also reveals how in certain allele groups, an allele considered rare initially it actually presents a common occurrence while the lowest numbered allele is not the most frequent (e.g. *HLA-B\*35:01:01:01* allele represents only 3.6% of this allele group whereas *HLA-B\*35:01:01:02* allele represents 96.4% of this allele group found in this Spanish population healthy cohort). Rare alleles in Spanish population such as *HLA-C\*12:166* (AF=0.2%) and *HLA-B\*15:220* (AF=0.4%), as well as null alleles such as *HLA-C\*04:09N* allele (AF=0.4%) and *HLA-DRB4\*01:03:01:02N* allele (AF=1.8%), were also detected and described in the present NGS HLA Spanish healthy population study.

1c) Two novel HLA alleles (*HLA-B\*38:20:02* and *HLA-DRB3\*02:71*) were identified during this Spanish population study using a NGS-based HLA genotyping method. One individual presents a single base mismatch with *HLA-B\*38:20* reference allele sequence in exon 3 (codon 99), which leads to a synonymous substitution (Tyr or Y (TAC) to Tyr or Y (TAT)). Whereas, another subject shows a single base mismatch with *HLA-DRB3\*02:02:01:01* reference allele sequence in exon 3 (codon 166), which leads in this case to a non-synonymous substitution and, therefore, to an amino acid (aa) change (Arg or R (CGG) to Gln or Q (CAG)). This latter observed non-synonymous amino acid change (with a plausible associated alteration of aa side chains interactions and corresponding bonds) could potentially mean a certain level of variation of protein folding and configuration within the encoded  $\beta 2$  domain by exon 3. Thus, respective HLA class II  $\alpha 2/\beta 2$  domain for binding the given CD4 T cell co-receptor may be partially or minimally conditioned by this amino acid change.

1d) Just as an initial and tentative analysis, Ewens-Waterson Homozygosity (EWH) test of neutrality was used for analysis of selective processes based on HLA allelic diversity at the 3-/4-field allele resolution level of this Spanish population cohort. Here, all HLA loci analyzed show levels of observed homozygosity ( $F_o$ ) that are below the expected homozygosity under neutrality ( $F_e$ ) with the exception of *HLA-DPBI* locus. These initial observations, however, would need to be further confirmed on a larger Spanish population cohort.

1e) A comparison of HLA allele distributions was carried out, by calculating Nei genetic distances ( $D_A$ ) and constructing respective Neighbor-joining (NJ) dendrogram specifically based on allele frequencies found at *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* loci, between the 3 different major geographical Spanish regions established (Northern-Central, Eastern and Southern Spain) as well as between the 10 Spanish locations included in the present study. In general, these major Spanish population regions and different individual local sub-groups, which were compared according to these *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* allele distributions in this NJ relatedness analysis, also clustered according to their geographical location, thus illustrating the existing HLA regional variation within Spanish general population. Moreover, despite of limitations in the sample size shown by these different Spanish population sub-groups in the present study thus taking these observations with caution; it can be observed the present entire Spanish population healthy cohort shows a Mediterranean genetic substrate that seems to be represented more predominantly by Eastern and Central regions/locations situated within the Central Plateau. Whereas the most Northern and Southern regions/locations (which are mountainous areas that are more isolated geographically unlike this Central Castilian Plateau region in mainland Spain; or even being very unique island areas such as Canary Islands) diverge from this aforementioned Mediterranean Spanish HLA genetic background. As a notable example, the striking

divergence in the NJ clustering observed in Malaga and Gran Canaria population subsets may be explained by the reported historical genetic contribution from North African Berber and Muslim Arab population ancestries in the Iberian Peninsula and Canary Islands, as a very relevant demographic event with a great impact in the HLA allele diversity observed in modern-day Spanish population and in particular in certain Spanish neighboring regions close to North Africa.

2) At the HLA haplotype level, very interesting, novel and informative 3-/4-field associations have been described for both HLA 2-locus and extended class I and class II haplotypes:

2a) We observed unique 2-locus haplotype associations in non-coding regions at the 4-field allele resolution level that are not apparent when testing at the 2-field level (e.g. non-coding *HLA-DQA1\*05:01:01* variants, *HLA-B\*18:01:01* variants, *HLA-C\*05:01:01* variants or *HLA-C\*06:02:01* variants). In contrast, HLA loci pairs such as *B\*07:02:01~C\*07:02:01:03*, *DQA1\*01:01:01:02~DQB1\*05:01:01:03* and *DQB1\*02:02:01:01~DRB1\*07:01:01:01* are some examples of 4-field highly conserved associations found in this Spanish population cohort. Also very interestingly, the present NGS HLA study allowed us to describe that a very common allele such as *HLA-B\*51:01:01:01* displays a very broad distribution in relation to its association with *HLA-C* alleles (i.e. 7 different associated *HLA-C* alleles were observed in the present study), thus it may represent a primary negative predictive factor when searching for a full-match unrelated donor (URD), as it has been similarly observed in other populations of European ancestry.

2b) In relation to the distribution of extended HLA class I and class II haplotypes, and similarly to what it was found in 2-locus haplotypes, it can be observed very distinctive extended haplotype associations in non-coding regions at the 4-field level that are not apparent, and indeed unattainable, at lower allele resolution level (2-field or 3-field) results

that are obtained when using legacy methodologies. Most common haplotypes (i.e. for example, haplotype frequencies (HF) higher than 5.0%) that were identified in the present Spanish population cohort include:

*HLA-A\*01:01:01:01~C\*07:01:01:01~B\*08:01:01:01~DRB3\*01:01:02:01~*

*DRB1\*03:01:01:01~DQA1\*05:01:01:02~DQB1\*02:01:01* (HF=7.8%);

*HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~*

*DRB1\*07:01:01:01~DQA1\*02:01:01:01~DQB1\*02:02:01:01* (HF=7.8%); and

*HLA-A\*02:01:01:01~C\*07:02:01:03~B\*07:02:01~DRB5\*01:01:01~DRB1\*15:01:01:01~*

*DQA1\*01:02:01:01~DQB1\*06:02:01* (HF=5.2%).

Most of these common extended HLA haplotypes are also relevant haplotypes in worldwide registries and population datasets of predominant European ancestry. Nevertheless, there are also certain common extended haplotypes in our Spanish population cohort, which are not as frequent as in other foreign (mostly of predominant European descent) registries or reported populations datasets. For instance, this occurs with *HLA-A\*30:02:01:01~C\*05:01:01:01~B\*18:01:01:01~DRB3\*02:02:01:01~DRB1\*03:01:01:01~DQA1\*05:01:01:01~DQB1\*02:01:01* (HF=4.7%); and, even more steeply, in the case of *HLA-A\*25:01:01~C\*12:03:01:01~B\*18:01:01:02~DRB5\*01:01:01~DRB1\*15:01:01:01~DQA1\*01:02:01:01~DQB1\*06:02:01* (HF=2.1%) as well as for *HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~DRB1\*07:01:01:01~DQA1\*02:01:01:01~DQB1\*02:02:01:01* haplotype (HF=7.8%). Therefore, this remark illustrates the importance of development of local donor registries.

2c) At the same time, from an anthropological standpoint, interrogation of HLA diversity via NGS with the description of very high-resolution 3-/4-field extended haplotypic associations

and specific patterns displayed within the present Spanish population cohort have permitted the detection and characterization of genetic imprints from both:

-Early ancestral contributions throughout the history, where it is found a complex European-Mediterranean overall genetic substrate made up of North-Central European (such as *HLA-B\*44:02:01:01*, *HLA-B\*18:01:01:01*, *HLA-B\*07:02:01* and *HLA-B\*08:01:01:01* bearing extended haplotypes), North African Berber-Muslim Eastern Arab (*HLA-B\*51:01:01:01*, *HLA-B\*49:01:01*, *HLA-DQB1\*03:19:01* and *HLA-DRB1\*01:03~DQB1\*05:01:01:03* bearing extended haplotypes) and Sephardic Jew (*HLA-B\*35:02:01~DRB1\*11:04:01* and *HLA-B\*38:01:01~DRB1\*13:01:01:01* bearing extended haplotypes) genetic components, in addition to a remarkable presence of still relatively isolated Romani (“Gypsy”) genetic ancestry in a portion of the Spanish general population (*C\*15:02:01:01~B\*40:06:01:02~DRB1\*14:04:01~DQB1\*05:03:01:01~DPB1\*02:01:02* bearing extended haplotypes).

-And also from some other more recent and current demographic events, mainly Latin American (presenting Amerindian genetic background such as *HLA-A\*68:17~B\*40:02:01~DRB1\*04:04:01* bearing extended haplotypes) and Eastern European-Mediterranean (especially from Romania) ethnic groups migrating to Spain.

2d) As previously reported in other recent studies, many identical haplotypes across 7 loci (comprising *HLA-A~B~C~DRB3/4/5~DRB1~DQA1~DQB1*, and excluding *HLA-DPA1* and *-DPB1*) become extremely divergent in terms of the multiplicity of *HLA-DP* alleles with which they associate. This seems to be especially due to the weak LD between *HLA-DP* and the rest of the class II haplotype since existing hotspot of recombination is present between *HLA-DQ* and *-DP* loci. We also observed this pattern at the 3-/4-field level in the present study. Therefore, these observed prominently increased multiplicity and haplotype diversity when evaluating 3-/4-field allele resolution and, even more, when including *HLA-DP* loci

(which, as an example, critically contribute to the increase of mismatches in the donor-recipient transplantation setting) inside haplotype distributions may have direct implications, for example, in relation to the lesser likelihood of finding a full-matched URD in HSCT.

2e) Moreover, as one of the main dissimilarities observed between reported Iberian populations (including the present NGS HLA Spanish population study) and other populations of Northern-Central-Eastern European descent, it should be noted the very striking findings that are particularly related to the respective *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* carrying extended HLA haplotype frequency distributions detected (and, indeed, being found inverted between these two broad population groups). Relative to *HLA-B\*44:02:01:01* carrying extended HLA haplotype frequency distributions, in Spanish population these specific extended haplotypes are found in much lower relative frequencies and in a more spread distribution than in other reported populations of European ancestry. Conversely, *HLA-B\*44:03:01:01* carrying extended HLA haplotypes are found in much higher relative frequencies in comparison to those frequency distributions described in other reported populations of European ancestry.

II) Secondly, referring now to the HLA-MS association study in these aforementioned Spanish population cohorts and related analyses performed as part of the current thesis work, exemplifying the great potential of NGS HLA data for the fine-mapping of allele and haplotype associations:

In summary, the refined *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* signal was significantly associated with predisposition as expected. A second independent risk allele *HLA-DPB1\*03:01:01* was also identified, being in consonance with previous studies in populations of European descent. Protective effects from several distinctive HLA class II signals (including *HLA-DRB1\*04:01:01:01*, *-DRB1\*04:02:01* and *-DRB1\*04:04:01*, which all are tightly associated with the secondary DRB *HLA-DRB4\*01:03:01:01*; and, separately, *HLA-*

*DRB5\*01:02~DRB1\*15:02:01:02~DQB1\*06:01:01* signal) were attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele. While the HLA class I alleles *HLA-B\*38:01:01*, previously identified in other studies, and newly *HLA-B\*58:01:01:01* showed moderately protective effects independently from each other and from the HLA class II associated factors. Furthermore, we did not find a clear Bw4 (relative to only HLA-B alleles and according to motif subgroups NLRIALR, DLRTLLR and NLRTALR, respectively) protective association by itself with MS susceptibility. On the other hand, we described that the Bw6 epitope (SLRNLRG), encoded by the respective group of HLA-B alleles analyzed here, shows a risk association that cannot be attributed simply to LD patterns in relation to the highly predisposing allele *DRB1\*15:01:01:01*.

## Conclusions

I) Firstly, in relation to the Spanish population healthy cohort (as part of the 17<sup>th</sup>-IHIW) studied here:

1) To the best of our knowledge, this is the first and largest study performed using NGS for the genomic characterization of HLA diversity found in Spanish population. In the present NGS study, we were able to describe allelic diversity at the 3-/4-field resolution of major HLA genes *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* (enabling full sequencing of class I loci and extended coverage of class II loci) with minimum level of ambiguities and also to estimate extended haplotype frequencies.

2) NGS HLA sequencing in the present Spanish population cohort has shown striking and highly informative 3-/4-field genotyping results including the description of previously unknown haplotype associations in non-coding regions up to the 4-field allele resolution level, the detection of rare, null and novel polymorphisms as well as the more accurate evaluation of allele and haplotype distributions and prevalence in Spanish population.

3) Overall, results of the present study may contribute as a useful and first reference source for future population studies, for HLA-disease association and pharmacogenetics studies as a healthy control group dataset, for improved virtual panel reactive antibody (vPRA) calculations in Spanish population and for improving donor recruitment strategies of bone marrow and umbilical cord blood registries. Moreover, fine ultra-high allele resolution by NGS and determination of 3-/4-field haplotypic associations have allowed us to identify more accurately specific patterns displayed within Spanish population (including a significant regional variation and population substructure) and to better detect genetic imprints and substrates of either more ancient demographic events or some other more recent or stable throughout history. Data from the present and from future larger NGS studies may also contribute to establish strategies for improving the efficacy of both current and novel immunotherapies and selection criteria of personalized therapeutic approaches. Lastly, knowledge of the most common extended HLA haplotypes at the 3-/4-field resolution in Spanish population may also serve to construct the most representative Spanish HLA haplo-homozygous bank for allogeneic transplantation of induced pluripotent stem cells (iPSC) derived cell therapies such as novel cellular adoptive therapies based on genetically engineered T cells.

II) Secondly, referring now to the HLA-MS association study:

1) Overall, very high-resolution HLA genotyping data allows fine-mapping of susceptibility and protective factors and exclusion of bystander (“hitchhiking”) alleles from contiguous loci.

2) The refined *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* signal was significantly associated with predisposition.

3) Nominal and stratified analyses identified a second significant MS risk signal relative to *HLA-DPBI\*03:01:01* allele, being independent from the highly predisposing *HLA-DRB1\*15:01:01:01* factor.



4) Protective effects from several distinctive HLA class II signals (several *HLA-DRB1\*04-* and the *HLA-DRB1\*15:02:01:02*-bearing haplotypes) were attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele.

5) *HLA-B\*38:01:01* and *-B\*58:01:01:01* alleles confer protection and operate independently of the presence of *HLA-DRB1\*15:01:01:01* risk factor.

6) In the present dataset, we did not find a clear Bw4 (relative to only HLA-B alleles and according to motif subgroups NLRIALR, DLRTLLR and NLRTALR, respectively) protective association by itself with MS susceptibility. On the other hand, we described that the Bw6 epitope (SLRNLRG), encoded by the respective group of HLA-B alleles analyzed here, shows a risk association that cannot be attributed simply to LD patterns in relation to the highly predisposing allele *DRB1\*15:01:01:01*.





***RESUMEN***



## Introducción

El estudio del sistema HLA, incluyendo la descripción completa de su secuencia y de la diversidad de este complejo HLA a nivel poblacional, es de gran importancia de cara a poder entender los mecanismos moleculares y funciones del sistema inmune así como su regulación en individuos sanos y enfermos. Además, la caracterización exhaustiva de la diversidad de alelos y haplotipos HLA de cada población humana es esencial en el campo de la inmunología de trasplante e histocompatibilidad al igual que en las áreas de farmacogenética e inmunoterapia. El inmenso polimorfismo y gran complejidad que presenta el sistema HLA han sido hasta ahora importantes barreras de cara a poder caracterizarlo en gran detalle (por alta resolución) y sin ambigüedades mediante métodos de genotipaje HLA tradicionales disponibles (como son SSP, SSO o incluso SBT). La reciente aplicación de la novedosa tecnología de secuenciación masiva NGS para el genotipaje molecular HLA por alta resolución ha posibilitado obtener secuencias completas o mucho más extendidas para genotipos de los principales genes de HLA, superándose así estas previas limitaciones.

## Objetivos

I) Caracterización de la diversidad alélica y haplotípica de los principales genes HLA (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* y *-DRB3/4/5*) mediante la aplicación de NGS en una primera cohorte representativa de la población española que, igualmente, constituirá una población control de referencia para estudios de asociación de HLA y enfermedades. También, respectivos análisis estadísticos se realizaron para estos resultados de genotipaje HLA.

II) Caracterización de la diversidad alélica y haplotípica de los principales genes HLA (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* y *-DRB3/4/5*) mediante la aplicación de NGS en una correspondiente cohorte de pacientes con esclerosis múltiple (EM) de la población española (reclutados y procedentes del Departamento de Neurología del Hospital Clínic

(Barcelona, Cataluña)). Un primer estudio de asociación HLA tomando casos (pacientes EM) frente a controles sanos se llevó a cabo para examinar la asociación de genes HLA y la enfermedad de EM en estas cohortes de población española antes mencionadas. Así se buscaba realizar un mapeo fino de las respectivas asociaciones alélicas y haplotípicas de HLA mediante la gran resolución alélica proporcionada por esta metodología de secuenciación masiva. De modo adicional, y como un segundo ejercicio de análisis en este estudio de asociación HLA, se utilizó un grupo control sano alternativo al previo, que incluía individuos procedentes de la región de Cataluña (situada al noreste de España) exclusivamente en este caso, para evaluar así posibles diferencias dadas en la asociación de HLA con EM debido a la probable variación genética en HLA existente a nivel regional dentro del territorio de España.

### **Materiales y Métodos**

Se llevó a cabo el estudio de asociación entre alelos y haplotipos extendidos de los 11 principales genes HLA de clase I y II (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*; genotipados mediante secuenciación masiva NGS por alta resolución del 3- al 4-campo) y enfermedad de EM en unas cohortes de población española, siendo de origen étnico similar, que comprendían muestras (anónimas y codificadas para no identificarse) de 238 pacientes españoles de EM (reclutados y procedentes del Departamento de Neurología del Hospital Clínic (Barcelona, Cataluña)) y 282 sujetos sanos no emparentados (procedentes de diferentes regiones del territorio de España) como grupo control. Tras su recogida y recepción, todas las muestras de ADN genómico fueron genotipadas empleando kits de HLA MIA FORA NGS HLA FLEX Typing 11 (RUO) de 96 tests (Immucor, Inc. Norcross, GA, USA), siguiendo el protocolo semi-automatizado del fabricante. Para el análisis de las secuencias obtenidas y asignación de alelo/s de cada gen por muestra, se utilizó la versión 3.0 para investigación del software MIA FORA™ NGS FLEX para genotipaje HLA (la versión 3.25.0 (Julio 2016) de la base de datos IPD-IMGT/HLA

fue la disponible en el momento de este estudio). Posteriormente, a partir de estos resultados de genotipos HLA, se realizó una serie de análisis estadísticos. En primer lugar, mediante el software de Pypop v.0.7.0 se llevo a cabo el test de Hardy-Weinberg Equilibrium, el test de homocigosidad Ewens-Waterson, determinación de frecuencias alélicas y de haplotipos de 2-locus así como la estimación de sus respectivos valores de desequilibrio de ligamiento. También el software Haplo-Mat v.1.1 se empleó para la estimación de frecuencias de haplotipos a través del algoritmo de Expectation-Maximization (EM). El cálculo de distancias genéticas (Nei genetic distances ( $D_A$ )) y construcción de respectivos dendogramas se realizó mediante POPTREEW, la versión web del software POPTREE. Los análisis estadísticos correspondientes al estudio de asociación HLA de casos EM frente a controles sanos se llevaron a cabo utilizando el paquete BIGDAWG v.2.1 software en lenguaje R. Como segunda parte de este estudio de asociación HLA, se escogió un dataset de genotipos HLA (correspondiente en este caso y sólo disponible para los genes *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1*, *-DRB1* y *-DRB3/4/5*, establecido al 2-campo de resolución alélica y en la versión 3.35.0 (Enero 2019) de la base de datos IPD-IMGT/HLA) de un grupo control alternativo constituido por 196 sujetos (anónimos y codificados para no identificarse) sanos, no emparentados, étnicamente similares y procedentes exclusivamente en este caso de la región de Cataluña. Con la finalidad de verificar las asociaciones HLA encontradas inicialmente en el primer análisis (antes mencionado) de casos EM españoles frente a controles sanos del grupo representativo de una gran parte del territorio español, y evaluar así el posible efecto de la variación de distribuciones genéticas de HLA a nivel regional dentro del territorio de España en las asociaciones encontradas en EM y HLA dentro de este estudio concretamente.



## Resultados

I) En relación al estudio en la cohorte de población Española sana (siendo parte del 17<sup>th</sup>-IHIW):

1) A nivel alélico, y definido en alta resolución vía NGS, estos fueron los principales hallazgos:

1a) Al 3-/4-field de nivel de resolución alélica estudiado aquí, no se observaron desviaciones respecto al equilibrio de proporciones de Hardy-Weinberg (HWEP) en ninguno de los HLA loci estudiados salvo en el caso de *HLA-DPA1* locus (p-value = 0.0104). Todos los datos de genotipaje de HLA por NGS de este estudio coincidieron totalmente con los datos de tipaje molecular disponibles que habían sido obtenidos a una resolución alélica menor y siendo procedentes de los diferentes laboratorios participantes de origen en España.

1b) Respectivamente, 36 *HLA-A*, 53 *HLA-B*, 40 *HLA-C*, 14 *HLA-DPA1*, 29 *HLA-DPBI*, 23 *HLA-DQA1*, 24 *HLA-DQB1*, 37 *HLA-DRB1*, 5 *HLA-DRB3*, 5 *HLA-DRB4* and 3 *HLA-DRB5* distintos alelos (k) fueron identificados en el presente estudio. En relación a la región de HLA clase I, la aplicación de esta metodología de NGS para el tipaje HLA nos permitió poder detectar una significativa diversidad genómica siendo mayor a nivel nucleótido en los genes *HLA-A* y *HLA-C* y en comparación con el *HLA-B* locus (mucho más diverso a nivel proteína). Asimismo, en el caso de los genes HLA de clase II, se pudo observar un elevado nivel de heterocigosidad a este 3-/4-field de resolución en aquellos genes (*HLA-DPA1* y *HLA-DQA1*) que codifican para la respectiva subunidad alpha del correspondiente heterodimero siendo además mayor que en el caso de aquellos genes HLA de clase II que codifican por su parte la subunidad beta (*HLA-DPBI*, *HLA-DQB1* o *HLA-DRB1*). También, la resolución al 3-/4-field vía NGS nos permitió revelar como en ciertos grupos alélicos, aquellas variantes que se podrían considerar “raras” dada su mayor numeración al 4-field en la nomenclatura son en realidad más comunes que aquellas otras variantes con una numeración menor de acuerdo con

la nomenclatura actual definida (como ejemplo particular dentro de este estudio en población Española, el alelo *HLA-B\*35:01:01:01* representaba el 3.6% dentro de este grupo alélico mientras que el alelo *HLA-B\*35:01:01:02* correspondía al 96.4%). El genotipaje de genes HLA vía NGS también nos permitió asimismo poder detectar alelos raros en población Española como por ejemplo *HLA-C\*12:166* (AF=0.2%) y *HLA-B\*15:220* (AF=0.4%). Igualmente, con este genotipaje HLA vía NGS se pudo caracterizar alelos nulos en esta cohorte Española como el alelo *HLA-C\*04:09N* (AF=0.4%) y el alelo de clase II *HLA-DRB4\*01:03:01:02N* (AF=1.8%).

1c) Se identificaron y confirmaron dos nuevos alelos, *HLA-B\*38:20:02* y *HLA-DRB3\*02:71*, durante este estudio gracias a la metodología NGS. Así, dentro de esta cohorte, uno de los individuos estudiados presentaba una única diferencia en el codón 99 del exón 3 respecto a la secuencia de referencia para el alelo *HLA-B\*38:20*, donde se encontró una mutación puntual con una sustitución sinónima (de Tyr ó Y (TAC) a Tyr ó Y (TAT)). Mientras que en otro individuo dentro de esta misma cohorte, se observó una única diferencia en el codón 166 del exón 3 respecto a la secuencia de referencia para el alelo de clase II *HLA-DRB3\*02:02:01:01*, donde se tenía una mutación puntual con una sustitución en este caso no sinónima que conllevaba un cambio de amino ácido (de Arg ó R (CGG) a Gln ó Q (CAG)). En este último caso, se podría pensar que a priori este cambio de amino ácido (con un muy probable efecto en las interacciones entre cadenas laterales dado el cambio de carga y de enlaces asociados) supondría cierto nivel de variación respecto a la configuración y plegamiento proteico del dominio  $\beta 2$  de la molécula HLA dada, que se encuentra codificado por este exón 3. Como consecuencia, y desde un punto de vista funcional, la región  $\alpha 2/\beta 2$  de la respectiva molécula HLA de clase II que se sabe que interacciona con la molécula co-receptor CD4 expresada en

células T podría verse condicionada dado este cambio de amino ácido descrito para este nuevo alelo.

1d) Únicamente como un test estadístico preliminar e inicial, se realizó el test de neutralidad de Ewens-Waterson Homozygosity (EWH) para poder evaluar los posibles procesos selectivos que estarían dirigiendo la diversidad alélica encontrada en la presente cohorte de población Española estudiada a este 3-/4-field de nivel de resolución. Así, en el caso de todos los genes HLA estudiados aquí y con la única excepción del locus *HLA-DPBI*, se vio que bajo condición de neutralidad los valores de homocigosidad observada ( $F_o$ ) eran siempre menores que aquellos expresados para la homocigosidad esperada ( $F_e$ ). Aun así, estos resultados preliminares dados en el contexto de este estudio de NGS necesitarían ser posteriormente confirmados en futuros estudios a una mayor escala.

1e) También se llevó a cabo un análisis estadístico para la comparación de distribuciones alélicas de HLA (considerando en este caso los genes de mayor diversidad *HLA-A*, *-B*, *-C*, *-DQB1* y *-DRB1*) entre las 3 principales regiones geográficas diferenciadas de la presente cohorte de población Española (regiones Norte-Centro, Este y Sur de España) así como de las 10 diferentes localidades que se cubrieron individualmente en el presente estudio. Donde se calcularon las respectivas distancias génicas de Nei ( $D_A$ ) y se construyó el correspondiente dendograma Neighbor-joining (NJ). En general, se observó que todos estos subgrupos comparados se situaban en el dendograma construido no solo de acuerdo con estas distribuciones de frecuencias alélicas analizadas sino que también esta misma disposición coincidía de acuerdo a su distribución geográfica, mostrándose así esta variación de HLA a un nivel también regional en población Española. Asimismo, a pesar de las limitaciones de tamaño de muestra del presente estudio especialmente en relación a cada uno de los subgrupos evaluados y por tanto teniendo que tomarse con cautela las posibles interpretaciones derivadas

de los resultados de este análisis; se puede decir que la presente cohorte de población Española sana estudiada mostraba un sustrato genético principalmente Mediterráneo dada también una mayor representatividad de regiones/localidades del Este y Centro y fundamentalmente dentro de la Meseta Central. Mientras que aquellas regiones localizadas más en los extremos Norte y Sur (caracterizadas por ser zonas generalmente más montañosas y aisladas de la Meseta Central así como de zonas geográficas tan singulares como aquellas que son insulares siendo el caso, por ejemplo, de las Islas Canarias) divergen de manera significativa de este sustrato genético típicamente Mediterráneo en relación a los genes HLA. Como ejemplo representativo, se tiene la llamativa divergencia observada en el dendograma NJ representado del clúster que engloba los subgrupos de Málaga y Gran Canaria; pudiéndose deberse esto muy probablemente a la contribución genética (como un evento demográfico muy relevante para entender la diversidad de genes HLA observada hoy día en población Española) muy bien documentada y bien conocida históricamente procedente del Norte de África por la influencia de Musulmanes Bereberes/Árabes especialmente en regiones próximas de la Península Ibérica y también del territorio de las Islas Canarias.

2) Respecto a la distribución de haplotipos HLA observada, y gracias de nuevo a la aplicación de la tecnología NGS para la secuenciación de genes HLA, se pudieron describir asociaciones al 3-/4-field de resolución muy interesantes, novedosas e informativas tanto para asociaciones haplotípicas de 2-locus como para haplotipos completamente extendidos de clase I y clase II.

En resumen:

2a) Pudimos observar singulares asociaciones haplotípicas de 2-locus respecto a regiones no codificantes al 4-field de resolución en NGS que no eran evidentes cuando se testaba con metodologías más tradicionales limitadas al 2-field de resolución (como, por ejemplo, variantes no codificantes de *HLA-DQA1\*05:01:01*, *HLA-B\*18:01:01*, *HLA-C\*05:01:01* y

*HLA-C\*06:02:01*). Por el contrario, pares haplotípicos de genes HLA como, por ejemplo, *B\*07:02:01~C\*07:02:01:03*, *DQA1\*01:01:01:02~DQB1\*05:01:01:03* y *DQB1\*02:02:01:01~DRB1\*07:01:01:01* fueron casos característicos en esta cohorte de población Española donde se encontró un significativo nivel de conservación al 4-field de resolución. Asimismo, y de modo muy interesante, el presente estudio de HLA mediante secuenciación por NGS nos permitió observar que un alelo muy común en población Española como el *HLA-B\*51:01:01:01* establece una muy amplia distribución haplotípica en su asociación a alelos *HLA-C* (hasta 7 diferentes asociaciones se pudieron describir en este caso), representando así un importante factor predictivo negativo en la búsqueda de un donante totalmente compatible, tal y como también se ha visto en otras poblaciones de ascendencia Europea.

2b) En relación a la distribución de haplotipos extendidos HLA de clase I y clase II, y en la misma línea que con lo hallado anteriormente para haplotipos de 2-locus, se pudieron describir muy distintivos y singulares asociaciones haplotípicas respecto a regiones no codificantes al 4-field de resolución en NGS que no eran evidentes cuando se testaba con metodologías de genotipaje HLA más tradicionales limitadas a una menor resolución alélica (al 2-field o 3-field de resolución). Los haplotipos extendidos más comunes (por ejemplo, aquellos con una frecuencia alélica mayor al 5%) identificados en esta presente cohorte de población Española fueron los siguientes:

*HLA-*

*A\*01:01:01:01~C\*07:01:01:01~B\*08:01:01:01~DRB3\*01:01:02:01~DRB1\*03:01:01:01~DQA1\*05:01:01:02~DQB1\*02:01:01* (HF=7.8%);

*HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~DRB1\*07:01:01:01~DQA1\*02:01:01:01~DQB1\*02:02:01:01* (HF=7.8%); y

*HLA-A\*02:01:01:01~C\*07:02:01:03~B\*07:02:01~DRB5\*01:01:01~DRB1\*15:01:01:01~DQA1\*01:02:01:01~DQB1\*06:02:01* (HF=5.2%).

Así pues, y en general, la mayoría de estos haplotipos HLA extendidos más comunes observados en población Española también son generalmente frecuentes en registros y bases de datos poblacionales de todo el mundo, y especialmente en aquellos de predominante ascendencia Europea. Por otro lado, también se observaron ciertos haplotipos relativamente comunes en población Española que sin embargo no eran tan frecuentes en otros registros del extranjero, aun siendo todavía preferentemente de ascendencia Europea. Por ejemplo, este era el caso del haplotipo *HLA-A\*30:02:01:01~C\*05:01:01:01~B\*18:01:01:01~DRB3\*02:02:01:01~DRB1\*03:01:01:01~DQA1\*05:01:01:01~DQB1\*02:01:01* (HF=4.7%); e incluso de modo más acentuado en el caso del haplotipo *HLA-A\*25:01:01:01~C\*12:03:01:01~B\*18:01:01:02~DRB5\*01:01:01~DRB1\*15:01:01:01~DQA1\*01:02:01:01~DQB1\*06:02:01* (HF=2.1%), al igual que en el caso del haplotipo *HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~DRB1\*07:01:01:01~DQA1\*02:01:01:01~DQB1\*02:02:01:01* (HF=7.8%). Por lo tanto, estos últimos ejemplos mostrados aquí subrayan la importancia del desarrollo de registros de donantes locales dentro del país que cubran de manera efectiva y representativa la diversidad alélica y haplotípica más común de población Española en este caso dado.

2c) Al mismo tiempo, y en este caso en el contexto de un estudio antropológico, la evaluación de la diversidad del sistema HLA a través de la aplicación de genotipaje por NGS (con la descripción a altísima resolución al 3-/4-field de asociaciones haplotípicas y determinados patrones de desequilibrio de ligamiento) permitió poder detectar y caracterizar contribuciones

de sustratos genéticos (debidos a eventos demográficos de trascendencia) que han persistido en la población actual moderna Española:

-Tanto de muy temprano origen en la historia, como es el componente general y complejo Europeo-Mediterráneo constituido por un bloque del Norte-Centro de Europa (representado por haplotipos conteniendo *HLA-B\*44:02:01:01*, *HLA-B\*18:01:01:01*, *HLA-B\*07:02:01* y *HLA-B\*08:01:01:01*); un bloque también del Norte de África de origen Bereber-Árabe (representado por haplotipos conteniendo *HLA-B\*51:01:01:01*, *HLA-B\*49:01:01*, *HLA-DQB1\*03:19:01* y *HLA-DRB1\*01:03~DQB1\*05:01:01:03*); así como de un bloque correspondiente al origen Judío Sefardita (representado por haplotipos conteniendo *HLA-B\*35:02:01~DRB1\*11:04:01* y *HLA-B\*38:01:01~DRB1\*13:01:01:01*); y adicionalmente al también llamativo bloque genético de origen Romaní-Gitano dentro de una porción de la población general de España (caracterizado por haplotipos conteniendo *C\*15:02:01:01~B\*40:06:01:02~DRB1\*14:04:01~DQB1\*05:03:01:01~DPB1\*02:01:02*).

-Como de eventos demográficos más recientes e incluso que están transcurriendo todavía en la actualidad. Principalmente, dado el importante y muy reciente movimiento migratorio a España procedente de Latino América (presentando un perfil singular y característico como es el Amerindio, con haplotipos que contienen por ejemplo *HLA-A\*68:17~B\*40:02:01~DRB1\*04:04:01*) y de regiones del Este de Europa (especialmente Rumanía) y del Mediterráneo.

2d) También, como ya se había observado en otros estudios con anterioridad, haplotipos que son totalmente idénticos en relación a los 7 principales HLA loci (comprendiendo *HLA-A~B~C~DRB3/4/5~DRB1~DQA1~DQB1*) pasan a ser realmente divergentes al incluir y dada la multiplicidad vista con los loci *HLA-DP* en su respectiva asociación haplotípica. Se

entiende que esto se debe al muy bajo desequilibrio de ligamiento presentado en la región *HLA-DP* con respecto al resto de la región HLA de clase II dada la existencia de “puntos calientes” de recombinación específicamente ente los loci *HLA-DQ* y *HLA-DP*. Este patrón se pudo así también ver al 3-/4-field de resolución en el presente estudio. Esta elevada diversidad haplotípica observada al 3-/4-field, y en adición a esta multiplicidad dada por *HLA-DP*, sin duda constituye un factor importante en el número de incompatibilidades a poder encontrar en la pareja donante-receptor en trasplantes y así en la menor probabilidad de encontrar un donante totalmente compatible, siendo esto especialmente crítico en trasplante alogénico de células madre hematopoyéticas.

2e) Asimismo, y como una de las principales y más notorias diferencias encontradas entre poblaciones Ibéricas (incluyendo la cohorte de población Española del presente estudio) y el resto de Europa (incluyendo regiones del Norte-Centro-Este), destacan las respectivas distribuciones haplotípicas relativas a *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* encontrándose invertidas entre estos dos amplios grupos poblacionales. Así, en relación a haplotipos que contienen *HLA-B\*44:02:01:01*, y de forma opuesta a lo observado en poblaciones de ascendencia Europea, en población Española estos haplotipos se encuentran con frecuencias relativas menores y mostrando una distribución aún más dispersa. Mientras que en el caso contrario de haplotipos conteniendo *HLA-B\*44:03:01:01*, estos presentan frecuencias relativas mucho mayores en población Española y con respecto a otras poblaciones de ascendencia Europea.

II) En segundo lugar, en relación ahora al estudio de asociación de genes HLA y Esclerosis Múltiple (EM) en población Española, donde la aplicación de secuenciación masiva NGS en el genotipaje HLA permitió el mapeo de gran precisión de aquellas variantes alélicas y haplotípicas asociadas:



Resumidamente, la señal refinada *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* asociada a predisposición con EM fue estadísticamente significativa tal y como se esperaba en base a previos estudios publicados. Una segunda señal significativa de susceptibilidad a EM, en este caso independiente del factor principal de riesgo *DRB1\*15:01:01:01*, se encontró asociada al alelo *HLA-DPB1\*03:01:01*, concorde también con resultados procedentes de estudios previos realizados en otras poblaciones de origen Europeo. Por otro lado, se detectaron efectos de protección a EM correspondientes a varias diferentes señales HLA de clase II (incluyendo *HLA-DRB1\*04:01:01:01*, *-DRB1\*04:02:01* y *-DRB1\*04:04:01*, en fuerte desequilibrio de ligamiento (DL) con el DRB secundario *HLA-DRB4\*01:03:01:01*; y, separadamente, la señal correspondiente a *HLA-DRB5\*01:02~DRB1\*15:02:01:02~DQB1\*06:01:01*). Todos los cuales se debían a un patrón negativo de DL con respecto al factor principal de riesgo *DRB1\*15:01:01:01* y haplotipos asociados. Mientras que alelos HLA de clase I *HLA-B\*38:01:01*, previamente identificado en anteriores estudios, y la nueva señal de *HLA-B\*58:01:01:01* mostraron un efecto protector moderado siendo independiente entre ellos y también independiente de factores asociados de HLA clase II. Además, no encontramos un claro efecto protector a EM atribuido para el Bw4 por si solo (considerando alelos HLA-B y los correspondientes subgrupos NLRIALR, DLRTLLR y NLRTALR respectivamente). Mientras que, por otra parte, el epitopo Bw6 (SLRNLRG, codificado por los respectivos alelos HLA-B estudiados aquí) mostraba una asociación con riesgo a EM que no podía explicarse simplemente en relación a patrones de DL con respecto al factor principal de riesgo *DRB1\*15:01:01:01*, siendo así más bien independiente.

## Conclusiones

I) En primer lugar, relativo a la cohorte de población Española sana (siendo parte del 17<sup>th</sup>-IHIW) estudiada en el presente trabajo de tesis:

1) Desde nuestro conocimiento y habiéndose revisado ampliamente la literatura publicada hasta la fecha, el presente trabajo de tesis constituye el mayor estudio realizado por primera vez haciendo uso de la tecnología NGS para el genotipaje de genes HLA al 3-/4-field de resolución en el caso de población Española. Donde se ha podido describir la diversidad alélica de todos los principales genes HLA: *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*, con la caracterización de la secuencia completa de genes HLA clase I y una cobertura muy extendida de la secuencia de genes HLA de clase II consiguiendo así tener un mínimo nivel de ambigüedades y donde también se ha podido así estimar las frecuencias haplotípicas

2) La secuenciación de los genes HLA mediante NGS en la presente cohorte de población Española ha permitido obtener unos resultados interesantes y muy informativos al 3-/4-field de resolución incluyendo: la descripción de asociaciones haplotípicas (no conocidas hasta ahora) en el contexto de regiones no codificantes al 4-field de resolución; la detección precisa y rápida de variantes raras, nuevos alelos y alelos nulos; así como una evaluación mucho más fehaciente de las distribuciones alélicas y haplotípicas y su prevalencia en población Española.

3) En definitiva, y como principales posibles aplicaciones a tener en cuenta, los resultados del presente estudio podrían contribuir como una primera base de datos HLA de referencia para: futuros estudios poblacionales; para estudios de asociación a enfermedades y para el campo de la farmacogenética como un grupo control sano; para actualizar y mejorar las estimaciones del virtual panel-reactive antibody (vPRA) en relación a población Española; y también para mejorar las estrategias y protocolos de reclutamiento de donantes en los registros de médula ósea y de sangre de cordón umbilical. Asimismo, gracias a la muy elevada resolución alélica proporcionada por la tecnología NGS para el genotipaje HLA, la determinación de asociaciones haplotípicas al 3-/4-field de resolución nos ha permitido poder identificar en mucho mayor detalle patrones característicos de distribución HLA dentro de esta cohorte de población en

España (incluyendo una significativa variación regional observada dada una muy posible estratificación de la población respecto a esta variabilidad genética en HLA); así como una mejor detección de aquellos sustratos genéticos debidos a eventos históricos demográficos relevantes del pasado o también de aquellos más recientes o estables a lo largo de la historia en la Península Ibérica y, en concreto, en población Española. Por otro lado, datos del presente estudio, así como también de futuros estudios relacionados, podrían tener una contribución relevante en la eficacia de nuevas y presentes inmunoterapias y en los criterios de selección de terapias personalizadas. Por último, el conocimiento de los haplotipos HLA al 3-/4-field de resolución más comunes en población Española podría ser de gran utilidad para el diseño de haplo-bancos representativos de determinadas líneas celulares homocigóticas para el trasplante alogénico de células madre pluripotentes inducidas (iPSC) y su respectiva terapia como en el caso de la terapia adoptiva con el uso de células T diseñadas genéticamente (como, por ejemplo, las células CAR-T con un receptor de antígeno quimérico).

II) En segundo lugar, en relación al estudio de asociación de genes HLA y Esclerosis Múltiple (EM) en población Española:

- 1) La aplicación de secuenciación masiva NGS en el genotipaje HLA permitió el mapeo de gran precisión de aquellas variantes alélicas y haplotípicas asociadas, pudiendo excluir al mismo tiempo aquellas asociaciones que no estaban directamente asociadas (únicamente por DL) con riesgo o protección a EM.
- 2) La señal refinada *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* asociada a predisposición con EM fue estadísticamente significativa.
- 3) Análisis nominales y por estratificación identificaron también una segunda señal significativa de riesgo a EM relativa al alelo *HLA-DPBI\*03:01:01*, siendo en este caso independiente del factor principal de riesgo *DRB1\*15:01:01:01*.

- 4) Efectos de protección asociados a varias diferentes señales HLA de clase II (varios haplotipos que contenían respectivamente *HLA-DRB1\*04-* y *HLA-DRB1\*15:02:01:02*) se debían a un patrón negativo de DL con respecto al factor principal de riesgo *DRB1\*15:01:01:01*.
- 5) Alelos HLA de clase I *HLA-B\*38:01:01*, previamente identificado en anteriores estudios, y la nueva señal de *HLA-B\*58:01:01:01* mostraron un efecto protector moderado siendo independiente entre ellos y también independiente de factores asociados de HLA clase II.
- 6) En el presente estudio, no se encontró un claro efecto protector a EM atribuido para el Bw4 por si solo (considerando alelos HLA-B y los correspondientes subgrupos NLRIALR, DLRTLLR y NLRTALR respectivamente). Mientras que, por otra parte, el epitopo Bw6 (SLRNLRG, codificado por los respectivos alelos HLA-B estudiados aquí) mostraba una asociación con riesgo a EM que no podía explicarse simplemente en relación a patrones de DL con respecto al factor principal de riesgo *DRB1\*15:01:01:01*, siendo así más bien independiente.



## ***INTRODUCTION***



## **I. HUMAN LEUKOCYTE ANTIGEN (HLA) SYSTEM**

### **1. DEFINITION, DISCOVERY AND GENERAL ASPECTS**

The human Major Histocompatibility Complex (MHC), also denominated as the Human Leukocyte Antigen (HLA) system, is a genomic region located on the short arm of chromosome 6, band p21.3 (6p21.3). The classic human MHC region comprises approximately 3.8 million base pairs (Mbp) of DNA representing 0.13% of the human genome [1]. More recent studies have defined the concept of the extended human MHC (xMHC), which includes a total of 7.6 Mbp, based on the finding that linkage disequilibrium (LD) and MHC related genes exist outside the boundaries of this classic human MHC region [2].

The classic 3.8 Mbp MHC region is among the most gene-dense segments of the human genome in which about 40% of the protein-coding human MHC genes have immune-related function in both innate and adaptive immunities. Importantly, this classic human MHC region includes the highly polymorphic HLA genes that encode the HLA glycoproteins, which play a key role in immune recognition and regulation [3]. These cell-membrane-bound glycoproteins known as HLA classical class I and class II molecules regulate the immune response by presenting peptides of fragmented proteins to different types of effector cells. In the context of adaptive immunity, HLA class I and II serve as the structures that present self and foreign peptides to cytotoxic (CD8+) and helper (CD4+) T lymphocytes, respectively [4]. At the same time, HLA class I molecules also participate in aspects of innate immunity by acting as ligands that interact with receptors located on the surface of natural killer (NK) cells [5].

The MHC region is now known to be part of the genome of all the jawed vertebrates studied so far, presenting also a highly complex and large group of linked genes where many of them are involved functionally with the adaptive and innate immune responses [6]. The MHC system was



first discovered in 1936 by Peter Gorer (inspired by John B. S. Haldane's hypothesis about possible resistance factors to the growth of allogeneic tumors that might be associated with some blood group antigens) during his pioneering studies of antigenic responses to transplanted sera by inbred mouse strains in which he identified an antigen (named as antigen II) responsible for tumor rejection [7-9]. Even before then, in the early 1900s, based on previous Carl Jensen and Leo Loeb's work on allogenic tumor transplantations in different mice strains, geneticists Ernest Tyzzer and Clarence Little arrived at the conclusion that susceptibility/resistance to the growth of allogeneic tumors was somehow genetically determined [10]. In 1948, the MHC in mice (named as the *H-2* locus or H-2 complex) was genetically defined more precisely by George Snell (initially in collaboration with Peter Gorer) in his studies of tumor resistance genes, which he called histocompatibility or H genes [11][12]. The human MHC system was initially discovered and defined as an antigenic system by Jean Dausset in 1958, as he found the first iso(allo)antibodies in the blood of transfusion patients which were specific against antigens expressed by human leukocytes from certain individuals [13]. During that same year, there were two other research groups (headed by Jon van Rood and Rose Payne, respectively) that independently also noted how sera from multiparous women or from previously transfused individuals contained antibodies that agglutinated leukocytes from many but not all individuals who were tested [14][15]. As a consequence of the observations made about these first leukocyte antigens described on these initial and other subsequent studies, the human MHC system was termed as the Human Leucocyte Antigen (HLA) complex. During the 1960s, 1970s and 1980s there was a very remarkable progress of knowledge in the field thanks to significant instrumental and technical innovations (especially in serological typing methods and microlymphocytotoxicity assays), including their standardization, as well as an extensive international collaboration represented by the organization of International Histocompatibility Workshops (IHIWs) [16]. In that period of time, many different

groups of investigators carried out studies that contribute to describing and defining the genomic organization of the HLA chromosomal region and its six different major and very polymorphic series of determinants (A, B, C, DR, DQ and DP antigens) (from the studies of respective groups led by Dausset, van Rood, Payne, Amos, Ceppellini, Bodmer, Kissmeyer-Nielsen and Thorsby among others); the evidence that HLA class I and II molecules are important histocompatibility antigens in transplantation (from the studies of respective groups led by Dausset, van Rood, Amos, Ceppellini, Kissmeyer-Nielsen, Terasaki, Thorsby, Leiden, Ting and Morris among others); the first associations of HLA antigens with diseases (from the studies of respective groups led by Amiel, Brewerton and Schlosstein among others); the immunobiological function of these HLA antigens as peptide-presenting molecules (from the studies of respective groups led by Benacerraf, McDevitt, Tyan, Doherty and Zinkernagel among others); the molecular structure of HLA antigens that also help to understand the phenomenon of MHC/HLA restriction where both the peptide and the presenting HLA molecule comprise a complex ligand that interact with the corresponding receptor located on the surface of certain effector cells (from the studies of respective groups led by Doherty, Zinkernagel, Ziegler, Unanue, Townsend, Strominger, Wiley, Bjorkman, Brown and Engelhard among others) [16][17].

The basis of the initial discovery of the HLA molecules was that they constitute histocompatibility transplantation antigens defining rejection response to grafted tissue. However, it was not until later (in the 1970s and 1980s) when it became known their primary biological function and pivotal role in the regulation of the immune system [16]. The main function of both HLA class I and class II molecules is to bind peptides derived from self or nonself antigens and then traffic to the cell surface, where these peptides are presented for recognition by the appropriate effector T cells. Conversion of antigens from pathogens or transformed (tumor) cells into HLA-I and HLA-II bound peptides is critical for mounting protective T cell responses (engaging the key

elements of adaptive immunity: specificity, memory and diversity), and similar processing of self proteins is necessary to establish and maintain tolerance [18]. These peptides are products of proteolysis, and there are two major proteolytic systems operating within the cell that contribute to HLA-dependent T cell recognition depending on the antigen source of the displayed peptides:

- HLA class I molecules (expressed on the surface of all nucleated somatic cells) present peptides (8-11 amino acids in length) from intracellular antigens (coming from the cell's own proteome (e.g. tumor cell) or from foreign intracellular pathogens (e.g. virus or bacteria)) to T cell receptors of specific CD8+ cytotoxic T cells (see references in [16]). Additionally, HLA class I proteins can also act as ligands for killer-cell immunoglobulin-like receptors (KIRs) [19] that regulate the cytotoxic activity of CD8+ cytotoxic T cells (CTLs) and natural killer (NK) cells and leucocyte immunoglobulin-like receptors (LILRs) expressed on myelomonocytes and other leucocyte lineages [20].

- Meanwhile, HLA class II molecules (expressed normally on the surface of a subgroup of immune cells that includes B cells, activated T cells, macrophages, dendritic cells, and thymic epithelial cells) present peptides (15-25 amino acids in length) from exogenous antigens (coming from foreign extracellular pathogens (e.g. parasite or bacteria)), which are degraded in the endocytic pathway, to T cell receptors of specific CD4+ helper T cells (see references in [16]).

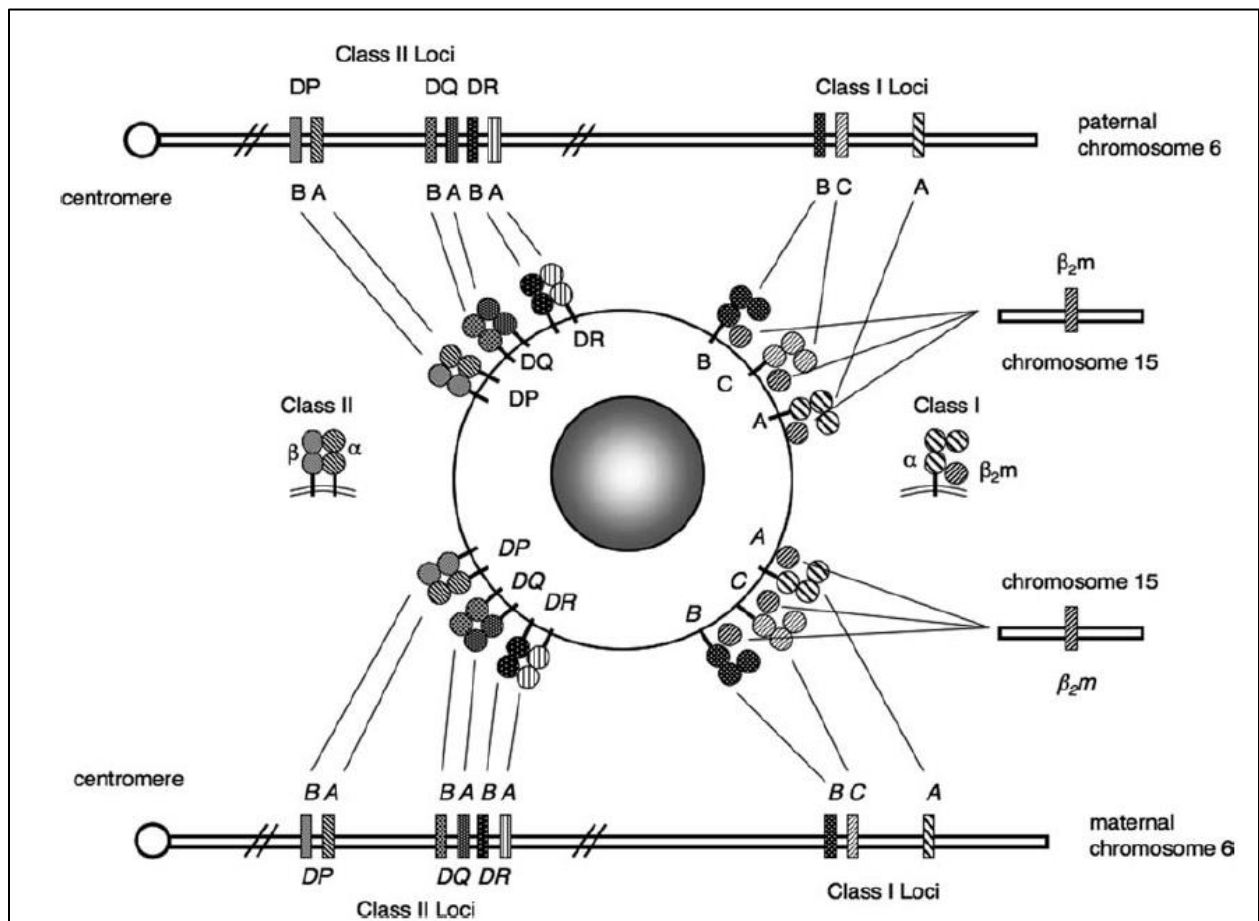
The HLA's immunobiological function in immune responsiveness is also intrinsically reflected in its genetic polymorphism. Within the human MHC, the classical HLA class I and class II genes or loci are among the most polymorphic in the human genome (as well as their close MHC homologs in other organisms [21]) [1], presumably (as this is still an unproven hypothesis) to preserve the variability of the antigen-presenting ability. As this reflects the evolutionary advantages of a diverse immunological response to defend against and survive the natural selection

pressure from a wide range of infectious pathogens [22][23]. Where this extensive HLA allelic sequence diversity affects peptide binding and recognition of the HLA-peptide complex by the T cell receptor [3].

In brief, the main properties of the HLA system, which define its complex and diverse nature, are:

-*Polygenic*: it contains several different HLA class I and class II genes, so that every individual possesses a set of HLA molecules with different ranges of peptide-binding specificities [18].

-*Diallelic/Codominant expression*: in an individual for each HLA gene, the alleles inherited from both maternal and paternal chromosomes are codominantly expressed on the cell surface, maximizing the number of HLA molecules available to bind peptides for presentation to T cells [18][24] (see **Figure I-1**).

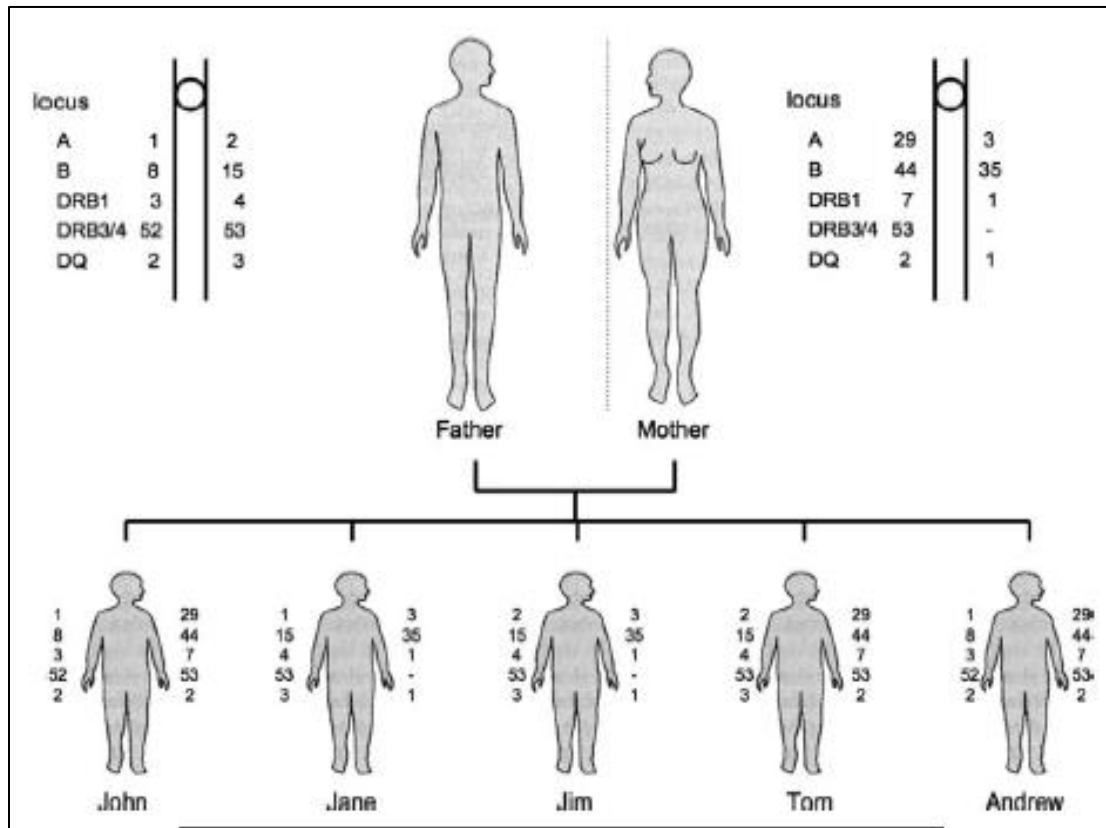


**Figure I-1.** Codominant expression of HLA gene products encoded by the major histocompatibility complex. The *HLA-A*, *-B*, and *-C* class I loci, and the linked *HLA-DR*, *-DQ*, and *-DP* class II loci are located on the short arm of chromosome 6 (6p21), and the class I light-chain locus,  $\beta 2M$ , is encoded on chromosome 15. HLA genes and their respective proteins are shaded to reflect the different loci encoding these proteins and the inheritance of different alleles from the two parental chromosomes. The separate HLA class II  $\alpha$ - and  $\beta$ -chain loci are also shown. The products of both maternal and paternal chromosomes are codominantly expressed on the surface of antigen-presenting cells, resulting in expression of up to six distinct class I allotypes. The number of expressed class II gene products can be even greater, because some haplotypes have extra *HLA-DRB* loci that produce additional  $\beta$  chains capable of assembling with  $DR\alpha$ . In addition, pairing of certain  $DQ\alpha$  molecules from the *HLA-DQA* locus encoded on one chromosome, with  $DQ\beta$  chains derived from the other chromosome, can result in expression of new DQ cis–trans isotypes. The HLA class I and class II loci are separated by the class III region of the major histocompatibility complex (not shown). HLA class II molecules are constitutively expressed only on B cells, macrophages, and dendritic cells, whereas class I molecules are found on nearly all nucleated cell types. Figure and respective footnote are obtained and adapted from [24].

*-Polymorphic:* there are multiple allelic variants, relatively with high frequency, of each HLA gene within the population as a whole. HLA genes exhibit a high degree of polymorphism, and a number of different mechanisms may contribute to the generation and maintenance of this polymorphism. Among these are the selective advantage of a heterozygous pool of antigen presenting elements in a given individual that might allow the binding and presentation of antigenic peptides derived from a wide variety of environmental pathogens [24]. 2005). It has been widely described how nucleotide substitutions are shared by more than one HLA gene, and thus indicating a patchwork nature of HLA sequence polymorphism suggestive of segmental exchanges. Therefore, it is well-accepted that the extensive allelic diversity at these HLA loci is generated by recombinations and gene conversions (involving relatively short fragments of DNA leading to single or short amino acid motives substitutions) as well as by point mutations [25].

*-HLA Haplotype:* an important concept (first introduced by Cepellini and his associates in 1967 [16]) commonly used is “haplotype,” which refers to the linked set or string of particular alleles at distinct neighboring loci (relatively located physically close) that occur and are inherited as

a group on a parental chromosome. This existing linkage of HLA loci on chromosome 6 means that a given individual will usually inherit (in a Mendelian fashion) a set of nonrecombined HLA alleles encoded at linked HLA loci from each parent [16][24]. Thus, the complete HLA genotype of an individual comprises the corresponding paternal and maternal HLA haplotypes (see **Figure I-2**).



**Figure I-2.** Segregation of HLA haplotypes in a nuclear family. Each of these sets (haplotypes) of neighboring HLA polymorphisms is co-transmitted or segregated on a single parental chromosome in the absence of recombination during meiosis (in that case, crossing-over events occur between HLA loci generating newer haplotype/s different from original parental one/s) When meiotic recombination appears to have occurred within a family, interpretation of HLA typing results as well as related phasing and assignment of HLA haplotypes can be difficult. Nevertheless, before concluding that recombination explains the results observed, HLA histocompatibility clinical laboratories should also consider the possibilities of typing error and false paternity as possible alternatives. Figure and respective footnote are obtained and adapted from [24][90].

-*Linkage Disequilibrium (LD)*: Genes within the HLA region demonstrate extensive linkage disequilibrium observed even among relatively distant genomic regions. Linkage

disequilibrium is a phenomenon where alleles at linked loci (organized in Haplotypes) segregate more commonly together than predicted by chance. Existing data suggest that positive selection is operating on the haplotype and that the linked loci confer a particular selective advantage for the host [26]. Moreover, common haplotypes within a given population appears to reflect functional interdependencies of the HLA gene alleles.

## **2. GENOMIC ORGANIZATION OF THE HUMAN MHC OR HLA SYSTEM**

The human MHC complex, or HLA system, is located on the short arm of chromosome 6, in the distal portion of the region 6p21.1-6p21.3. The MHC system is among the most gene-dense segments and polymorphic regions of the human genome, where the defined classic MHC region at around 4 Mbp occupies 0.13% of the human genome ( $3 \times 10^9$  bp), but contains about 0.5% (which is more than 150) of the approximately 32,000 known protein-coding genes [27].

### **2.1 Human Major Histocompatibility Complex (MHC) or HLA Genomic Map**

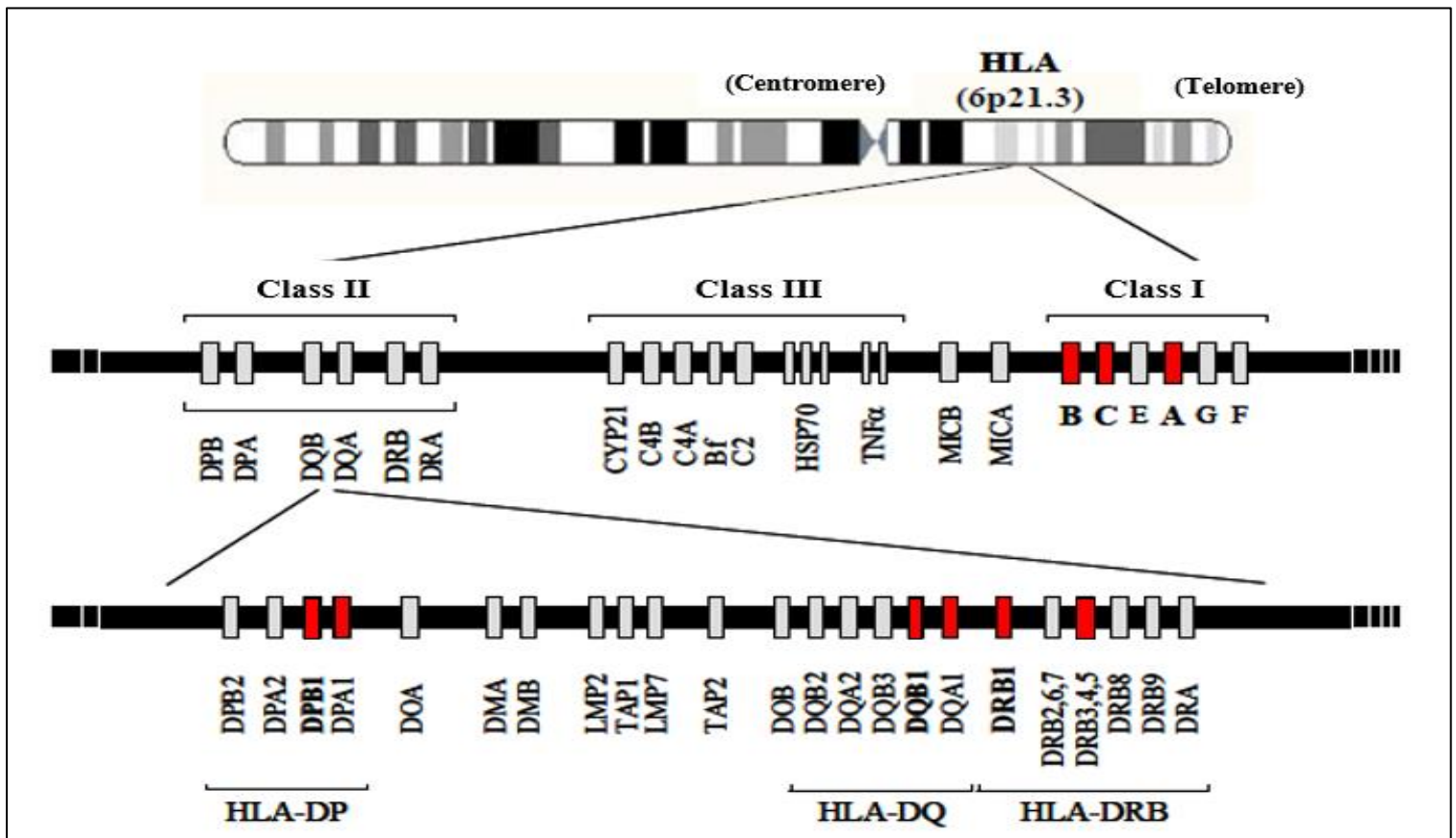
In 1991, Trowsdale et al. reported the first genetic map of the human MHC [28]. Few years later, in 1999, the first-sequenced based map (derived from many individuals of unknown HLA type, defining a “mosaic” human MHC haplotype) comprised 3.6 Mbp DNA, described 224 gene loci of which 57% were predicted to be expressed and about 40% of the expressed genes were estimated to have immune system function. It also divided the human MHC into three regions: class I (located at the telomeric end), class III, and class II (located at the centromeric end) [1]. Later on, that virtual MHC haplotype of 1999 was replaced by a single reference sequence of a homozygous haplotype derived from sequencing the PGF cell line [29]. When the entire sequence of the chromosome 6 was described [30], linkage disequilibrium (LD) and MHC related genes were found outside the boundaries of this classic 3.6 Mbp human MHC region [2]. Thus, this evidence has led to the concept of the extended MHC (xMHC) region, which covers a total of 7.6 Mbp on the short arm of this chromosome (corresponding to the updated region 6p22.2-6p21.3),

and it is divided into 5 subregions: extended class I (telomeric side of *HCG4P11*; ~3900 kbp), class I (ranges from *HLA-F* to *MICB*; ~1900 kbp), class III (ranges from *PPIAP9* to *BTNL2*; ~700 kbp), class II (from *HLA-DRA* to *HLA-DPA3*; ~900 kbp) and extended class II (centromeric side of *COL11A2*; ~200 kbp) regions. In this human xMHC, of the 421 loci described, 60% are considered to be expressed and about 28% of the expressed transcripts are potentially associated with immunity [2]. Whereas the human MHC class I and class II genomic subregions encode the highly polymorphic gene complex of the HLA class I and HLA class II genes, the class III subregion is the most gene dense subregion (containing 58 (23%) of the expressed genes in a 0.7 Mbp segment) of the xMHC and of the human genome [31]. Additionally, within the xMHC there are also 139 loci classified as pseudogenes (since xMHC region is very rich in paralog genes as a result of genomic duplications), representing a 33%, and the remaining 7% includes tRNA genes [2]. In 2008, the MHC Haplotype Project study, which described over 44,000 variations, both substitutions and indels (insertions and deletions) by comparing eight different haplotype sequences from homozygous cell lines, also confirmed the presence of more than 300 loci, including over 160 protein-coding genes within the human MHC region. Combined analysis of the variation and annotation datasets revealed 122 gene loci with coding substitutions of which 97 were non-synonymous [32]. Currently, the defined classic human MHC genomic map comprises 3.8 Mbp and includes 158 protein-coding genes and 86 pseudogenes of unknown functionality, spanning from the gamma-aminobutyric acid receptor (*GABBR1*) gene on the telomeric side of the region to the kinesin family member C1 (*KIFC1*) gene toward the centromere, based on the Genome Reference Consortium Human Build 38 patch release 7 (GRCh38.p7) in the National Center for Biotechnology Information (NCBI) database (<http://www.ncbi.nlm.nih.gov/gene/>) (ENSEMBL 86 GRCh38. p7 coordinates chr6: 29555629-33409924) [27][33] (see **Figure I-3 and Figure I-4**).





GRCh38.p7 primary assembly of the NCBI map viewer. The regions separated by arrows show the HLA sub-regions such as extended class I, class I, class III, classical class II and extended class II regions from telomere (left and top side) to centromere (right and bottom side). Family pedigree analysis has shown that recombination occurs at specific locations within the MHC, leading to a structure of four major genomic blocks (in general, each block defines a set of loci without an intervening hotspot of recombination). Dawkins and colleagues referred to these blocks (delimiting segments between these respective pair of genes) as the Alpha ( $\alpha$ ; *HOG4P11* and *HLA-J*), Kappa ( $\kappa$ ; *TRIM26BP* and *HLA-E*), Beta ( $\beta$ ; *HCG27* and *MICB*), Gamma ( $\gamma$ ; *PPIAP9* and *HCG23*) and Delta ( $\delta$ ; *BTNL2* and *HLA-DMA*) blocks [459]. The Alpha block contains *HLA-A*; the Beta block contains *HLA-C* and *HLA-B*; the Gamma block (which lies between the Beta and Delta blocks, it contains >60 genes; which were also used to characterize Conserved Extended Haplotypes (CEH), or also known as Ancestral Haplotypes (AH), in the past [486][530]) contains the tumor necrosis factor (*TNF*) gene and the complement proteins *C2*, *C4*, and factor B (*Bf*); and the Delta block contains the *HLA-DR* and *HLA-DQ* genes, where the *HLA-DP* genes could be also considered part of this latter Delta block as an exception, despite the presence of a hotspot of recombination that creates a weak LD between the *HLA-DPBI* and the *-DQB1* loci [459][485]. Blue and pink boxes and arrows show the spans of  $\alpha$ ,  $\kappa$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  blocks and framework gene (well-conserved non-MHC genes in mammalian species) blocks, respectively. White, grey, dotted and black boxes show protein-coding genes, non-coding RNAs (ncRNAs), small nucleolar RNAs (snoRNAs) and pseudogenes, respectively. Red and blue letters indicate *HLA* class I/*MIC* and *HLA* class II genes, respectively. Figure and respective footnote are obtained and adapted from [34] and [35].



**Figure I-4.** Gene map of the HLA genomic region (located on the short arm of chromosome 6, band p21.3 (6p21.3)) showing, from telomere (right side) to centromere (left side), main relevant genes of interest in the immunogenetics and histocompatibility field. Highlighted in red color are the 11 major classical HLA genes characterized via next-generation sequencing (NGS) in the present thesis work. In bold black color letters, the most polymorphic major classical HLA loci (*HLA-A*, *-C*, *-B*, *-DRB1*, *-DQB1* and *-DPBI*) are indicated. Original figure and respective footnote are obtained and adapted from [531].

## 2.2 Human xMHC Gene Clusters

Segmental duplication that results in the formation of gene clusters is a particular hallmark of the human MHC complex [2] [36]. This existing linkage or clustering of immune-system genes along the HLA system is thought to be functionally advantageous. For instance, in order to ensure that the protein components will be co-expressed in quantities appropriate for the formation of heterodimers (e.g. *HLA-DQA1* and *HLA-DQB1*) or those products involved in antigen processing, including peptide transporters (*TAP1/2*) which are needed to provide peptide antigens for loading onto HLA class I molecules [36]. These clusters (three or more paralogous genes or pseudogenes that are present within a 1 Mbp stretch,) and superclusters (clusters with additional related gene(s) outside the core cluster, but within the xMHC) are described according to their main position on the chromosome from telomere to centromere [2]:

*-Histone supercluster:* Histones are basic proteins involved in nucleosome formation. They occur in five classes, H1 (linker histone), and H2A, H2B, H3 and H4 (core histones). With a total of 66 loci (55 expressed genes, 11 pseudogenes), mostly located in the extended class I subregion, they present the largest histone cluster in the human genome and the largest protein encoding supercluster within the xMHC [37].

*-Solute carrier cluster:* Solute carrier (*SLC*) genes, located in both the extended class I and extended class II subregions, are part of a diverse family with significant physiological roles in solute and nutrient transport [38].

*-HLA class I supercluster:* The HLA class I supercluster comprises the classical class I genes (*HLA-A*, *-B* and *-C*), the non-classical class I genes (*HLA-E*, *-F*, *-G*, *HFE* and 12 pseudogenes) and the class I-like genes (*MICA*, *MICB*, and 5 pseudogenes) [27][39]. Although structurally very similar, classical HLA class I molecules are distinguished by their extraordinary

polymorphisms, whereas the non-classical HLA class I genes, HLA-E, HLA-F and HLA-G, are distinguished by their tissue-specific expression and limited polymorphism [27].

*-tRNA supercluster:* tRNA genes are only 75–90 bases long and are crucial as the molecular adaptors in mRNA-mediated protein synthesis [40]. The tRNA supercluster found within the xMHC (located in the extended class I subregion), it is the largest tRNA cluster in the human genome, comprising 157 tRNA loci.

*-Butyrophilin supercluster:* Butyrophilin (BTN) genes are members of the immunoglobulin superfamily (IgSF). Although it is still unknown, the function of the BTN genes is thought to be related with lipid metabolism [41].

*-Vomeronasal-receptor cluster:* Vomeronasal-receptor (VNR) genes, also located in the extended class I subregion, are members of the pheromone receptor family, which is involved in the subconscious perception of volatile substances such as pheromones. Nevertheless, there is only exclusive presence of pseudogenes in the human VNR cluster [42].

*-Olfactory-receptor supercluster:* it contains 34 olfactory-receptor loci, 14 of which are potentially functional as they provide the basis for odor perception. Similar to immune genes that provide protection from pathogens, they provide an essential survival tool in behavioral processes, including reproduction and predation [43].

*-Zinc-finger supercluster:* Genes that encode zinc-finger proteins are grouped according to the presence of particular zinc-finger domains rather than overall genomic sequence similarity. Zinc finger gene products have diverse functions and can act as enzymes, storage proteins, replication proteins and transcription factors [44].

*-Tumor necrosis factor cluster:* The tumor necrosis factor (TNF) cluster contains genes for three cytokines (*TNF*, *LTA* and *LTB*). All three genes belong to the TNF superfamily and are involved in various inflammatory pathways [45].

*-Lymphocyte antigen cluster:* Lymphocyte antigen 6 (LY6) genes encode glycosyl-phosphatidyl-inositol (GPI) anchored cell-surface proteins with putative immune function [46]. They present the largest gene cluster within the MHC class III subregion.

*-Heat shock cluster:* Heat shock protein (HSP) genes are upregulated by cellular stress such as heat shock and act as chaperones in the synthesis, folding, assembly, transport and degradation of proteins [47]. The cluster of three HSP genes in the MHC class III subregion is involved in stress-induced signaling for immune system mediated elimination of damaged, infected or malignant cells [48].

*-Complement factor genes:* the *C4A*, *C4B*, *C2*, *CFB* genes in the HLA class III region are involved in cascade activation of the classical complement pathway and consequently interact with proteins encoded by genes from outside the MHC [49].

*-HLA class II cluster:* The HLA class II cluster comprises the classical class II genes (*HLA-DP*, *-DQ*, *-DR* and pseudogenes) and the non-classical class II genes (*HLA-DM* and *-DO*). The classical class II genes are expressed on the cell surface as heterodimers consisting of corresponding  $\alpha$  and  $\beta$  chains that present antigens to CD4<sup>+</sup> T cells. The non-classical class II genes are not expressed on the cell surface, but form heterotetrameric complexes involved in peptide exchange and loading onto classical class II molecules [50]. No class II-like gene has yet been found elsewhere in the human genome [2].

### **3. HLA CLASS I REGION**

This region is located at the telomeric end of the human MHC complex in chromosome 6 and spans approximately 1.9 Mbp of DNA [2].

#### **3.1 HLA Class I Genes**

The HLA class I region includes the classical HLA class I genes (*HLA-A*, *HLA-B* and *HLA-C*), the class I-related (like) genes (*MICA* and *MICB*), the non-classical HLA class I genes (*HLA-E*, *HLA-F* and *HLA-G*), and a group of pseudogenes (<http://hla.alleles.org/genes/index.html>). Although structurally they are very similar and some of their functions seem to be coordinated, HLA class I molecules present differences in aspects such as tissue-specific expression and level of polymorphism.

##### **3.1.1. Classical HLA Class I Genes (HLA Class-Ia)**

The classical HLA class I genes include *HLA-A*, *HLA-B* and *HLA-C* loci (of ~4-5 kb of DNA sequence length per gen). Each of these classical HLA class I loci encode a corresponding heavy  $\alpha$  chain that is highly polymorphic. The classical class I molecules are ubiquitously expressed (in all nucleated cells and in platelets) membrane-bound glycoproteins that associate non-covalently with the “light”  $\beta 2$  microglobulin (encoded by its respective non-polymorphic gene (*B2M*) on chromosome 15) to present, on the cell surface, intracellularly processed self/nonself peptide antigens to T CD8<sup>+</sup> cytotoxic lymphocytes, thereby controlling cell-mediated immune response. Additionally, intact *HLA-A*, *HLA-B*, or *HLA-C* molecules are also ligands for KIR receptors (encoded by the KIR cluster that maps to chromosome 19q13.4 within the leukocyte receptor complex (LRC)) located on the surface of NK cells, regulating their development, tolerance and response [51].

### 3.1.2. Non-Classical HLA Class I Genes (HLA Class-Ib)

The non-classical HLA class I genes *HLA-E* (~8 kb of length), *-F* (~6 kb of length) and *-G* (~5 kb of length) encode, respectively, molecules E, F, and G. These molecules have a similar protein structure (presenting an  $\alpha$  chain associated to the  $\beta$  chain,  $\beta 2$  microglobulin) to that of the classical HLA class I counterparts and also require a bound peptide in the binding groove to form a stable complex. However, non-classical HLA class I molecules are characterized by few allelic polymorphisms and play a more tolerogenic role in regulating both innate and adaptive immune responses. Furthermore, they present a more limited tissue distribution, where their expression patterns are often related to their function, such as HLA-G expression in the extravillous trophoblasts at the placenta where it interacts with maternal effector cells [52]. In addition, although the gene *HFE* is also included as another non-classical HLA class I gen, the function of its product is in iron metabolism rather than in antigen processing and presentation [27].

### 3.1.3. HLA Class I-like Genes

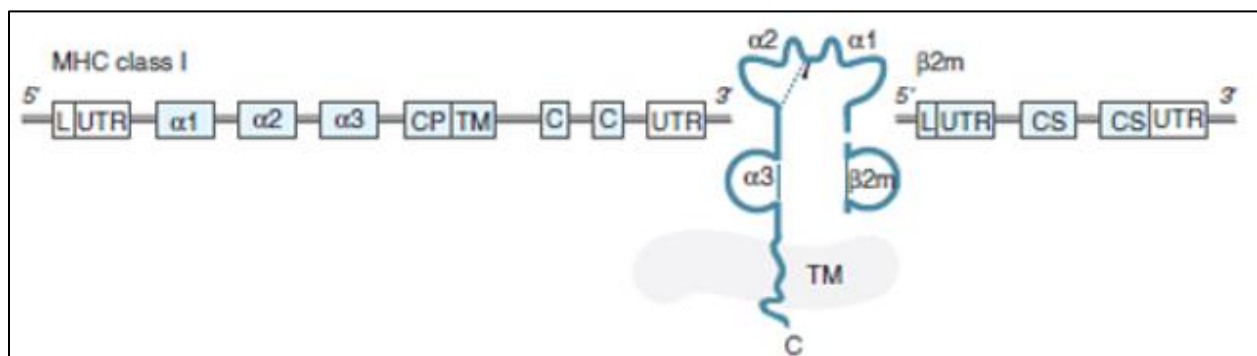
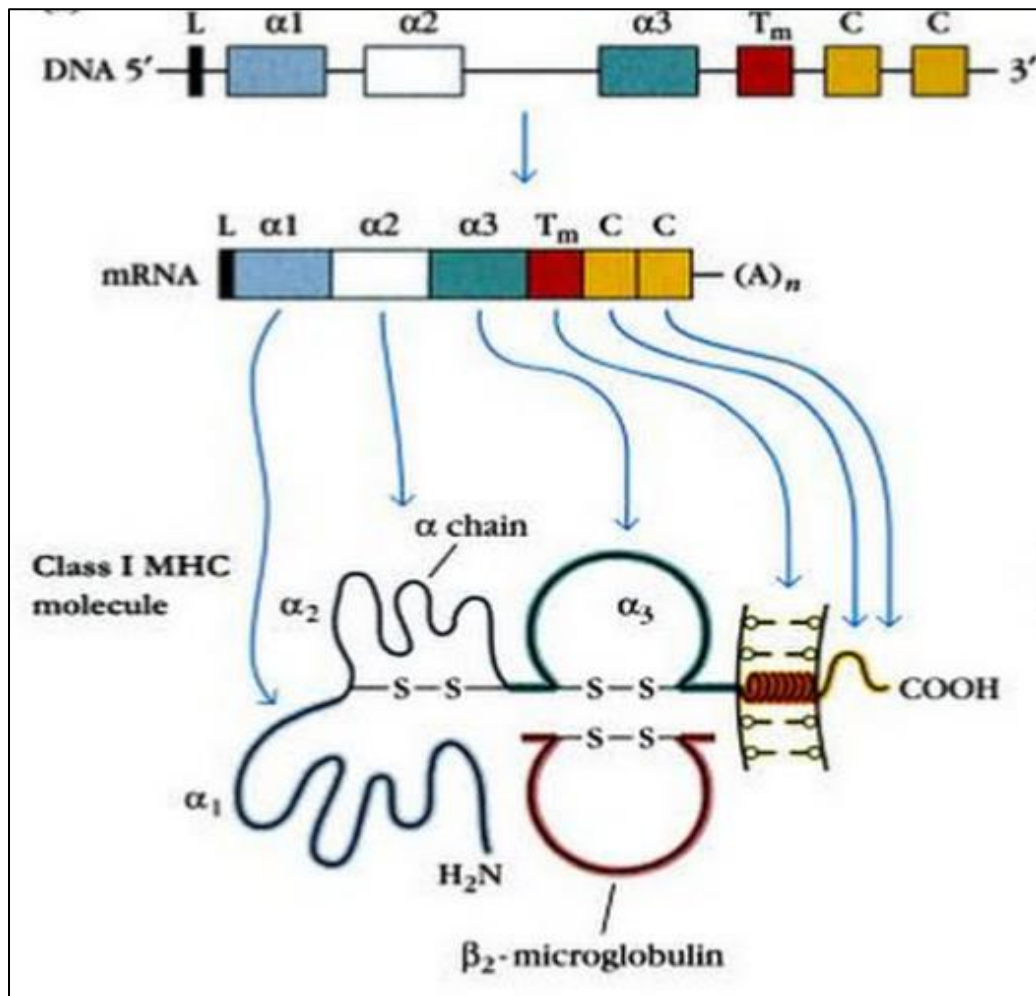
Within the HLA class I region, there is also a group of genes called HLA class I-related polypeptide sequence A (~11 kb of length) and B (~13 kb of length) (*MICA* and *MICB*) that encode molecules presenting other functions different from antigen processing and presentation. The products of these genes are more distantly related members of the class I family that neither associate with  $\beta 2$  microglobulin nor bind peptides. These molecules are expressed as “danger signals” by virus-infected or otherwise stressed/transformed (i.e. tumor) cells. Thus, MICA and MICB are ligands for the natural killer group 2D (NKG2D) receptor located on the surface of memory-effector  $T\gamma\delta$  cells or NK cells, where this interaction activate their effector cytolytic response [35][53].

### 3.2 Genetic Organization of HLA Class I Genes

HLA class I genes are each composed of a series of eight exons delineated by intervening seven introns in addition to the untranslated regions (UTR) located at the 5'UTR and 3'UTR ends. Each HLA class I exon encodes for a specific region of the  $\alpha$  chain molecule: exon 1 encodes the leader peptide sequence; exons 2, 3 and 4 encode the  $\alpha 1$ ,  $\alpha 2$  and  $\alpha 3$  domains, respectively; exon 5 encodes the trans-membrane portion, and exons 6, 7 and 8 encode the C-terminal cytoplasmic tail [54][55] (see **Figure I-5**). The leader peptide sequence, which presents a central hydrophobic region, is located at the N-terminus of the new translated protein and plays a role as a targeting signal to direct the immature protein into the endoplasmic reticulum (ER). Exons 2 and 3 present the most polymorphic nucleotide sequences, where the respective encoded  $\alpha 1$  and  $\alpha 2$  domains, which are the most distal to the cell membrane, are responsible for the peptide binding specificity of each HLA class I molecule [56]. For *HLA-A* and *-C* genes, the cytoplasmic tail domain is encoded by exons 6, 7 and 8 whereas for *HLA-B* gene it is encoded by exons 6 and 7 since the termination codon is contained on that exon 7. Similar to *HLA-B*, although shorter in length due to an existing deletion that causes a reading frame shift establishing an earlier termination codon, the cytoplasmic tail domain of the non-classical HLA-E molecule is also encoded by exons 6 and 7 [57]. In contrast, exon 5 (partially) and exon 6 generate the intracellular cytoplasmic tail of the HLA-F protein. The cytoplasmic tail of HLA-F is much shorter than those of the other HLA-I molecules because of the in-frame translation termination codon located at codon 2 in exon 6. Thus, exons 7 and 8 are not translated. Moreover, the length of the cytoplasmic tail of HLA-F molecule varies, which can lead to the generation of different HLA-F isoforms (splicing variants) [58]. Regarding the HLA-G molecule, a total of seven isoforms exist due to mRNA alternative splicing and differential association with  $\beta 2$  microglobulin. Four of them are found on the cell surface (HLA-G1, -G2, -G3, and -G4), while the other three are soluble forms



released from the cell (HLA-G5, -G6, and -G7), due to the lack of the transmembrane and intracellular domains of membrane-bound HLA-G [59].



**Figure I-5. (Upper Image)** Schematic diagram of *HLA* (or *MHC*) class I gene, respective messenger RNA (mRNA) transcript (after transcription) and respective assembled protein molecule (after translation and maturation). There is a correspondence between exons (represented in colored boxes) and the domains of the *MHC* class I molecule in the gene products [exon 1 encodes the leader peptide sequence; exons 2, 3 and 4 encode the  $\alpha 1$ ,  $\alpha 2$  and  $\alpha 3$  domains (in the N-terminal side of the protein), respectively; exon 5 encodes the trans-

membrane (TM) portion (embedded in the lipid bilayer of the cell surface), and exons 6, 7 and 8 encode the C-terminal cytoplasmic tail (C)] with the exception of the leader (L) exon (encoding leader peptide, which is removed in a post-translational reaction after leading the pre-mature peptide to the endoplasmic reticulum (ER) and before the MHC class I molecule is expressed on the cell surface). Note that (during the process of transcription) the mRNA transcript is spliced to remove introns sequences (represented as black thin lines between exons). **(Lower Image)** Schematic diagram of *HLA* (or *MHC*) class I gene similar to previous upper one (showing in addition here non-polymorphic gene (*B2M*)), showing again correspondence between exons (represented in colored boxes) and the domains (L, leader sequence; UTR, untranslated region; CP, connecting peptide; TM, transmembrane region; C, cytoplasmic region; CS, coding sequence) of the MHC class I molecule. Note that  $\beta 2$  microglobulin is encoded by its respective non-polymorphic gene (*B2M*) on chromosome 15 different from chromosome 6. Original figures and respective footnotes are obtained and adapted from [532] and [533].

### 3.3 Structure of HLA Class I Molecules

The HLA class I molecule consist of a membrane-spanning glycosylated heavy  $\alpha$  chain (masses 45 kilodalton (kDa) and is 362-366 amino acids long) bound non-covalently to extracellular light chain  $\beta 2$  microglobulin (12 kDa), which does not span the membrane [35]. The  $\alpha$  chain has three extracellular domains ( $\alpha 1-3$ , with alpha  $\alpha 1$  being at the N terminus), a transmembrane region and a C-terminal cytoplasmic tail which is enriched in serine and tyrosine amino acids (see **Figure I-6**).

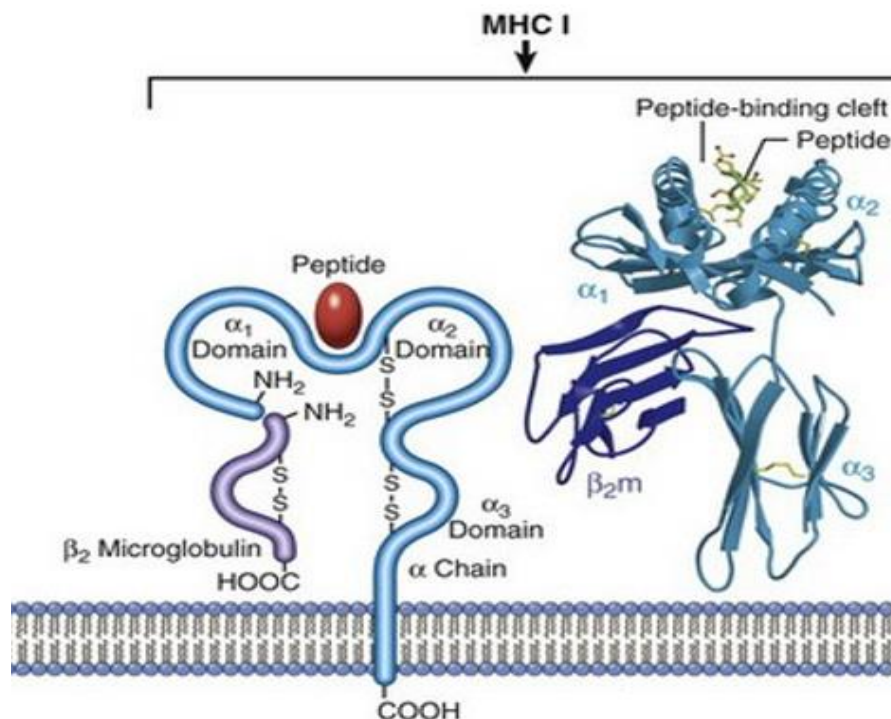
The  $\alpha$  chain folds into three domains:  $\alpha 1$ ,  $\alpha 2$ , and  $\alpha 3$ . The first two  $\alpha$  domains ( $\alpha 1$  and  $\alpha 2$ ) are the most distal to the cell membrane. They fold together into a single structure consisting of two segmented  $\alpha$  helices lying on a sheet of eight antiparallel  $\beta$  strands. Folding of the  $\alpha 1$  and  $\alpha 2$  domains creates a peptide-binding groove (where the floor is composed of symmetric strands of  $\beta$  pleated sheet), or cleft, that is flanked by a surface that interacts with a T cell receptor (TCR) or a NK KIR receptor. In the class I peptide-binding groove, the ends of the helices of the  $\alpha 1$  and  $\alpha 2$  domains converge to close the groove and fix the peptide's orientation. Thus, the class I groove accommodates peptides that average nine amino acids in length (6-16 aa). The  $\alpha 3$  domain and  $\beta 2$ -microglobulin show similarities in amino acid sequence to immunoglobulin C domains and have

similar folded structures. Together they create a structure that supports the peptide-binding domain and, with the transmembrane domain of the  $\alpha$  chain, attaches the molecule to the cell surface [60].

Sequence polymorphism is concentrated in the  $\alpha 1$  and  $\alpha 2$  domains that are encoded by exon 2 and 3, respectively. Some of these polymorphic HLA class I residues (mostly found in the  $\alpha 1$  domain) determine the binding specificities for peptides by forming structures called pockets that interact with complementary residues of the bound peptide, called anchor residues. In consequence, the anchor residues of peptides that bind to each allelic HLA variant are different. Typically, a HLA class I molecule has six pockets (A through F) distributed along the length of the groove. But only two or three of the pockets (in HLA-A, B, or C molecules, usually pockets B and F) are particularly influential in determining which peptide a molecule binds, because the side chains that fit into them serve as the peptide's anchors [61]. Moreover, other polymorphic HLA class I residues (mostly found in the  $\alpha 2$  domain) and some residues of the peptide define the structure that interacts with the TCR [62][63]. The  $\alpha 3$  domain interacts through an acidic loop with the CD8 co-receptor in the T cell (where CD8 uses its two dimeric Ig-like domains to clamp onto the CD loop of the MHC class I  $\alpha 3$  domain in a fashion similar to an antibody binding to an antigen) [64].

Although with important exceptions and a much lower level of polymorphism, the non-classical HLA class I molecules present a very similar protein structure to their classical HLA class I counterparts, showing three globular domains heavy-chain non-covalently bound to  $\beta 2$ -microglobulin and a nonapeptide [65]. In the HLA-E molecule, its groove preferentially presents a restricted subset peptides derived from leader sequences of other classical class I molecules (where mature HLA class I molecules expressed on the cell surface are encoded by exons 2–7). Transporters associated with antigen processing (TAP), in cooperation with tapasin, transfer leader peptides to the endoplasmic reticulum (ER), where they can be associated to HLA-E molecules

permitting their expression on the cell surface [66]. HLA-F molecule can be found in at least three conformational forms: complexed with the light chain  $\beta_2$ -microglobulin, in open conformation, and complexed with a HLA I heavy chain (HC). However, an amino acid-peptide has never been eluted from its peptide-binding groove. Studies have suggested that HLA-F glycoprotein would be capable to escape the ER lumen and reach the cell surface independently from TAP, tapasin, and peptide binding, but using an alternative ER signal encoded in its cytoplasmic tail [65][67]. The extracellular structure of HLA-G1 and HLA-G5 molecules is identical to the well described structure of classic HLA class I molecules. The other HLA-G isoforms are simpler structures with only one or two globular domains, not binding to  $\beta_2$ -microglobulin neither presenting peptides [68].



**Figure I-6.** Schematic diagram and crystal structure of MHC class I molecule. Figure and respective footnote are obtained and adapted from (<https://veteriankey.com/diseases-of-immunity/>) being originally from Dr. P. Bjorkman (California Institute of Technology, Pasadena, CA, USA).

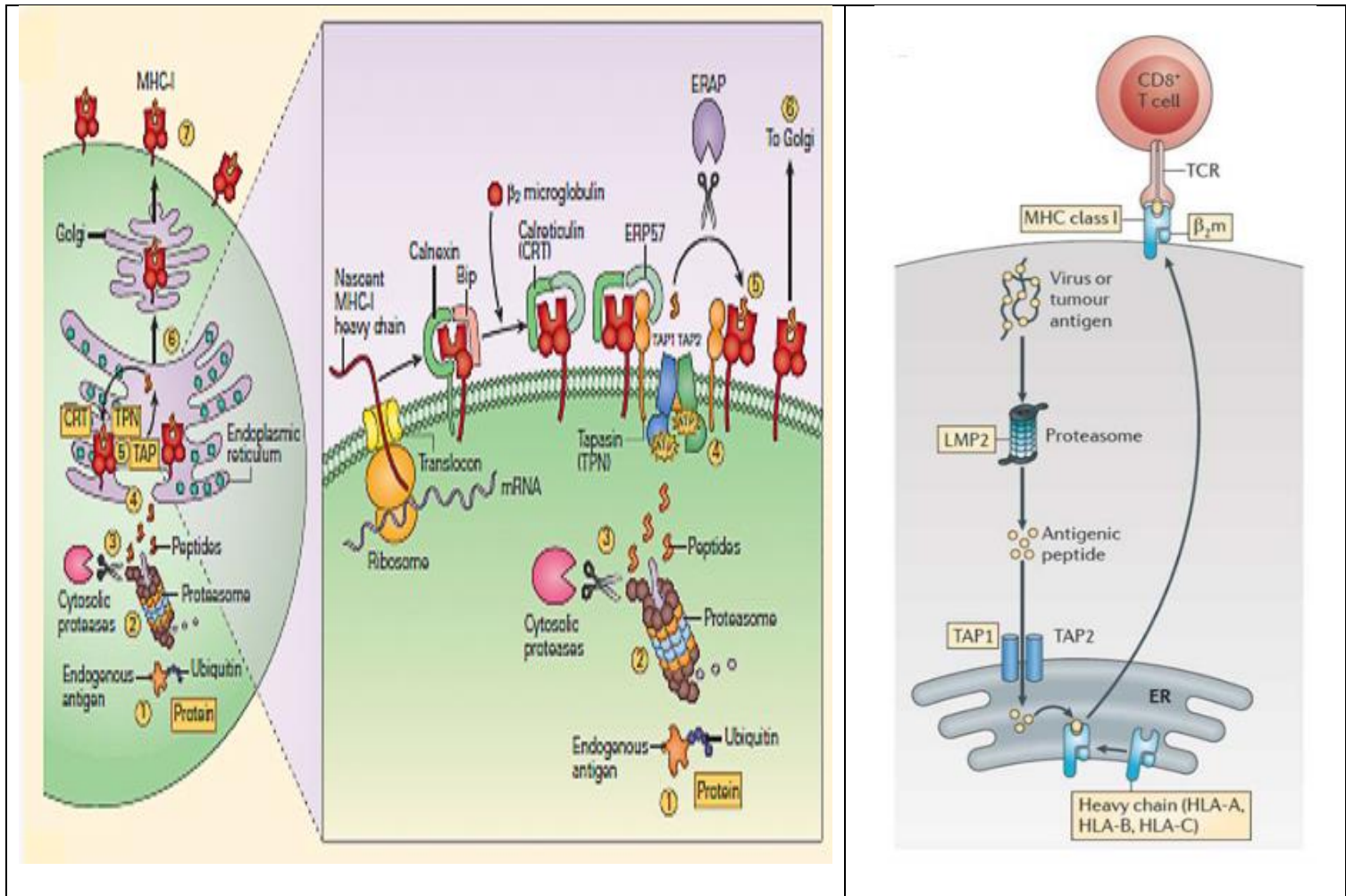
### **3.4 Biological Function of HLA Class I Molecules**

HLA class I molecules are ubiquitously expressed in all nucleated cells and in platelets. HLA class I molecules bind endogenously synthesized peptide fragments of proteolytically degraded proteins originally coming from an intracellular pathogen or other self/nonself antigen. The processing of antigens and the peptide binding to HLA class I molecules is accomplished by a complex series of intracytoplasmic events involving antigen-processing machinery [69] (see **Figure I-7**). Briefly, self/non-self cytosolic proteins are processed primarily by the action of the proteasome generating peptides. Exceptionally, in certain antigen presenting cells, particularly dendritic cells (DCs), exogenous proteins can also be fed into this pathway by retrotranslocation from phagosomes (vacuolar pathway) or can be exported into the cytosol after phagocytosis (cytosolic pathway), a phenomenon known as cross-presentation. At the same time, HLA class I molecules are assembled and stabilized by chaperone proteins (calreticulin, Erp57, protein disulfide isomerase (PDI) and tapasin) in the endoplasmic reticulum (ER). Tapasin interacts with the transport protein TAP (transporter associated with antigen presentation) which translocates peptides from the cytoplasm into the ER. Translocated 8-16 amino acids may require additional trimming by aminopeptidases in the ER before binding to MHC class I molecules. Once on the cell surface, HLA class I molecules present peptide antigens to peripheral CD8<sup>+</sup> T cells, whose main function is cytolysis of pathogen-infected cells or transformed tumor cells. The phenomenon that T lymphocytes recognize, via their TCR, a foreign peptide antigen only when it is bound to a particular allelic form of a self MHC molecule is called the MHC restriction. Thus, the recognition of antigens by CD8<sup>+</sup> T cells is restricted by self MHC class I alleles (see **Figure I-7**). In addition to the induction and regulation of immune responses in the context of CD8<sup>+</sup> T cells, self HLA class I molecules are involved in the selection of the engaged CD8<sup>+</sup> T cell repertoire at the thymus establishing also self-tolerance [70].

Intact expressed HLA-A, HLA-B, or HLA-C proteins are also ligands for KIR molecules located on the surface of NK cells, regulating their development, tolerance and response [51]. Essentially, upon encounter with infected or stressed/tumor cells with decreased expression of these HLA class I molecules (the denominated “missing-self”), NK cells are no longer subject to inhibitory signals initiated by the engagement of HLA class I-specific KIR receptors, promoting NK cell cytotoxicity and cytokine production [71]. Different from HLA-TCR interaction, KIR receptors do not interact with the whole top area of the HLA molecule. Instead, they interact with one end of the top of the molecule, where dimorphisms in the HLA class I  $\alpha$  domains are the major determinants for this interaction (binding motifs are referred to as C1 and C2 in HLA-C and Bw4 in HLA-B and HLA-A). Interaction between HLA class I and their inhibitory KIR receptors is thus thought to play a major role in the mechanisms of self tolerance during NK cell effector phases [72].

The non-classical HLA molecules have low expression levels, are less polymorphic and present more limited tissue distribution compared with their classic HLA class I counterparts. HLA-E (is expressed in all nucleated cells and frequently overexpressed in tumor cells and virus-infected cells) primarily presents self peptides to the TCR of CD8<sup>+</sup> T cells. The diversity of these self peptides is limited and includes the leader peptide of classic HLA class I molecules. The binding of HLA-E to inhibitory receptors in NK cells, such as CD94/NKG2A, is an important part of the surveillance mechanism for missing-self. Although HLA-F is predominantly expressed in an intracellular, unstable, and immature form, high level of HLA-F surface expression has been observed in activated B, T, and NK cells. HLA-F molecule has a small binding cleft that does not contain peptide, and its functions are not well known (although it is thought that it may present tolerogenic and immunomodulatory properties). HLA-G is primarily expressed by placental trophoblast cells, the thymus, the cornea, and some erythroid and endothelial precursor cells. HLA-G has a peptide groove, binds a nonamer peptide, and is recognized as an MHC-peptide complex

ligand by the leukocyte Ig-like inhibitory receptors (LIR-1 and LIR-2) and KIR receptors. HLA-G seems to be important in the modulation of the maternal immune system during pregnancy and thereby the maternal acceptance of the semiallogenic fetus [65].



**Figure I-7. (On the Left Image)** Peptide loading of MHC-I molecules. Panel shows the synthesis and peptide loading of MHC-I through the endogenous pathway. Endogenous proteins (e.g., a self-protein or a viral protein) synthesized in the cytoplasm are modified initially by ubiquitin (1), following which they are processed by the proteasomes (with protein subunit LMP2 shown here) (2). After trimming by cytosolic proteases (3), the peptides enter the endoplasmic reticulum via the TAP 1 and TAP 2 transporters (4). The MHC-I alpha chain, which is initially formed as a linear peptide in the ER, is then folded with the help of several chaperones (calnexin, calreticulin [CRT]). Binding immunoglobulin protein (BiP) and endoplasmic reticulum protein 57 (ERP57), during which the  $\beta$ 2 microglobulin is added to the  $\alpha$  chain, complete the synthesis of the complete MHC-I molecule (right inset). The complex is held together by tapasin (TPN), which facilitates transfer of the peptide to the antigen-binding cleft (5). The peptide-loaded MHC-I complex is then transferred to the Golgi (6) and then transported to the surface of the cell (7). **(On the Right Image)** It is shown a simplified version of peptide loading of MHC-I molecules and, once expressed on the cell surface, the final recognition of antigens by CD8+ T cells restricted by self MHC class I alleles. Figures and respective footnotes are obtained and adapted from (<https://www.immunopaedia.org.za/immunology/basics/4-mhc-antigen-presentation/>), [534] and [535].



## **4. HLA CLASS II REGION**

The HLA class II cluster comprises the classical class II genes (*HLA-DP*, *-DQ*, *-DR* and respective pseudogenes) and the non-classical class II genes (*HLA-DM* and *-DO*) (<http://hla.alleles.org/genes/index.html>). This region is located at the centromeric end of the human MHC complex in chromosome 6 and spans approximately 0.9 Mbp of DNA [2].

### **4.1 HLA Class II Genes**

The classical class II genes (~10-17 kb of DNA sequence length per gen) are selectively expressed on the cell surface of antigen presenting cells (APCs), including primarily dendritic cells (DCs), B cells, macrophages and activated T cells but also, under IFN- $\gamma$  stimuli, by mesenchymal stromal cells, fibroblasts and endothelial cells, as well as by epithelial cells and enteric glial cells. These class II molecules form heterodimers consisting of corresponding  $\alpha$  (encoded by respective *HLA-DPA*, *-DQA*, *-DRA* genes) and  $\beta$  (encoded by respective *HLA-DPB*, *-DQB*, *-DRB* genes) chains that bind and display antigens to CD4<sup>+</sup> T cells. The non-classical class II genes (*HLA-DM* and *-DO*) are not expressed on the cell surface, but form heterotetrameric complexes involved in peptide exchange and loading onto classical class II molecules [2][50].

#### **4.1.1. Classical HLA Class II Genes (HLA-IIa)**

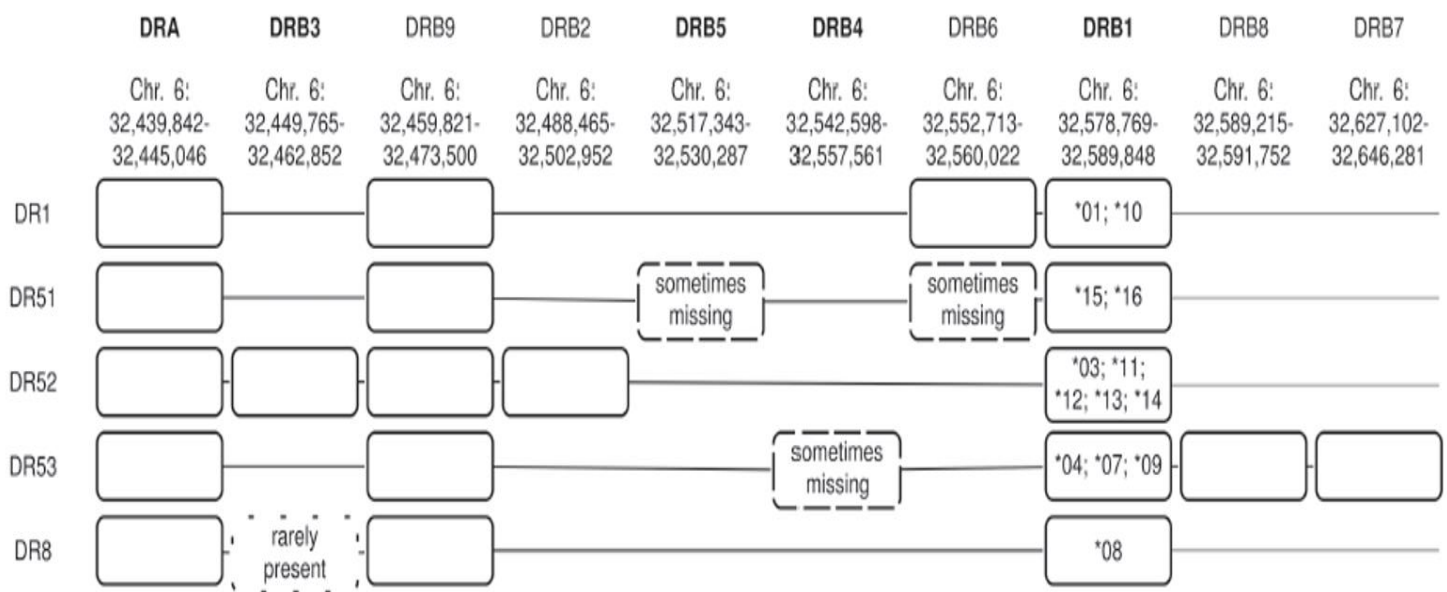
Within the *HLA-DP* region (located in the centromeric end of human MHC classical class II subregion), it has been described a total of five genes, three genes “A” (*HLA-DPA1*, *-DPA2* and *-DPA3*) and two genes “B” (*HLA-DPB1* and *-DPB2*). Respective encoded *HLA-DPA1* ( $\alpha$  chain) and *-DPB1* ( $\beta$  chain) molecules form functional heterodimers. In contrast, nucleotide sequence determination of the *HLA-DPA2*, *-DPA3* and *-DPB2* shows that these are pseudogenes which are not expressed [73].



Similarly, the *HLA-DQ* region contains two genes “A” (*HLA-DQA1* and *-DQA2*) and three genes “B” (*HLA-DQB1*, *-DQB2* and *-DQB3*), where only heterodimers DQA1-DQB1 are functional. Whereas *HLA-DQA2*, *-DQB2*, *-DQB3* could encode respective  $\alpha$  and  $\beta$  chains but are not known to be expressed [74].

The unique *HLA-DR* region (located in the telomeric end of human MHC classical class II subregion) is characterized by the presence of one invariant *DRA* gene, encoding the  $\alpha$  chain, and multiple *DRB* genes, encoding the  $\beta$  chain, on different haplotypes, which display, in addition, copy number variation (usually according to Andersson’s *HLA-DRB* haplotype rule) [344]. Although the *HLA-DRA* gene is highly conserved, nine different *HLA-DRB* genes have been described. *HLA-DRB1* (highly polymorphic), *-DRB3*, *-DRB4* and *-DRB5* (presenting lower polymorphism) encode functional gene products, whereas *-DRB2*, *-DRB6*, *-DRB7*, *-DRB8*, and *-DRB9* represent pseudogenes as manifested by various insertions/deletions (indels) and deleterious mutations. In humans, the non-polymorphic *HLA-DRA* gene is linked to a varying number of *HLA-DRB* genes defining different *HLA-DR* haplotypes. Based on this variation, five different *HLA-DR* haplotypes (denoted as *DR1*, *DR51*, *DR52*, *DR8* and *DR53*) have been identified, where all these described haplotypes present at least the *DRB1* and *DRB9* loci [56][75][344] (see **Figure I-8**). At the same time, *DRB1* sequences from the 13 allelic lineages (defined by phylogenetic analyses) cluster within the five haplotypic groups from where they are encoded. The allelic lineages are denoted: \*01 and \*10 (the *DR1* group), \*08 (the *DR8* group), \*15 and \*16 (the *DR51* group), \*03, \*11, \*12, \*13, and \*14 (the *DR52* group), and \*04, \*07, and \*09 (the *DR53* group). Thus, *DRB1* allelic lineages can be further grouped into five families which correspond to the five main haplotypic groups. Moreover, linkage constraints between the *DRB3/4/5* loci and the *DRB1* locus are based on these *DRB1* allele families. Where *DRB3*, *DRB4* and *DRB5* loci are exclusive to the *DR51*, *DR52* and *DR53* haplotypes, respectively [56][75]. In

closer detail, and according to literature of reported human population studies [75], alleles of the *HLA-DRB3/4/5* loci occur within a specific *HLA-DRB1* context, being present in some haplotypes and absent in others. Haplotypes with *HLA-DRB1* always carry the pseudogene *HLA-DRB9*, which is located downstream of *HLA-DRB1* and that consists of two exons. *HLA-DRB1* \*01, -*DRB1* \*08 and -*DRB1* \*10 are not found with any *HLA-DRB3/4/5* allele. Haplotypes with *HLA-DRB1* \*03, \*11, \*12, \*13 and \*14 are found with *HLA-DRB2* and -*DRB3*. *HLA-DRB1* \*04, \*07, \*09 are found with *HLA-DRB4* as well as -*DRB7* and -*DRB8*. Finally, *HLA-DRB1* \*15 and \*16 are reported to be located on the same haplotype as *HLA-DRB5*. Relatively infrequent exceptions to this rule have been described for *HLA-DRB1* \*15 and \*16, where especially in African Americans *HLA-DRB5/6* can be missing. *HLA-DRB1* \*08 has also been previously identified together with -*DRB3* \*03:01.



**Figure I-8.** Known architecture of *HLA-DRB3/4/5*: HLA haplotypes that usually contain a specific *HLA-DRB1* allele (*HLA-DRB1* column) are shown. 1-field alleles are denoted. All loci are depicted in order of their genomic location. *HLA-DRA*, *HLA-DRB1* and *HLA-DRB9* coincide with all haplotypes. The remaining loci are present or absent depending on the haplotype. The most prevalent haplotypes with the known exceptions are shown in the rows below. Exceptions are sometimes seen for *HLA-DRB1* \*08, -*DRB1* \*07, -*DRB1* \*15 and -*DRB1* \*16. *HLA-DRB1* \*08 can occur with *HLA-DRB3*; *HLA-DRB1* \*07 can occur without an expressed form of *HLA-DRB4* and *HLA-DRB1* \*15 and -*DRB1* \*16 can occur without *HLA-DRB5/6*. Loci that usually occur together are joined by a line. The name of the corresponding serotype is shown on the left and haplotypes are ordered by serotype name. Information for this figure was retrieved from [177] and, in turn, originally from Robbins et al., Holdsworth et al. and Bontrop et al. [75][86][536].

#### 4.1.2. Non-Classical HLA Class II Genes (HLA-IIb)

Genes in the DO (*DOA* and *DOB* genes)(~5 kb and ~8 kb of length per gen respectively) and DM (*DMA* and *DMB* genes)(~20 kb and ~6 kb of length per gen respectively) regions of the HLA complex encode polypeptides that closely resemble classical HLA class II  $\alpha$  and  $\beta$  chains. HLA-DM and HLA-DO molecules are non-peptide binding class II MHC-II homologs, which function to edit the peptides presented by classical MHC class II molecules. These non-polymorphic non-classical class II molecules HLA-DM and HLA-DO are localized not on the cell surface but within the endosomal compartment where classical MHC class II molecules are loaded with peptides. HLA-DM appears to facilitate and regulate peptide loading onto classical MHC class II molecules (thereby shaping MHC-II immunopeptidomes). Whereas HLA-DO has recently been found to associate HLA-DM with in the endosomal compartment of B cells, specifically, and to negatively regulate HLA-DM activity [35][50].

#### 4.2 Genetic Organization of HLA Class II Genes

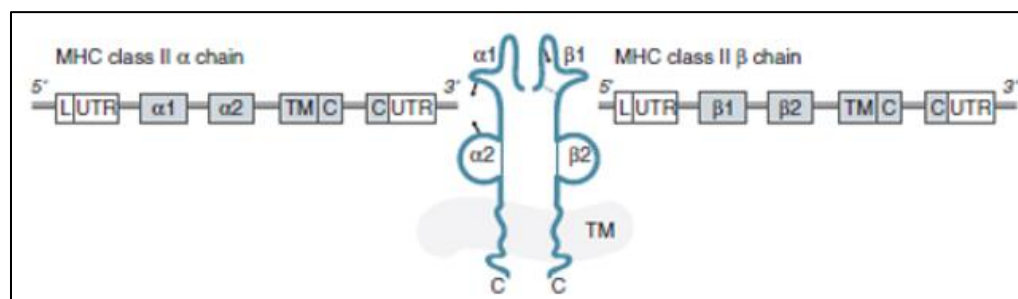
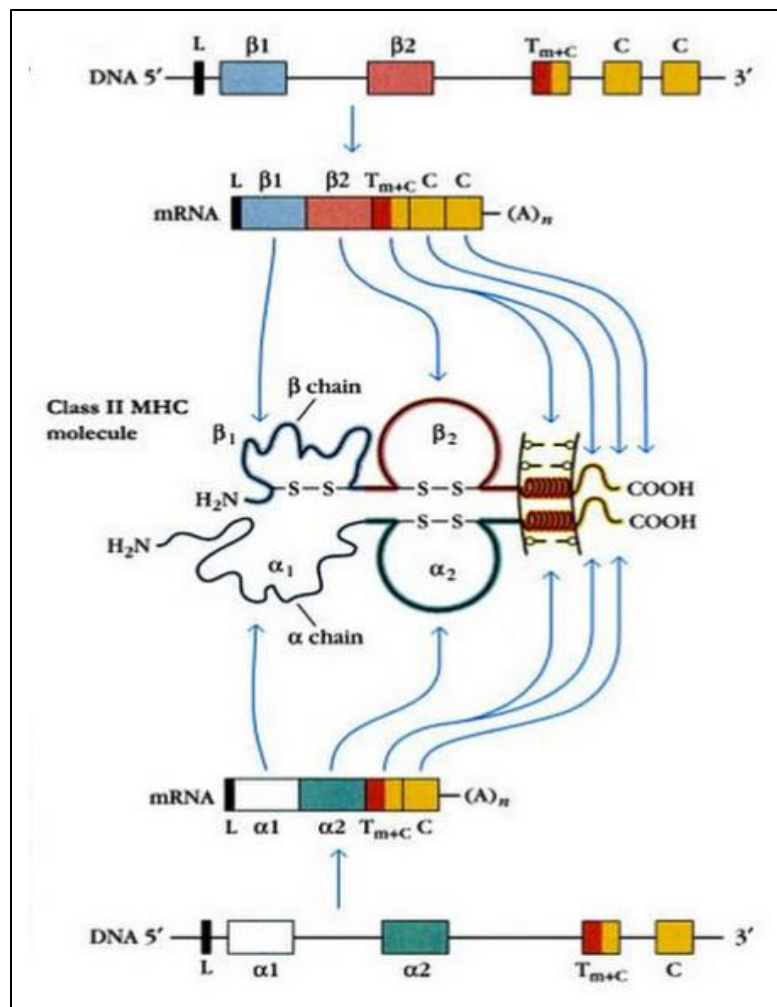
In relation to functional HLA class II heterodimeric molecules, classical HLA class II “A” (*DPA1*, *DQAI*, *DRA*) and “B” genes (*DPB1*, *DQB1* and *DRB1/3/4/5*) encode  $\alpha$  and  $\beta$  chains, respectively (see **Figure I-9**). Thus, the HLA class II molecule is made up of  $\alpha$  and  $\beta$  protein transmembrane subunits, encoded by these separate genes within the human MHC complex, which are non-covalently associated post-translationally.

Each classical HLA class II “A” (*DPA1*, *DQAI*, *DRA*) gene is composed of a series of five exons delineated by intervening four introns in addition to the untranslated regions located at the 5’UTR and 3’UTR ends. Exon 1 encodes the leader peptide sequence; exons 2 and 3 encode, respectively,  $\alpha 1$  and  $\alpha 2$  domains constituting the extracellular portion; while exon 4 encodes a

short transmembrane region and the cytoplasmic domain; finally, exon 5 encodes part of the 3'Untranslated (UTR) region.

On the other hand, each classical HLA class II “B” (*DPB1*, *DQB1*, *DRB1/3/4/5*) gene is composed of a series of six to seven exons delineated by intervening five to six introns. Where exon 1 encodes the leader peptide sequence; exons 2 and 3 encode, respectively,  $\beta 1$  and  $\beta 2$  domains constituting the extracellular portion; while exon 4 encodes a short transmembrane region and part of the cytoplasmic domain; finally, exons 5 to 7 encode the rest of the cytoplasmic domain and the 3'UTR region depending on the HLA class II molecule [76]. For instance, and as an exception, the  $\beta$  chain of the HLA-DQ molecule (being specific for certain *HLA-DQB1* alleles only) is shorter by eight amino acid residues than other major histocompatibility complex class II  $\beta$  chains due to elimination of the fifth exon coding for part of the cytoplasmic domain. This elimination is caused by one base substitution in the splice acceptor site of the exon [77].

Although,  $\alpha 1$  and  $\beta 1$  domains form the peptide-binding groove. Polymorphisms of HLA class II molecules occur mainly in the first amino terminal  $\beta 1$  domain (encoded on exon 2) of *HLA-DRB1/3/4/5*, *DQB1*, and *DPB1* gene products. Whereas  $\alpha 1$ ,  $\alpha 2$  and  $\beta 2$  have limited polymorphisms, with the single exception of the  $\alpha 1$  domain of HLA-DR molecule (encoded on *HLA-DRA*) which is not polymorphic [56].



**Figure I-9. (Upper Image)** Schematic diagram of *HLA* (or *MHC*) class II gene “A” or  $\alpha$  (bottom) and “B” or  $\beta$  (top) pair, respective messenger RNA (mRNA) transcripts (after transcription) and respective assembled protein chain ( $\alpha$  and  $\beta$ ) molecules (after translation and maturation). There is a correspondence between exons (represented in colored boxes) and the domains of the *MHC* class II chain molecules in the gene products, respectively:

*MHC* class II gene “A” or  $\alpha$ : exon 1 encodes the leader peptide sequence; exons 2 and 3 encode, respectively,  $\alpha 1$  and  $\alpha 2$  domains constituting the extracellular portion (in the N-terminal side of the protein chain); while exon 4 encodes a short transmembrane region and the cytoplasmic C-terminal domain; finally, exon 5 encodes part of the 3’Untranslated (UTR) region.

MHC class II gene “B” or  $\beta$ : exon 1 encodes the leader peptide sequence; exons 2 and 3 encode, respectively,  $\beta 1$  and  $\beta 2$  domains constituting the extracellular portion; while exon 4 encodes a short transmembrane region and part of the cytoplasmic domain; finally, exons 5 to 7 encode the rest of the cytoplasmic domain and the 3'UTR region depending on the HLA class II molecule.

Here again, with the exception of the leader (L) exon (encoding leader peptide, which is removed in a post-translational reaction after leading the pre-mature peptide to the endoplasmic reticulum (ER) and before the MHC class II molecule is expressed on the cell surface. Note that (during the process of transcription) the mRNA transcript is spliced to remove introns sequences (represented as black thin lines between exons).

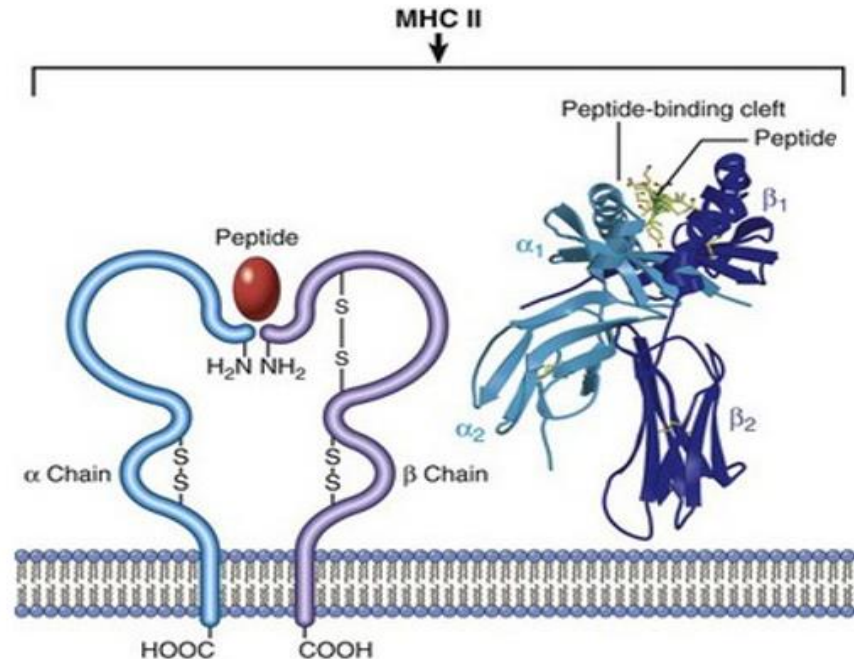
**(Lower Image)** Schematic diagram of *HLA* (or *MHC*) class II gene similar to previous upper one, showing again correspondence between exons (represented in colored boxes) and the domains (L, leader sequence; UTR, untranslated region; TM, transmembrane region; C, cytoplasmic region) of the MHC class II molecule. Original figures and respective footnotes are obtained and adapted from [532] and [533].

### 4.3 Structure of HLA Class II Molecules

Classical HLA class II molecules are heterodimers that consist of two transmembrane glycoprotein  $\alpha$  (33-35 kDa) and  $\beta$  (26-29 kDa) chains that are non-covalently associated post-translationally (see **Figure I-10**). Each chain has four regions: the peptide-binding region (formed by  $\alpha 1$  and  $\beta 1$  domains); the immunoglobulin-like region proximal to the membrane ( $\alpha 2$  and  $\beta 2$  domains), the transmembrane region and the C-terminal cytoplasmic tail. At the N-terminus, the extracellular  $\alpha 1$  and  $\beta 1$  domains of class II molecules form the peptide-binding groove or cleft. Unlike HLA class I, the class II binding cleft is open-ended and thus can bind longer peptides ranging in length from 12 to 24 amino acids (aa), but longer ones are not uncommon. In relation to its structure, four  $\beta$  strands of the floor of the cleft and one of the  $\alpha$ -helical walls are formed by the  $\alpha 1$  segment, and the other four  $\beta$  strands of the floor and the second wall are formed by the  $\beta 1$  segment. HLA class II molecules bind their peptides in an extended conformation with about a third of the peptide surface being accessible for interaction with the TCR. The termini of class II-bound peptides are not ligated by the same network of H-bonds that bind class I peptides so they may hang over the end of the cleft. Similar to HLA class I molecules, the docking of a peptide with the peptide-binding cleft of MHC class II protein is facilitated by peptide-binding pockets, which include polymorphic residues mostly located

in the first amino terminal  $\beta$ 1 domain. In this sense, several studies have described peptide binding to class II MHC protein as a combination of discrete anchor residue preferences for pockets 1, 4, 6, 7 and 9, where both negative and positive cooperative effects between both pocket and solvent exposed residues are involved during the peptide binding process. The  $\alpha$ 2 and  $\beta$ 2 segments of class II molecules, like class I  $\alpha$ 3 and  $\beta$ 2-microglobulin, are folded into Ig domains. Both the  $\alpha$ 2 and  $\beta$ 2 domains of class II molecules, proximal to the membrane, contribute to a concavity which accommodates a protrusion from the CD4 co-receptor in the T cell (where CD4 solely uses its N-terminal domain for insertion into a hydrophobic cave-like structure formed by the two membrane-proximal domains of the MHC class II molecule). The C-terminal ends of the  $\alpha$ 2 and  $\beta$ 2 segments continue into short connecting regions followed by approximately 25-amino acid stretches of hydrophobic transmembrane residues. In both chains, the transmembrane regions end with clusters of basic amino acid residues, followed by short hydrophilic cytoplasmic tails [63][78-80].

The structure and sequence of HLA-DM proteins is very similar to other MHC class II molecules. However, HLA-DM differs in that it lacks a transport signal N-terminus and does not have the capability to bind peptides. This is due to lack of a deep peptide binding groove; instead, it contains a shallow, negatively charged indent with two disulfide bonds. On its  $\beta$  chain cytoplasmic tail, a tyrosine based motif YTPL regulates trafficking to specific endosomal compartments called MHC class II compartments (MIICs) from the ER. In complex with HLA-DM, HLA-DO adopts a classical HLA class II structure, with alterations in the N-terminus. The structure of the free HLA-DO protein, however, remains to be elucidated [81].



**Figure I-10.** Schematic diagram and crystal structure of MHC class II molecule. Figure and respective footnote are obtained and adapted from (<https://veteriankey.com/diseases-of-immunity/>) being originally from Dr. P. Bjorkman (California Institute of Technology, Pasadena, CA, USA).

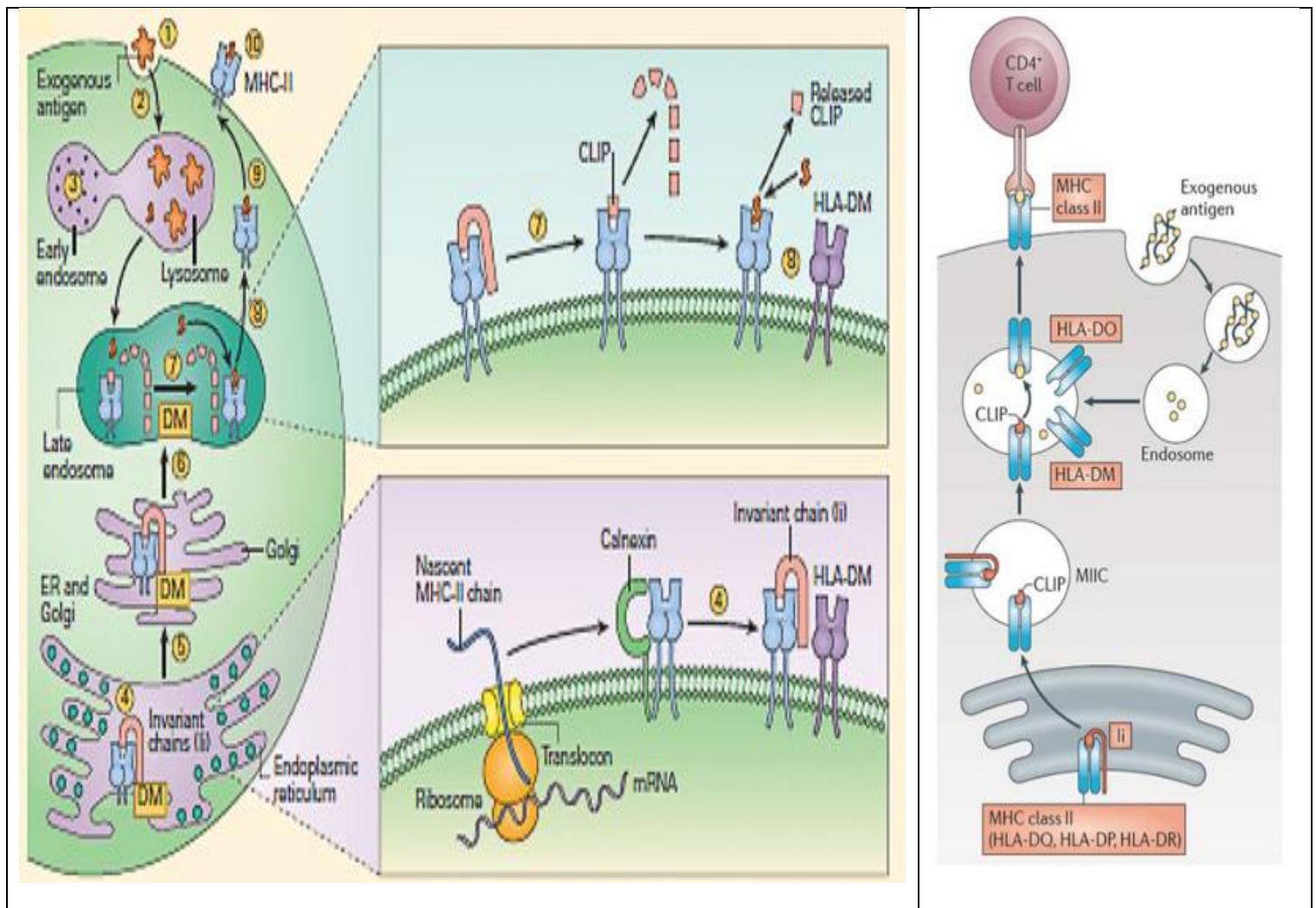
#### 4.4 Biological Function of HLA Class II Molecules

Human MHC class II molecules are expressed by APCs, including dendritic cells (DCs), macrophages and B cells. In addition, under IFN- $\gamma$  stimuli, they are also expressed by mesenchymal stromal cells, fibroblasts and endothelial cells, as well as by epithelial cells and enteric glial cells. Regarding antigen processing (see **Figure I-11**) in the context of HLA class II. APCs use specialized receptors to bind and internalize exogenous antigens (derived from extracellular bacteria or parasites) in vesicles called phagosomes which may fuse with lysosomes to produce phagolysosomes or secondary lysosomes. Additionally, though this occurs less often, cytoplasmic and membrane proteins may be processed and displayed by HLA class II molecules. In this case, cytoplasmic proteins are trapped within membrane bound vesicles called autophagosomes. These vesicles fuse with lysosomes, and the cytoplasmic proteins are degraded



by proteolysis. In both cases, degraded proteins are then able to bind to HLA class II molecules in the ER. Thus, HLA class II molecules bind to peptides that are derived from proteins degraded in the endocytic pathway. Regarding peptide-binding in the context of HLA class II. Firstly, HLA class II  $\alpha$  and  $\beta$  chains assemble in the ER with a non-polymorphic protein called invariant chain (Ii). The interaction with the Ii has the effect of stabilizing the structure of the HLA class II molecule while preventing the binding of peptides within the peptide-binding groove. Ii protein is anchored in the ER membrane, and the cytosolic portion of the molecule directs intracellular sorting of class II molecules through the Golgi to the late endosomal HLA class II compartment (MIIC). Within the MIIC, proteolytic enzymes such as cathepsins S and L generate peptides from internalized proteins and also act on the Ii to degrade it and leave only a 24 amino acid remnant called class II-associated invariant peptide (CLIP), which sits in the peptide-binding groove. Later, the CLIP is exchanged for an antigenic peptide derived from a protein degraded in the endosomal pathway. This process requires the chaperone HLA-DM, and, in the case of B cells, the HLA-DO molecule. HLA-DO binds to HLA-DM at the same sites implicated in HLA class II interaction, and kinetic analysis demonstrates that HLA-DO acts as a competitive inhibitor. After DM-editing of the highest-affinity peptides, a cohort of HLA class II molecules that has stable peptides bound is exported to the cell surface for presentation to CD4<sup>+</sup> T cells [82]. Thus, class II-bearing cells can activate CD4<sup>+</sup> T cells. CD4<sup>+</sup> T cells after being activated and differentiated into distinct effector subtypes play a major role in mediating immune response through the secretion of specific cytokines. The CD4<sup>+</sup>T cells carry out multiple functions, ranging from activation of the cells (e.g. macrophages) of the innate immune system, B-lymphocytes, cytotoxic CD8<sup>+</sup> T cells, as well as nonimmune cells, and also play critical role in the suppression of immune reaction (see **Figure I-11**). [83]. In addition, various HLA class II-bearing thymic APCs subsets are strategically

positioned in particular microenvironments of the thymus and orchestrate the selection of a functional, self-restricted and self-tolerant CD4<sup>+</sup> T cell repertoire [84].



**Figure I-11. (On the Left Image)** Peptide loading of MHC-II molecules. Panel shows the uptake of protein and peptide loading of MHC-II through the exogenous pathway. Exogenous proteins are taken up (1) and processed in the early endosomal compartment (2) and cleaved into peptides by cathepsins and other acid proteases (3). MHC-II molecules are formed in the endoplasmic reticulum (ER) with the help of the chaperone calnexin (4) and are held ready by the invariant chain (Ii); the complex is later fused with the HLA-DM (DM) (right lower inset in panel B). After passage of the Ii-loaded MHC-II-DM complex through the Golgi (5) into the late endosomes (6), the invariant chain is cleaved by acid proteases, leaving a residual peptide referred to as the class II-associated invariant chain peptide (CLIP) (7) in the MHC-II cleft (right upper inset in panel B). The HLA-DM facilitates the insertion of the peptide in the MHC-II cleft replacing CLIP (8). The MHC molecule loaded with peptide is transported (9) and expressed on the cell surface (10). **(On the Right Image)** It is shown a simplified version of peptide loading of MHC-II molecules and, once expressed on the cell surface, the final recognition of antigens by CD4<sup>+</sup> T cells restricted by self MHC class II alleles. Figures and respective footnotes are obtained and adapted from (<https://www.immunopaedia.org.za/immunology/basics/4-mhc-antigen-presentation/>), [534] and [535].

## **5. IMPORTANT ASPECTS OF THE HLA SYSTEM**

Polygeny, codominant expression (including a high heterozygosity) and extensive polymorphism are the main intrinsic features that contribute to the vast diversity presented by the HLA system and, ultimately, maximizing the peptide repertoire presented to T cells both in a given individual and in the population as a whole. Because of the polygeny of the HLA system (as it contains several different HLA class I and MHC class II genes), every individual possesses a set of HLA molecules with different ranges of peptide-binding specificities on the cell surface. HLA alleles are also expressed in codominant fashion from both HLA haplotypes in an individual, with the protein products of both the alleles at a locus being expressed in the cell, and both gene products being able to present antigens to T cells. Moreover, extensive polymorphism at each locus thus has the potential to increase the number of different HLA molecules expressed in an individual and thereby rises the diversity already available through polygeny. Particularly, classical HLA class I and class II genes exhibit a high degree of polymorphism, where a number of different mechanisms may contribute to the generation and maintenance of this polymorphism. It is believed among these are the selective advantage of a heterozygous pool of antigen-presenting elements in a given individual that might allow the binding and presentation of antigenic peptides derived from a wide variety of environmental pathogens [27].

Another unique characteristic of the HLA system is the extensive linkage disequilibrium (LD) observed among the very distant genomic regions of *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DR*, and *HLA-DQ* genes, but not the *HLA-DP* genes. LD is the phenomenon whereby particular alleles of gene loci on the same strand of DNA are inherited together more often than would be expected by chance. Anthropological population studies have suggested that the particular combinations of alleles of the different genes, as distant as they may be, provide a survival advantage, perhaps reflecting functional interdependence in antigen-specific immune responses [26].

## **5.1 Polymorphism**

The HLA system is one of the most highly polymorphic regions known to date of the human genome, where the majority of this polymorphism has been found at the classical HLA class I and class II loci [56].

Polymorphisms in these HLA loci were first defined phenotypically using serological typing, which was based on serologic and cell proliferation methodologies that use defined human alloantisera or monoclonal antibodies (mAbs) to identify HLA antigens (alloantigens) on the cell surface. Initial human alloantisera were able to identify an unprecedented number of antigenic determinants or HLA antigens at different HLA loci with unique sequence motifs or epitopes defining private specificities. As the field evolved, new antisera were discovered that could “split” some HLA antigens into narrower specificities. At the same time, it was observed how some antigenic determinants are shared by many HLA antigens. Consequently, these public specificities lead to antigenic cross-reactivity. Due to epitope sharing, those HLA antigens were also arranged in cross-reactive groups or CREGs. Overall, these serologic and cell-based methodologies were successfully utilized to initially characterize the HLA system. However, in spite of their broad application, these methodologies presented important limitations in terms of reproducibility and accuracy. In fact, most of these allotypic epitopes recognized by alloantisera were found to be three-dimensional conformational determinants commonly located on the most external part of the HLA molecules (mainly in the helices of the  $\alpha 1$  and  $\alpha 2$  domains as well as external portions of the  $\beta$ -pleated sheet in class I molecules; and  $\alpha 1$  and  $\beta 1$  domains in class II molecules). Whereas other polymorphic amino acid residues on the HLA molecules (e.g. all allelic variation located deep within the cleft) were not accessible to alloantibodies and, thus, remained undetectable [85]. To date (January 2020), a variety of serological and cellular HLA specificities have been described for classical HLA genes: 28

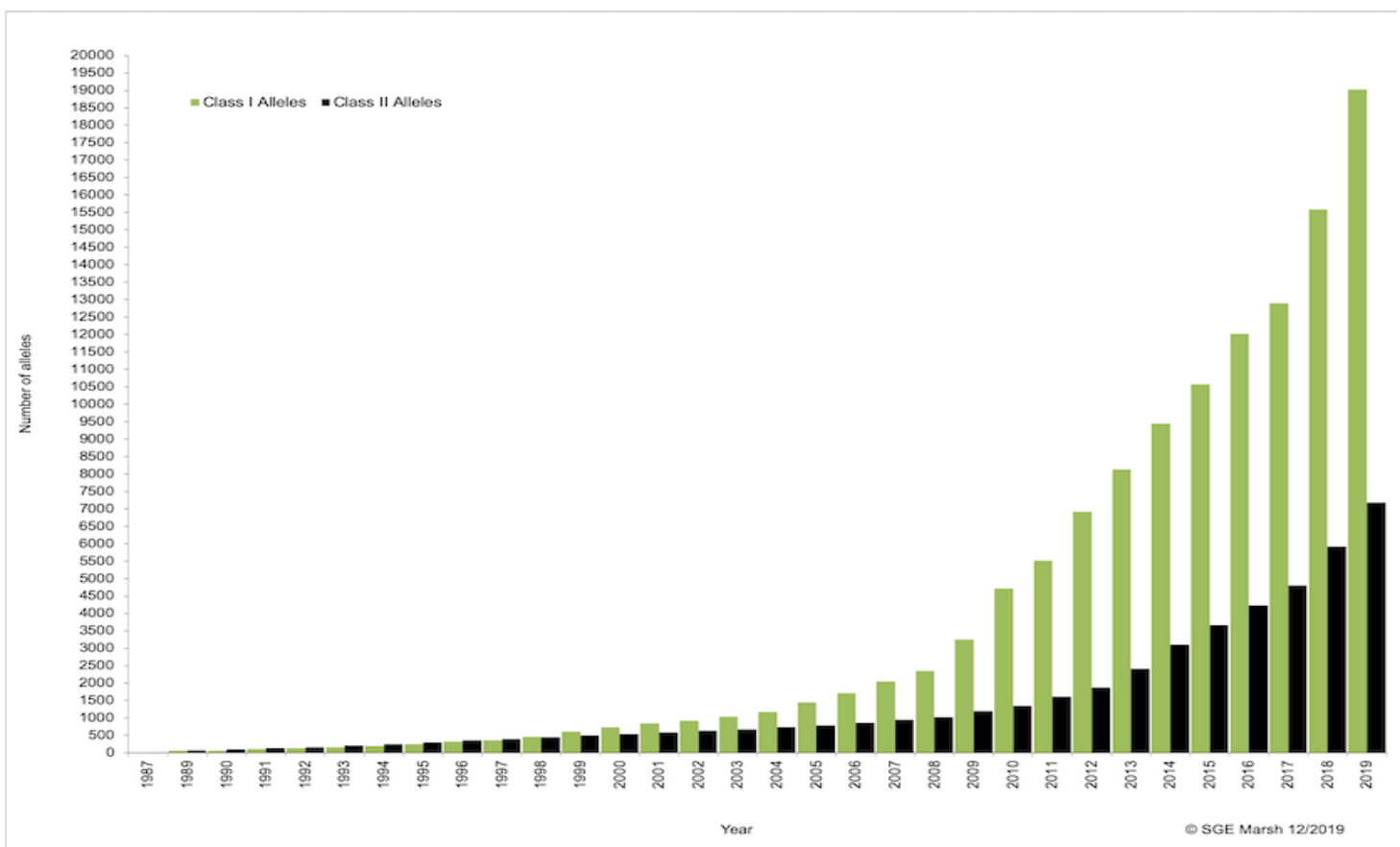
different specificities for *HLA-A* locus; 62 for *HLA-B* locus; 10 for *HLA-C* locus; 24 for *HLA-DR* genes; 9 for *HLA-DQ* genes; and 6 for *HLA-DP* genes [74].

The application of polymerase chain reaction (PCR)-based HLA genotyping methods to histocompatibility testing made apparent the fact that the extent of HLA protein phenotypic polymorphism greatly underestimates the true degree of HLA polymorphism found at the genomic level. Furthermore, detailed analyses of HLA genes lead to the determination that serologically defined antigens included multiple allelic variants that differed at one or more nucleotide residues. Thus, several alleles could encode proteins, all recognized as a single serologic specificity [86]. In the last decades, wide application of molecular methods has allowed the characterization of a vast number of alleles in all HLA class I and II loci. Up to date, almost 26,214 HLA alleles have been described and registered in the IPD-IMGT/HLA Database (Release 3.39.0 January 2020), with some genes currently having over 5,000 known allelic variants (see **Figure I-12 and Figure I-13**)[87].

| Numbers of HLA Alleles         |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
|--------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|------------|------------|------------|--------|
| HLA Class I Alleles            |             |             |             |             |             |             |             |             |             |            |            |            |            | 19,031 |
| HLA Class II Alleles           |             |             |             |             |             |             |             |             |             |            |            |            |            | 7,183  |
| HLA Alleles                    |             |             |             |             |             |             |             |             |             |            |            |            |            | 26,214 |
| Other non-HLA Alleles          |             |             |             |             |             |             |             |             |             |            |            |            |            | 298    |
| Number of Confidential Alleles |             |             |             |             |             |             |             |             |             |            |            |            |            | 6      |
| HLA Class I                    |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
| Gene                           | <i>A</i>    | <i>B</i>    | <i>C</i>    | <i>E</i>    | <i>F</i>    | <i>G</i>    |             |             |             |            |            |            |            |        |
| Alleles                        | 5,907       | 7,126       | 5,709       | 84          | 44          | 69          |             |             |             |            |            |            |            |        |
| Proteins                       | 3,720       | 4,604       | 3,470       | 15          | 6           | 19          |             |             |             |            |            |            |            |        |
| Nulls                          | 308         | 244         | 243         | 1           | 0           | 3           |             |             |             |            |            |            |            |        |
| HLA Class I - Pseudogenes      |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
| Gene                           | <i>H</i>    | <i>J</i>    | <i>K</i>    | <i>L</i>    | <i>N</i>    | <i>P</i>    | <i>S</i>    | <i>T</i>    | <i>U</i>    | <i>V</i>   | <i>W</i>   | <i>Y</i>   |            |        |
| Alleles                        | 25          | 9           | 6           | 5           | 5           | 5           | 7           | 8           | 5           | 3          | 11         | 3          |            |        |
| Proteins                       | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0          | 0          | 0          |            |        |
| Nulls                          | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0           | 0          | 0          | 0          |            |        |
| HLA Class II                   |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
| Gene                           | <i>DRA</i>  | <i>DRB</i>  | <i>DQA1</i> | <i>DQA2</i> | <i>DQB1</i> | <i>DPA1</i> | <i>DPA2</i> | <i>DPB1</i> | <i>DPB2</i> | <i>DMA</i> | <i>DMB</i> | <i>DOA</i> | <i>DOB</i> |        |
| Alleles                        | 29          | 3,331       | 229         | 38          | 1,795       | 168         | 5           | 1,537       | 6           | 7          | 13         | 12         | 13         |        |
| Proteins                       | 2           | 2,357       | 98          | 11          | 1,194       | 65          | 0           | 1,006       | 0           | 4          | 7          | 3          | 5          |        |
| Nulls                          | 0           | 141         | 6           | 0           | 77          | 3           | 0           | 80          | 0           | 0          | 0          | 1          | 0          |        |
| HLA Class II - DRB Alleles     |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
| Gene                           | <i>DRB1</i> | <i>DRB2</i> | <i>DRB3</i> | <i>DRB4</i> | <i>DRB5</i> | <i>DRB6</i> | <i>DRB7</i> | <i>DRB8</i> | <i>DRB9</i> |            |            |            |            |        |
| Alleles                        | 2,690       | 1           | 340         | 165         | 123         | 3           | 2           | 1           | 6           |            |            |            |            |        |
| Proteins                       | 1,889       | 0           | 254         | 109         | 95          | 0           | 0           | 0           | 0           |            |            |            |            |        |
| Nulls                          | 89          | 0           | 16          | 19          | 17          | 0           | 0           | 0           | 0           |            |            |            |            |        |
| Other non-HLA Genes            |             |             |             |             |             |             |             |             |             |            |            |            |            |        |
| Gene                           | <i>HFE</i>  | <i>MICA</i> | <i>MICB</i> | <i>TAP1</i> | <i>TAP2</i> |             |             |             |             |            |            |            |            |        |
| Alleles                        | 6           | 159         | 109         | 12          | 12          |             |             |             |             |            |            |            |            |        |
| Proteins                       | 4           | 92          | 35          | 6           | 5           |             |             |             |             |            |            |            |            |        |
| Nulls                          | 0           | 4           | 2           | 1           | 0           |             |             |             |             |            |            |            |            |        |

**Figure I-12** Summary Table of IPD-IMGT/HLA database v.3.39.0 (Release in January, 2020) [295]. The numbers above represent the number of named alleles for each gene. This number includes alleles which have been named but whose sequences are still held as confidential. This means that the number of sequences found in some files may differ to the numbers printed in this table. Details of which alleles are still confidential can be seen in the latest version report. The numbers below do not include the names of any deleted alleles - details of deleted alleles can be found in a separate report (<https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/deleted.cgi>). The

main HLA Reference Sequence Repository is a centrally curated repository for the sequence data of the hyperpolymorphic genes of the HLA system and it is provided by the IPD-IMGT/HLA database [87][295][362]. Every 3 months, it releases a fresh snapshot of all publicly available sequences of the HLA system, officially named by the WHO Nomenclature Committee for Factors of the HLA System [74]. The IPD-IMGT/HLA Database is updated every three months, and the number of named HLA gene and pseudogene sequences increases with each update. As such it provides the canonical reference sequences against which HLA genotyping is performed. For historical reasons the IPD-IMGT/HLA database is populated to a large extent by alleles where only certain gene features have been characterized. For many entries (~80-90%) in the database, only the antigen recognition domains (ARD) have been characterized, which are encoded by exons 2 and 3 for the class I genes and exon 2 for class II genes. Figure and respective footnote are obtained and adapted from [35] and [295].



**Figure I-13.** Gene map Graph showing the number of HLA alleles (green colored bars corresponding to HLA class I, and black colored bars corresponding to HLA class II alleles respectively) reported and officially named by year from 1987 to the end of September 2019 (<http://hla.alleles.org/nomenclature/index.html>). Figure and respective footnote are obtained and adapted from [295].

A general feature of the highly polymorphic classical HLA class I and II genes is that the distal membrane domains present a high degree of variability, whereas the proximal membrane domains as

well as the transmembrane and cytoplasmic domains have limited or no polymorphism within each locus. The differences among classical HLA proteins are localized primarily to the amino-terminal region of these molecules, which bind peptides and interact with T cell receptor molecules. Class I polymorphisms are predominantly found in the first 180 amino acids of the heavy chain, at the extracellular  $\alpha 1$  and  $\alpha 2$  domains of the protein that comprise the peptide-binding groove. While class II polymorphisms are found in the first 90 to 95 amino acids of the  $\alpha$  and/or  $\beta$  chains [56]. In detail, for the  $\beta$  chain genes (*HLA-DPB1*, *HLA-DQB1*, *HLA-DRB1/3/4/5*), polymorphisms are concentrated within the  $\beta 1$  domain of the molecule. Whereas polymorphism within the  $\alpha$  chain genes varies by locus, with the *HLA-DRA* gene being nonpolymorphic and, on the other hand, the *HLA-DQA1* and *HLA-DPA1* molecules show polymorphic residues at the  $\alpha 1$  domain. In addition, unlike class I, where the peptide-binding domain is encoded by  $\alpha 1$  and  $\alpha 2$  domains in the same gene, cis- and trans-arrangement and configuration of  $\alpha$  and  $\beta$  chains derived from the two different haplotypes of the same or even different isotypes permit combinatorial polymorphism in class II [88].

Analysis of the nucleotide sequences indicates that exons or segments of the gene that encode these polymorphic residues are exons 2 and 3 for HLA class I, and exon 2 for HLA class II genes (encoding both  $\alpha$  and  $\beta$  chains). Since most HLA polymorphisms are located in positions that interact with antigenic peptides or the T cell receptor, it is widely believed the high degree of polymorphism likely has been positively selected for during evolution to promote diversity in the repertoire of peptides that can be presented by HLA class I and class II molecules [89]. Therefore, it is widely thought that HLA polymorphisms provide a major evolutionary survival benefit, since they equip the species with a large number of very specific, but alternative, HLA molecules that differ in their binding pockets, are most efficient in presenting different peptides, and selecting different T cell repertoires.

Many studies have also pointed out how the pattern of allelic sequence diversity for both the class I and class II loci is unusual. Most alleles differ from their closest neighbor by multiple substitutions,



with some alleles differing in the second and third exons by as much as 15 %. This pattern is suggestive of segmental exchange of nucleotide motifs between alleles of the same locus. Thus, different HLA alleles of a locus are patchwork (i.e., mosaic) combinations of polymorphisms [90][91]. Moreover, the extensive allelic diversity at these HLA loci is thought to have been generated by intra- and intergenic recombination, by gene conversion and also by single-point mutation events [56][91]. In detail:

- Recombinant meiotic events can occur within human families. Chromosomal crossover is the bidirectional sequence exchange between homologous chromosomes (exchange of interhomologous arms) that results in recombinant chromosomes. Crossover events have been described between nearly all the neighboring HLA loci (*A-C*, *C-B*, *B-DR*, *DQ-DP*) except for between *HLA-DR* and *-DQ* [26][501]. Genetic recombination or crossing over in the HLA region is a relatively rare event, occurring for the most part no more than 1% per meiosis between *HLA-A* and *HLA-B* and between *HLA-B* and *HLA-DR*. Recombination also can occur between *HLA-A* and *HLA-C* and between *HLA-B* and *HLA-C* (0.6% and 0.2%, respectively). The frequency of recombination between *DQ* and *DP* is 0.74%. Recombination provides a means of generating novel haplotypes, which may eventually prove to be beneficial to a population against a recently introduced pathogen. This crossing over in the human MHC is thought to have played a role in generation of novel alleles at various HLA loci. It is also responsible for the diversity observed at the haplotype level, although the functional consequences of this activity are not clear [92].

- Gene conversion refers to the unidirectional transfer of a DNA segment from a ‘donor’ sequence to a highly homologous ‘acceptor’ sequence that can happen during meiosis but also as a mechanism for repair of double-strand breaks caused by DNA damage. Interallelic or intralocus gene conversion (conversions between alleles of same HLA locus) appears to be more common than intergenic or interlocus gene conversions (conversions between alleles of different HLA loci,

where most are the result of *HLA-B* and *HLA-C* recombination) in the contribution of generating the extensive allelic diversity [93].

- Point mutation can occur as substitution or insertion/deletion of a single nucleotide. A nucleotide substitution can lead to synonymous (without amino acid altering) or non-synonymous (change of amino acid coding) nucleotide exchange. Insertion/deletion of one nucleotide or more (other than three or the multiple of three) nucleotides often causes a frame shift with subsequent generation of a premature stop codon [56].

## **5.2 Nomenclature**

The highly polymorphic nature of the HLA system has required an enormous effort from the international histocompatibility scientific community during these last decades in order to establish a systematic and common uniform nomenclature that could define the complexity of HLA genes and their products [515].

Initially, a standard nomenclature for expressing serologically defined antigens was established by the World Health Organization (WHO) Nomenclature Committee for Factors of the HLA System [74]. The WHO Nomenclature Committee for Factors of the HLA system undertook the first systematic approach for the naming of HLA alleles in 1968. Currently, for serological and cellular HLA specificities, HLA refers to the entire genetic region whereas A, B, C, DR, DQ, and DP each refer to a particular locus. A small "w" is included in HLA-C allele nomenclature. This connotation was originally a designation of alleles in workshop status and it is retained to distinguish it from the C designation of the complement genes [94]. Another feature of the nomenclature at the serological level is the Bw4 and Bw6 specificities. These latter constitute public epitopes, where HLA-B antigens (as well as some HLA-A (23, 24, 25, 32) and HLA-C (1, 3, 7, 8, 9, 10, 12, 14, 16:01) antigens that also bear the Bw4 and Bw6 epitopes, respectively

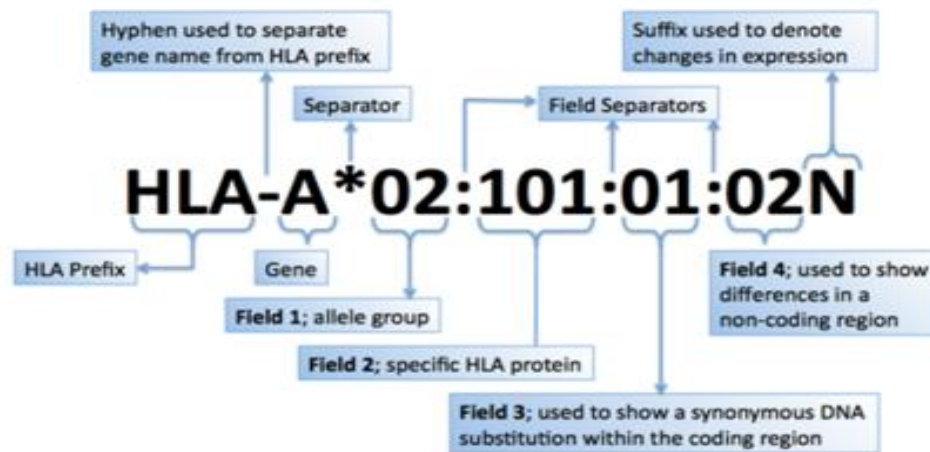
[537][538]) present one of these two possible epitopes located on the  $\alpha 1$  helix. The “w” prefix is retained in this case to distinguish them from true HLA alleles. In addition to the WHO Nomenclature Committee for Factors of the HLA system, the standardization of HLA antigenic specificities has been controlled by the exchange of typing reagents and cells in the International Histocompatibility Workshops and the UCLA International Cell Exchange Program [95].

With the introduction of molecular biology techniques in the 1980s, HLA typing at the DNA level resulted in a rapid increase in newly identified alleles and eventually created a need to further refine the HLA nomenclature. In 1998, the IPD-IMGT/HLA database (<http://www.ebi.ac.uk/imgt/hla>) became the official centrally curated public repository for HLA sequences data [87]. This HLA repository and database is the primary source of DNA sequences and protein sequences for all known HLA alleles. In addition to the physical sequences, this repository also contains analysis tools, data submission pipeline and a database with detailed information concerning the material from which the sequence was derived and data on the validation of the sequences. The HLA naming convention at the DNA level underwent a substantial number of iterations because the earlier naming conventions were unable to address the growing numbers and complexity of new alleles being discovered.

Finally, the most recent nomenclature was introduced in 2010 by the Harmonization of Histocompatibility Typing Terms Working Group to address these challenges and to reduce naming complexity and confusion [96]. The changes added colons (:) into the allele names to act as delimiters of the separate fields (field separator). Also an asterisk in an allele name indicates that the allele has been typed using molecular methods. Thus each HLA allele name has a unique number corresponding to up to four sets of digits separated by colons. The first field (1-field) following the asterisk in the allele name (XX:xx:xx:xx) describes the allele family and generally corresponds to the serological assignment carried by the allele (traditionally referred to as “low-

resolution typing”). The second field following the first colon (xx:XX:xx:xx) is assigned sequentially as new alleles are determined (e.g., 01, 02, 03....101, etc...). Together, these two fields (XX:XX) indicate one or more nucleotide substitutions that change the HLA protein coding sequence. HLA typing defined at the second field is often referred to as “2-field typing”, that distinguishes alleles based on the sequence of the peptide-binding region of the HLA molecule (and also, traditionally referred to as “high-resolution typing”). The third field (3-field) (xx:xx:XX:xx) is for designating synonymous nucleotide substitutions within the coding sequence that do not change the amino acids of the protein. Lastly, the fourth field (4-field) (xx:xx:xx:XX) identifies sequence polymorphisms in introns, or in the 5’ and 3’ untranslated regions and it is currently referred to as “4-field typing”. All alleles receive a name that includes at least the first two fields. At the end of the allele name, specific characters have been added to designate unique characteristics for an allele (N = null expression; L = low cell surface expression; S = secreted, expressed as a soluble; C = present in cytoplasm and not on the cell surface; A = aberrant expression, where there is some doubt as to whether a protein is actually expressed; Q = questionable expression, given that the mutation seen in the allele has been shown to affect normal expression levels in other alleles). For ambiguous allele strings, the codes “P” and “G” were introduced (and also defined by the HLA nomenclature system). A group of alleles having nucleotide sequences that encode the same protein sequence for the peptide binding domains (exon 2 and 3 for HLA class I and exon 2 only for HLA class II alleles) will be designated by an upper case “P,” which follows the two-field allele designation of the lowest-numbered allele in the group. For example, *HLA-A\*01:01:01:01*, *HLA-A\*01:01:01:03*, or *HLA-A\*01:37* could be named *HLA-A\*01:01P*. A group of alleles that have identical nucleotide sequences across the exons encoding the peptide-binding region (PBR) or antigen recognition domain (ARD) (exons 2 and 3 for HLA class I and exon 2 for HLA class II) were named after the first allele in the sequence and given the

code “G” as a suffix. The upper case “G” follows the first three fields of the allele designation. For example, *HLA-A\*01:01:01:01*, *HLA-A\*01:01:01:03*, or *HLA-A\*01:37* could be named *HLA-A\*01:01:01G* [35]. HLA nomenclature and convention of HLA allele naming are shown in **Figure I-14**.



| Nomenclature                | Indication  |
|-----------------------------|---|
| <i>HLA</i>                  | The HLA region and prefix for an HLA gene   |
| <i>HLA-DRB1</i>             | A particular HLA locus, i.e., DRB1  |
| <i>HLA-DRB1*13</i>          | A group of alleles that encode the DR13 antigen or sequence homology to other DRB1*13 alleles   |
| <i>HLA-DRB1*13:01</i>       | A specific HLA allele with a unique protein sequence  |
| <i>HLA-DRB1*13:01:02</i>    | An allele that differs by a synonymous mutation from DRB1*13:01:01  |
| <i>HLA-DRB1*13:01:01:02</i> | An allele that contains a mutation outside the coding region from DRB1*13:01:01:01  |
| <i>HLA-A*24:09N</i>         | A “null” allele, an allele that is not expressed  |
| <i>HLA-A*30:14L</i>         | An allele encoding a protein with significantly reduced or “low” cell surface expression  |
| <i>HLA-A*24:02:01:02L</i>   | An allele encoding a protein with significantly reduced or “low” cell surface expression, where the mutation is found outside the coding region   |
| <i>HLA-B*44:02:01:02S</i>   | An allele encoding a protein that is expressed as a “secreted” molecule only  |
| <i>HLA-A*32:11Q</i>         | An allele that has a mutation that has previously been shown to have a significant effect on cell surface expression, but where this has not been confirmed and its expression remains “questionable” |

**Figure I-14.** Nomenclature system of HLA alleles. Figure and respective footnote are obtained and adapted from (<http://hla.alleles.org/nomenclature/naming.html>) and [90]. The digits before the first colon describe the type, which often corresponds to the serological antigen carried by an allotype. The next set of digits are used to list the subtypes, numbers being assigned in the order in which DNA sequences have been determined. Alleles whose numbers differ in the two sets of digits must differ in one or more nucleotide substitutions that change the amino acid sequence of the encoded protein. Alleles that differ only by synonymous nucleotide substitutions (also called silent or non-coding substitutions) within the coding sequence are distinguished by the use of the third set of digits. Alleles that only differ by sequence polymorphisms in the introns, or in the 5' or 3' untranslated regions that flank the exons and introns, are distinguished by the use of the fourth set of digits.

### 5.3 Linkage Disequilibrium

Linkage disequilibrium is the phenomenon whereby particular alleles at adjacent and linked HLA loci on the same strand of DNA segregate together more often than would be expected by chance (based on random association of single locus frequencies). The HLA system is known to be one of the most dense, clustered and linked gene regions of the human genome [27]. In this context, human MHC presents an extensive linkage disequilibrium (LD) observed among the very distant genomic regions of *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DR*, and *HLA-DQ* genes, but not the *HLA-DP* genes. In detail, it is well described that LD is strongest between *HLA-B* and *HLA-C* (these two loci are situated within a 90-kb region at chromosome 6p21.33) and, also, between *HLA-DR* and *HLA-DQ* (*HLA-DRB3/4/5*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* genes within HLA class II region are located in a 150–210-kb region at chromosome 6p21.32), most likely because of their physical proximity [29]. However, there is no strong LD between *HLA-DP* and the rest of the class II haplotype because of existing hotspot of recombination between *DQ* and *DP* loci, even though these two loci are relatively proximal to each other (belonging to the same  $\delta$ -block) [92]. Also, in contrast to the tight LD between *HLA-DRB1*, *DQA1* and *DQB1* loci, the monomorphic *HLA-DRA* locus is separated from *HLA-DRB1* by a recombination hotspot region. It is widely believed that LD can be created by various evolutionary factors: selection (including disease) either directly on one or both loci, or indirectly via a hitchhiking event, migration and admixture, inbreeding, or genetic drift [97]. Furthermore, anthropological population studies have suggested that the particular combinations of alleles of the different genes, as distant as they may be, provide a survival advantage, perhaps reflecting functional interdependence in antigen-specific immune responses [26]. Different measures of LD strength have been established [97]. The basic definition of the LD parameter  $D_{ij}$  (also written as  $D_{A_iB_j}$ ) of nonrandom association between a pair of alleles  $A_i$  and  $B_j$  at two loci (A and B) is the difference between the observed (or estimated)

haplotype frequency ( $f_{ij} = f(A_iB_j)$ ) and that expected under random association of the two alleles (linkage equilibrium) [97]. In addition, there are several normalized measures of the strength of LD for 'bi-' and 'multi-allelic data' [98][99][780][787].

As mentioned previously, a particular combination of alleles of different loci in LD on the same strand of DNA is called an haplotype. This set of neighboring polymorphisms is co-transmitted or segregated on a single parental chromosome in the absence of recombination (in that case, crossing-over events occur between HLA loci). Population studies have extensively showed how the frequency of a given haplotype varies among different ethnic groups and geographical regions, reflecting distant effect of these aforementioned evolutionary factors on the haplotype frequency distributions and that are intrinsically related to these LD patterns [97]. In this sense, due to LD, the number of haplotypes observed in populations is much restricted than theoretically expected, indicating that segregation of these alleles at various loci is not random [100]. There are three main approaches to determine haplotypes [101]. The first approach includes several molecular methods that allow for the construction of the haplotypes in unrelated individuals, such as allele-specific polymerase chain reaction (AS-PCR) and somatic cell hybrids. A second approach relies on statistical methods, such as the expectation-maximization (EM) algorithm, for the inference of haplotype frequency distributions in unrelated individuals from large population HLA genotype datasets. The third method involves family based studies, which is considered the most efficient and robust current tool (i.e. being considered the gold standard for assessing HLA haplotype segregation patterns) to establish phase and determine haplotype segregation.

#### **5.4 Diversity and Evolution**

As mentioned previously, the vast diversity (at both the protein and DNA level) of the HLA system is believed to be intrinsically related to its function in the immune system. The majority of

HLA polymorphisms are concentrated in the domains that form a groove-like structure known as the peptide-binding region (PBR) of HLA class I and II molecules, which engages the peptides to be presented to T cells respectively [60]. At the DNA level, the PBR codons (exons 2 and 3 for classical HLA class I genes; and exon 2 for classical HLA class II genes) exhibit striking features regarding their diversity, including a high heterozygosity [56] and high rates of non-synonymous substitutions [102]. Thus, HLA variation often deviates significantly from neutral expectations towards an excess of genetic diversity [103]. Considering HLA diversity in human populations, although a very large number of alleles can be found in the global population, a much smaller number is present in most individual populations. Importantly, different populations tend to have different frequency distributions of alleles and exhibit different patterns of LD (thus, presenting also different frequency distributions of haplotypes). This variability has been widely reported among both different ethnicities and groups from different geographical regions [90]. Altogether, these extensively reported characteristics of HLA diversity have convincingly shown that classical HLA loci bear signatures of natural selection, where certain evolutionary mechanisms have been proposed [103].

The HLA system presents a complex evolution where not only deterministic (including positive natural selection, negative or purifying selection, neutral selection, and balancing selection (in the form of frequency-dependent selection and heterozygote advantage)) but also stochastic (e.g. genetic drift) forces appear to be involved [103][104]. Important research advances in knowledge on HLA function, HLA protein and genomic diversity and in theoretical population genetics have allowed evolutionary hypotheses to be proposed and tested [104]. Currently, there are several key ideas which are firmly established regarding HLA evolution:

- High degree of allele polymorphism, excess of nonsynonymous variants and higher heterozygosity than expected at classical HLA loci are hallmarks of balancing selection. As



balancing selection can enhance variation by maintaining alleles in a population for longer than expected under neutrality. Furthermore, balancing selection maintains a larger number of alleles than expected from genetic drift [105]. At the same time, balancing selection encompasses a broad range of selective regimes (heterozygote advantage or overdominance; variable selection over time and space; and negative frequency-dependent selection (or rare allele advantage), which favors maintaining less frequent alleles since pathogens are likely to evolve escape mutations to the most common alleles), all of which generate high levels of adaptive genetic variation [104]. Moreover, sequence data suggest that balancing selection in HLA is asymmetric as distinct heterozygote genotypes display greater fitness than others and, thus, a frequency-dependent selection with fluctuation in the fitness of specific heterozygotes may be operating over time [106].

-Host-pathogen co-evolution has also been proposed to lead to balancing selection that maintains high levels of HLA allelic diversity within populations (the model is known as the pathogen-driven balancing selection or PDBS). There are several lines of support for a role of pathogen-driven selection in shaping HLA variation: HLA genes are associated with susceptibility (or risk, or predisposing) and resistance (or protection) to infectious diseases [107][513]; experimental studies show that pathogen pressure influences MHC variability [108]; and HLA polymorphism is correlated with pathogen diversity [89]. Therefore, it is assumed that HLA genotypes with more divergent alleles allow for broader antigen-presentation to immune effector cells, by that increasing immunocompetence and, then, being a critical selective advantage against pathogens. Furthermore, in addition to the generation of certain promiscuous HLA alleles capable of binding an exceptionally large set of epitope peptide segments, incessant arise of new HLA alleles in human populations with specific peptide binding repertoires may be importantly contributing to cope with the also substantial

emergence of novel pathogens [502-504]. To further comprehend this complex relationship of the genetic diversity found in the HLA system and in pathogens, studies based on ancient DNA sample cohorts can also shed light on how certain past (as well as present or future) outbreak, epidemic and pandemic events may have had shaped the immunogenetic diversity of human populations which is currently observed [505][514][520].

-Recent availability of new genomic data (at the 4-field resolution) at the scale of populations is contributing to understand in-depth other additional functional evolutionary processes that may be involved in the generation and maintenance of this vast HLA diversity, for instance: infer the selection pressures that operate on the observed high LD and its effect on haplotype frequency distributions; selection on HLA genes can be identified at various timescales, detecting long-term selection or recent selection events as well as ancient and modern admixture demographic events; epistatic interactions can be better evaluated between HLA loci and other genes (e.g. KIR genes) to consider co-evolutionary events; regulation of expression levels of the HLA genes can be also incorporated into evolutionary analyses [104].

- Finally, the role of stochastic forces such as population bottlenecks (sharp reduction in the size of a population due to environmental events), genetic drifts (random fluctuation of allele frequencies over time; thus, adaptive alleles may be lost and deleterious alleles could be fixed in the population), inbreeding and/or certain demographic events (e.g. migrations) in shaping HLA variation over time has been also described and particularly detected in small and isolated human populations [103].

Therefore, complex selection mechanisms appear to be acting simultaneously (in general terms, led by both natural selective forces as well as stochastic major environmental and demographic events) on the human MHC system, including among others: pathogen-driven selection for

antigen-binding breadth and expansion of the HLA gene clustered system; associated autoimmunity/tolerance trade-offs; hitchhiking of deleterious mutations linked to the human MHC; geographic subdivision in the context of major environmental and demographic events; and even genetic influence of ancient traces in the form of adaptive introgression of archaic HLA alleles (i.e. incorporation of a given foreign HLA variant from primitive ancient hominids which may have led to an increase of the fitness of the recipient modern-day human (*Homo sapiens*) populations pool) [501].

## **6. ROLE AND RELEVANCE OF HLA IN MEDICINE AND POPULATION GENETICS**

Histocompatibility testing is most often equated with the determination of the human leukocyte antigen (HLA) phenotype (referred to as serological HLA typing) and/or genotype (referred to as molecular HLA typing) of an individual. Histocompatibility testing became a very specialized area of clinical laboratory science with particular relevance to transplant programs since the discovery of HLA antigens as strong and important histocompatibility antigens [16]. In addition, the description of the diversity and biological function of HLA genes and molecules have been found to be also important in non-transplant settings such as disease association [109] and pharmacogenetics [110]. Furthermore, HLA phenotype and genotype characterization performed in hundreds of populations worldwide and their resulting allele and haplotype frequency distributions have been most useful to describe signatures of demographic events and/or selective pressures within and across the different continents. Thus, this has provided an abundant source of information to infer human migration history based on this extensive HLA genetic variation geographically and between different ethnic groups [111].

## **6.1 Clinical Transplantation**

Transplantation of solid organs (kidney, liver, pancreas, lungs, heart, intestine, cornea, etc...) is an important medical therapy that has made possible significantly saving lives in patients affected by terminal organ failures and substantially improving quality of life. However, the immune system poses a significant barrier to successful organ transplantation when tissues/organs are transferred from one individual to another (allotransplantation, which is the most common type). Rejection of solid organ allografts is the result of a complex series of interactions involving coordination between both the innate and adaptive immune system with T cells central to this process. The ability of recipient T cells to recognize donor-derived antigens, called allorecognition, initiates allograft rejection. Once recipient T cells become activated, they undergo clonal expansion, differentiate into effector cells, and migrate into the graft where they promote tissue destruction. In addition, CD4<sup>+</sup> T cells help B cells produce alloantibodies [112]. On the other hand, allogeneic hematopoietic stem cell transplantation (allo-HSCT) is an established therapy for a broad range of hematological malignancies, bone marrow failure states, and genetic diseases, where sustained engraftment of donor stem cells represents a fundamental prerequisite for a good outcome. However, bone marrow failure (BMF) syndromes are severe complications of allo-HSCT, including: the graft failure (GF), as the result of a classical alloreactive immune response mediated by residual host immunity persisting after the conditioning regime; and the poor graft function (PGF), which is the consequence of more complex and less well-defined interactions between the immune system and the hematopoietic compartment (niches) [113]. Furthermore, donor-derived T cells can also result in a series of critical events for the outcome of HSCT, such as: the graft-versus-host disease (GVHD), a life-threatening complication in patients undergoing HSCT, induced by the reaction of donor T cells to recipient alloantigens; or the graft-versus-

leukemia (GVL) effect, that is a result of donor T cells capable of recognizing residual tumor cells from the recipient and rejecting these cells, resulting in dramatically reduced risk of relapse [114].

The HLA system plays a crucial role in the regulation of immune function in the determination of self from non-self. As HLA molecules interact with T cell receptors in the thymus to modulate the immune response and determine what antigens and cells are recognized as self, thus, shaping immune tolerance [18]. As a consequence of this *in vivo* function of HLA molecules, there is now a clear understanding that HLA antigens (particularly, the most polymorphic classical HLA class I and II molecules) constitute an important and major biological barrier to a successful transplantation and has substantial impact on the prolongation of graft survival. As HLA molecules can elicit an immune response either by presentation of variable peptides or by recognition of polymorphic fragments of foreign HLA molecules [18]. Thus, it has been extensively reported how HLA disparity or HLA mismatch between donor and recipient is clinically relevant as it is associated with rejection, in turn, graft failure and, ultimately, resulting in considerable morbidity and mortality of the recipient. In the mid-1960s, this was first reported by Kissmeyer-Nielsen [115] and Terasaki [116] groups who showed in clinical kidney grafting that hyperacute rejection resulted from recipient sensitization specific for mismatched donor HLA antigens, and that a pre-transplant lymphocytotoxicity crossmatch test could predict such rejection. Similarly, among the many factors that influence the outcome of HSCT, polymorphism of the classical HLA genes represents the most important barrier [117]. As HLA compatibility or matching affects not only the ability to achieve sustained engraftment following HSCT but also the risk of developing acute and chronic GVHD [114]. In fact, many studies have also indicated how even single amino acid differences can result in immunologic responses against donor antigens, where also allelic level differences in HLA antigens have been shown to affect graft survival and especially for the successful engraftment in HSCT [117]. Moreover, studies have reached different conclusions

regarding the relative contributions of HLA class I and class II mismatching because of population-based differences in the specific HLA-mismatch combinations between patients and donors. Furthermore, recent observational studies in HSCT have described how not only the high expression loci (*HLA-A*, *-B*, *-C*, and *-DRB1*) are strongly associated with transplant outcome but also, and especially in combination with mismatches in other loci, considered low expression loci (*HLA-DRB3/4/5*, *-DQ*, and *-DP* loci) appear to have a contribution on the outcome as well [118]. Nevertheless, further and larger studies still need to be carried out to accurately define these contributions.

Although advances in immunosuppression modalities have been shown to improve transplanted graft outcomes, accurately determining both the donor and recipient HLA types and minimizing HLA mismatches is also of utmost importance to maximizing graft and patient survival. Consequently, accurate HLA typing has become essential in solid organ transplantation (SOT) and in HSCT. In the field of HSCT, bone marrow transplant graft survival has been highly dependent mostly on the degree of HLA matching (e.g. a registry donor matched for *HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1* at 2-field resolution, designated as 10/10 matched, is usually considered optimal for unrelated donor (URD) transplant) [112-114][117][119]. Furthermore, application of 4 field-resolution DNA-based HLA typing in different populations has significantly contributed for improving donor recruitment strategies of bone marrow registries. While in the context of SOT, knowledge of the donor and recipient HLA genotypes (allele-level) and phenotypes (antigen/epitope-level) is also needed (in addition to optimize HLA matching) to monitor recipients for development of donor-specific anti-HLA antibodies (presence of donor-specific anti-HLA antibodies (DSAs) before transplantation (performed/preexisting DSAs) leads to a hyperacute rejection; whereas DSAs may be also formed after transplantation (de-novo DSAs) and are associated typically with acute and chronic graft rejection). Since, in the majority of antibody-

mediated transplant rejection, the antibodies are directed against donor HLA antigens expressed by the transplanted organ but not present in the recipient [119]. Another important clinical scenario that requires the evaluation of HLA matching is platelet (which bears HLA antigens) transfusion from random donors in, for example, patients with aplastic anaemia (AA) who typically become alloimmunized and develop immunological platelet refractoriness (IPR) [119].

As previously mentioned, HLA match of donor and recipient is crucial as it increases the success rate of grafts. Nevertheless, perfect HLA matching is possible mostly when donor and recipient are related (e.g. 25% likelihood of having a full HLA-matched sibling donor for patients with one sibling) and, even in these cases, genetic differences at other non-HLA genes may still trigger rejection mechanisms (e.g. minor histocompatibility antigens, killer cell immunoglobulin-like receptor (KIR) genes and/or several other groups of genes) [117].

## **6.2 Disease Associations**

The MHC has been established as the region of the genome that is associated with the greatest number of human diseases. The majority of these diseases have an immunological component, which is consistent with the enrichment for key immune genes within the MHC region. During the last several decades, a large number of studies have reported strong associations between certain diseases (particularly those with an autoimmune component, on the basis of the fundamental role played by HLA molecules on orchestrating thymic selection as well as peripheral anergy of T cells) and individuals (where, in certain instances, there are also age and gender related differences) carrying particular HLA alleles [120]. Main reported HLA-disease associations are shown in **Figure I-15**.

---

**LIST OF MAIN EXAMPLES OF HLA-DISEASE ASSOCIATIONS REPORTED**


---

**AUTOIMMUNE DISEASES**

| <u>Disease</u>                    | <u>Main (HLA-) locus associated in susceptibility</u> | <u>Main Review Reference</u> |
|-----------------------------------|---|------------------------------|
| Celiac Disease (CD)               | HLA-DQ2 and -DQ8                                      | [90]                         |
| Diabetes Mellitus Type 1 (DMT1)   | <i>DRB1*03, DQA1*03, DQB1*02</i>                      | [539]                        |
| Ankylosing Spondylitis (AS)       | <i>B27 (B*27:04, B*27:05)</i>                         | [123]                        |
| Behçet's Disease (BD)             | <i>B51 (B*51:01)</i>                                  | [90]                         |
| Birdshot Retinochoroidopathy (BR) | <i>A29</i>  | [90]                         |
| Rheumatoid Arthritis (RA)         | Shared Epitope <i>DRB1</i> alleles                    | [395-397]                    |
| Graves Disease (GD)               | <i>C*07, B*08, DR3, DRB1*08</i>                       | [540]                        |
| Psoriasis in skin                 | <i>Cw6 (C*06:02)</i>                                  | [383]                        |
| Pemphigus Vulgaris (PV)           | <i>DQB1*05:03 and DRB1*04:02</i>                      | [541]                        |

**INFECTIOUS DISEASES**

| <u>Disease</u>                     | <u>Main (HLA-) locus associated in susceptibility</u> | <u>Main Review Reference</u> |
|------------------------------------|---|------------------------------|
| Human Immunodeficiency Virus (HIV) | <i>A23, B37, B49, B35-Cw*04, Class I Homozygosity</i> | [513][539]                   |
| Leprosy                            | <i>A*02, A*11, B*40, B*51, Cw*04, Cw*07</i>           | [513][539]                   |
| Tuberculosis (TB)                  | <i>DQB1*05</i>  | [513][539]                   |

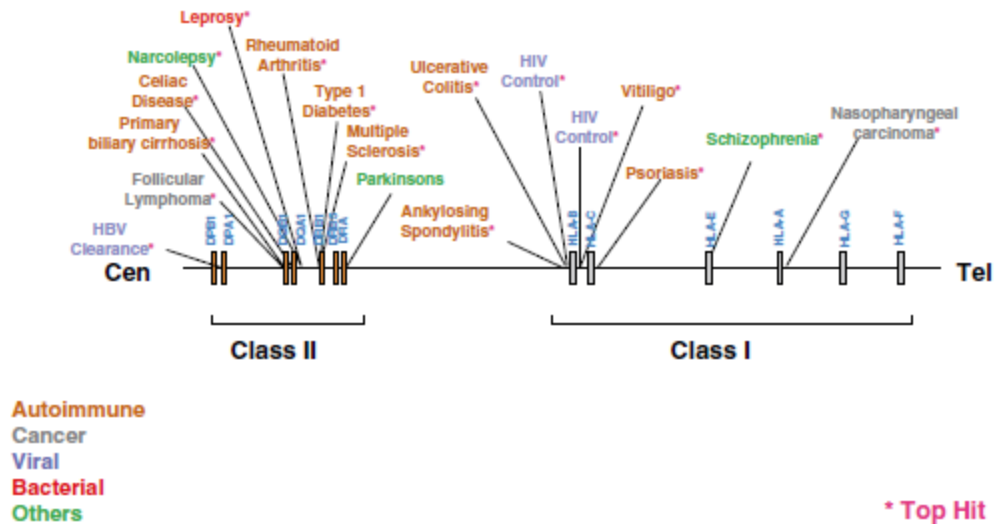
**CANCER**

| <u>Cancer Disease</u> | <u>Main (HLA-) locus associated in susceptibility</u> | <u>Main Review Reference</u> |
|-----------------------|---|------------------------------|
| Cervical Carcinoma    | <i>DQB1*03</i>  | [383]                        |
| Lymphoid Tumors       | <i>A2, A24, B12, Cw3</i>                              | [383]                        |

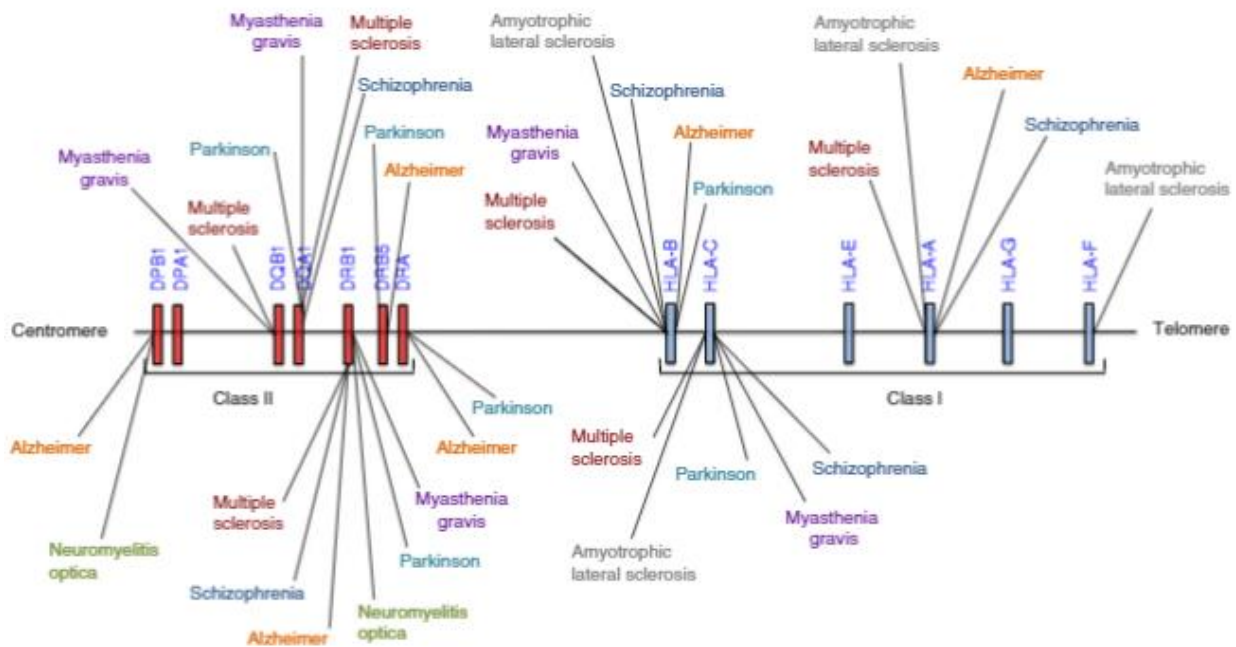
**NEUROLOGICAL DISEASES**

| <u>Disease</u>                      | <u>Main (HLA-) locus associated in susceptibility</u> | <u>Main Review Reference</u> |
|-------------------------------------|---|------------------------------|
| Narcolepsy                          | <i>DQB1*06:02 (and DQA1* pair)</i>                    | [90]                         |
| Multiple Sclerosis (MS)             | <i>DRB1*15</i> and associated class II haplotype      | [290]                        |
| Neuromyelitis Optica (NMO)          | <i>DRB1*03, DPB1*05</i>                               | [290]                        |
| Parkinson's Disease (PD)            | <i>DRB1*04</i>  | [290]                        |
| Alzheimer's Disease (AZD)           | <i>DRB1*03</i>  | [290]                        |
| Myasthenia Gravis (MG)              | <i>DRB1*03, *04</i>                                   | [290]                        |
| Schizophrenia Disease (SCZD)        | <i>DRB1*01, *03</i>                                   | [290]                        |
| Amyotrophic Lateral Sclerosis (ALS) | <i>A*03, A*02, A*28; B*40, B*35, C*04</i>             | [290]                        |





(Image originally from Mary Carrington, NIH, MD, USA)



(Image originally from Jill Hollenbach, UCSF, CA, USA)

**Figure I-15. (Upper Image)** Main examples of HLA-disease associations reported in scientific literature. These HLA-disease associations may vary among ethnic/regional groups and worldwide populations between other co-factors. Some of the reasons attribute for such variation are occurrence of population stratification based on geographical location and level of isolation, consequences of founder effect, racial admixture or selection pressures due to environmental factors. Hence certain HLA alleles that are predominantly associated with disease susceptibility or resistance in one certain population may or may not show any association in other populations

for that same given disease. Despite of these limitations, HLA associations are widely studied across the populations worldwide and are found to be important in prediction of disease susceptibility, (**Lower Image**) Schemes of gene map of the HLA genomic region and related mapped disease associations (top image including the main group of diseases; bottom image including only neurological diseases). Respective tables/figures are obtained and adapted from [90][123][290][383][395-397][539-541].

In general, although immune mechanisms appear to be involved in their pathogenesis, such diseases have unclear etiology. Moreover, these diseases have familiar recurrence and are supported by polygenic and environmental factors. Thus, their relation with HLA alleles identifies only one of the predisposing genetic factors, where the development of these diseases involves a genetic predisposition resulting from a combination of factors at HLA and at other genes [121]. There are at least 100 Online Mendelian Inheritance in Man (OMIM) identifiers concerning the HLA region loci, mostly of expressed genes, that can be accessed through <http://www.ncbi.nlm.nih.gov/> or through links from other sites, including Entrez Gene database at NCBI [27].

HLA disease associations are of diagnostic importance. Nevertheless, no situation has been identified yet where all individuals with a particular allele develop a disease, but some diseases have been identified in which most affected individuals have a particular HLA allele. Based on this, it is widely believed the inheritance of a disease-associated HLA allele may increase a patient's likelihood of developing the disease but is only a risk factor and does not guarantee that the disease will occur (i.e. incomplete penetrance). For instance, one of the very first reported and best known HLA-disease association is *HLA-B27* with ankylosing spondylitis (AS) [122]. In this sense, AS is strongly associated with *HLA-B27*, but the lifetime risk of a *HLA-B27*-positive individual developing it is only about 2%. This illustrates that in diseases with HLA associations, it is not the disease itself, but the predisposition to it that is inherited.

It is also noteworthy that initial HLA-disease associations defined at the serological level have been refined as HLA molecular testing has improved and become more widespread [119]. Continuing this previous example of *HLA-B27* association with AS, HLA molecular typing at higher resolution (making genetic testing more accurate over and above serological testing) has shown that while most *HLA-B27* alleles, such as the common *HLA-B\*27:05*, are associated with increased risk of AS, other alleles such as *HLA-B\*27:06* and *-B\*27:09* are not associated with this disease [123]. In addition, association between risk of AS and other HLA loci and alleles have been identified, including *HLA-B\*40:01* and *HLA-DRB1\*04:04* [124]. Therefore, the high-resolution typing of classical class I and class II HLA genes and the identification (e.g. via genome-wide association studies (GWAS) by analyzing an enormous number of genetic variants called single-nucleotide polymorphisms (SNPs) as markers) of other genes (both within and outside the human MHC region) have significantly increased the definition of the genetic basis of these HLA-associated diseases of unknown etiology.

The ultimate goal of all studies on HLA-linked diseases is to determine the molecular and cellular bases for disease. Even though it is commonly difficult to reconcile with epidemiologic data and functional characteristics of HLA molecules, three general categories of hypotheses have been proposed to explain these diseases associations with the HLA region [35][125]:

-The first category invokes linkage disequilibrium between a particular HLA allele that is associated with a given disease and another neighboring genomic element on the haplotype that is actually causative of the disease and does not involve HLA molecules directly. This can occur because the genes within the MHC are in extensive LD. An example of this type of associations include hereditary hemochromatosis, where an apparent association with *HLA-A* alleles (A3 and A29) results from mutations in a non-classic HLA class I gene, *HFE*, which is in LD with *HLA-A* locus.

-A second category implicates antigen presentation by the HLA allele, especially in the case of diseases that have a strong immunological component. Under this second category, different hypothesis have been proposed: immune reactivity to self-antigens due to aberrant T cell repertoire selection; immune cross-reactivity with foreign antigens; immune attack of “altered-self” antigens; or differences in the expression levels of certain HLA alleles that secondarily influence the course of certain diseases (including infectious diseases and cancer).

-The MHC cusp theory represents the other third hypothesis. In this case, the MHC codes for allele-specific ligands in the cusp region of the molecule, which interact with non-MHC receptors and activate various pathways. Aberrations in these pathways could cause MHC-associated diseases. According to this hypothesis, the cusp region constitutes a peculiar three-dimensional shape that has been conserved on both class I (in the  $\alpha 2$  domain) and class II (in the  $\beta 1$  domain) molecules through evolution, not dependent on antigen presentation, and is a hub for signal transduction ligands that interact with a variety of receptors and activate important biological functions.

### **6.3 Pharmacogenetics**

Adverse drug reactions (ADRs) are a significant cause of morbidity and mortality and represent a major burden on the healthcare system. Besides classical reactions that are related to pharmacologic activity of the drug (on-target ADRs or type A), some reactions are unpredictable, not dose dependent, and seem to occur in genetically predisposed individuals (off-target ADRs or type B). Moreover, it has been better understood how a significant group of this type of reactions is immunologically driven and they are referred to as hypersensitivity reactions (immune mediated ADRs or IM-ADRs). The expanding field of pharmacogenetics aims to understand these genetic factors that influence the outcomes of drug therapy, both beneficial and adverse. More recently, a

growing number of studies have provided clear evidences that specific HLA alleles increase the risk of developing hypersensitivity drug reactions [110][126].

As previously mentioned, extensive polymorphism of HLA molecules enables presentation of a wide range peptide ligands (termed as the immunopeptidome) thus maximizing immune surveillance of evolving pathogens diversity. A consequence of the diversification of the HLA peptide-binding pocket is the enhanced opportunity for off-target binding/interaction of small drugs or metabolites by HLA molecules, with subsequent immune reactivity. As these potential off-target interactions can generate T cell-mediated adverse drug reactions even though the precise mechanisms of most HLA-drug interactions are still poorly understood [127]. As it is considered unlikely that small molecule drugs (similar in size to one to two amino acids) would be able to stabilize HLA molecules alone, three core hypotheses have been proposed to explain the interaction of small molecule drugs or metabolites with the HLA-peptide-TCR axis to induce drug-specific T cell responses [127]: 1) Small molecule drugs or metabolites can react with specific amino acid side chains to covalently modify proteins. This covalent modification of cellular proteins can generate haptenated HLA ligands via antigen processing or direct haptentation of HLA-peptide complexes at the cell surface; 2) they may interact non-covalently with immune receptors either at the HLA-TCR interface or allosteric, creating neoantigens that engender HLA-TCR ligation and T cell activation; 3) the denominated “altered repertoire model” establishes how a non-covalent interaction between these small molecule drugs or metabolites and the HLA molecule antigen-binding cleft can modify the peptide-binding motif, allowing the entry of a novel array of self-peptides into the immunopeptidome and creating a new different conformation of the original HLA-peptide complex that is recognized now as foreign by circulating T cells. This latter proposed model has allowed to understand, as an exception, the association between abacavir hypersensitivity syndrome and *HLA-B\*57:01* at a molecular and mechanistic level.

The associations reported (see **Figure I-16**) between drug hypersensitivity and specific HLA alleles have been a recent finding and this has led to the possibility that hypersensitivity reactions may be predictable and preventable, as it is already the case of abacavir hypersensitivity and *HLA-B\*57:01* association, established into the clinic as a routine screening test [126][127].

**LIST OF MAIN EXAMPLES OF  
HLA-ASSOCIATED DRUG HYPERSENSITIVITIES REPORTED\***

| <b><u>Drug</u></b> | <b><u>Main (HLA-) allele associated in susceptibility</u></b> | <b><u>Main Review References*</u></b> |
|--------------------|---|---------------------------------------|
| Abacavir           | <i>B*57:01</i>  | [126][127]                            |
| Allopurinol        | <i>B*58:01</i>  | [126][127]                            |
| Carbamazepine      | <i>B*15:02</i>  | [126][127]                            |

**Figure I-16.** Main examples of HLA-associated alleles to adverse drug reactions (ADRs) or drug hypersensitivities. Respective table is obtained and adapted from [126][127].

## 6.4 Population Genetics and Anthropology

Knowledge of the genetic diversity of the human populations has expanded considerably in recent decades, especially thanks to the rapid progress in genomic research, where application of DNA markers (e.g. mitochondrial DNA (mtDNA), Y chromosome, SNPs, microsatellites, molecular HLA genotyping data or molecular KIR genotyping data) has become predominant in comparison to the use of serological/protein based genetic markers (including serological HLA typing data) [128].

The high and functional polymorphism, tight linkage among the different loci and the perpetuation of allelic lineages over time makes the HLA genes a very informative DNA marker

and a very useful tool for population genetics, phylogenetic and anthropological studies [128]. Thus, the frequency distributions of HLA alleles and haplotypes of human populations have been used to track human evolutionary processes, such as migration, admixture and selection. In addition, the accurate description of allelic and haplotypic HLA profiles and the identification of both common and rare or new HLA variants in human populations is also of great importance for the fields of clinical transplantation, epidemiology, pharmacogenetics and characterization of the genetic predisposition to many diseases that may enable a more predictive, preventative and personalized medicine [120].

During these last three decades, significant advances in molecular HLA typing methods (including the very recent development of ultra high-resolution typing at the 4-field, minimizing significantly the level of ambiguities thanks to full-length sequence analysis [152]) and accumulation of large population datasets by the international community (in the context of the International Histocompatibility and Immunogenetics Workshops (IHIWS) [129][487]; and the online Allele Frequency Net Database (AFND; <http://www.allelefrequencies.net>) [130][464]) have been most useful to radically increase our knowledge on the HLA polymorphism and diversity in human populations. Thus, it has been more extensively reported that different populations exhibit different HLA genetic profiles, where common allele and haplotype frequency distributions (and, thus, also patterns of LD) vary between population groups. Moreover, this worldwide HLA variation exists among both racial and ethnic groups as well as geographically between different regions [111][136]. Consequently, the improved and more precise description of HLA polymorphism at the DNA-level (mostly based on the distribution of allele and haplotype frequencies defined at the allele-level of resolution as well as distribution of HLA lineages) has been contributing to an in-depth understanding of the two main types of genetic signatures that appear to shape this HLA genetic diversity found among different worldwide human populations:

1) Human peopling history (genetic signatures related to demographic events). Indeed, patterns and distributions of this genetic variation of HLA genes among populations can help to better correlate genetic profile of populations and their past migrations in the determination of their origin as well as the geographic dispersal and recent admixture of modern human populations throughout the world [131][132].

2) The interaction of populations with their environment in a pathogenic context (genetic signatures related to natural selection). As previously mentioned, description of the extensive HLA polymorphism through population studies has been also contributing to unraveling the mechanisms of its molecular evolution, where strong selection appears to be operating at various levels. In this sense, the current level of diversity and the variation in observed allelic and haplotypic distributions for different populations probably also may result from evolutionary forces that have changed as human populations have encountered new environments in their spread around the globe [133].

Therefore, whereas the molecular evolution of these polymorphisms has most likely been subject to natural selection, principally driven by host-pathogen interactions [107], their patterns of genetic variation worldwide also show significant signals of human geographic expansion, demographic history and cultural diversification events [131].

To date, there are some key findings and proposed concepts from the main HLA modern population studies that have been carried out so far:

- At the allelic variation level, for most classical HLA loci that have been studied, allele frequency distributions are usually even (except in some cases) and populations achieve very high heterozygosity levels. Where certain alleles have identical protein sequences and are distinguished at the nucleotide sequence level by silent substitutions or substitutions in non-



coding segments, and thus these may be related by descent from a common ancestral sequence group [132]. As a result, both ancient and recent allelic diversification and strong selection for haplotype diversity may have sustained a maximized available peptide-binding repertoire [132]. Also, LD patterns displayed between neighboring loci may have had also an imprint on the evolutionary relationship between parental (more ancestral) and novel (more recent) alleles [132]. In addition, most classical HLA loci experience asymmetric balancing selection, a pattern that is found across diverse human populations. The main exception from these observations has been detected at the *HLA-DPBI* locus, which appears to fit more into a neutral model of evolution [131]. At the same time, in parallel and conversely to allele diversification events being led by selective pressure mechanisms, there are also a group of alleles that appear to have arisen independently through convergent evolutionary mechanisms [132].

-Furthermore, genetic distances between open populations estimated from frequency data (at both the allele and haplotype level) for all HLA loci are highly significantly correlated with geographic distances, and for migrant populations with their regions of origin. In contrast, the genetic distance measurements are larger than expected between inbred populations of the same region, very likely because of the existence of a large number of unique alleles in a small number of lineages as the result of limited founder polymorphism [131-134].

- LD patterns between alleles of various HLA loci can certainly provide significant insight with regard to the history of a particular allele. The examination of both LD and structural features may help elucidate possible evolutionary relations between alleles (e.g. rapid or recent diversification of an allelic lineage or selection for specific cis combinations). For instance, some alleles are found in multiple populations with distinctive haplotypic associations, suggesting that convergent evolution events may have taken place as well [132]. Thus, alleles differing only by silent substitutions either can be related by descent from a common ancestral

sequence or arise independently (convergent evolution) as they are selected on the basis of its functional capabilities. In addition, as the HLA region is characterized by strong LD between neighboring loci, long (extended) haplotypes encompassing several HLA loci have been preserved in present-day populations. Therefore, the sharing of HLA haplotypes among populations can be used for inferring genetic background and ethnical composition to evaluate relatedness between modern human populations and also for tracing migrations of modern or more ancient populations. In this sense, although a given population presents distinctive HLA allele frequency distributions in comparison to other populations, even more clear distinctions arise through the analyses of distribution of extended haplotypes between human populations, ethnic groups or regions. Hence, allelic diversity in HLA needs to be analyzed in the context of HLA haplotypes and blocks and in conjunction with other genetic markers to accurately track the migrations of modern humans [131]. At the same time, information from HLA genes as human migration markers for anthropological population studies needs to be interpreted with caution since undetected convergent evolution events may be confounding in the investigation of population relatedness, leading to erroneously close relations between populations [132][542][543]. Thus, it may not be always feasible to totally distinguish selective pressure events from demographic events, and vice versa, in the context of HLA [104].

- Some studies [135][136] have also proposed that, while many HLA lineages are old and have been inherited in a trans-specific fashion, the alleles within these lineages may be the result of a recent diversification. One illustrative example is the complex evolutionary history of the *HLA-DRB1* locus as this locus seems to be composed of segments with different levels of nucleotide diversity. On one hand, sites (exon 2 in this case of *HLA-DRB1* locus) encoding for amino acids involved in antigen binding (located in antigen recognition sites (ARS), or also known as the peptide-binding region (PBR) or antigen recognition domain (ARD)), where

evidence of selection has been observed, appear to have a more ancient origin, defining the different main allelic lineages. Whereas the polymorphism found at coding sites that are not involved in antigen binding (synonymous substitutions), which are thought to be evolving under (or close to) neutral conditions, suggests a more recent origin of the intra-lineage diversity. Furthermore, the lower level of intron diversity at the *HLA-DRB1* locus appears to also reflect a recent origin of most alleles from the same allelic lineage rather than being an effect of homogenization. In fact, this is consistent with a relatively rapid generation of novel alleles by gene conversion like events.

-Nevertheless, until recently the majority of reported population studies have presented important limitations in the most available traditional molecular typing methods (while being very useful for the clinical routine practice, they only capture a part of the molecular diversity information and generate a lot of ambiguities) and from a statistical perspective (low sample size to be representative of the given population of interest). Therefore, still more and larger high-resolution (ideally at the 4-field) HLA genotyping studies at a worldwide geographic scale need to be performed for obtaining complete HLA sequence data to allow a detailed comparison of different regions (coding and non-coding) of HLA genes in many populations [137].

## **II. MOLECULAR HLA GENOTYPING METHODOLOGIES**

### **7. MOLECULAR HLA GENOTYPING METHODOLOGIES**

Since the discovery of the HLA system over 50 years ago, there has been a concerted effort from the international histocompatibility scientific community [16] to properly and accurately define the high and complex polymorphism shown by the HLA loci. Both description and

understanding of the complexity and polymorphic nature of the HLA genes has been substantially improved as the HLA typing technologies have advanced and have been constantly innovated.

Initially, serological and cellular testing (antibody and mixed lymphocyte culture [MLC]) was first developed in the 1960s [115][116]. HLA serological typing was able to define HLA antigenic types based on cytotoxicity and antibody responses. Basically, panels of antisera (obtained from individuals sensitized to HLA antigens via pregnancy, transfusion, or transplantation or with monoclonal antibodies) were incubated with viable lymphocytes (to be HLA typed) in the presence of complement and a vital dye. These panels were designed containing sera with antibodies to a multitude of known HLA alleles and, thus, detect most of the HLA antigenic types defined at that time (including latter discovered “splits” that defined HLA antigens into narrower specificities). Thus, when a cell expresses a specific HLA antigen that is defined by a particular antiserum from the panel, the antibodies will bind and activate the complement, resulting in cell death that is detected by the addition of a viability dye. A positive result (meaning cell death indicating recognition of the unknown HLA antigen by the typing serum) is indicated by greater than 20% cell death in the reaction. In spite of refinement and standardization of extensive panels of antisera and complement were established, these serologic methods still could only resolve a very small and limited number of HLA antigenic specificities (e.g. typing sera are often polyspecific, requiring the use of larger numbers of antisera to be able to clearly define the presence of a HLA antigen). Therefore, despite of being initially useful for clinical transplantation purposes, significant limitations (relying on the availability of viable lymphocyte preparations and a battery of antisera to recognize HLA polymorphisms) of serologic HLA typing methods represented an important handicap in order to fully understand the diversity of the HLA system [88].

In the 1970s and 1980s, first molecular DNA-based methods applied to HLA typing (e.g. two-dimensional electrophoresis and restriction fragment length polymorphism (RFLP)) started to

show, although still insufficiently, a higher degree of HLA polymorphism that was still not properly revealed [138]. In the mid-1980s, the development of the polymerase chain reaction (PCR) and the application of (PCR)-based HLA genotyping methods to histocompatibility testing have allowed to clearly characterize this complex polymorphism by sequencing HLA genes:

- Since molecular HLA typing allows the definition of HLA type based on DNA sequence in addition to the amino acid sequence and serologic reactivity. It was able to confirm that the HLA phenotypic polymorphism defined by HLA serological typing methods had greatly underestimated the true degree of HLA polymorphism. Indeed, sequencing analyses of HLA genes lead to the determination that alleles with different DNA sequences can encode proteins with similar serologic reactivity (even all can be recognized as a single serologic specificity) [88].

-At the same time, immunologic studies demonstrated that these small nucleotide differences, which implied in some cases only a single amino acid difference between two HLA molecules, although not serologically distinguishable, could be recognized by the cellular arm of the immune system as foreign and lead to cellular immune reactivity. Thus, identification of HLA genes to the allele-level by molecular-based DNA typing methods have become extremely important in the clinical setting, being critical and essential in the transplantation field (as HLA allele-level matching contributes to minimize complications such as acute rejection, graft failure or GvHD) [112-119].

-Sequencing of HLA genes by molecular typing methods resulted in a rapid increase of newly identified alleles that also promote and significantly help to refine and standardize nomenclature of the HLA system [87][96].

First (PCR)-based HLA genotyping methods utilizing sequence-specific oligonucleotide probes (SSOPs, or SSO), [139] among others, and sequence-specific primers (SSPs), [140] among others, provided the means for more directly evaluating the highly variable sequence motifs within the HLA genes. Subsequently, Sanger sequence-based typing (SBT), [141] among others, in the 1990s significantly advanced tissue typing and transplantation genetics by providing an unprecedented molecular view of HLA polymorphism in the context of variation at coding and non-coding regions. Most recently, application of “next-generation sequencing” (NGS) [142] appears to have revolutionized the field by addressing the HLA typing complexity in a very definitive way. As NGS-based HLA typing provides complete and phased (allele-level or locus-level) HLA genomic sequence characterization of these highly polymorphic genes [76].

Consequently, clinical HLA typing over the past decade has transitioned from a combination of serological and cellular-based methods to more direct, faster, more affordable, and more informative DNA-based techniques. Even though serological typing may continue to have some clinical or research-based testing in determining the expression of the HLA molecule at the cell surface (a function that DNA-based testing cannot always verify), direct DNA-based typing techniques have all but replaced serological methods in routine HLA typing [76].

## **8. TRADITIONAL MOLECULAR HLA GENOTYPING METHODOLOGIES: ADVANTAGES, LIMITATIONS AND UNRESOLVED AMBIGUITIES**

Over the last decades, and still in use today, SSO, SSP, and SBT have been the most prevalent and well-developed DNA-based HLA typing techniques. Despite each of them presents a different basis and approach, their respective designs and chemistries (with the exception of SBT until certain extent) have optimized only the analysis of exons 2 and 3 of HLA class I genes and exon 2 of HLA class II genes. Since these regions include most of the known polymorphisms and encode

those domains that interact with bound peptides being directly related with the functional biology of the HLA system and clinically relevant [56][89][112-119]. Nevertheless, it is well known how this limited genomic characterization, given by these traditional or legacy techniques, still cannot describe all the HLA polymorphism (also found in other coding and non-coding regions) and, in turn, causing this many typing ambiguities.

### **8.1 Sequence-Specific Oligonucleotide (SSO) Probes**

The SSO or SSOP technique interrogates polymorphic differences using panels or “pools” of short (~20-30 bp) individual DNA oligonucleotide probes that differentially hybridize to the target of interest (the probe either perfectly matches or mismatches the target’s polymorphic sites) found in a HLA PCR product or amplicon. In this sense, control of hybridization conditions and the stringency of the washing process are critical and challenging aspects of this methodology. The hybridization pattern of the oligonucleotides is compared to a reference pattern, based on the sequence database of HLA alleles, and is interpreted as a HLA type or allele call. After the early development and application of, firstly, forward and, later, reverse SSO “dot-blot”/“line-blot” typing methods on a solid membrane system as the hybridization medium. The newer microbead-based reverse SSO typing is one of the most widely used methods currently as it is able to provide low-cost, rapid and low-to intermediate-resolution typing. In which, the microbeads presenting complementary probes to the amplicon will hybridize. Furthermore, both the microbeads’ fluorescent signature (thereby identifying the specific attached SSO oligonucleotide probe) and the fluorescent hybridization signal generated by the biotinylated HLA PCR product labeled with a streptavidin-phycoerythrin conjugate are detected by a flow cytometer system in an automated manner (interrogation of HLA amplicons by up to 500 sequence-specific probes in a single tube) [143]. Thus, software analysis program for allele assignment translates the probe reaction patterns into HLA genotypes. Nevertheless, because of its nature and design, SSO mostly provides low-

resolution typing expressed commonly as serological equivalents and rarely produces unambiguous two-field high-resolution typing. In addition, short oligonucleotide probes can only provide phasing information for polymorphisms underlying the probe, making linkage across an exon difficult, and between exons impossible. This technique also requires continual development as new probes often have to be developed to detect the newly defined alleles, otherwise undetectable [76][85].

## **8.2 Sequence-Specific Primers (SSP)**

The SSP method uses panels of specific primer sets that target and overlap polymorphic sites. Perfectly matched primers produce an amplification product, while mismatched primers do not. Thus, the pattern of amplification from multiple primer sets determines the HLA allele. In detail, resulting amplification products are separated and imaged by standard gel electrophoresis or separated by polymer-based capillary separation, allowing HLA typing to be inferred from the pattern of amplification. SSP method offers a short turnaround time (“STAT”) (2-3 h) making it a good choice for STAT HLA typing cases as in deceased donor typing for organ transplantation. Also, a significant advantage over the SSO method is its ability to provide some phasing information, leading to fewer ambiguities. But it is still limited to only those instances when is possible to design PCR primers that overlies polymorphisms on both the forward and reverse strand, therefore linking the phase of the two polymorphic sites together. Furthermore, the SSP method can be used to provide both low-(1-field) and, although challenging, high-resolution (2-field) typing results. However, similar to SSO method, SSP does not provide comprehensive coverage of all known and unknown alleles. Thus, if there are no primers that target the position differentiating the novel/rare/null allele, then that allele will go unassigned and may lead to false reporting of the HLA genotype [76][85].



### **8.3 Real-Time PCR (RT-PCR)**

Traditional real-time PCR (RT-PCR) is an extension of SSP typing. This automated method utilizes real-time PCR to amplify (amplification processes similar to those of the SSP method) specific HLA gene regions (overlapping known polymorphic sites) and an intercalating dye which has two main functions. Firstly, the intercalating dye signal is measured to verify that amplification has occurred (endpoint identification of amplification). Secondly, the melting temperature of any given amplicon is reached, the double-stranded PCR product denatures and the intercalating dye is released, resulting in a drop in fluorescence, and this defines a specific melting curve profile. Allele specificities are evaluated through differences in melting temperature (specific melting curve profile analysis). Thus, software analysis program converts the specific amplification/melting-curve profile into allele calls, which are then used to determine genotype based on expected assay behavior and known IPD-IMGT/HLA database alleles. RT-PCR not only has the fastest turnaround time but it has also a unique set of SSP assays that can identify the majority of known alleles, including null alleles (or non-expressed alleles, which are often due to “InDels” (insertions/deletions) causing a frameshift mutation resulting in a premature stop codon or single point mutations causing a premature stop codon). However, RT-PCR still shows similar limitations (e.g. limited phasing information, limited identification of novel alleles, limited allele-level resolution) as in SSO and SSP methods [76].

### **8.4 Sanger Sequence-Based Typing (SBT)**

As a robust and comprehensive method of determining high-resolution HLA typing results, Sanger SBT became the gold standard for HLA molecular typing, being able to overcome many of the limitations presented by these previous methods (SSO or SSP). With expanded coverage of the HLA genes, an improved accuracy and resolution in a bidirectional sequencing fashion, SBT generates high-resolution typing results (not only at the 2-field, but also at the 3- and, even, 4-

field) and, thus, enables the detection of known and many novel/rare/null alleles. In SBT, specific gene regions are amplified and sequenced through a process of polymerase-based extension of specific sequencing primers using, in addition to normal nucleotides (dNTPs), distinguishable fluorescently labeled 2', 3'-dideoxynucleotides (ddNTPs), indicating allelic differences base by base. In general, SBT requires two amplification steps: a first amplification that serves to expand the copy number of a particular gene region (exons or an entire gene), in which balanced and specific amplification for every possible pair of alleles is crucial; then there is an intermediate cleanup step, that is essential to remove or inactivate excess of thermal cycling reactants and amplification primers; and, finally, a second amplification step to perform the cycle sequencing reactions (called dideoxynucleotide or chain termination sequencing) that produce the fluorescently terminated sequences for subsequent capillary analysis. The sequences of the fluorescently labeled, single-stranded DNA (ssDNA) fragments are determined by performing electrophoresis using a capillary genetic analyzer, in a process called dye terminator sequencing (based on the detection of the respective fluorescent dye attached to the terminal ddNTP of each of the ssDNA fragments that go through the capillary sequencer). Thus, the entire DNA sequence of interest can be determined as the pool of fragments, one nucleotide longer than the previous, sequentially reveal their base content (determination of sequence from terminal dideoxy labeling). Where the sequencing data is represented as an "electropherogram" (showing a combination of the DNA sequences of two alleles of a given locus, which are amplified and sequenced together). Finally, HLA typing software compares electropherograms against a reference database of all HLA alleles in order to provide the HLA allele calls per locus. Nevertheless, while Sanger sequencing is considered the gold standard for HLA typing, it still has important limitations. From a technical point of view, SBT is more complex, more laborious and more time-consuming than SSO or SSP methods, presenting many steps (one or several primary amplifications, cleanup steps, sequencing

reaction setup per exon/region and each direction (forward/reverse), and further cleanup steps) and, consequently, it has increased costs comparatively. Furthermore, although SBT typically yields higher resolution results than the SSO and SSP methods, it also presents HLA typing ambiguities. The primary cause is that since SBT generates heterozygous electropherograms (i.e. SBT can produce long contiguous reads (up to ~1000 bp) but it mixes the signals from the two chromosomal strands), the phase of the polymorphic sites (i.e. which nucleotides are linked together forming an allele) cannot be determined in many cases (known as cis/trans ambiguities or phasing ambiguities). In addition, due its technical complexity and in order to meet the clinical demand and expected turnaround times routinely, existing SBT kits have been primarily designed to characterize only exons 2, 3, and 4 for class I and exons 2 and 3 for class II, and, thus, polymorphisms outside of these sequenced regions cannot be resolved [76][85].

### **8.5 Limitations and Unresolved Ambiguities of Traditional HLA Typing Methods**

Overall, these traditional molecular methods (SSO, SSP, RT-PCR and SBT) have been successfully established and prevalent in the HLA typing field, being able to interrogate and characterize an important part of HLA polymorphism that, in addition, is clinically relevant. Each of these traditional DNA-based methods continues to be used for routine low- and high-resolution HLA typing. No one method has supplanted the others as each has its unique benefits and limitations. However, common limitations shared by all these legacy methods certainly still represent an important restriction for advancement in both research and clinical applications [143][144]:

- 1) Scalability: despite development of sophisticated automated systems (e.g. including, liquid- and plate-handling robotics and pipeline software to operate, collect and analyze sequencing

data), the sequencing chemistry basis and instrumentation of these legacy HLA typing methods only permit low and moderate test volumes (low-throughput, of the order of dozens of samples/HLA loci per run) not very compatible with the growing demand of clinical HLA typing or large-scale population studies. In addition, significant amount of DNA per sample is required for testing.

2) Cost: Once again, the chemistries and design of these legacy HLA typing methods make them labor intensive and not very cost-effective, considering also their low scalability.

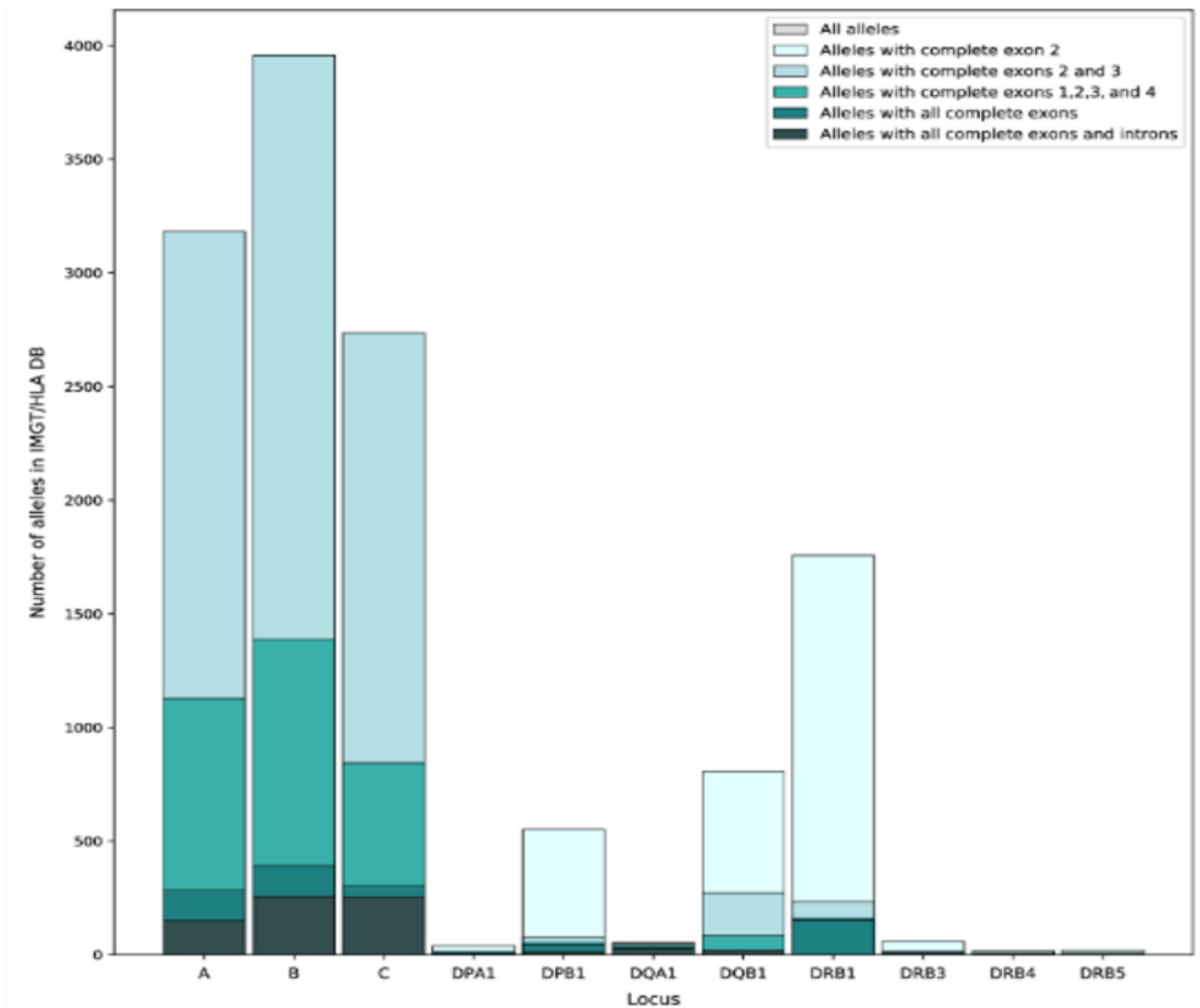
3) Time: Despite RT-PCR and SSP methods offer very rapid HLA typing (of great importance for deceased donor typing in solid organ transplantation), traditional molecular HLA typing methods are low-throughput and mostly produce low-resolution typing with significant ambiguities, which to be resolved they require additional typing tests extending the turnaround time.

4) Coverage and Ambiguities: Most importantly, and again based on the chemistries and design, traditional DNA-based HLA typing approaches are limited in their ability to discriminate between all possible alleles and combinations of alleles. Three different types of ambiguities are recognized [85]:

a) Single alleles, which cannot be discriminated because the nucleotide differences (polymorphisms) between the alleles are located outside the region amplified. Using traditional molecular HLA typing methods it has been attempt to overcome this limitation by expanding coverage of the HLA genes (e.g. incorporation of additional primers and probes as well as allele-specific sequencing primers) but still this has not been sufficient considering the high and complex HLA polymorphism. In addition, this limitation contributes to the inability to directly detect novel alleles.

b) Polymorphism at the HLA loci is clustered in a distinctive patchwork pattern of sequence motifs, which results in the extensive allelic diversity observed for these loci [145]. A consequence of the patchwork pattern of sequence polymorphism is that a large number of alleles share the same gene segments and therefore cannot be easily distinguished. In this context, the sequencing chemistry, nature and coverage of traditional molecular HLA typing methods shows an inability or very low ability to determine phasing (which nucleotides are linked together forming an allele) between enough number of proximal and/or distal polymorphic sites, generating this type of ambiguities termed as “phasing” or “cis/trans” ambiguities. Thus, several allele pairs are equally possible as a HLA typing result for a given locus as the different allele combinations present identical heterozygous sequences. Different strategies have been used in order to resolve these ambiguous heterozygous combinations (e.g. allele-specific amplification, family based HLA haplotype segregation studies, application of cloning techniques, or single-allele amplification), however all these approaches are usually time-consuming, labor-intensive and costly.

c) Finally, many other HLA genotyping ambiguities also exist since only about 10% of known HLA class I and class II alleles have been fully sequenced (full genomic sequences) [146][463] and that, in turn, are available as HLA sequence references in the official database (IPD-IMGT/HLA) [87] (see **Figure I-17**). Thus, HLA genotype calling software programs are limited by this incomplete sequence database, which (until recently) have been built using only these traditional HLA typing methods that interrogate HLA genes only partially. In addition, this also contributes to the inability to detect more novel alleles.



**Figure I-17.** Bar plot showing the number of available exons and introns (presumably, including also here, under this category, the 5' and 3'UTR regions) sequences per HLA locus in the IPD-IMGT/HLA database (according to release v.3.36.0; April 2019). As it can be observed, only a small portion of HLA alleles has been fully sequenced (colored in darkest green at the bottom of each bar represented). Currently, most (~80-90%) of the HLA alleles sequences reported and officially named and incorporated in this IPD-IMGT/HLA reference database only cover the domains that encode for the antigen recognition domain (ARD; exons 2 and 3 for the HLA class I genes; and exon 2 for HLA class II genes). Thus, HLA polymorphism in rest of coding regions (beyond these exons that encode the ARD) and most of the non-coding regions is considerably underrepresented in the current IPD-IMGT/HLA database. Figure and respective footnote are obtained and adapted from [463].

Considering the fact that resolution of HLA ambiguities (caused by both the new allele sequence and the unknown phasing of the additional polymorphisms that can lead to additional allele possibilities) is a time-consuming, labor-intensive and costly process. In the clinical transplantation setting, HLA typing reporting needed a set of guidelines where alleles could be categorized based on the prevalence in the population to resolve ambiguities accurately as possible, but in an optimized and prioritized way, given the technical limitations in these traditional molecular HLA typing methods. Thus, through a detailed analysis of large datasets and pertinent literature review (originally it had the goal to provide guidance for external proficiency testing but it rapidly became a reference for clinical decision making and a critical tool for testing development and laboratory clinical decisions and registry policy), an initial Common and Well-Documented (CWD) allele catalogue was reported by the Ad-Hoc Committee of the American Society of Histocompatibility and Immunogenetics (ASHI) (v.1.0.0.)[147] (v.2.0.0.)[148]. This CWD catalogue defined “common” (HLA alleles found at known high frequencies), “well-documented” (replicated using sequence-based typing and HLA haplotype data) and “rare” (HLA alleles found at known low frequencies) alleles applying a specific criteria that is based on the frequency of the alleles in a given population and on the polymorphisms located within exons 2 and 3 of class I and exon 2 of class II alleles. The CWD categories also extend to ‘G’ and ‘P’ designations for different alleles with identical nucleotide and protein sequences, respectively. Furthermore, this CWD catalogue became a worldwide reference to infer (being used as a statistical approach) the most likely allele in a string of possible alleles or ambiguous combinations to assist with resolving HLA genotyping ambiguities. Thus, this allele-prevalence information allowed clinical testing laboratories to establish a reporting convention criteria where, for instance, rare alleles (based on their low frequency in the population tested) may be discounted as well as certain ambiguities,

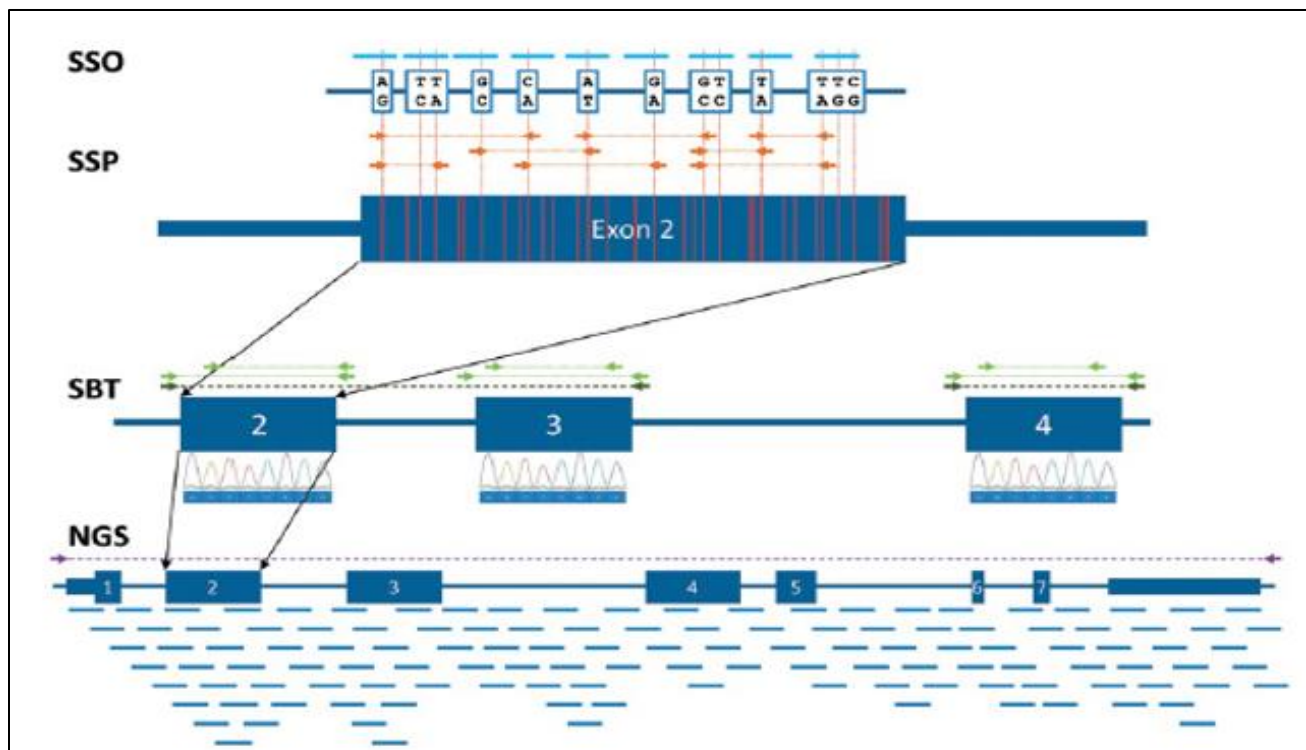
including alleles with identical antigen recognition sites (ARS) are left unresolved, as they are presumed to be clinically irrelevant. Nevertheless, technological and scientific innovations provide new criteria for considering allele prevalence and CWD catalogues need to be constantly updated [146][463][479]. In addition to the U.S. population, studies from European and Chinese population HSCT donor registries have shown the importance of developing local tables of CWD alleles according to each population/region due the differences found in allele frequency distributions between human populations, since CWD catalogues based on different population sets do not overlap completely [149][150][480]. It is also noteworthy that precise definitions of common and well-documented differed somewhat among all these studies [147-150][480], in general, alleles were classified as common if they were observed in multiple population groups with frequencies greater than 1 in 1000 in groups of at least 1500 individuals. Well-documented alleles were more restricted in their distribution with unclear frequencies but were observed at least five times by DNA sequencing or three times in a shared haplotype. The remainder of the alleles were classified as not-CWD.

At the same time, in order to report HLA genotyping results with unresolved ambiguities (having one or two allele lists instead of one or two true final allele calls per HLA locus, in situations such as: when the full nucleotide sequence of a given HLA gene is not characterized or when the quality of data does not allow a clear-cut interpretation), certain HLA ambiguities reporting systems have been established by international immunogenetics organizations and working groups [296]. A common form of representing such ambiguous genotypes is by using Multiple Allele Codes (MAC), also known as NMDP (National Marrow Donor Program) codes (<https://bioinformatics.bethematchclinical.org/hla-resources/allele-codes/>). With this MAC system, any currently known combination of allele numbers is encoded as 2–4 letters which are added to an allele name following the first field. As an example, an allele reported



as DQB1\*02:GKDU can be any allele from the DQB1\*02: 02 group of alleles or DQB1\*02:97 [296]. While MAC codes are widely used in particular for communicating donor registry genotyping data, they are not part of the official HLA nomenclature system as it is the case of “P” or “G” allele groups as previously explained [35][74].

Overall, because of all these limitations found in traditional molecular HLA typing methods (as well as in many other genetic testing fields), a need for a more accurate, more informative and high-throughput DNA sequencing strategy stimulated and led to the recent development of novel technologies of multiplex DNA sequencing denominated “next-generation sequencing” (NGS) and its implementation in HLA genotyping, being able to obtain full length and phased (locus-level) genomic sequences of HLA loci with minimum ambiguities (see **Figure. I-18**).



**Figure I-18.** Scheme of main molecular HLA typing techniques and their respective genotyping approaches/coverage capacity for interrogating the HLA gene polymorphism. For any given HLA gene sequence (dark blue rectangles (*exons*) and adjunct lines (*introns*)):

-Sequence-Specific Oligonucleotide (**SSO**) Probes of an average segment size ~20 bp (light blue lines) can only provide single-nucleotide resolution of HLA haplotype differences (polymorphic differences, represented as red lines in exon 2 rectangle). This technique requires a complex panel of oligonucleotide probes to discern differences between specific HLA alleles. Moreover, this (commercially available or prepared in-house) set of probes is mostly static and therefore cannot adjust to novel alleles (it is unable to identify them).

-Sequence-Specific Primers (**SSP**) (orange arrows) can provide HLA haplotype-specific and/or allele-specific resolution of nucleotide differences and additionally provide some level of phasing between polymorphic sites. As with the SSO probes, these oligonucleotide sets for performing SSP HLA typing are complex and static, being thus very limited to describe HLA diversity.

-Sanger Sequence-Based Typing (**SBT**) most commonly (especially in the case of commercial versions) only provides whole-exon information on the polymorphic content of the HLA allele (amplification primers [dark green] and sequencing primers [light green arrows]) but cannot discern phasing, as this sequencing method generally does not rely on allele-specific primers for amplification as a first step.

-Next-generation sequencing (**NGS**) provides whole-gene (exon, intron and UTR regions) amplification (amplification primers, purple arrows) and detection of polymorphic content for any HLA allele (known or unknown) and provides significant phasing between polymorphic sites that are within the read lengths of the system being used (usually between 200-600 bp for short-read sequencing platforms and over 1000 bp for long-read sequencing platforms). In the case of short-read sequencing strategies, this is accomplished through the alignment of thousands of short overlapping reads that are combined to form a single consensus sequence (blue lines).

Figure and respective footnote are obtained and adapted from [35][76].

## **9. NEXT GENERATION SEQUENCING (NGS)-BASED HLA GENOTYPING STRATEGIES: IMPACT AND RELEVANCE**

### **9.1 Three Generations of Sequencing Technologies and Its Application to High-Resolution HLA Typing**

I) The Sanger method (considered the “first generation” sequencing technology) was the primary sequencing technology between 1975 and 2005. Sanger sequencing produces relatively long (500-1000 bp) high quality DNA sequences. The implementation of Sanger based sequencing of targeted HLA amplicons (Sequencing Based Typing (SBT)) has allowed the analysis of all the sequences within the amplicon, giving rise to SBT as the gold standard for high-resolution HLA typing. Nevertheless, as previously mentioned, SBT presents important limitations such as: genomic regions not targeted by the primers or not full-coverage to minimize time-consuming and laborious workflow amenable to low and moderate test volume; and the inability to set phase for linked polymorphisms within the amplicon for heterozygous samples providing ambiguous genotyping results (phasing ambiguities) that are very difficult to resolve if not impossible [141][143][144].

II) Beginning in 2005, a new series of short-read (~25-600 bp) sequencing platforms emerged, which are based on massively parallel clonal sequencing and are commonly referred to as “second generation sequencing” or “next-generation sequencing” (NGS) [151]. Although these short-read sequencing platforms differ substantially (e.g. in terms of DNA template generation and immobilization, sequencing chemistry, engineering configuration, imaging system, read length and data analysis), they all are conceptually similar through the combination of a clonal sequencing chemistry and a high level of parallelism defined by its own engineering design. Clonal sequencing chemistry allows that single fragments of DNA (where each of the fragments is derived from a given single strand) are amplified and sequenced (one nucleotide base at a time) independently.

Therefore, this allows a potential complete distinction and characterization of both alleles for a given targeted gen. This is accomplished using methods that enable the isolation of single DNA molecules (originally generated as part of a prepared DNA library of the targeted region of interest), which are subsequently immobilized to a surface/medium and clonally amplified. During this clonal amplification, each original DNA template produces a clonal population or cluster that represents thousands of identical copies of that same original DNA fragment, which are in close proximity in a defined area ensuring that the sequencing signal can be distinguished from the background noise. In addition, creation of millions of spatially separated immobilization DNA template sites displayed in a microscale designed sequencer system allows this massive parallelization of clonal sequencing reactions (where hundreds of thousands to hundreds of millions of sequencing reactions can be performed simultaneously) in a single run and instrument. Consequently, with a “step-wise” cycling process of sequencing reactions of each of the templates copies of each cluster, of the order of millions of sequencing reads are generated (where a read refers to the consensus number of nucleotide bases or the sequence of a given cluster that is obtained after the end of the sequencing base calling process, which is ultimately the sequence of a section of that unique original single template molecule) and then analyzed. Thus, all these features of “second generation sequencing” technologies define them as deep-sequencing platforms presenting very high coverage (through the generation of such as high number of reads, where coverage refers to number of unique reads that include a given nucleotide in the final reconstructed sequence in order to provide a reliable base call) and, consequently, low error rates (ranging from 0.2% to 2.0% at the individual raw read level, where the per-read error rate is defined as the proportion of reads containing sequencing errors). Furthermore, through the use of multiplex identifiers (MID), a very high number of samples (hundreds or even thousands) tagged with unique identifiers (indexing) can be pooled and sequenced together in a single run increasing even more

the high-throughput capacity of these NGS sequencing systems [151]. It is also noteworthy that, in addition to the inherent high-throughput capacity of multiplexed NGS platforms, implementation of automation and robotics systems for plate- and liquid-handling steps have also contributed to the high scalability and reproducibility of NGS targeting and library preparation workflows. Therefore, in comparison to NGS technologies, the technology of capillary electrophoresis-based dye-terminator Sanger sequencing is very limited, especially in aspects such as: 1) Throughput (very limited number of capillary tubes presented by the Sanger sequencer system in comparison to the massive NGS template immobilization systems); 2) Accuracy (while Sanger sequencing presents inherent diploid cis/trans ambiguities (a single signal trace for both alleles delivered by Sanger sequencing does not allow putting heterozygous positions in phase)); NGS methods enable separate, parallel sequencing of multiple single strands of DNA, allowing base calls at heterozygous positions to be correctly assigned to the paternal or maternal allele and, thus, yielding a high level of phased data per targeted locus (referring here to NGS short-read sequencing platforms) in a single pass than Sanger sequencing); 3) Dynamic range in sensitivity for detection of both alleles (the signal-to-noise ratio intensity is proportional to the number of clusters in NGS, while Sanger sequencing presents a very low signal-to-noise ratio that directly depends on number of dye-terminated fragments per allele); and 4) Cost per sample (in comparison to the limited scalability offered by SBT, heavy multiplexing of genes and samples enabled by NGS technologies (based on massively parallel clonal sequencing of single DNA molecules) has led to a dramatic reduction in cost per sample).

The significant value of the unique and novel properties of short-read NGS platforms was rapidly recognized and this NGS technology was soon applied for developing targeted high-throughput high-resolution HLA typing approaches that could overcome the important limitations given by the current gold standard Sanger SBT [152]. Since the clonal sequencing property allows setting

phase for linked polymorphisms within each HLA gene, thereby reducing the level of genotyping ambiguity (cis/trans ambiguities); while the massively parallel property makes possible the sequencing of many different genomic regions (enabling an expansion of the coverage of HLA regions sequenced and, thus, facilitating the description of both novel exonic and intronic polymorphisms); as well as, with the multiplexing approach, testing simultaneously all classical HLA class I and II loci per sample and a large number of samples (of the order of hundreds and even thousands [154]) in a single run. Starting around 2009, many different approaches for using short-read NGS platforms for high-resolution HLA genotyping have been developed using a variety of targeting (including PCR-based or capture/hybridization-based methods) and library preparation strategies for DNA template isolation; sequencing platforms; and sequence analysis approaches to greatly enhance sequencing coverage depth and resolution of ambiguities [152]. Consequently, approximately since 2015, the development, clinical validation (regulated by accreditation organizations such as the American Society for Histocompatibility and Immunogenetics (ASHI) or the European Federation for Immunogenetics (EFI)) and implementation of vendor-supported (e.g. Omixon, GenDX, CareDX, ThermoFisher/One Lambda or Immucor) and/or in-house kits and HLA genotyping analysis software packages for high-throughput HLA genotyping by NGS have taken place in histocompatibility laboratories (some main examples in [153-157]).

Nevertheless, current second generation sequencing platforms present a primary inherent limitation on the read length (number of bases identified contiguously in the same read) as they generate short reads due to the nature and basis of their own main chemistries that have been developed so far: 1) Sequencing by hybridization and ligation (SBL): consisting on cycles of hybridization and ligation of various one/two-base-encoded probes to the template only allow read lengths up to ~ 80 bp; 2) Sequencing by synthesis (SBS) (with read lengths up to ~300-600 bp):

being the most widely adopted clonal chemistry among second generation sequencing platforms, here the sequence of a DNA template is determined by synthesizing the complementary DNA one nucleotide base at a time that is detected or “gated” in a “wash-and-scan” sequencing cycle process. Hence, these two main second generation sequencing “step-wise” and “ensemble-based” chemistries tend to present loss of DNA sequence quality with sequence length because the yield for the gated addition of each base is less than 100% due to uneven PCR amplification efficiency during either DNA ligation or DNA polymerization. Subsequently, this causes that sequence reads of each cluster gradually diverge in length (becoming asynchronous, as primers move out of synchronicity for any given processing cycle), if an extra base is added (leading-strand dephasing) or a base fails to incorporate (lagging-strand dephasing). With the many “wash-and-scan” repeated cycles, this loss of synchronicity amplifies into what is referred to as dephasing error, which represents the main limit of achieving long read lengths. The decrease of the signal-to-noise ratio per cluster associated to this dephasing error causes a decrease in quality (more sequencing errors) towards the ends of longer extending reads. This fact effectively limits the maximum reliable read length (to a range of only several hundred bases per read that is established by a maximum number of sequencing cycles possible to be performed) produced by these ensemble-based second generation sequencing systems while still maintaining a suitable signal-to-noise ratio intensity; and, consequently, being significantly less than the average read lengths achieved by Sanger sequencing [151][158]. Furthermore, in relation to the bioinformatics analysis of the raw sequencing data, short reads represent a very significant and major challenge. As the processing of short-read sequencing data requires multiple over-lapping of sequences to achieve full gene and even partial gene sequencing. This becomes very challenging (even impossible in many cases) when establishing mapping and alignment of targeted sequencing reads in regions highly homologous and yet polymorphic as it occurs in many regions (especially non-coding) along the

different HLA loci (e.g. some combinations of *HLA-DPB1* alleles cannot be unambiguously phased because of the low SNP diversity shown across intron 2) [159]. In addition, as sequencing errors (due to decaying signal intensity detection errors in later cycles) tend to accumulate towards the read end, longer reads are in general better than shorter ones and can be trimmed near the end [184]. Also, second generation technologies tend to have reduced or completely absent coverage over DNA regions (mostly non-coding regions) with repetitive and extensive low-complexity and imbalanced sequence composition (e.g. short tandem repeats (STRs) or high AT- or GC-rich regions), due to amplification-bias, making also this assembly of reads very difficult [160]. In the context of high-resolution HLA typing via short-read NGS, the major problem is that incorrectly aligned fragments could result in HLA typing errors. Thus, it is possible that in a system as polymorphic as the HLA genes, incorrect phasing of polymorphic positions (or SNPs) that are distant to each other across the gene but otherwise show complete sequence homology could result in an incorrect allele being assigned. At the same time, rare or novel alleles formed by a recombination event may be missed if the consensus sequence analysis tools are biased towards the more common alleles and/or as these rare or novel allele sequences are not or only partially characterized yet in the available IPD-IMGT/HLA sequence database [159]. Therefore, since the short reads from the second generation sequencing platforms tend to generate relatively fragmented genome assemblies. The need of longer reads, in order to generate closed and fully completed reference genomes, has led to the development and introduction of the so-called “third generation sequencing” platforms [160]. Also, importantly and aside from the aforementioned limitation given by the assembly and alignment of short reads. Main current chemistries and engineering designs of second generation sequencing platforms are based on sequencing a large ensemble of DNA molecules with “wash-and-scan” techniques in a massively parallel fashion, where tens of thousands of identical strands per cluster are anchored to a given



location to be read in a process consisting of successive cycles of washing and scanning operations. Due to the large number of scanning and washing cycles required for this type of sequencing process, the time-to-result of raw sequencing data for these second generation sequencing methods is generally long, of the order of a full day (17-24h) or even several days depending on the sequencing throughput capacity used. Moreover, this period of sequencing time does not include the required time for doing the previous step of DNA template isolation process (~6-16 hours) and, also, for the post-sequencing step of the genotyping software analysis process (of the order of minutes to hours depending on number of samples per test). Thus, in comparison to Sanger sequencing, where results can be delivered in 2-3 days, at least 4-5 days are needed from sample entrance to the final report by using a NGS-based HLA typing approach. As for the histocompatibility clinical setting, although in living donation the test-to-result time associated with these second generation sequencing workflows may be acceptable, there are applications like HLA genotyping for deceased donors that require much faster total turnaround times (less than 7 hours to minimize ischemia time) which are not compatible with current short-read sequencing methods [161]. In this sense, the new developed sequencing chemistries and engineering designs of the “third generation sequencing” platforms have also shown the capacity to optimize and to drastically decrease the time (of the order of few hours even minutes instead of days) required for generation of raw sequencing data, by not depending on time-consuming cycles scanning and washing steps but based on approaches that directly interrogate single molecules of DNA [158][161].

III) Quite recently, approximately since 2010 (and even earlier), the so-called “third generation sequencing” platforms or “single-molecule sequencing” technologies (or even, “long-read sequencing platforms”) have been developed with potential for dramatically longer read lengths (~1-100 kb or even more), shorter time-to-result (only hours or minutes) and lower overall cost

than the “first-“ and “second-generation sequencing” technologies [158]. Despite the existing broad range of chemistry sequencing approaches and engineering designs, single-molecule sequencing technologies can be defined into three different categories: (i) non-step-wise SBS technologies in which single molecules of DNA polymerase are observed as they synthesize a single molecule of DNA; (ii) nanopore-sequencing technologies in which single molecules of DNA are threaded through a nanopore or positioned in the vicinity of a nanopore, and individual bases are detected as they pass through the nanopore; and (iii) direct imaging of individual DNA molecules using advanced microscopy techniques [158]. Most importantly, the majority of these novel single-molecule sequencing approaches presents a series of common and related characteristics that have enabled to overcome important limitations from the previous generations of sequencing technologies. In summary:

- 1) Application of novel biology/chemistry/engineering systems that allow production of extremely long sequencing reads:

In comparison to first-generation (with reads up to 1000 bp but with significant cis/trans ambiguities) and second-generation (where short reads offer high quality base calling on each position and phasing over short distances (<600 bp)) sequencing technologies; third generation sequencing platforms or single-molecule sequencing technologies radically increase read length from tens of bases to tens of thousands of bases per read. The breakthrough of these single-molecule sequencing technologies is based on the application of a variety of novel systems of different nature. The two main approaches that have been developed so far are the following:

- a) In the context of single-molecule SBS technologies, it has been possible to achieve high catalytic rates and high replicative processivity of DNA polymerase systems by selection of novel DNA polymerases from specific microorganisms in addition to further improvement

via protein engineering or directed enzyme evolution as well as incorporation of newly modified nucleotides for the ease of signal detection. Thus, while DNA polymerase systems used in previous first- (*Escherichia coli* DNA polymerase I proteolytic (“Klenow”) fragment was originally utilized in Sanger’s dideoxy chain-terminating DNA sequencing chemistry due to its specific efficiency for the incorporation of 2,3-ddNTPs; and later improved with the use of specific family A/B DNA polymerases from mesophilic/thermophilic bacteria/archaea due to their efficient incorporation of bulky dye-terminator for the ease of signal detection); and second-generation sequencing (with the use of other specific family A/B DNA polymerases from mesophilic/thermophilic bacteria/archaea efficiently incorporating reversible dye-terminator nucleotides in this case) technologies are not extremely optimal. Single-molecule SBS technologies present DNA polymerase systems with a super-high replicative processivity (thousands of bases in length) such as the enterobacterial phage  $\phi$ 29 derived DNA polymerase. Indeed, this particular strand-displacing DNA polymerase is used in single-molecule real-time (SMRT) SBS technology, as it also possesses an intrinsic unique capability to efficiently incorporate terminal phosphate-labeled nucleoside polyphosphates whose dye disposition shows a very low interference to the DNA polymerase activity allowing the observation of DNA synthesis in real time. In addition, it enables the resequencing of closed circular templates, increasing even more the output of number of reads [151][158][162].

b) As for the rest of single-molecule sequencing technologies which are not SBS-based. This group of platforms presents very sophisticated and complex electronic and/or optic designs of nanoscale dimensions with inherent unique biological (e.g. biological nanopores) and/or engineering-based physical systems (e.g. solid-state synthetic nanopores, scanning-tunneling and transmission-electron microscopy-based approaches or transistor-based technology). Due to the nature of these systems, each of these platforms allows the direct interrogation of single

molecules of DNA by imaging/measuring a certain structural feature of every nucleotide and chemically/physically detecting and identifying the complete nucleotide bases sequence comprising a DNA template. Furthermore, this type of systems also enables the generation of very long sequencing reads not depending, for example, on the enzymatic efficiency and processivity of a DNA polymerase system or on a chemical labeling step [158]. In this sense, under this group of single-molecule sequencing strategies, the biological nanopore-based sequencing technology, driven by electronics and not optics, has been one of the most developed and optimized platforms and it provides a high long read length profile that is very similar to that of single-molecule real-time (SMRT) SBS technology [163].

2) Direct sequencing of single DNA molecules without the need for PCR amplification:

Unlike most widely used SBS-based second generation sequencing technologies (which rely on PCR to grow clusters of a given DNA template in order to generate that large number of DNA molecules and, consequently, to provide a high-coverage and high-throughput capacity), single-molecule sequencing does not require routine PCR amplification for the generation of reads as raw sequencing data. Therefore, third generation sequencing platforms can overcome issues related to the systematic biases introduced by PCR amplification:

a) Amplification bias: non-uniform amplification of DNA that leads to over- or under-representation of some complex sequences (e.g. repetitive DNA, high sequence homology, or extreme AT or GC content) and/or, even, one of the two alleles (allele imbalance or drop-out) of the gene tested [152][158].

b) Dephasing error: As step-wise and ensemble-based second generation sequencing technologies present dephasing error (sequence reads of each cluster gradually diverge in length due to uneven PCR amplification efficiency during either DNA ligation or DNA

polymerization), they rely on precisely gating the identical processing of many DNA molecules during sequencing. Consequently, production of limited short read lengths is required to assure a minimum level of length synchronicity and, thus, DNA sequence quality within each clonal population or cluster for the ultimate interrogation of the sequencing result (as an example see **Figure I-19**). In contrast, single-molecule sequencing technologies directly interrogate each DNA molecule independently and not in an uneven PCR clustered configuration, thereby avoiding this problem of loss of synchronicity or dephasing, allowing the generation of long read lengths with uniform DNA sequence quality [158].

3) Workflow and design bases of single-molecule sequencing technologies considerably decreases time-to-result:

In comparison to first- and second-generation sequencing technologies (see **Figure I-20**), nanoscale third generation sequencing platforms drastically decrease time-to-result as they require minimal amounts of input material (theoretically only a single DNA molecule may be required for sequencing) and sample preparation to carry out a run. In addition, there are no time-consuming scanning and washing steps (even during DNA synthesis in the case of SBS-based single-molecule technologies) and minimal use of biochemical reagents, enabling time-to-result in a matter of minutes or few hours as opposed to days [158].

4) Generation of long reads by single-molecule sequencing technologies enables the tentative highly improvement of bioinformatics analysis of raw sequencing data [158]:

Bioinformatics analysis of the NGS raw sequencing data is a complex process that comprises several main steps, in general [296]: i) Filtering of raw sequence reads based on quality metrics of the sequencing software instrument; ii) Sequence alignment or mapping of filtered raw reads against the given reference sequence database to capture all SNPs and structural variance; iii)

Sequence assembly (reference-based or de-novo) and phasing (based on polymorphic linkage) of mapped reads in order to build and resolve consensus sequences or contigs (that define the series of mapped, assembled and phased reads); iv) Final sequence alignment or mapping of these built phased consensus sequences against the given reference sequence database to determine the best fit and to provide the final allele/genotype variant call. In this bioinformatics analysis process, the read length is a critical factor for the assembly step. At this step, assembly algorithms are used to align overlapping mapped reads, which allows the original genomic region of interest to be assembled into contiguous sequences. Assembly algorithms can be reference-based and consider a reference sequence as input, or can be de-novo and blind to any data beyond the sequence reads. Long read length at the scale of the one generated by single-molecule sequencing technologies (generally limited only by the sample preparation process) significantly facilitates and enhances both reference-based and de-novo assembly bioinformatics algorithms of very long genomic regions and even full genomes and, thus, it can provide a very high consensus sequence accuracy. Since the larger reads are easier to assemble due to more overlap between reads. Importantly, single molecule sequences spanning several thousand bases can also span repeats in the DNA sequence, allowing unambiguous location of the read on a reference, or overlapping the flanking contigs to unambiguously resolve the contig order and orientation (a bioinformatics process known as scaffolding) to reconstruct the original DNA sequence. Whereas contigs assembled from short-read data alone cannot be unambiguously ordered (especially in very homologous regions) because they overlap but do not span a repeat region. Therefore, in this respect, third generation of sequencing technologies offer clear advantages over previous generations of sequencing technologies, that can be summarized as:

- a) Direct detection of haplotypes (including challenging regions that are highly homologous (low SNP diversity) and yet polymorphic, presenting very distal polymorphic positions) and even, potentially, whole chromosome phasing thanks to long-read enhanced assembly algorithmic process during the bioinformatics analysis.
  - b) Detection of extremely rare or novel variant thanks to very high consensus sequence accuracy.
  - c) Construction and completion (filling the gaps) of full genomic sequence references based on all these aspects previously mentioned.
- 5) Single-molecule sequencing technologies present a lower overall cost, when sequencing very long DNA regions or even complete whole genomes at high fold coverage in comparison to the “first-“ and “second-generation sequencing” technologies.
- 6) Single-molecule sequencing technologies can detect DNA base modifications as part of sequencing [158]:

Single-molecule sequencing technologies can also detect chemical modifications to nucleotide bases (that, for example, constitute important epigenetic markers or signals of DNA damage products), such as methylation, in the normal course of collecting sequence data since they are able to detect changes either in the kinetics of incorporation or by direct inspection of a specific physical property that is associated with a given DNA base modification. In contrast, the majority of these modifications cannot be sequenced with ensemble-based second generation sequencing methods, and where possible require very complex chemical steps.

On the other hand, and very importantly, current long-read single-molecule sequencing technologies are still highly limited and restricted by a much lower throughput (referring to number

of reads generated per run as well as the degree of sample multiplexing), much higher error rate (at the individual raw read level) and much higher cost per base in comparison to short-read sequencing platforms. Nonetheless, several considerations are worthy to be mentioned in relation to these current limitations [158]:

1) Low-throughput: The unique nanoscale single-molecule scanning approach of third generation sequencing platforms intrinsically limits the throughput of this type of technologies. In relation to its engineering design, there is an ongoing optimization to maximize the number and extension of reaction sites while still maintaining this crucial nanoscale dimension. In addition, some single-molecule sequencing platforms allow certain approaches (e.g. by overloading the template molecule or by increasing the number of sequencing passes) to increase the number of reads (higher coverage) but unavoidably causing a decreased read length and, thus, complicating the posterior assembly bioinformatics analysis process. Furthermore, the degree of sample multiplexing per instrument is still significantly limited.

2) High error rate: single-molecule sequencing platforms show an inherent relatively high error rate (>10%) at the individual raw read level. Unlike previous generations of sequencing technologies, the error model here is stochastic given the random fluctuations that result from interrogating a single-template DNA molecule. Thus, to compensate this initial high error rate, this type of technologies systems are still able to sequence the same template molecule more than once in order to construct a consensus read (by aligning all the sequences from each template molecule) that is much more accurate than the original raw read. Nevertheless, complex assembly and error correction algorithms need to be employed in order to produce high quality assemblies. In addition, this error rate is mostly biased towards indels (insertions and deletions) being still difficult to accurately call long homopolymer stretches (so far, only 9-mers of the T nucleotide [159]) and repetitive sequences in some cases. Thus, once a



homopolymer stretch reaches a certain length, the exact number of nucleotides at such positions still become impossible to ascertain with current sequencing technologies [296].

3) High cost: due to their very low-throughput capacity, current single-molecule sequencing technologies present a much higher cost per base relative to short-read sequencing platforms. This may be decreased with future developments, but, at the moment, only second-generation sequencing workflows have been able to operate on a cost-covering basis (in relation to both amount of data generated and number of samples to be tested in a single run) despite the required high costs associated to reagents, automated robotics instrumentation and bioinformatics systems and support-storage logistics.

In addition, as this third generation sequencing still comprises very evolving sequencing platforms (at varying stages of development), the streamline, standardization and validation of workflows, informatics infrastructure, associated primary, secondary and tertiary analysis tools and, ultimately, data generated of these platforms are still at a very early stage. Thus, many more research studies still need to be carried out in order to prove the feasibility of these single-molecule sequencing for their validation and implementation into the clinical routine while also meeting the quality control (QC) and quality assurance (QA) requirements of the respective official laboratory regulatory institutions (e.g. International Organization for Standardization (ISO)) [164].

In the context of high-resolution HLA typing, the power given by long-read single-molecule technologies has been rapidly considered, although just very recently developed and tested, as the possible new definitive sequencing approach with the potential for [159][200]:

1) Obtaining complete full-length consensus sequences for all HLA loci due to long sequence reads (20 kb and even longer) can allow full coverage to encompass whole HLA class I (~4-5

kb) and class II (~10-17 kb) genes sequences. Therefore, having the potential for obtaining the maximum allele-level resolution HLA genotyping at the 4-field unambiguously.

2) Consequently, thanks to the long-read enhanced assembly algorithmic process during the bioinformatics analysis, having the potential of resolving the important HLA ambiguities found in previous sequencing generation platforms:

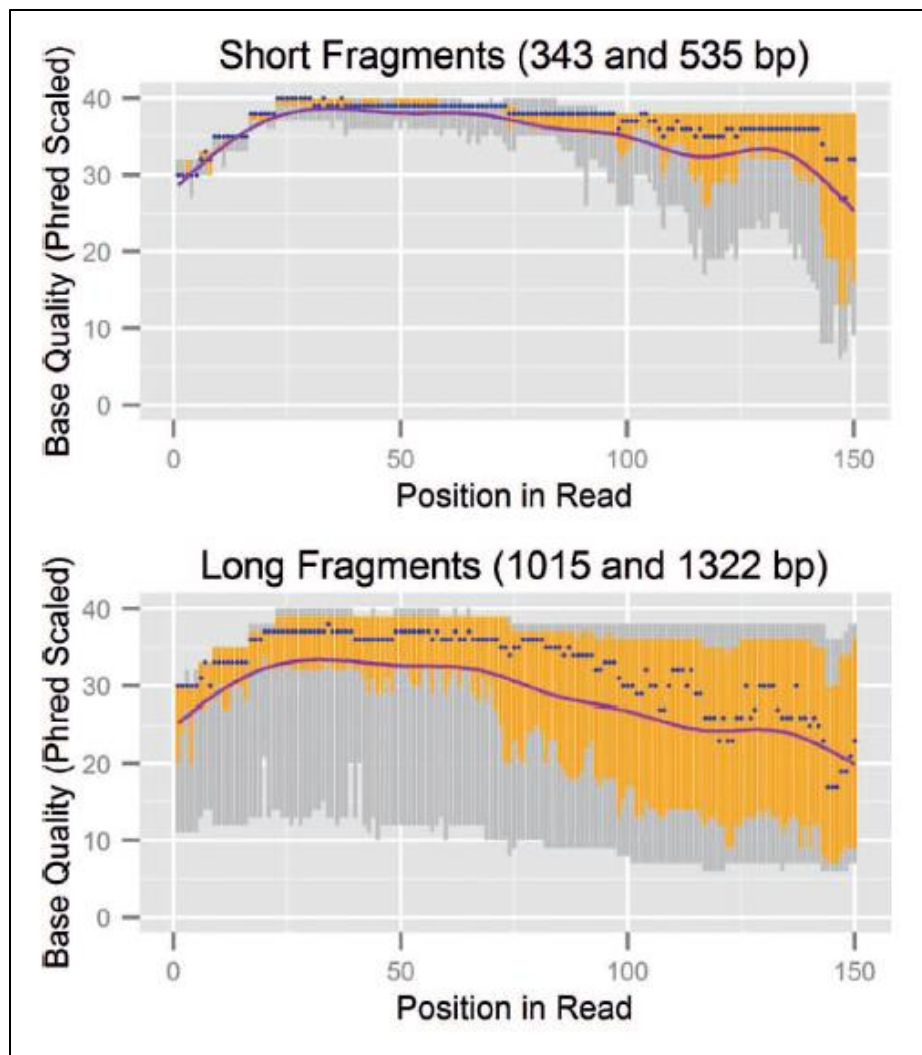
a) Sanger sequence-based typing (cis/trans phasing heterozygous ambiguities).

b) And short-read NGS platforms: presenting low coverage on DNA regions (mostly non-coding regions) with repetitive and extensive low-complexity and imbalanced sequence composition, such as: repetitive DNA regions (including homopolymer repeats poly(dA), poly(dT), poly(dG) and poly(dC)); regions of short-tandem repeats (STRs); or high AT- or GC-rich regions). Also in HLA sequence regions where de-novo assembly of shorts reads is very limited or even impossible to be done accurately, such as highly homologous sequences with very distal polymorphic positions (e.g. coding and non-coding sequences of HLA genes of the same class that are highly homologous between them; as well as nonfunctional HLA pseudogenes (e.g. *HLA-H*) with very similar sequences to functional HLA genes (e.g. *HLA-A*)). Thus, a HLA genotype may present two adjacent but distal SNPs or variants that cannot be linked by a single read, either because the distance between the variants is greater than the achievable read or read-pair length, or because they are located on different exons deriving from separate PCR reactions. To some extent, this phasing ambiguity can be mitigated by considering only exact matches to the allele reference repository and restrict results to those alleles during the bioinformatics analysis for HLA allele assignment. However, this approach will lead to spurious results if the sample harbors a so far undescribed allele that has evolved by recombination [296].


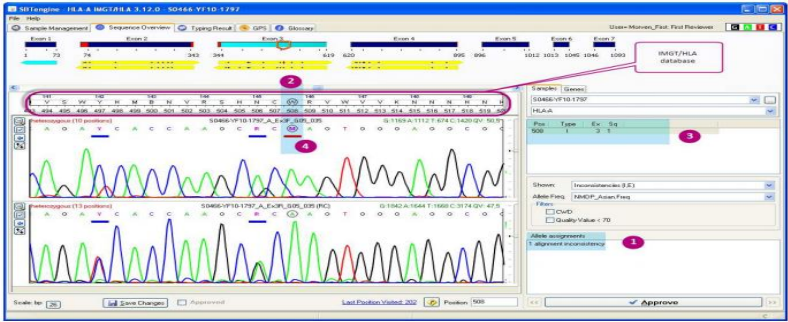

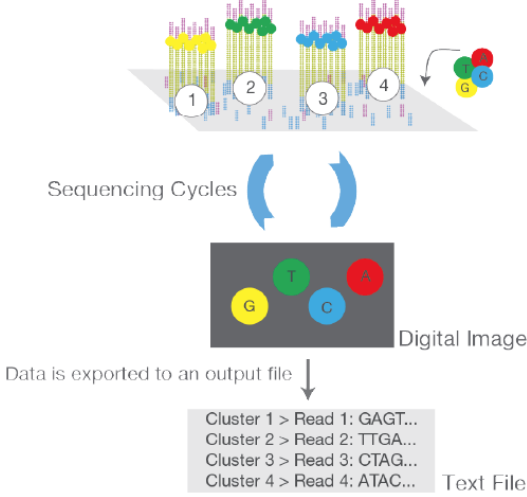

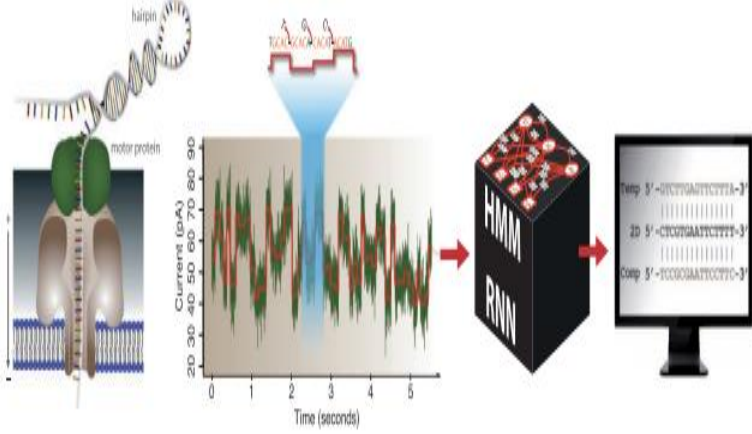
3) While second-generation short-read sequencing approaches have only the potential to obtain phased nucleotide sequence within each HLA locus, thanks to clonal sequencing property that allows setting phase for linked polymorphisms within each HLA gene (although with some important exceptions, especially in long HLA class II genes, such as *HLA-DPB1* and *HLA-DRB1*). Long-read single molecule sequencing platforms can provide very long sequences that can define entire full phased HLA haplotypes. Thus, allowing, for example, a better delineation of genes from pseudogenes and a potential absolute resolution of cis/trans ambiguities. Since long-read data, unlike short-read data, can overcome even challenging regions including those that are highly homologous (low SNP diversity) and yet polymorphic, presenting very distal polymorphic positions. It is expected that future analysis of the entire HLA haplotype region may allow more detailed understanding of the linkage of human MHC genes and its relation to their biological role (not only as individual genes but also as a single haplotype sequence) in different contexts such as pregnancy, transplantation or disease. Therefore, complete LD patterns could be totally described and understood within and, even, outside the human MHC region.

4) Also, importantly, with the potential to fill in the vast number of gaps of the current HLA allele sequence database, which is largely incomplete as only the 10% of submitted reference HLA alleles are completely sequenced [146][463]. Thus, a more comprehensive catalog of HLA allele reference sequences would be very beneficial for improving primer/probe design, current NGS bioinformatics HLA genotyping analysis software tools and, ultimately, in the evaluation of the role of HLA in all its different applications (such as mismatches in transplantation, studies of population genetics, the evolution of HLA system, regulatory mechanisms and HLA expression, genomic organization of the MHC, etc....).

5) In the clinical histocompatibility setting, the highly time-effective (of the order of minutes to hours) workflow of single-molecule sequencing technologies opens, for the first time, the possibility to consider a real “STAT” high-resolution HLA typing protocol for clinical needs as deceased donor typing for organ transplantation in a low-throughput manner [161].



**Figure I-19.** Base quality along the length of sequencing reads in the Illumina SBS system (as a representative example of 2<sup>nd</sup> generation of sequencing technologies). For each fragment size, box plots are generated showing the median (blue line), upper and lower quartiles (orange bands), and the 10% to 90% quantiles (grey bands) with a smoothed curve shown across the read length in purple. **(Top Graph)** Short (343–535 base pair [bp]) fragments demonstrate a better quality of base calling. **(Bottom Graph)** Long (1015–1322 bp) fragments demonstrate a comparatively lower quality of base calling, which drops further toward the end of the reads. Figure and respective footnote are obtained and adapted from [157].

| <u>Sequencing Technology</u>   | <u>Sequencing Instrument</u>  | <u>Sequencing Process/Analysis and Performance Basis</u>                             |
|--|---|--|
| <p><b>SBT</b></p>  |    |    |
| <p><b>Short-read Sequencing</b><br/><br/>(e.g.:<br/><b>Illumina SBS chemistry</b>)</p> |   |   |
| <p><b>Long-read Sequencing</b><br/><br/>(e.g.<br/><b>Oxford Nanopore</b>)</p>          |  |  |

**Figure I-20.** Scheme of the main three Generations of Sequencing Technology Systems:

\*Sanger Sequence-Based Typing (SBT) is considered the “first generation” sequencing technology. In the context of HLA genotyping, upon completion of sequencing, all the raw sequence files of the given sample are then loaded into the software analysis program and aligned with the IPD-IMGT/HLA reference sequences database. Specific positions in the sample sequence that do not align with the IPD-IMGT/HLA database are known as alignment inconsistencies. The technology of capillary electrophoresis-based dye-terminator Sanger sequencing is very limited, especially in terms of accuracy as it presents inherent diploid cis/trans ambiguities. Figure and respective footnote are obtained and adapted from:

<https://www.thermofisher.com/order/catalog/product/313001R#/313001R>

<https://www.gendx.com/downloads/IFU/GenDx%20SBTEngine%20IFU%20CE-IVD%20V3-2013-09%20M-13009%20EN.pdf>

\*Next-Generation Sequencing is considered the “second generation” sequencing technology and it is based on short-read sequencing approaches. Here, as an example, it is represented an overview of Illumina Sequencing By Synthesis (SBS) chemistry. Figure and respective footnote are obtained and adapted from:

<https://www.illumina.com/systems/sequencing-platforms/miniseq.html>

[https://www.illumina.com/documents/products/illumina\\_sequencing\\_introduction.pdf](https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf)

\*Single-Molecule Sequencing is considered the “third generation” sequencing technology and it is based on long-read sequencing approaches. Here, as an example it is represented an overview of Oxford Nanopore Technologies (ONT) chemistry. Figure and respective footnote are obtained and adapted from:

<https://nanoporetech.com/about-us/news/oxford-nanopore-announces-ps100-million-140m-fundraising-global-investors>

and [544]

Beginning around 2015, several research groups have been developing and reporting the first studies showing different approaches (depending on which are the amplicon-based targeting strategy and the commercial single-molecule sequencing platform used) for using long-read NGS platforms for high-resolution HLA genotyping [159][165-173]. Most of these studies have described a common targeted sequencing strategy (oriented to clinical practical purposes) based on an initial long-range PCR step (to specifically target and amplify these highly polymorphic HLA genes), followed by a simplified DNA library preparation step prior to the final single-molecule sequencing process in a very fast time-to-result workflow. Through this approach, all these studies have shown robust and accurate full-gene sequencing results for HLA class I loci [165-173][475]. However, mostly due to the inherent limitations given by this PCR-dependent approach (PCR associated errors/biases and limitations of coverage), full-length genomic typing

of HLA class II loci (substantially longer in length than the HLA class I genes) has not been achieved yet, being equivalent or even inferior to the one obtained by short-read NGS approaches. Thus, while a robust HLA class II typing strategy is still pending (although being already in the first stages of development [368]), some groups [172][173] have tried to compensate these current limitations (e.g. lack of full-length sequencing and complete phasing) by using a complex and laborious dual redundant sequencing strategy (e.g. using programs such as DR2S software, with the combined final analysis of HLA sequence genotyping data generated independently and in parallel by short-read and long-read sequencing platforms) that at least serves as a workflow, not for clinical routine HLA typing, but for submission of novel full-length alleles and characterization of sequences that are as yet incomplete (especially for the characterization of novel alleles harboring long intronic regions) [146][463]. Therefore, in contrast to short-read NGS-based HLA genotyping methods (which are relatively more simple, reproducible, robust, scalable, highly accurate and amenable to clinical and research testing), optimization of long-read NGS-based approaches is still required in order to be completely feasible for considering its future clinical validation and implementation [296][368].

## **9.2 Main Characteristics of NGS-based HLA Genotyping Workflow Approaches**

From a very broad general perspective, molecular- or DNA-based HLA genotyping workflow strategies can be categorized into two major groups:

- I) Targeted sequencing HLA genotyping strategies (based on direct and specific targeting of HLA genes from the genomic DNA to determine the HLA typing profile information) that, in turn, comprise:



1) Traditional legacy sequence-based typing methods: the previously described SSO, SSP, RT-PCR and SBT [76][85].

2) Novel NGS-based typing methods: including both short- and long-read sequencing platforms respectively and that have been previously mentioned in the introduction of the present thesis work [152][159].

II) Non-targeted sequencing HLA genotyping strategies (where HLA genotyping information is determined indirectly by whole genome/exome sequencing data approaches and inferred in-silico by bioinformatics analysis approaches using either genomic or exonic or SNPs data sources that potentially include or are linked to HLA sequence data) that, in turn, comprise:

1) HLA typing can be determined from the sequencing (applying NGS-based methods with very high-capacity instruments) of non-amplified genomic DNA using whole-genome sequencing (WGS) data [152][174]. Presently, as previously mentioned for single-molecule NGS sequencing approaches, this strategy does not generate sufficient amount of data to obtain a high enough depth of coverage for accurate HLA genotyping. Also, the current cost, turnaround time and logistical support required for WGS technologies are still not optimal and almost prohibitive for HLA typing, especially at the clinical setting [76].

2) HLA typing can be also determined from the sequencing (applying also NGS-based methods with very high-capacity instruments) of non-amplified complementary DNA (cDNA) (generated by reverse transcription of RNA into complementary DNA) using whole-exome sequencing (WES) data [152][174][175]. Nevertheless, the WES data is only comparable to the exon-based typing strategy (where polymorphism in non-coding regions is not defined) given by traditional legacy methods or even NGS-based methods. In addition, WES data shows relatively lower coverage expected for the HLA regions of interest causing

allelic imbalance or dropouts very often [76]. A similar approach as it is transcriptome sequencing (or the so-called RNA-Sequencing) also presents similar limitations.

3) In addition, HLA typing can be determined from dense SNP array data by using genotype imputation algorithmic methods and reference panels (datasets of which SNP variants are associated with which HLA alleles in previously genotyped samples of a given population and ethnic group) [528]. Where known comprehensive linkage disequilibrium information between single nucleotide polymorphism variants (SNPs) in the MHC region and specific HLA alleles is informative in predicting the HLA genotype. Thus, using a large number of SNPs in the neighborhood of classical HLA loci can potentially produce accurate inferences. In comparison to other approaches, the combined use of inexpensive array-based SNP genotyping and HLA imputation represents a very cost-effective strategy avoiding costs for wet lab-based HLA typing and thus renders association analyses of the HLA in large cohorts feasible (of special interest for large-scale disease-association studies and population studies) [176][177]. However, imputed HLA genotype calls are still much less accurate than those obtained by direct molecular sequencing approaches (both traditional and NGS-based) [142] especially due to important limitations presented by the imputation reference panels and the current knowledge of SNP data in the human MHC region. Since current imputation reference panels are based [176][177]:

- a) On LD data between the SNPs and the sequence variants located only (mostly) in the exons of HLA class I and class II genes (and not always for the totality of the classical polymorphic HLA genes), thus it is limited to 2-field allele resolution level.
- b) Only on the most common HLA alleles for a given population, where rare alleles are usually poorly represented and novel alleles cannot be detected.

c) On HLA data only from certain ethnic groups (mostly Caucasoid/European ethnicities). Thus, complete multi-ethnic HLA imputation reference panels have not been established yet.

d) Many imputation tools allow the imputation of *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* but only a few studies have reported on the imputation of the *HLA-DRB3*, *-DRB4* and *-DRB5* (*HLA-DRB3/4/5*) loci. Where, as previously mentioned, these genes can be present or absent in an individual depending on the *HLA-DRB1* genotype [56][75].

On the other hand, it is expected that high-resolution NGS-based HLA typing (including phased datasets from long-read technologies) data can significantly contribute for the development of more comprehensive and more accurate future imputation reference panels (since the discovery of variants via NGS will allow imputation-based analyses to take into account an increasingly extensive set of SNPs, including regulatory and intronic variants) [137].

Therefore, this second set of non-targeted sequencing HLA genotyping approaches can provide huge amounts of potentially valuable information more suitable for nonclinical applications such as large-scale population genetics studies, MHC-disease association studies and identification of de-novo MHC-linked histocompatibility loci [178]. But, at the moment, non-targeted sequencing HLA genotyping approaches are still less robust, less accurate, less comprehensive and less efficient time-wise and cost-wise than targeted sequencing HLA genotyping approaches. In contrast, these latter comprise methods that show high specificity and sensitivity for obtaining accurate HLA genotyping, with minimum level of ambiguities at the highest resolution (4-field), and that present a series of streamlined, standardized and validated lab-work protocols and data analysis bioinformatics systems of great importance in time constrained clinical diagnostics and applications. In this context, current targeted NGS-based HLA genotyping approaches have

become the considered new gold standard (especially for the clinical setting but also for HLA research purposes) for high-resolution HLA typing with minimum level of ambiguities, allowing also the detection of new HLA alleles as well as null and rare alleles [76][142][152][159][161][178].

All targeted NGS-based HLA genotyping workflows consist of the following general steps [76][157][296]: 1) DNA extraction and quantitation; 2) Template generation or preparation; 3) DNA sequencing library preparation; 4) Sequencing platform; and 5) HLA sequence data bioinformatics analysis process for assignment of HLA genotypes. In the following, the main technological/technical aspects and relevant metrics of each of these workflow steps are described:

1) DNA extraction and quantitation:

NGS-based HLA genotyping workflows are compatible with standard and conventional DNA extraction either automated or manual methods [76][157]. Importantly, NGS-based HLA genotyping workflows generally require less amount of genomic DNA (gDNA) per sample (ranging from 2.0 to 30.0 ng/uL) per test than other legacy sequence-based typing methods (SSO, SSP, RT-PCR and SBT) while covering and providing the genotyping data of more HLA loci in a single run. Nonetheless, many of these NGS-based methods are quite sensitive to low quality of gDNA sample (e.g. nucleic acid degradation, protein or RNA contamination or presence of anticoagulant chemical components). As this factor can interfere with the downstream PCR efficiency and data analysis affecting, consequently, the performance and robustness of this NGS-based HLA genotyping test. Nevertheless, several studies have shown how optimized NGS-based HLA genotyping methods can overcome these limitations to provide high quality and high-resolution HLA typing even in the most challenging cases such as when testing buccal swab samples (since buccal DNA is prone to nucleic acid degradation, presents

a much lower yield and is more contaminated with exogenous DNA from bacteria than that obtained from peripheral blood) [179].

## 2) Template generation or preparation:

Once the genomic DNA specimen is extracted with required optimal conditions, the first step of NGS-based HLA typing workflows is the targeted isolation and enrichment (referred here to as template generation or preparation) of the HLA-specific region of the genomic DNA (either the entire human MHC (currently only considered for large-scale research studies) or, more oriented for practical clinical purposes, full-length genes or only certain exons within HLA genes) prior to the library preparation for sequencing [76][152][157][178][296]. For target isolation and enrichment in highly polymorphic regions as the HLA genes, two major optimal approaches have been developed [76][184]:

a) Mid-/Long-range PCR-based target enrichment methods with the selection of target HLA sequence regions by PCR amplification (generating targeted templates termed as amplicons) that, in turn, can be divided into two main subgroups depending on the targeted amplicon size (which, at the same time, define the subsequent type of library preparation workflow and sequencing strategy that need to be carried out):

(i) Mid/Short-range amplicon-based (exon-based) sequencing strategy (when the targeted region is relatively small, amplicon size <500 bases) with the:

Multiplex PCR targeting of the clinically relevant and most highly polymorphic exons encoding for the antigen recognition domain (ARD) (exons 2 and 3 for HLA class I and exon 2 for HLA class II loci) and the surrounding genomic regions (e.g. exon 4 in HLA class I genes or exon 3 in HLA class II genes) [180-183]:

*Principle:* Designed “fused” region-specific primers with universal adapter/tag sequences are used to amplify target HLA regions by PCR. Thus, the resulting specific PCR HLA amplicon incorporates a universal adapter sequence at its ends. In a second PCR thermocycling step, there is an additional second primer pair that binds specifically to these universal adapter sequence ends. This second primer pair usually contains: a unique sample/amplicon-specific DNA barcoding sequence (or multiplex identifier tag (MID)), where the most common strategy used is the indexing by sample; also, possibly, a sequence that allows for specific final PCR enrichment of adapter-ligated DNA fragments only; and the specific-instrument sequencing adapters at the very ends. These latter instrument sequencing 5' and 3' adapters have important functions for the sequencing process, since they may hold forward/reverse primers (only required for paired-end sequencing platforms) and act as binding sequences for immobilizing the adapter-ligated library fragments to the respective sequencer hybridization sites system (flow cell or chip). To cover completely these most polymorphic exons of HLA genes with a maximum length of ~276 bp, these amplicon-based methods must be able to produce reads with at least 300-350 bp, and primers must be located close to the exons, which might be challenging (as these are high polymorphic sequence regions that could cause mismatches between primers and template easily) [184].

*Advantages:* In comparison to shotgun-based sequencing strategy, amplicon-based enrichment offers the advantage that DNA library is created without the need for subsequent manipulations such as fragmentation (only required for short-read sequencing platforms). Thus, their entire wet lab-workflow may be faster and simpler as well as the posterior bioinformatics HLA allele calling analysis [152].

*Disadvantages:* Apart from common PCR-based associated errors and the challenging design of primers to amplify the highly polymorphic genes of a multi-gene family with enough specificity (amplification of only one locus) while maintaining enough robustness (amplification of all alleles with comparable efficiency to minimize allelic imbalance or even complete allele “drop-out” due to inherent PCR preferential amplification or inefficiency). As with Sanger sequencing, mid/short-range amplicon-based approaches present important ambiguous genotyping results, due to genomic regions not targeted by the primers and to the inability to set phase for linked polymorphisms within the amplicons for heterozygous samples [152].

*Relevance:* Up to now, all PCR amplification (short-/mid-/long-range) strategies are the most suitable method for target enrichment in highly polymorphic regions like HLA genes even in comparison to capture hybridization-based strategies. In addition, although this particular mid/short-range amplicon-based approach may present the same level of allelic resolution as routine HLA Sanger sequencing (targeting most commonly only the ARS-exons), its workflow and NGS-nature allows a much more streamlined and higher sample/locus volume test than this legacy HLA typing method being suitable for the transplantation clinical setting [152].

(ii) Long-range shotgun-based sequencing strategy (when large targeted regions or amplicons of >500 bases in length are generated) with the:

Targeting of the full-length gene using long-range PCR, spanning most or the totality of the coding and non-coding genomic regions for each HLA loci [185-190]:

*Principle:* Large parts of the HLA genes or even the complete genomic sequence of the genes are specifically amplified in long-range PCRs followed by fragmentation of the

PCR product to sizes more appropriate for sequencing on NGS instruments (only required for short-read sequencing platforms). In this case, ligation of sequences that include instrument sequencing adapters and unique indexes MID-tagged adapters are added post-PCR prior to the final step of shotgun sequencing [184]. Alternatively, full HLA gene constructs, generated by long-range PCR, can be sequenced in whole on a given long-read sequencing platform [296].

*Advantages:* Long-range PCR method enables sequencing of (almost) the entire HLA genes providing a much higher resolution (lower ambiguity) genotyping than a targeted amplicon strategy for selected regions, being able to generate high-resolution typing results (not only at the 2-field, but also at the 3- and, even, 4-field) routinely [152]. Furthermore, this long-range PCR strategy gives the possibility to place primers in less polymorphic regions allowing for improved resolution of genetic differences and using only one set of primers per locus. Thus, exons of the same gene can be amplified in one fragment, decreasing variation in coverage as it happens in mid/short-range amplicon-based (exon-based) sequencing strategies [184]. Therefore, more polymorphic sites are sequenced to provide genotyping information of higher definition and the physical linkage between exons can be determined to resolve combination ambiguity.

*Disadvantages:* Apart from common PCR-based associated errors and the challenging design of primers with maximum coverage and minimal allelic imbalance or drop-outs (due to inherent PCR preferential amplification or inefficiency; being very common with long-range PCR, where PCR efficiency decreases with increasing amplicon length). Methods using long-range PCR require intact DNA at least several kilobases (kb) long because the amplicons for delineating the HLA genes can potentially range from 4 to 17 kb. Also, long-range PCR approaches require a more laborious and time-consuming



workflow (e.g. longer PCR times and a complex DNA library preparation with several steps and manipulations of long amplicons). Nevertheless, very recent developments, performed by different vendors now, have substantially increased the multiplexing capacity of this initial PCR amplification (e.g. single high-multiplexed PCR set-up for simultaneous amplification of all major HLA class I and class II genes per sample, removing the need for multi-amplicons pooling for each sample) and thus, consequently, simplifying the following DNA library preparation plus minimizing the overall protocol time (less than 3 days). This is due to all locus-specific primers are multiplexed in a single tube for amplicon generation, followed by an innovative library preparation process that allows the pooling of all samples into a single tube during the first steps of library preparation [191][483][484]. Moreover, although long-range PCR of HLA class I genes is relatively easy to design and perform, for HLA class II genes it is more difficult because of their large size and the relatively high GC content. For instance, *HLA-DRB1* locus has a very large intron 1 (about 10,000 base pair) that is difficult to amplify as a whole by PCR. To overcome these difficulties, 2 primer sets are usually designed to amplify the desired gene in sections or, alternatively, designing primers that skip intron 1 to amplify the gene starting with exon 2. However, this lack of coverage influenced by the experimental long-range PCR design limitations significantly increases the level of typing ambiguities for these loci [192]. Therefore, current long-range shotgun-based sequencing strategies do not present 100% unambiguous HLA genotyping results for all HLA loci (especially longer HLA class II genes) yet.

*Relevance:* Long-range PCR approach allows to optimize and maximize the entire capacity of NGS-based methods at a large-scale in terms of test volume. In contrast to amplicon-based strategies, long-range PCR approach has the full potential to provide

complete, minimally ambiguous and very high-resolution HLA typing in addition to complete genomic characterization of novel/rare/null HLA alleles and the completion of sequence for existing, partially sequenced alleles on the IPD-IMGT/HLA database. Therefore, at the moment, this strategy has become the most widely developed (encompassing the large majority of vendor-supported and/or in-house NGS-based HLA genotyping kits), standardized, validated and practiced showing a high feasibility and robustness for its routine use in clinical and research applications [192].

In the context of mid-/long-range PCR-based methods, during this step of template generation or preparation certain quality metrics and aspects need to be considered. Robust and reproducible amplification of every targeted locus and of every sample is required, where the efficiency and success of the used multiplexed set of primers need to be evaluated and confirmed after every PCR run. Thus, the common approach to confirming and quantifying amplicons is to use gel electrophoresis to detect amplification success/failure (usually per HLA gen per sample). After confirmation of the amplicons, there is a quantification step (using methods based on ultra-sensitive fluorescent nucleic acid stain for quantitating double-stranded DNA (dsDNA) from DNA amplification products) and, usually, a subsequent normalization step (either balancing by simple equimolar pooling of amplicons per sample or balancing by using paramagnetic beads for the collection of comparable amounts of DNA between amplicons for each sample). Since NGS is based on massive parallel sequencing of a vast number of clones with multiple loci of many different samples in a single run. This is why confirmation, quantitation and normalization of initial amplicons are important steps in order to optimize an equal representation (which is also essential during the posterior bioinformatics analysis stage) of all genes/amplicons in each of the samples that constitute the final DNA sequencing library [157]. In some NGS-based HLA typing workflows, there is

a bead-based clean-up step prior to the normalization step in order to get rid of primer-dimers (and/or adapter-dimers) and any other PCR sub-product or reagent that are not of interest and, thus, isolating the targeted HLA amplicons.

b) Hybridization-based capture methods, with the enrichment of target HLA sequence regions by complementary oligonucleotide (oligo-based) hybridization, represents an alternative to PCR-based target enrichment methods [193-196]:

*Principle:* A high number of biotinylated DNA or RNA oligonucleotides or baits of specific length (~55-120 bp) are designed specifically complementary to the target regions. As a specific probe panel, these baits are then hybridized (via either in-solution hybridization capture approach (known as region specific extraction, which is based on pull-down of hybridization oligomers attached to magnetic beads); or solid-phase hybridization capture approach that can be array-based or presenting high-density on-chip baits) to the adapter-ligated and fragmented (through nebulization, sonication, enzymatically using endonucleases or transposons-based methods) genomic DNA library and bind to their respective target sequences. Adapter-ligated fragments bound by the baits are then enriched by streptavidin coated micro magnetic beads. These enriched fragments are subsequently amplified by PCR and further processed/purified to obtain a final DNA sequencing library that is sequenced in a massively parallel fashion as well [184].

*Advantages:* It is generally applicable for NGS-based target sequencing of larger genomic regions and a larger number of genes than the PCR-based methods, presenting a high-multiplexing capacity [142][178].

*Disadvantages:* Target sequencing of the HLA genes using the sequence capture method has not been well developed compared with PCR-based HLA typing. Furthermore, PCR-

based strategies have the potential for higher throughput of sequencing reads and thus, lower cost per sample. This hybridization-based enrichment requires an expensive and high-quality-balanced probe-pool to cover all of the allelic variations of the targeted HLA genes. As some regions are better captured than others, the main difficulty of this method is to achieve a homogeneous coverage along all targeted genes and alleles per locus. In addition, design of probes is limited to the current knowledge of the described HLA sequence variation in the incomplete reference databases (IPD-IMGT/HLA) [146][463]. These drawbacks make it difficult for large-scale and routine high-resolution HLA typing [142][178].

*Relevance:* As previously mentioned, it is notable that the hybridization-based capture method is generally applicable for NGS-based target resequencing of larger genomic regions (e.g. entire human classic MHC region (~4 Mbp); or the ~1 Mbp Leukocyte Receptor Complex (LRC) located on chromosome 19 (19q13.4); or within this LRC, there is the KIR region (~150 Kb)) and a larger number of genes than the PCR-based methods. Therefore, currently, this strategy appears to be more suitable for nonclinical applications such as large-scale population genetics studies, MHC- or KIR-disease association studies and identification of de-novo MHC-linked histocompatibility loci as well as the evaluation of the LD at the KIR region.

### 3) DNA sequencing library preparation:

In the context of mid-/long-range PCR-based methods, after the generation, confirmation, quantitation and normalization of initial targeted HLA amplicons per sample there is a post-PCR DNA sequencing library preparation step that consists on several processes depending on the targeted sequencing strategy and the type of sequencing platform selected [76][157][192]:

a) Mid/Short-range amplicon-based (exon-based) sequencing strategy using short-read sequencing platforms:

(i) Consolidation of samples into a single tube creating a whole double-stranded DNA (dsDNA) sequencing library. Where each sample consists on a normalized pool of adapter-ligated (including unique sample (more used than by amplicon) barcoding sequence, sequence for specific final PCR enrichment and sequencing platform-specific sequence) cleaned amplicons of all the targeted HLA exonic regions. On the other hand, a very few recent protocols are based now on a single multiplex reaction which amplifies all major class I and class II loci, completely removing the need for amplicon pooling. Samples are transferred directly to library preparation and made ready for sequencing in a single workday [191][483][484].

(ii) Post-consolidation purification or clean-up of adapter-ligated dsDNA library: there may be possible additional steps of bead-based clean-up of the consolidated DNA library.

(iii) Size-Selection of adapter-ligated dsDNA sequencing library amplicon fragments: the proper selection of size of the amplicon fragments from the DNA library (obtaining a final adequate representations of both smaller and larger fragments (uniform size distribution) for optimal library preparation) secures optimization of the sequencing run, increases the number of samples sequenced, provides high-quality sequencing data and maximizes possible phasing. In short-read sequencing platforms, when the size of the fragment increases, consequently, the sequencing efficiency decreases. Thus, smaller fragments have higher quality sequencing data than larger fragments (although they provide distal phase information not available from smaller fragments). In general, for short-read sequencing platforms, fragments of 300 to 500 bases are sequenced efficiently. Whereas the very small

(<150 bp) and very large fragments (>0.7-1.3 kb) have the ability to interfere with the sequencing of the most optimal intermediate-sized fragments and should be excluded before sequencing. Also this size selection step allows to remove adapters and/or adapter dimers from the libraries as they decrease the availability of library hybridization sites for subsequent clonal amplification due to competitive binding. Commonly used size selection strategies include bead-based technologies or gel electrophoresis-based technologies. Bead-based methods have the advantage of simultaneously concentrating the pools, while electrophoresis-based methods provide better precision.

(iv) Final PCR enrichment of size-selected adapter-ligated dsDNA library: the size-selected DNA library is amplified by PCR with specific-sequencing platform primers that may contain the sequences necessary for cluster generation (if they have not been added previously). Here, there is also a post-PCR bead-based clean-up step.

(v) Final quantification of cleaned PCR-enriched size-selected adapter-ligated dsDNA library: generally, DNA library final quantitation can be done using a fluorometric measurement of DNA or more sensitively by real-time quantitative PCR (qPCR). The real-time qPCR assays can use the ligated adapter sequence as the priming site for amplification; therefore, only DNA molecules that have successfully incorporated adapters at both ends will amplify. Nonspecific intercalating fluorescent dyes such as SYBR green are then used to detect the amplification in “real-time.” Concentration of the library is determined by comparison of fluorescence to that of a standard curve. Thus, the advantage of using real-time qPCR is that the primers are complementary to the instrument specific adapters and, therefore, only adapter-ligated libraries are quantified thereby providing useful information about the robustness of the library preparation process (as only those fragments with both adapters will eventually be sequenced). Alternatively, automated electrophoresis-based

instruments for DNA library quality control can determine both size distribution of the library and concentration. However, the concentration measurement from these systems does not accurately represent the library concentration, as it measures all double-stranded DNA that is present and cannot differentiate between completely ligated fragments containing both adapters and fragments that are missing adapter sequences.

(vi) Cleaned PCR-enriched size-selected adapter-ligated dsDNA library template final preparation and loading onto the short-read sequencing platform instrument: a process that generally consists on a conditioning treatment including denaturation of dsDNA fragments to obtain single single-stranded (ssDNA) fragments followed by dilution with a sequencing buffer that optimizes the final ssDNA library concentration for a balanced loading and an efficient immobilization process of the ssDNA library fragments to the respective sequencer hybridization sites system (flow cell or chip). Since the overloading of the DNA library onto the short-read sequencing platform instrument results in poor template generation and low sequencing owing to the clusters being too compacted. While the underloading of final DNA library material wastes the flow cell/chip (not using all its optimal clustering capacity) and, consequently, fewer data can be generated.

b) Long-range shotgun-based sequencing strategy using:

b.1) Short-read sequencing platforms:

(i) Fragmentation of cleaned normalized pools of long dsDNA amplicons per sample (although a few very recent protocols start directly with a single tube [191][483][484]): the large PCR amplicons must be fragmented to sizes more appropriate and compatible for sequencing on short-read NGS instruments (since they cannot sequence fragments larger than about 700 bases, which are not cloned and sequenced as efficiently, presenting

also a too high dephasing error). Breaking large amplicons into smaller fragments is accomplished through either nebulization or sonication or enzymatically using two endonucleases system (the first enzyme randomly generates nicks on dsDNA and the second one recognizes the nicked sites and cuts the opposite DNA strand across the nick, producing dsDNA breaks) or using transposons-based methods. Fragmentation typically results in dsDNA fragments with short overhangs of 5'-phosphates and 3'-hydroxyl groups. DNA fragmentation needs to be performed in a measured way (optimizing fragmentation treatment conditions to avoid over- or underfragmentation), so there are adequate representations of both smaller and larger fragments for optimal library preparation.

(ii) End-repair (blunt-ending and dA-tailing): the ends of the fragmented dsDNA amplicons are then enzymatically blunt-ended and in some cases 3' adenylated in preparation for adapter ligation.

(iii) Adapter ligation: this step involves attaching adapter sequences (with an overhang of a single T base) to the A-tailed ends of the fragmented end-repaired HLA dsDNA amplicons per sample. An adapter sequence will potentially consist on three components with different purposes:

-Unique sample/amplicon barcoding sequence or index: the use of indexed adapters enables pooling of multiple samples and loci in a single run. There are two primary indexing strategies: indexing by locus or indexing by sample. In the first strategy, a single library is prepared with a unique index for each HLA locus for a single sample. The sequences obtained with this index can then be definitively assigned to a particular locus (having more reads per locus than in the indexing by sample strategy). The primary



benefit of this strategy is that software programs used to align sequence data and assign HLA genotypes can be aided by this information and can guard against misalignment of reads to an incorrect locus. The second approach, indexing by sample, offers a much simplified sample preparation process that also reduces cost. Thus, the massively parallel sequencing ability of NGS platforms is best exploited by the sequencing of many samples (each loci for each sample has the same index) in a single run through the use of indexed adapters. That is why indexing by sample is most commonly used instead of indexing by amplicon. In this strategy, amplicons of all loci from a single sample are normalized and pooled prior to library preparation. The indexing here is simply used to distinguish one sample from the next. Nevertheless, this amplicon pooling strategy requires sophisticated bioinformatics HLA software programs that accurately obtain genotyping when cross-mapped reads from different loci may be present.

-Sequence for specific final PCR enrichment: which allows for specific final PCR enrichment of adapter-ligated DNA fragments only.

-Sequencing platform-specific sequence: platform-specific sequencing 5' and 3' adapters have important functions for the sequencing process, since they may hold forward/reverse primers (only required for paired-end sequencing platforms) and act as binding sequences for immobilizing the adapter-ligated library fragments to the respective sequencer hybridization sites system (flow cell or chip).

(iv) Most, but not all, shotgun sequencing workflows include one or more purification or clean-up steps (typically accomplished with bead-based systems) to remove enzymes and other reactants between the fragmentation, end-repair, and adapter ligation steps. Nonetheless, recent developments have made possible to integrate these three enzymatic

processes (fragmentation, end-repair and adapter ligation) into the same thermocycling reaction and thus, consequently, simplifying the following DNA library preparation plus minimizing the overall protocol time [197].

(v) From this point on, the same NGS experimental workflow steps (i) to (vi) occur as previously described in a) Mid/Short-range amplicon-based (exon-based) sequencing strategy using short-read sequencing platforms. It is noteworthy that when using shotgun sequencing, it is also important to specifically select for library sizes that are ideal for clonal amplification on the short-read NGS platforms (300 to 500 bp) while, at the same time, being as long as possible for phasing distal polymorphic HLA positions.

b.2) Long-read sequencing platforms:

(i) In this context and in comparison to the workflow described before in b.1) Long-range shotgun-based sequencing strategy using short-read sequencing platforms:

Workflows of long-range shotgun-based sequencing strategy using long-read sequencing platforms show many similarities and share all the main steps as previously described in b.1), in general:

- Initial long-range PCR generating targeted HLA amplicons.
- Confirmation, quantitation and normalization (using typically the amplicon pooling strategy, with the equimolar pooling of the HLA amplicons per loci of the same sample).
- A series of steps (with also intermediate bead-based clean-up steps in between) including: barcoding of samples with unique indexes, end-repair (blunt-ending and dA-tailing), consolidation of samples into a single tube and adapter ligation (adding to the end-repaired amplicon fragments the long-read sequencing platform-specific

sequences) creating a whole adapter-ligated double-stranded DNA (dsDNA) sequencing library.

-Thus, steps such as fragmentation or PCR enrichment of adapter-ligated dsDNA library are not required in this case. This is because long-read sequencing platforms enable direct sequencing analysis of entire long single DNA molecule where no cluster generation is required.

-At a certain stage of this workflow (generally prior to consolidation of samples), a size-selection step (using bead-based technologies or gel electrophoresis-based technologies) may be included in order to obtain a final size-selected adapter-ligated dsDNA library that only presents very long fragments to maximize the potential offered by these third-generation single-molecule long-read sequencing platforms (that are able to produce long read lengths of 20 kb and even longer).

#### 4) Sequencing platform:

Up to now, the sequencing platforms that have been used in NGS-based HLA genotyping workflows can be broadly classified depending on the fragment-length that is read and phased [76][152][157][158][159][163][184][192][198][199]:

a) Short-read (ranging from ~25 to 600-800 bp) sequencing platforms or second-generation sequencing technologies for both amplicon-based and shotgun-based DNA libraries, which mainly differ in the following characteristics:

In general, DNA libraries contain the targeted HLA sequences and “adapter” sequences that allow capture of single molecules by an oligonucleotide immobilized to a bead/sphere particle or a sequencer slide/chip surface system in a massively parallel fashion. The captured single

molecules are then clonally amplified and the parallel clonal populations or clusters sequenced. Second-generation short-read sequencing application to HLA genotyping, by virtue of massively parallel sequencing and clonal DNA amplification, is able to provide high-resolution allele-level typing with minimal ambiguity, high coverage of HLA sequences for several loci and a high-throughput capacity for testing a large number of samples simultaneously.

a.1) Roche 454 systems (started in 2005):

(i) Clonal amplification: DNA library fragments are clonally amplified on beads upon an emulsion PCR (emPCR). In emPCR, the DNA library is diluted and stoichiometrically added to capture beads such that a single ssDNA molecule binds to a single bead. The ssDNA binds to beads via adapter sequence with homology to oligonucleotides that are bound to the bead surface. The DNA-bound beads, along with biotinylated primers and other amplification reagents, are placed into an oil emulsion and are shaken in a controlled way such that a single DNA-carrying bead becomes encapsulated in a single micelle droplet. Amplification is thereby performed in isolation within each micelle (microreactor), producing a bead covered with many copies of DNA with identical sequence. After emPCR, the beads are recovered and enriched using streptavidin-coated magnetic microparticles.

(ii) Sequencing: Roche 454 sequencing platform is based on single-end (library fragment is sequenced exclusively from one direction) sequencing by synthesis (SBS) pyrosequencing reactions that take place in a picotiter plate with wells containing each a previously isolated single DNA-carrying bead (in turn, each bead contains millions of copies of the respective original single-stranded DNA molecule). Although this system

presents single-end sequencing, in order to yield phased data: from each of the original dsDNA library molecules that are loaded, one strand is captured by one set of beads and the other strand is captured by another set of beads so that the two strands of the target DNA are both sequenced but in different wells. During a given pyrosequencing reaction, once a nucleotide is incorporated by the DNA polymerase, the released pyrophosphate (used as an indicator of specific base incorporation) is enzymatically (with the use of a sulfurylase and a luciferase) transformed into a light signal (photon) that can be recorded by a charge coupled device (CCD) sensor.

(iii) Current relevance for HLA-sequencing: despite this technology is able to generate reads of up to ~550 bases for amplicon sequencing and it presents a maximum multiplexing capacity for the analysis of 380 samples in one GS FLX instrument run. Due to a number of limitations (e.g. low throughput of number of reads, high reagent costs and high error rate related to insertion-deletion (indel) variants and long homopolymer regions (since the signal strength (based on optical detection) does not increase linear with growing homopolymer length)) the Roche company ended the production of these 454 platform systems in 2016.

a.2) Illumina platforms (started in 2006):

(i) Clonal amplification: in Illumina platforms, instead of emPCR, clonal amplification is based on a solid-phase (since all of the enzymatic processes and imaging steps of the Illumina technology take place in a flow cell slide system) process called bridge PCR amplification. In which, firstly, denaturated adapter-ligated ssDNA library molecules are hybridized or captured onto the flow cell surface system through one of the two oligonucleotides (forward and reverse oligos for amplification (one with a cleavable site))

that are complementary to the adapter sequences introduced during the DNA library preparation step. These two oligonucleotides are pre-bound onto the flow cell surface at a concentration that allows spatial separation between captured immobilized single original adapter-ligated ssDNA library molecules. These capture oligonucleotides also act as amplification primers, priming each captured ssDNA library fragment for an initial amplification in each immobilization site. Thus, a respective initial copy of each of the individual sequencing template molecules is generated, obtaining respective dsDNA products, one per immobilization site. Then, each of these dsDNA molecules is denatured and each of the initial original DNA strand library molecules is removed. After that, each of the remaining copied flow cell-attached strands is clonally amplified through bridge PCR amplification in an isothermal amplification program, generating, each of them, a localized clonal cluster of identical template molecules. In this isothermal amplification program, multiple cyclic alternations of three specific buffers take place and they mediate, respectively, the denaturation, annealing and extension steps at 60 °C. In closer detail, during this bridge PCR amplification process, performed in a cyclic parallel fashion, for each of the initial remaining copied flow cell-attached strands:

-Firstly, the initial copied strand folds over and the 3' adapter region hybridizes (annealing step) to the nearby second type of oligonucleotide on the flow cell creating a bridge structure.

-Then, DNA polymerase generate the complementary strand forming a double stranded bridge (extension step).

-After that, this double stranded bridge is denatured, resulting in two single stranded copies of the molecule that are tethered to the flow cell (denaturation step).

-This process is then repeated over and over, and occurs simultaneously for millions of clusters resulting in clonal amplification of all the fragments.

-Upon completion of all the cycles of denaturation-annealing-extension of this bridge PCR amplification process. The final step (right after the last extension step) consists on, firstly, the removal of one of the two strands of each of the generated dsDNA fragments through the cleavable site found in one of the two flow cell surface oligonucleotides (thus, reverse strands are cleaved and washed off, leaving only the forward strands); and, secondly, the blocking of all 3' ends of the remaining forward strands with ddNTP to prevent unwanted priming (the otherwise open 3' ends to act as sequencing primer sites on adjacent library molecules). At this point, for each of the clusters, all the generated forward ssDNA library strands are ready to be sequenced.

The clonal generation of clusters is necessary to generate sufficient signals for detecting the sequencing reaction, as a minimum of signal intensity is necessary before it is detected by the camera of the instrument. Optimal cluster density depend not only on the concentration of the library, but also on the length of the molecules. Short molecules yield clusters with a small area that are denser and therefore generate more intense signals. Loading a wide fragment size distribution will generate clusters varying widely in size and signal strength which may impair the number of passing filter reads. These final clonal clusters, each produced from different, single DNA fragments, are then sequenced.

(ii) Sequencing: the Illumina sequencing by synthesis (SBS) technology is based on a cyclic reversible termination (CRT) sequencing chemistry (using reversible terminator (RT) nucleotides) with a unique paired-end sequencing mode. The RT nucleotides consist on modified dNTPs protected at 3'-OH groups (2-cyanoethyl) and that are fluorescently

labeled with a specific single dye (Illumina 4-channel SBS detection method) or a mix of dyes (Illumina 2-channel SBS detection method). In this Illumina sequencing process, each of the attached ssDNA library fragments from each of the clusters on the flow cell is read simultaneously (in a massively parallel process), one nucleotide at a time per cycle in repetitive cycles. Thus, the number of cycles determines the length of the read in each direction. As Illumina SBS technology presents a unique paired-end sequencing mode, this means that each attached ssDNA library fragment is sequenced from both the ends (bidirectional read through for a target region), defining Read # 1 (R1) or Forward read and Read # 2 (R2) or Reverse read. For instance, paired-end 2x150 bp reads in a 300 cycles sequencing run means that the given Illumina SBS technology kit/flow cell is able to generate sequencing paired-end reads (in both directions, forward and reverse) of 150 bp in length of each attached ssDNA library fragment that is sequenced. Where the first 150 sequencing run cycles (one nucleotide at a time per cycle) correspond to the sequencing of the forward ssDNA library fragments (Read #1); whereas the second left 150 sequencing run cycles (one nucleotide at a time per cycle) correspond to the sequencing of the reverse ssDNA library fragments (Read #2). In closer detail, Illumina SBS paired-end sequencing by reversible termination consists of these following steps:

- For the generation of the Read #1 or Forward read of each of the clusters:
  - First step: incorporation of the complementary RT nucleotide by DNA polymerase to each of the respective forward ssDNA templates attached to the flowcell.
  - Second step: imaging detection of the different fluorescence signals for a given cycle used as an indicator of specific base incorporation through two possible detection approaches: a separate dye for each base (Illumina 4-channel SBS detection method,



presented in Illumina MiSeq or HiSeq sequencers); or a mix of dyes for some of the bases (Illumina 2-channel SBS detection method, presented in Illumina MiniSeq or NextSeq 500/550 sequencers, that allows shorter sequencing (also reducing fluorophore usage) and data processing times comparatively; nevertheless, 2-channel systems present a more ambiguous base discrimination which causes slightly higher error profile and underperformance for low-diversity samples).

-Third step: restoration of free 3'OH group of all the incorporated RT nucleotides in a given cycle by cleaving the terminating moiety and reporter dye molecule. As a result, the template strands are ready for the next incorporation cycle. Thus, the sequence of each forward DNA template is read by following the fluorescent signal per cycle extension step repeatedly for each cluster.

- For the generation of the Read #2 or Reverse read of each of the clusters:

-After the completion of the first read, the read product is denatured and washed away. Then, the 3' ends of the respective forward ssDNA templates (already sequenced) are deprotected. This allows that a single bridge PCR amplification event can occur on each of the deprotected forward ssDNA templates. Thus, forward ssDNA templates fold over and the 3' adapter region hybridizes (annealing step) to the adjacent second type of oligonucleotide on the flow cell creating a bridge structure in each case. DNA polymerase extends (extension step) these second flow cell oligonucleotides forming double stranded bridges. Double stranded DNA molecules are then denatured and linearized and the 3' ends are blocked. The original forward strands (already sequenced) are now cleaved off and washed away leaving only the complementary reverse strands.

-After this single bridge amplification PCR event and isolation of reverse ssDNA library strands. Generation of the Read #2 or Reverse read of each of the clusters take place through the same series of sequencing by reversible termination steps as previously described for generation of Read #1, completing the paired-end Illumina sequencing process.

-Although it is not described in detail here, during the Illumina SBS paired-end sequencing process and immediately after the generation of Read #1 of each of the clusters as well as after generation of Read #2 of each of the clusters, the different respective specific indexing sequences of each sequenced fragment are also detected by the sequencer system to later identify and assign the reads generated with each of the barcoded samples/HLA amplicons (demultiplexing). In some instances this identification process of the reads is not done directly by the sequencer but by the posterior HLA calling bioinformatics software tool.

Unlike Sanger sequencing (based on dideoxynucleotide or chain termination sequencing chemistry), Illumina sequencing by synthesis (SBS) technology prevents multiple extension events by using exclusively this reversible termination chemistry during the sequencing process. Consequently, although this reversible termination chemistry may slow the sequencing process, it provides very accurate base detection, reducing the intrinsic error rate, at the individual raw read level (ranging from 0.2% to 0.8% error rate). Furthermore, homopolymer and indel associated sequencing errors are minimized by this technique due to highly controlled incorporation of single base at a time (“letter-space” based nature), as the terminating moiety needs to be removed first before the addition of another base. Therefore, in comparison to other short-read sequencing platforms (Roche 454 (1.0% error rate) or Ion torrent (1.78% error rate)) and long-read sequencing platforms

(>10% error rate), the Illumina sequencing by synthesis (SBS) technology currently shows a very high sequence quality and a very low error rate [198].

Also, a distinctive feature of the Illumina SBS chemistry is the inherent capacity (via bridge PCR amplification) to perform paired-end sequencing runs. Importantly, paired-end sequencing allows to sequence each DNA fragment from both the ends resulting in high coverage, high numbers of reads and more data as compared to single-end sequencing systems (such as Roche 454 and Ion Torrent platforms). At the same time, paired-end sequencing may cause that the two distinct sequencing reads performed (one from each end of the template DNA library fragments) may be separated by a stretch of unsequenced DNA of distinct length termed insert. This is because in short-read sequencing platforms the read length is limited and most typically shorter than the actual DNA fragment size of the library. Nevertheless, this is not a drawback. Since insert size distributions can provide valuable information to infer structural variations or in de-novo genome sequencing (by producing longer contigs for de-novo sequencing by filling gaps in the consensus sequence). Alternatively, the paired-end reads may overlap to join the two distinct reads together in order to form a longer, continuous read. This joined read can then for example be employed in amplicon sequencing to generate reads that encompass the complete amplicon length. Furthermore, paired-end sequencing facilitates detection of genomic rearrangements (insertions, deletions, and inversions), repetitive sequence elements and gene fusions.

(iii) Current relevance for HLA-sequencing: based on its very informative paired-end sequencing reads, high sequence quality and very low error rate. The Illumina sequencing by synthesis (SBS) technology is predominantly used in NGS-based HLA genotyping workflows as it sequences both amplicon-based and shotgun-based DNA libraries very

efficiently in comparison to other second-generation or, even, third-generation sequencing systems developed so far. In relation to high-resolution NGS-based HLA genotyping, and in comparison to other short-read platforms, Illumina paired-end sequencing reads data can offer a much higher phasing within a given targeted HLA locus (whose respective amplicon has been generated via long ranged PCR amplification) over longer distances for resolution of potential ambiguities associated to cis/trans polymorphisms. At the same time, insert size distributions, created by this paired-end sequencing process, can be a useful tool because one long fragment can effectively anchor 2 distant polymorphisms to establish phasing. The average length of HLA antigen recognition site coding exons is 270 base pairs. Thus, polymorphisms that span exons and even into the introns can be phased. From a systems standpoint, the primary disadvantage of the Illumina systems is the relatively long (~17h to 39 h, depending on flow cell and reagents cartridge capacity) run time and onboard data analysis of the instrument. Therefore, as it happens with other short-read sequencing platforms, Illumina SBS technology is clearly not a HLA sequencing system for rapid turnaround of HLA results that in certain contexts is required in the clinical setting. For high-resolution NGS-based HLA genotyping, MiniSeq (with multiplexing capabilities of a total of 96 distinguishable samples per run; and 2x150 bp paired-end sequencing runs in about 17h run time) and NextSeq 500/550 (with multiplexing capabilities of a total of 384 distinguishable samples per run; and 2x150 bp paired-end sequencing runs in about 29h run time) instruments are among the most suitable.

a.3) ThermoFisher Ion Torrent platforms (started in 2010):

(i) Clonal amplification: Ion Torrent platforms (ion sensing) present a very similar process as in Roche 454 systems (fluorescence sensing), where DNA library fragments are also

clonally amplified on beads/sphere particles upon an emPCR-based process in a single-nucleotide addition (SNA) approach. This approach consists on the addition of four nucleotides iteratively and the scanning of a signal after each addition in order to record an incorporated nucleotide.

(ii) Sequencing: Ion Torrent platforms are also based on the principle of SBS with sequential flows of nucleotide triphosphates (dNTPs) as in the Roche 454 platform. However, Ion Torrent platforms present a semi-conductor silicon chip with sensor wells (each well can adapt a single ion sphere particle on which a single DNA template molecule has been clonally amplified in an emulsion PCR). This semi-conductor chip enables the detection of the specific incorporated bases by the DNA polymerase during sequencing, not based on imaging of fluorescent signals but based on the release of a hydrogen ion  $H^+$  during extension of each nucleotide (termed as pH mediated sequencing, silicon sequencing or semiconductor sequencing). Since the incorporation of a dNTP into a growing DNA strand involves the formation of a covalent bond and the release of pyrophosphate and also a positively charged hydrogen ion. The release of  $H^+$  is detected as a change in the pH within the sensor wells on this semi-conductor chip. Thus, the Ion Torrent platform sequencing process is not dependent upon altered bases, enzymes nor optical detection, thereby simplifying the overall sequencing process, dramatically accelerating the time to result, reducing the overall footprint of the instrument and lowering cost. Indeed, Ion Torrent has the shortest sequencing run times of the short-read platforms (2–4 h for 200–400 bp). Nevertheless, similar to Roche 454 systems, homopolymer and indel associated sequencing errors as well as AT-/GC-bias are significant owing to the “flow-space” based nature of its sequencing chemistry. At the same time, although it still yields phased data (as Roche 454 systems), the limited single-

end sequencing is the most common approach in Ion Torrent technologies. There are three different Ion Torrent sequencing chips available with an increasing number of sequencing wells to enable different scales of throughput: the 314 chip enables an output of up to 10 Mb, the 316 chip up to 100 Mb, and the 318 chip up to 1 Gb. The chips cannot be further partitioned, yet there are 96 MIDs available for multiplexed sequencing.

(iii) Current relevance for HLA-sequencing: similar to Illumina platforms, Ion Torrent platforms (whose main strengths are the very high sequencing speed and an average read length (~400 bp) that encompasses the average length of a single HLA exon (~280 bp) [184][200]) are currently used in NGS-based HLA genotyping workflows for sequencing both amplicon-based and, although less efficiently than other platforms (e.g. Illumina), shotgun-based DNA libraries as well.

b) Long-read (20 kb or more) sequencing platforms or third-generation sequencing technologies for shotgun-based DNA libraries, which mainly differ in the following characteristics:

Instead of sequencing clonally amplified templates, single HLA DNA library templates are sequenced directly with minimal use of biochemicals and at nanoscale dimensions. Thus, third-generation sequencing technologies achieve long read lengths (>20 kb) by interrogating the nucleotide sequences at the single molecule level in contrast to second-generation sequencing platforms. Furthermore, third-generation sequencing technologies are characterized by a very significant decrease in sequencing run time (ranging from minutes to hours) and, not relying on PCR to generate clusters, a decrease in or elimination of sequencing biases introduced by PCR. In addition, third-generation platforms can directly

target single DNA molecules in real time so that the sequenced reads are ready for analysis immediately.

b.1) Pacific Bioscience (PacBio) platforms (started in 2010):

(i) Sequencing: Pacific Bioscience (PacBio) platforms (RSII and Sequel) are based on single-molecule real-time (SMRT) SBS technology. In this PacBio technology, instead of immobilizing ssDNA library strands, the high fidelity  $\phi$ 29 derived DNA polymerase along with a single strand DNA library template is immobilized at the bottom of zero mode waveguides (ZMW). Zero-mode waveguide (ZMW) is a nanophotonic confinement structure that allows the accurate detection of the activity of a single DNA polymerase incorporating a single dye-labeled nucleotide (where a fluorescence signal is detected by instrument optics) as it synthesizes the complementary strand of DNA. The entire sequencing process is performed on a chip (SMRT Cell) containing around 150,000 of these ZMWs. In addition, adapter sequences added to the target DNA library fragment by ligation allow formation of a dumbbell-shaped circular double stranded molecule, the SMRTbell library. Within each ZMW, the polymerase synthesizes a complementary strand to the denatured strand of the circular template and can generate strands of up to 40 kb with the median length around 10 kb. With a relatively short library insert, the polymerase can go around the circle multiple times (increasing the number of sequencing passes) thus, the error rate can be reduced by increasing the number of subreads generated. Whereas with a longer library insert, only one sequence may be generated per ZMW, having here a higher error rate and lower number of reads. While each individual read has an important inherent non-systematic error rate (11%-15%) in Pac Bio systems, a reliable consensus sequence is bioinformatically (assembly algorithm that relies on error correction) derived from the multiple reads within a ZMW (only when there are smaller

inserts) and the multiple reads from other ZMWs. The PacBio run time is fast (30–240 min per SMRT Cell).

b.2) Oxford Nanopore Technologies (ONT) platforms (started in 2014-2015):

(i) Sequencing: Oxford Nanopore Technologies (ONT) platform is based on direct, electrical detection of single DNA molecules. This ONT system consists on a biological nanopore which is constructed from a modified  $\alpha$ -hemolysin pore that has an exonuclease attached on the normally extracellular face of the pore. A synthetic cyclodextrin sensor is also covalently attached to the inside surface of the nanopore. This 3-component system is contained in a synthetic lipid bilayer, so that when DNA is loaded onto its exonuclease-containing face and a voltage is applied across the bilayer by changing the concentration of salt, the exonuclease can cleave off individual nucleotides. The individual nucleotides are detected once they are cleaved based on their characteristic disruption of the ionic current flowing through the pore. Therefore, this type of sequencing technology does not require polymerase enzyme, there is no need for DNA polymerization (no sensitive to temperature changes) and incorporation of nucleotides as well as no need for pH alteration detection. For instance, the ONT MinION instrument (of the size of a USB stick) comprises 512-2000 of these nanopores with each nanopore having the sequencing speed of 120-1000 bases per minute. The read length profile of MinION is very similar (~10 kb or more) to that of PacBio, but the error rate is even higher (12%-38 %) though it has been improving with recent advances in chemistry. Unlike PacBio, the error rate cannot be improved by increasing coverage since the MinION is limited to two sequencing passes by design. Similar to PacBio, complex assembly and error correction algorithms need to be employed in order to produce high quality assemblies. The current throughput is low



and not very stable and the default run time is 48 h though data can be analyzed in real-time as the reads pass through the sequencer.

Current relevance of long-read sequencing platforms for HLA-sequencing: as previously described, long-read HLA sequencing data (although presenting low-throughput, high error rate and a high cost) has the potential of resolving, in a very fast sequencing mode, all the important remaining HLA ambiguities found in previous sequencing generation platforms by providing full coverage for all HLA sequence genes. Therefore, having the potential for obtaining whole-gene consensus sequences at the maximum 4-field allele-level resolution unambiguously and completely phased HLA haplotypes. However, so far all long-read sequencing HLA genotyping approaches have been long-range shotgun-based sequencing strategies and, thus, still based on an initial long-range PCR with its inherent limitations of coverage (especially for longer HLA class II genes) and associated biases. In addition, long-read data quality (with a high error rate) so far lacks behind other current short-read NGS technologies. Therefore, improved bioinformatics algorithms are needed to generate HLA genotyping results at the accuracy required to be feasible and reliable for clinical and research applications. Furthermore, to be economically and practically viable, long-read sequencing technologies still need to improve their throughput capacity (number of reads generated and degree of sample multiplexing). At the same time, long-read sequencing approaches can be a good complementary HLA sequencing data source to generate a scaffold of long reads where shorter high quality reads, produced by short-read technologies, could be overlaid and, thus, to maximize coverage of all HLA loci [172][173][204][359].

5) HLA sequence data bioinformatics analysis process for assignment of HLA genotypes:

Bioinformatics analysis of the NGS-based HLA sequencing data is a critical step for obtaining accurate high-resolution HLA genotyping information. The nature of these bioinformatics approaches and the achievable level of typing resolution are dictated by the sample preparation methods and sequencing platforms used [296]. Many different software packages (both in-house and commercial) have been developed, validated (as they need to meet both quality QC and QA requirements) and released. All programs have the capacity to handle NGS data and utilize the IPD-IMGT/HLA allele database (even some programs present an additional in-house database that includes sequences of genomic regions lacking references in the current IPD-IMGT/HLA database [187]) genotype assignment. With most of the commercial programs also offering user interfaces that enable the visualization of sequences, coverage, genotype assignment, and mismatch information. While there has been significant development of software programs dedicated to NGS-based HLA data, these programs continue to evolve alongside the sequencing technology and approaches (e.g. long-read sequencing platforms) for HLA template generation [167]. At the same time, the majority of NGS-based HLA sequence data bioinformatics software analysis programs require major capital investments to the user/laboratory facility. As complex and costly informatics logistics and infrastructures are required to ensure optimal data storage and data processing routinely, especially in the clinical setting [76][156][192]. Since the amounts of data generated by NGS platforms are typically magnitudes larger compared to Sanger sequencing, where high volume and complex nature of NGS datasets necessitate non-trivial IT infrastructure and expert bioinformatics skills [296].

In detail, bioinformatics analysis of the NGS-based HLA sequencing data is a complex process, based on a combination of algorithmic and statistical methods (overlapping between the NGS sequencer software program and the respective external HLA allele calling and genotype

assignment bioinformatics software program), that comprises three main analysis stages [76][152][157][184][188][192][296]:

a) Primary analysis (generation of raw reads sequencing data and quality scores) by the respective NGS sequencing platform:

After the sequencing and base calling processes are completed by the respective sequencer system. The sequence information or read from each fragment/cluster is recorded by the sequencer's software system along with the quality scores. Quality scores evaluate the probability of a base calling error quantified using Phred (e.g. Sanger sequencing technology) or Q (e.g. Q30 in Illumina short-read platforms; QV50-QV70 in long-read PacBio SMRT technology; or Q5-Q13 in long-read ONT Nanopore platforms) values as sequencing and base calling quality metric scales. For instance, Q30 score is used as a metric and indicates that a base has a 1:1000 probability of being called incorrectly (or 99.9% base call accuracy). Therefore, the higher percentage of reads presenting the Q30 score, the more accurate the sequencing data is [192]. In this sense, each sequencing platform is unique not only in how quality scores are calculated but also in relation to the data storage format used (e.g. .sff or .bam). Regardless, most NGS data formats can be converted into the FASTQ format, which is a standard representation of NGS data, containing entries for each read's haploid DNA sequence and corresponding quality scores. For paired-end sequencing (Illumina platforms), two FASTQ files are generated, with reads from each end (Forward/Read #1 and Reverse/Read #2) of the DNA fragment placed into separate files and linked (demultiplexing) through the sequence identifier (sample or, less commonly, amplicon).

b) Secondary analysis (of the reads per each respective barcoded sample and HLA locus) by the respective HLA allele calling and genotype assignment bioinformatics software program

using a certain set of algorithms for: first, alignment of cleaned reads to a reference sequence; and, second, identification of variants and subsequent sequence assembly and phasing. Although the bioinformatics algorithms/pipelines (where their parameterization is often a compromise between speed and accuracy of the analysis) for interpreting mid/short-range amplicon-based sequencing data (in which the primer sequences are used to assign the reads, which all start at a defined position, to a specific genomic region) are different from the ones used for interpreting long-range shotgun-based sequencing data (which deal with a more diverse set of overlapping reads and typically assemble them based on a reference sequence), this NGS data secondary analysis stage may include all or several of these following main steps:

- i) Read cleaning/filtering: at this step, there is a filtering of raw sequence reads based on quality metrics of the sequencing software instrument as well as the removal of adapter sequences (which are useless and misleading for HLA allele calling). Removal of low-quality reads, too short in length (e.g. due to PCR errors/artifacts), and trimming of the ends of the reads (as for most of the NGS platforms, the quality of base calls declines toward the ends of the reads), including the adapter sequences, by bioinformatics tools is important to improve the quantity of reads useful for final HLA genotype assignment. If the data may derive from HLA-untargeted sequencing libraries or if amplicons/captured sequences from different genomic locations have been pooled for sequencing, reads first need to be correctly attributed to a specific genomic location prior to sequence alignment or mapping step [296].
- ii) Sequence alignment or mapping of cleaned reads against the given HLA alleles reference sequence database: competitive sequence alignment or mapping of cleaned raw reads against the given known (and any additional in-house described one (e.g. [187])) HLA alleles reference sequence database allows to capture all SNPs and structural variance of the

targeted HLA regions. Thus, reads are compared across a reference sequence database and attempts to locate the proper position of the read, or pairs of reads, given certain reference allele sequence. Where read mapping typically involves two sub-steps: a fast initial heuristic is used for identifying one or more top mapping locations on the reference for each read, followed by local alignment optimized for accuracy [296]. At the same time, different levels of reference sequences may be applied (establishing a hierarchical scoring strategy [296]), considering: partial cDNA (exons 2 and 3 for HLA class I loci, and exon 2 for HLA class II loci), full cDNA (all exon sequences) and full gDNA (both intron and exon sequences). A ranking list of candidate alleles is generated indicating also exon/intron mismatch in allele assignment information.

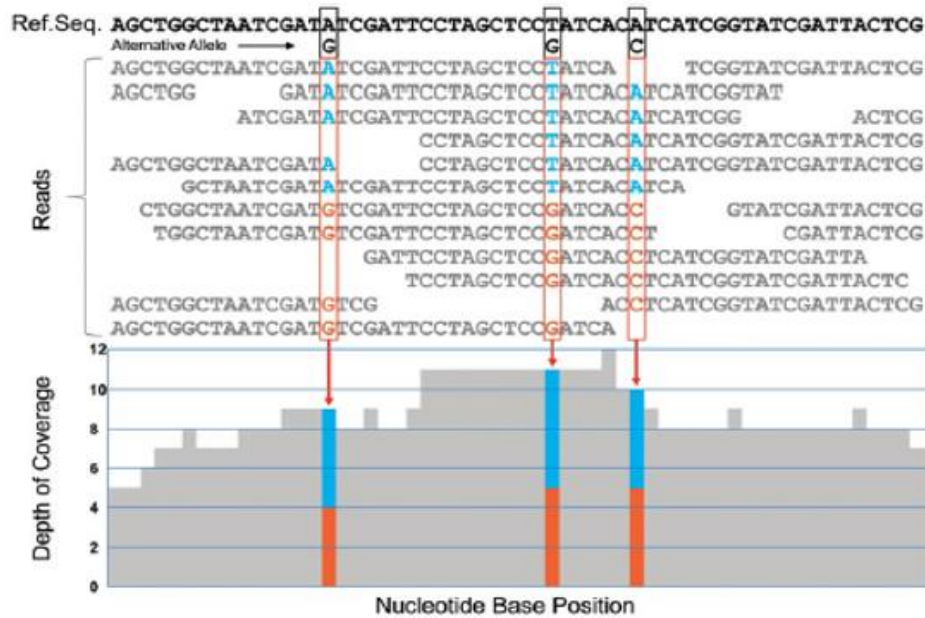
iii) Sequence assembly (reference-based or de-novo) and phasing (based on polymorphic linkage) of cleaned, mapped reads building one (homozygous sample) or two (heterozygous sample) consensus sequences or contigs (where a contig defines the series of mapped, assembled and phased reads per HLA allele of each HLA locus in a given tested sample): at this step, there is the identification of variants at different positions (polymorphic sites or SNPs) together with their phased assembly (as the process for annotation/interpretation of the identified variants) for building the consensus sequence/s per targeted HLA locus of a given sample. After alignment, firstly, the variant positions are determined at locations where the aligned sequencing data (reads) are different from the reference, and can be either homozygous or heterozygous in nature. In this variant calling step, two crucial sequencing data parameters/metrics (depth of coverage and the quality score) are used by the analysis program's algorithms and statistical methods to decide whether a position is truly different from the reference. The depth of coverage refers to the number of reads aligned to a given position in the reference sequence. Thus, aligned bases with lower quality scores indicate

lower confidence that the nucleotide call is correct and are down-weighted during variant calling at a given position (with preference given to aligned bases with higher quality scores) (see **Figure I-21**). Whereas a higher number of reads present at a given position (higher depth of coverage) makes it easier to discriminate the signal (actual bases present, high quality scores) from the noise (errors during sequencing, base calling, and alignment, often with lower quality scores) during variant calling. Moreover, the uniformity/extent of coverage (i.e. depth of coverage throughout the length of the amplified/targeted region of a given HLA gene) is also a very important metric so that all targeted regions of both HLA alleles per locus are equally represented. In this sense, as both the number of HLA loci (up to 11 classical loci) and the number of samples (of the order of hundreds) tested tend to be usually high for a given NGS-based workflow run, coverage uniformity becomes even a more relevant parameter to take into account. Since running fewer samples per run results in greater coverage per allele/locus, but at a higher cost per sample. Conversely, running more samples means less coverage per allele/locus and possibly less confidence in the final genotyping result. In addition, and intrinsically related to the depth and uniformity of coverage, a minimum total number of raw reads of a certain minimum average length (e.g. if the reads are 150 bases long, then proportionally more reads per allele/locus/sample are needed to obtain the same depth of coverage as a sequencing run with reads of 250 bases) must be initially obtained for HLA genotyping software analysis in order to: optimize a balanced representation metric of cleaned mapped reads of each allele, of each locus and of each sample that can ensure a final reliable HLA genotyping result; and, at the same time, to compensate (until certain extent) the technical/chemical/enzymatic-related limitations of these PCR-based NGS HLA genotyping workflows (e.g. PCR-related limitations/errors that cause allele imbalance/drop-outs as well as imbalanced representation of HLA gene

amplicons of different length (more reads are required for longer class II genes than for the shorter HLA class I genes)). Therefore, a careful optimized balance of all these factors (e.g. number of samples/loci per run to be tested and, also, minimum number of obtained raw reads of certain average length per locus/sample to be analyzed) must be established in the design of both the experimental protocol and the bioinformatics data analysis stage of any given PCR-based NGS HLA genotyping workflow to ensure minimum optimal sequencing data analysis metrics (such as quality score, coverage depth and coverage uniformity) in order to have a robust analysis for a reliable HLA allele calling and genotype assignment. At the same time, HLA genotyping software programs present built-in default filters to define the thresholds of each of these sequencing data analysis metrics (such as quality score (e.g.  $Q30 \geq 75\%$ ), coverage depth (e.g. 30-40 reads (30-40x)); coverage uniformity (e.g. 30-40 reads (30-40x)); and balanced representation metric of reads to define heterozygosity/homozygosity at a given position (e.g. to call a heterozygous position, ratio of bases must be at least 20%)) in order to establish credible genotyping. Importantly, the depth of coverage (at a given position and throughout the length of the given targeted HLA region with a minimum required balanced representation of reads per allele-per locus-per sample) has a significant and direct impact on the reliability of the consensus sequences that are built, as this consensus building process is based on all these identified variant positions (polymorphic sites or SNPs). During this same identification process of variants at different positions, assembly algorithms are used to align overlapping mapped reads, which allows the original targeted HLA genomic region of interest to be assembled into contiguous sequences. In general, assembly algorithms can be reference-based and consider a reference sequence as input, or can be de-novo and blind to any data beyond the sequence reads (i.e. algorithmic reconstruction of the source sequences from read fragments without recourse to

a reference). So far, analysis of HLA sequence regions without or only partial reference sequences in the IPD-IMGT/HLA database (as only about 10% of known HLA class I and class II alleles have been fully described (full genomic sequences)) [146][463] requires suitable NGS bioinformatics software that performs de-novo assembly of sequence reads. In the same de-novo assembly algorithmic process, proximal/distal identified polymorphic sites are linked using phasing algorithms and based on polymorphic linkage information according to the read length available from the given single-end/paired-end short-read (with the potential of phasing alleles within a given HLA locus) or long-read (with the potential of phasing entire haplotypes encompassing different HLA loci) sequencing data. Haplotypic linkage can, however, only be reliably reconstructed if two heterozygous variants that need to be phased are covered by reads or read pairs [296]. The combination of these two approaches (de-novo assembly plus phasing) is a very useful feature that also facilitates not only overlapping the flanking and gapped contigs but also resolving the contig order and orientation (scaffolding) to reconstruct the original targeted HLA DNA sequence. At the end of this sequence de-novo assembly and phasing combined algorithmic process, one (homozygous sample) or two (heterozygous sample) phased consensus sequences or contigs are built. Importantly, quality scores for each consensus base generated need to be considered at the final HLA allele assignment [296].





**Figure I-21.** Depth of coverage as derived from NGS reads. NGS reads (whether single or paired-end) are aligned to a consensus reference sequence (Ref.Seq.) to form an overlapping, stacked view of the coverage at any given nucleotide position. The number of reads at any given position represents the depth of coverage and is derived by the number of overlapping reads at that nucleotide position. Read depth at polymorphic positions is split between the reference sequence and the alternative allele (blue and orange bars) and indicates allelic representation. Figure and respective footnote are obtained and adapted from [76].

iv) Final sequence alignment or mapping of these built phased consensus sequences against the given HLA alleles reference sequence database: mapping of these built phased consensus sequences against the given known (and any additional in-house described one (e.g. [187])) HLA alleles reference sequence database allows now to determine the best fit and to provide the final allele/genotype variant call or assignment of each targeted HLA locus for a given tested sample. In addition to this “traditional” combined approach for HLA allele assignment (just described above and based on competitive mapping to a reference allele sequence repository (followed by varied scoring and refinement schemes to hone typing accuracy and precision) plus de-novo assembly and phasing consensus sequence generation from sequence reads followed by a final reference allele matching); it is

noteworthy a very novel alternate strategy (although it is still evolving) which consists on read mapping to graph-based structures that succinctly encode nucleotide and structural variation in a single rich reference structure (i.e. an alternative to the linear model of representing sequence data emerged by imagining a set of genotypes as a graph, a network with nodes denoting nucleotides and edges connecting each nucleotide to all its possible predecessor and successor nucleotides) [296].

c) Tertiary Analysis: bioinformatics interpretation of results to determine the best fit and to provide the final HLA allele/genotype variant call per sample. As previously described, haplotype frequencies for alleles in strong LD have been established for specific ethnic groups, racial categories and certain worldwide populations (in both family and unrelated subjects studies) [131][132][201]. Thus, although the haplotype LD information is not used (most commonly) to determine computed sample genotypes through these NGS HLA allele calling and genotype assignment bioinformatics software programs. This haplotype LD information (from the publicly available information and from analysis of internally typed samples of a particular given HLA genotyping analysis software) can be used manually or with recently developed software programs [201] to validate observable haplotypes in NGS-based HLA genotype data (to check if alleles identified in different loci match expected linkage disequilibrium tables) as well as helping to resolve situations of true homozygosity from allele dropout or vice versa (although it is still required to be confirmed by an alternative genotyping method) [157]. In addition to report the final HLA genotyping results per sample, NGS HLA allele calling and genotype assignment bioinformatics software programs are expected to identify and report the following events that also indicate the level of accuracy achieved by the given NGS-based HLA genotyping workflow (a critical aspect, especially in the clinical setting):

i) Novel alleles:

Novel alleles (exon variants) can be also identified as discrepancies from the coding sequence (exons) of the reference alleles as well as the identification of new non-coding variants (allele variants detected in introns and untranslated regions). Nevertheless, while new alleles that are possible due to novel single nucleotide polymorphisms can be detected by current NGS methods with a high level of accuracy, detection and characterization of new alleles that are hybrids of two or more known alleles is still a major challenge for the so far and currently developed bioinformatics tools [76]. Furthermore, most of the introduced HLA genotyping software analysis programs do not report a nucleotide sequence but yield the best matching alleles in the IPD-IMGT/HLA database [87]. As such, the genotyping accuracy and benefit for clinical applications rely highly on the current status of correctness and completeness of the database [296].

ii) Unresolved ambiguities and mistyping situations in NGS-based HLA genotyping:

So far, and as previously mentioned, all targeted sequencing HLA genotyping approaches developed (including both general groups: hybridization-based capture methods; and, the most widely used, PCR-based target enrichment methods with all their different strategy versions (mid/short-range amplicon-based or long-range shotgun-based using short-read/long-read sequencing platforms accordingly)) have inherent limitations avoiding a 100% unambiguous characterization at the 4-field and full phasing of HLA alleles yet and, thus, certain genotyping ambiguities still remain. In relation to the most widely applied PCR-based target enrichment methods, while some of their inherent PCR-related limitations/problems can be addressed or minimized during either the initial PCR amplification (e.g. optimizing reaction components (such as amplification conditions,

inclusion of co-solutes, molar ratios of reagents, primers design and targeted genomic location (primers binding sites)) for a robust and reproducible amplification of every targeted locus and of every sample) or at particular steps of the DNA sequencing library preparation (e.g. normalization and balancing (equimolar pooling) of amplicons in order to optimize an equal representation of all genes/amplicons per sample). Also, the NGS HLA allele calling and genotype assignment bioinformatics software programs can be able (although only until certain extent) to detect and address or minimize these observed PCR-related ambiguous events that can lead to incorrect HLA genotype determination. At the same time, other observed ambiguous events, that also compromise the accuracy and reliability of HLA genotype determination by NGS methods, are due to additional limitations coming from different current sources of ambiguity and mistyping (some of them closely related) such as: the incomplete/partial HLA allele sequence references in the official database (IPD-IMGT/HLA) [87][146][463]; inefficiency of the sequencing process; partial targeting strategy of NGS-based HLA genotyping approaches; NGS bioinformatics algorithm-related ambiguities; and the observed patchwork pattern of sequence polymorphism along the HLA system. In summary, these types of ambiguities (ambiguous events) and their corresponding sources as well as the possible troubleshooting options can be categorized and described as follows [202][296]:

- PCR-related ambiguities: considering the major role that PCR amplification plays in the majority of NGS-based HLA genotyping workflows (e.g. template generation based on PCR-based target enrichment; and short-read sequencing platforms based on clonal amplification (either emPCR or bridge PCR amplification)), it is important to be aware of possible inherent errors and artifacts that can be originated from PCR amplification, as

these can greatly affect the accuracy of HLA genotyping. PCR-related ambiguities are usually caused by two main issues:

- Signal loss caused by inefficient and imbalanced amplification having HLA allele or locus imbalances or drop-outs that can make consensus assembly difficult or can cause low coverage, both of which can increase ambiguity. Most common causes of this signal loss are: sample-related issues (e.g. input/quality DNA issues or false homozygous HLA typing results for HSCT in cancer patients due to chromosome 6 loss in cancer affected cells (e.g. acute lymphocytic leukemia)); technical-related issues (e.g. thermocycler/instrument malfunction); protocol-related issues (e.g. primer design problems or human user error); and/or uncharacterized HLA polymorphism-related issues (e.g. novel variant (coding or non-coding) in primer binding site). Samples or HLA loci affected by dropt-outs/imbalances need to be re-processed and re-sequenced in most cases, which can be very time-consuming and it delays the turnaround time. On the other hand, there are circumstances during the HLA genotype assignment bioinformatics analysis and review process in which going below the default thresholds of sequencing data analysis metrics may be acceptable in order to address these dropt-out/imbalance situations. This is the case when the polymorphisms of the two alleles of a locus are phased, the typing does not have a lot of noise from sequencing errors, and the locations with a low depth of coverage occur in a region that does not affect critically the allele call genotyping as well as the determination of homozygosity versus heterozygosity at a given locus.

- Mixed signals caused by PCR crossover artifacts or PCR stutter artifacts basically create a mix of artificial alleles in vitro that makes allele selection difficult and not very reliable. PCR crossover artifacts can be generated by incomplete primer extension

(premature extension stops), mostly due to non-targeted annealing of already existing partially amplified complementary sequences (e.g. partial length product becomes primer for another locus or allele; thus, there may be annealing between homologous HLA loci and/or between the two alleles within the same HLA locus), presenting higher concentrations especially at the end (during the last cycles) of the PCR amplification process. Thus, out of the PCR products generated, an important percentage of re-annealed partial amplicons are amplified (instead of the designated primer-amplicon template annealed pairs) which do not follow the original primer-designed specific targeting strategy. At the experimental protocol level, decreasing the number of amplification cycles or adjusting initial template concentration can greatly reduce the amount of PCR crossover artifacts. Also during the HLA genotype assignment bioinformatics analysis (if there is available an enough amount of initial raw reads at certain length to be analyzed), PCR crossover reads can be identified as systematic noise and, thus, filtered-out or down-weighted thanks to the phasing process. In which the algorithms can determine the correct base combination for each consecutive variant pair as long as the majority of the reads generated support the given real correct combination. On the other hand, PCR stutter products (differing from the original template by multiples of the repeat unit length) are a common artifact in the PCR amplification (and also in SBS-based sequencing approaches) of DNA regions (mostly non-coding regions) with repetitive and extensive low-complexity and imbalanced sequence composition, and that are present along the HLA system, especially at non-coding regions, such as: homopolymer repeats poly(dA), poly(dT), poly(dG) and poly(dC) (composed of eight or more nucleotides); regions of short-tandem repeats (STRs; comprised of 1–6 bp per repeating unit); or high AT- or GC-rich regions (that often contain mononucleotide

repeats of 10 or more bases). All these particular DNA regions can establish complex folded structures in the DNA molecule that, in turn, are prone to mutation via slipped-strand mispairing (termed also as “slippage”) by the DNA polymerase during in vitro PCR-mediated DNA replication, as well as during in vivo DNA replication [203]. Mutations at these particular DNA regions during in vitro enzymatic replication are usually the result of insertion or deletion of repeats in the extending, or nascent, DNA strand sequence. In order for slipped-strand mispairing or slippage to occur, the DNA polymerase enzyme first stalls (when it reaches a complex folded structure that constitute a physical barrier for continuing the replication) and dissociates from the dsDNA complex during replication of the repeated motif. If base pairing is disrupted after polymerase dissociation, then a loop of one or more repeat units may form in either the nascent or the template strand prior to re-association and cause the insertion or deletion of one or more units, respectively, in the newly formed DNA strand during replication. Deletion mutations are believed to be more common as they require fewer nucleotides of the dsDNA to dissociate and therefore are more energetically favorable than insertion mutations. Thus, these mutations may confound the delimitation of the true repeat number, as stutter products can be generated (during PCR-mediated DNA replication) in similar or even greater proportions than the true product in length (where slippage rates increase with the length of these particular DNA regions). A particular variation on DNA polymerase slippage is when it is not the length of the homopolymer that is changed, but a base surrounded by two homopolymers such as CCCCACCCC changing to CCCCCCCC and, thus, omitting the A in the middle position (i.e. short homopolymers in low complexity motifs as for instance GGGGCCGG (A\*68:01:02:02) versus GGGCCCCG (A\*68:142N) may not be reliably differentiated) [296]. During the

HLA genotype assignment bioinformatics analysis, the consensus assembly of these particular DNA regions with repetitive and extensive low-complexity and imbalanced sequence composition is itself difficult, and reads containing PCR stutter artifacts exacerbate this problematic analysis. Consequently, as it is very hard to identify PCR stutter artifacts as systematic noise and very challenging to filter them out, it is almost impossible to resolve (either with short-read or long-read sequences) the ambiguity of length polymorphisms, in this case between alleles that differ only in the length of these very repeats. Some common representative examples are the following:

(a) PCR-induced STR length mutation has been shown to lead to a loss of heterozygotic genotyping and to confound the discrimination of allelic differences in the HLA region in particular cases. In a well optimized PCR setup, however, those artefacts will only affect particular, mostly intronic, regions or a low fraction of samples [296]. STRs of considerable length and diversity exist, for instance, in intron 2 of *HLA-DPBI* as (AAGG)<sub>(4-17)</sub> tetranucleotide repeats or in intron 2 of various *HLA-DRB* genes as (GT)<sub>(7-27)</sub>(GA)<sub>(5-30)</sub> dinucleotide patterns [296]. As an example, *HLA-DRB1\*03:01:01:01* and *HLA-DRB1\*03:01:01:02* alleles differing only in an SNP in intron 1 and the length of GT repeats in intron 2. When the whole (very long) intron 1 of *HLA-DRB1* is not sequenced (as for most of the available NGS HLA genotyping kits) these two alleles are practically indistinguishable. This type of ambiguity is commonly found in other *HLA-DRB1* alleles as well, since there is an extensive low complexity region with STRs at the border of *HLA-DRB1* exon 2 and intron 2.

(b) The length of homopolymers can reach up to 30 nucleotides in classical HLA genes (*HLA-DRB1\*15:02:01:02*) and up to 45 in other clinically relevant genes in the MHC (e.g. *MICB*) [296]. Thus, classical HLA alleles differing in the length of the



homopolymer can be displayed as ambiguities, such as the null allele *HLA-A\*03:21N* where there is an insertion in the originally 7 bases-long C homopolymer in exon 4 of the allele compared to *HLA-A\*03:01:01:01*. Or *B\*51:01:01:01* and *B\*51:11N*, which differ by a 6C versus 7C homopolymer stretch. Another example is the case of alleles of the pseudogene *HLA-H* (sequence stretch harboring 8 C nucleotides) that differ from the corresponding *HLA-A* alleles only in the length of homopolymers.

- Incomplete/partial HLA allele sequence references in the official database (IPD-IMGT/HLA)-related ambiguities: as previously mentioned, only about 10% of known HLA class I and class II alleles have been fully sequenced and are publicly available as HLA sequence references in the official database (IPD-IMGT/HLA) [87][146][463]. Thus, HLA genotype calling software programs are limited by this incomplete sequence database for comparing alleles and selecting the most probable alleles identified in the sample data. Since most of the alleles are defined only partially, these comparisons cannot be always done accurately and unambiguously. De-novo assembly of sequence reads generated by short-read sequencing approaches or long-read sequencing approaches try to compensate and/or work through this lack of reference sequence data. In addition, in those instances where more than one combination of alleles is equally possible based on the references sequences only if the phase information is available and reliable, these kinds of ambiguities can be resolved reassuringly. Another way to deal with some of the ambiguity introduced by the incomplete HLA reference space is to employ some hierarchical scoring strategy. By penalizing mismatches in ARD-encoding exons higher than in other non-ARD-encoding exons and by penalizing mismatches in other non-ARD-encoding exons higher than in non-coding sequences, the best possible full resolution alleles may be recursively identified [296].

- Inefficiency of the sequencing process-related ambiguities: regardless of the sequencing signal detection technology (e.g. sequencing errors caused by the “flow-space” based nature of the sequencing chemistry, where the signal strength does not increase linear with growing repetitive DNA sequence regions) or of the DNA polymerase enzymatic system (e.g. sequencing errors caused by slippage events) applied by both short-read and long-read sequencing platforms; there are known current cross-platform problems and limitations during the sequencing process (especially of these particular DNA regions with repetitive and extensive low-complexity and imbalanced sequence composition) that can lead to HLA allele calling ambiguity as well. Even though very long single DNA molecules are sequenced or big amounts of sequencing reads of certain optimal length are generated and analyzed by the respective HLA genotype assignment bioinformatics analysis program. Also, it is possible that sequencing platforms generate low-quality reads (e.g. too short in length) and random artifacts reads (“orphan” or “off-target” reads which do not map at all to any known reference sequence used and/or are not similar to any other reads in the sequencing data) that may be identified (although only until certain extent) as systematic/random, respectively, noise and filtered-out by the given HLA genotype assignment bioinformatics analysis program tools.

- Partial targeting strategy of NGS-based HLA genotyping approaches-related ambiguities: due to inherent technical limitations (referring to those shown by the most widely applied PCR-based target enrichment methods, whose coverage ability resides on the primer design possibilities and the PCR amplification capacity) and the very high cost-coverage-time efficiency requirements involved (especially for its highly demanded clinical application), the large majority of current NGS PCR-based HLA genotyping approaches still characterize a limited number of genomic regions (excluding some

introns, UTRs and, even, exons) along the different HLA genes (especially in the long HLA class II genes, such as *HLA-DPB1* and *HLA-DRB1*). This is usually a compromise between accuracy and throughput depending on the given clinical/research application. Thus, the ambiguity introduced by partial targeting depends on the selection of the non-characterized regions, where polymorphisms located outside of these sequenced regions cannot be resolved. Some common relevant examples are the following:

- Some HLA class II loci have very long introns (>5,000 bp) which also present repetitive and extensive low-complexity and imbalanced sequence composition and, consequently, low SNP diversity. All these unique structural characteristics make very challenging the primer design and selection of primer binding sites while trying to span the maximum length of the targeted genomic sequence. Two major cases are noteworthy:

(a) For the *HLA-DRB1/3/4/5* loci (presenting a very large intron 1, about 10,000 bp), the designated binding sites of the targeting primers are usually not in the UTR region, but skipping the long intron 1 and, even, exon 1. Thus, there are two main long-range amplification strategies:

- \*One consists on two separated reactions: exon 1 amplified region and, skipping intron 1, exon 2 to exon 6 amplified region respectively.

- \*Or, alternatively, just a single reaction: skipping exon 1 and intron 1 to amplify the HLA-DR loci starting from exon 2 to exon 6.

This latter case makes space for ambiguities such as *HLA-DRB1\*12:01:01* versus *HLA-DRB1\*12:10* that are differing in a single SNP on exon 1.

(b) Another HLA class II locus that is particularly notorious for generating ambiguities is the *HLA-DPBI* locus. In this case, there are two main types of ambiguity possible for the *HLA-DPBI* locus:

\*The first one is, as previously mentioned with the *HLA-DR* loci, because the current NGS PCR-based HLA genotyping assay approaches do not characterize neither exon 1 nor the very long intron 1, where targeted amplification starts from exon 2 to exon 4 or exon 5. Therefore, *HLA-DPBI* alleles whose only difference is in exon 1 are not distinguished.

\*The second type of *HLA-DPBI* ambiguity is due to the inability (in particular when using only short-read sequencing data) to set phase between the heterozygous positions found within the exons sequenced (exons 2 to 5) based on the distinctive genomic organization found at the *HLA-DPBI* locus. Where many of the different described *HLA-DPBI* alleles, especially at their non-coding regions, are highly homologous (present low SNP diversity) and yet polymorphic, although presenting very distal polymorphic positions that are difficult to be phased. For instance, phase is often broken in intron 2, which is approximately 4 kb in length, and can be either sparsely or densely populated with heterozygous positions, depending on the combination of alleles. Consequently, different *HLA-DPBI* allele combinations may satisfy a same set of heterozygous positions but in different cis/trans combinations. Therefore, these combinations of *HLA-DPBI* alleles (combination of alleles that as pairs share the same exon 2) cannot be unambiguously phased. For example, the allele combination *HLA-DPBI*\*654:01 + *HLA-DPBI*\*417:01 have the same set of heterozygous positions as the *HLA-DPBI*\*01:01:01 + *HLA-DPBI*\*11:01:01 allele combination. The difference between the two allele combinations is the arrangement

of a polymorphic position located in exon 4 (encoding the transmembrane domain) with the polymorphic positions in exon 2 (encoding the antigen recognition domain) that are 4.8 kb apart, and phase is broken by a 1.6-kb homozygous region in intron 2. Therefore, unless all positions across the sequenced region are phased, the two possible combinations of alleles cannot be discerned from each other. Nonetheless, very recent bioinformatics data analysis strategies based on the combination of third-generation sequencing data (error prone but long reads) and second-generation sequencing data (high-quality data with short reads) have shown the potential to allow the resolution of these current very challenging cis/trans combination ambiguities when genotyping *HLA-DPBI* locus and, even, aid in novel allele discovery within this locus (especially for the characterization of novel alleles harboring long intronic regions) [172][173][204][359]. This novel combining short-read and long-read data approach allows the creation of highly accurate (e.g. solving the phasing between these heterozygous positions) allele sequences out of targeted HLA sequencing. Where the long reads are used for providing a phased guide alongside which the short reads can be mapped to receive an accurate sequence signal (coverage) [296].

-Also, untranslated regions of HLA class I loci are rarely targeted, although numerous alleles are differing from each other in a single base in these UTRs and these regions may influence to the gene expression after transcription [205]. Prime examples are *HLA-A\*02:01:01:01* and *HLA-A\* 02:01:01:02L* (differentiating SNP in the middle of the 5'UTR sequence) or *HLA-B\*35:01:01:01* and *HLA-B\*35:01:01:02* (differentiating SNP is at the end of the 3'UTR).

• NGS bioinformatics algorithm-related ambiguities: as previously described, most HLA genotyping bioinformatics algorithms incorporate reference alignment, assembly and phasing strategies that reconstruct, based on the application of a complex series of statistical methods, the DNA sequence of the respective HLA targeted regions. All these statistical methods always include some assumptions to avoid extremely high computation needs and very unlikely genotype assignment situations (e.g. bioinformatics phasing may assume that a potential genotype assignment with two novel alleles (i.e. both alleles absent from the database) is extremely unlikely; or several existing approaches rather arbitrarily reduce the search space by ignoring putatively rare alleles). When these assumptions fail this leads to ambiguity in the HLA genotyping results. Thus, since the alignment execution is essentially independent for each read/read pair, alignment algorithms often miss the capability of differentiating between random noise (e.g. “off-target” reads) and systematic noise (e.g. PCR artifacts or low-quality reads). Meanwhile, random noise is not disturbing the statistical methods (e.g. variant calling, coverage profile analysis, etc...), usually applied after the alignment step, the systematic noise introduces significant error that might prevent unambiguous genotype resolution due to not enough reliable information available to discriminate between candidate alleles. Both assembly and phasing algorithmic strategies have to consider only well-supported assembly-phasing paths to connect reads to each other to avoid the situation when artifacts mislead the assembly and phasing processes and to avoid multiple separate contigs (which hinder phasing and where the distance separation in-between contigs becomes unknown) of the targeted region (where continuous consensus of sequence parts is preferred, even if there are regions where the amount of reads is relatively low). A particular challenging situation is the identification of novel alleles. For alignment-based algorithms where input data is

processed read by read (or read pair by read pair in case of paired data), the differentiation between mismatches imposed by the novel allele and mismatches related to random noise is not possible during the alignment. For assembly-based algorithms, when the final consensus sequence is built, including the novelty, it is always required an exhaustive review and further investigation by the user of the respective sequencing data analysis metrics (such as quality score (e.g.  $Q30 \geq 75\%$ ), coverage depth (e.g. 30-40 reads (30-40x)); coverage uniformity (e.g. 30-40 reads (30-40x)); and balanced representation metric of reads to define heterozygosity/homozygosity at a given position (e.g. to call a heterozygous position, ratio of bases must be at least 20%)) to support the authenticity of the novel coding/non-coding variant detected.

- Patchwork pattern of sequence polymorphism along the HLA system-related ambiguities: polymorphism at the HLA loci is clustered in a distinctive patchwork pattern of sequence motifs, which results in the extensive allelic diversity observed for these loci [145]. At the same time, a consequence of the patchwork pattern of sequence polymorphism is that a large number of alleles share the same gene segments and therefore cannot be easily distinguished, showing a high level of homology between different HLA genomic regions (also known as HLA gene homology). This characteristic high homology pattern of HLA polymorphism often leads to an important level of ambiguity because: on one hand, makes the primer design difficult and not specific enough (where primer binding sites are not sufficiently unique and specific); and, at the same time, it is also very challenging for HLA genotype assignment bioinformatics software analysis programs (where more reliable mapping positions result in more accurate local alignments of reads, especially in repetitive regions or in the large HLA sequence space). In relation to the latter, two general circumstances are noteworthy:

- The consensus assembly and phasing of highly homologous sequences that present very distal polymorphic positions (e.g. as previously described, being common in long HLA class II genes and notably in *HLA-DPBI* locus) is itself extremely difficult (especially when they rely on short-read sequencing data: where, once polymorphisms (i.e. heterozygous positions) are separated by a long homozygous stretch of more than about 1000 bp, resolution of phase between the two chromosomes tends to fail using short-read technologies as read pairs no longer reliably span this distance) and it prevents unambiguous genotype resolution in most cases [296]. Generic examples may include: coding and non-coding sequences of HLA genes of the same class that are highly homologous between them; as well as nonfunctional HLA pseudogenes (e.g. *HLA-H*) with very similar sequences to functional HLA genes (e.g. *HLA-A*).

-At the same time, and also as a consequence of this patchwork pattern of HLA sequence polymorphism, this HLA gene homology can easily lead to the so-called “cross-mapping” events during HLA genotype assignment bioinformatics analysis. Where generated sequencing reads are mapping to multiple HLA loci at same sequence positions, being almost impossible to determine unambiguously their specific HLA genomic origin or “ownership/identity”. Which, in turn, creates a significant imbalance of the representation of specific reads per allele/locus in each tested sample. Common examples of these “cross-mapping” events are the following:

(a) The conserved exons of HLA genes, such as those coding cross-membrane and intracellular components, are very similar to each other and, consequently, are associated with significant cross-mapping events. It is especially true for *HLA-DRB1* and *HLA-DQB1* loci, where there is also a significant homology between intronic segments of *HLA-DRB1/3/4/5/7* loci (also between these loci) and *HLA-DQB1* locus.



A particular and very challenging case is when trying to determine unambiguously not only the *HLA-DRB1* allele (e.g. *HLA-DRB1\*07:01:01:01*) but also its respective associated *HLA-DRB3/4/5* allele, where several aligned candidate alleles can be extremely similar (e.g. *HLA-DRB4\*01:01:01:01* versus *-DRB4\*01:03:01:01* versus *-DRB4\*01:03:01:02N*) based on the generated reads. Moreover, important cross-mapping events are very common here (e.g. in the true presence of *HLA-DRB1\*07:01:01:01* allele, reads generated from its exon 3 sequence (in particular from the differential base position G at codon 135) provoke the automatic over-calling of *HLA-DRB4\*01:03:01:01* allele sequence (containing also a base G at codon 135) to the detriment of the possible true but under-called *HLA-DRB4\*01:01:01:01* allele (presenting instead a base A at codon 135)).

(b) Weaker cross-mapping can be also observed among HLA class I genes and between some HLA class I and class II sequences coming from conserved exons. In addition, as previously mentioned, cross-reads generated from physically closed located HLA pseudogenes (e.g. *HLA-H*) may interfere/contaminate the specific reads generated from contiguous functional loci (e.g. *HLA-A*).

Therefore, as these reads covering highly homologous sequences are not very informative, the HLA genotype assignment bioinformatics software analysis programs should be able to identify and filter them out as non-uniquely mapping. However, this identification process is quite complicated especially taking into account the currently used partial targeting strategy of NGS-based HLA genotyping approaches and the actual incomplete HLA allele sequence references database. In this situation, instead of using “mapping-uniqueness” some bioinformatics tools use a phred-scaled mapping probability (i.e. reference-based read mapping followed by various quality filtering steps

to discard spurious reads with low alignment scores). Using this metric, excluding/involving reads that are mapping to multiple genes can be assessed more objectively [202]. On the other hand, some other algorithms simply discard these reads, risking coverage holes in homologous regions.

iii) Reporting NGS-based HLA genotyping results and related metadata [296]:

As previously mentioned, the use of CWD catalogues [147-150][479][480], MAC or NMDP codes system [296], “P” and “G” groups (these latter as part of the official HLA nomenclature defined by the WHO Nomenclature Committee for Factors of the HLA System) [74][94][296] have been so far the most common forms of representing and reporting HLA ambiguous genotypes. However, NMDP codes, P allele groups (alleles that encode identical ARD protein sequences) and G allele groups (alleles that share identical ARD-encoding exon sequences) very often lead to loss of accuracy as not all alleles encoded in a certain group may be valid results from the given HLA genotyping of a sample tested and, thus, these more traditional HLA ambiguities reporting approaches can be considered inadequate (even obsolete) to address the currently described HLA allelic diversity which is vast and dramatically increasing via NGS application [295][296]. Therefore, in this recent era of rapid HLA allele discovery and ultra-high allele polymorphism (at the 4-field) described by application of NGS technologies and in order to accurately report genotyping ambiguity and allele variation, more suitable alternatives have been developed such as the so-called Genotype List (GL) Strings system [324]. In detail, the GL string text format uses a hierarchical set of operators to describe precisely the relationships between alleles, lists of possible alleles, phased alleles, genotypes, lists of possible genotypes, and multilocus unphased genotypes, without losing typing information or increasing typing ambiguity. Furthermore, phasing ambiguities can be systematically reported using GL strings

[296][324]. Moreover, in addition to report HLA allele assignments and ambiguities from NGS-based HLA data (e.g. IPD-IMGT/HLA reference context or consensus sequence), the international histocompatibility and immunogenetics scientific and clinical community (organized in this particular case as the Immunogenomic NGS Consortium) has also recently defined the Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING) reporting guidelines for ensuring the long-term portability, common standardized format when reporting and exchanging genotyping data (Histoimmunogenetics Markup Language (HML) with XML structures (i.e. hla.xml files) and GL strings) [325], and broad application (e.g. data collection, data processing, and interpretation) of this NGS HLA genotyping data [296][326]. Lastly, while the current four colon-delimited field HLA nomenclature system (which enumerates non-synonymous, synonymous and non-coding nucleotide variants in the second through fourth fields of an allele name) [74][94] provides insight into the types of polymorphism that distinguish alleles, this current nomenclature annotation does not identify the patterns and location of polymorphism across HLA gene features (GFs) at a given locus (since the extent of the nucleotide sequence represented by a HLA allele name cannot be inferred from that name annotation system). In this context, a novel gene feature enumeration (GFE) notation has been recently proposed as a supplement to this current HLA nomenclature for the purposes of: cataloging nucleotide sequence polymorphisms for non-ARD-encoding GFs; defining HLA alleles in the context of polymorphism distributed between GFs; and capturing novel nucleotide sequences for non-ARD-encoding GFs generated via NGS technologies [327].

### 9.3 Impact and Relevance of NGS Technologies on HLA Research and Clinical Applications

Within the human genome, the HLA system (namely the classic human MHC, ~4 Mbp in length) represents one of the most highly complex genomic regions with a mosaic nature of unique features, such as [206]:

- 1) Vast allele and haplotype polymorphism.
- 2) Very high gene-dense regions (e.g. class III region).
- 3) A high number of genes (~220 gene loci, encoding molecules participating in immune and inflammatory pathways but also molecules with nonimmunological roles) that display a complex and extensive LD structure (with, at the same time, certain known recombination hotspot regions that interrupt it).
- 4) Regions with numerous SNPs but also many other regions showing repetitive and extensive low-complexity and imbalanced sequence composition (e.g. STR regions, homopolymer stretches and AT- and GC-rich regions).
- 5) As well as regions with structural variation (e.g. inversions, deletions and duplications) and the presence of closely related genes (e.g. presence of paralogs) in these regions (e.g. two such regions contain the complement (*C2*, *C4A* and *C4B* in a copy-number variation (CNV) region) and *HLA-DRB* (*B1-B9*) genes (a segmental duplication region)).

In contrast to limited sequencing approaches such as SNP genotyping (only describing reference-based polymorphisms with limited allele resolution and until certain extent) [176][177][206] or exome-sequencing (ignoring non-coding regions known to be involved in control over expression, translation, regulation and protein presentation) [207]. New recently

developed NGS-based HLA genotyping methods (referring here to all its different targeted long-range deep-sequencing approaches based on both short-read (although presenting limitations of sequence coverage) and long-read (although currently not being very accurate) sequencing technologies) have shown the potential to describe, at a single-molecule level [207]:

1) Almost full-length (up to 4-field allele resolution) variation, with minimum level of ambiguity, in both coding and non-coding HLA regions thanks to extended genomic sequence coverage.

2) As well as to phase variants via phased-sequencing (in the case of long-read sequencing) and/or via de-novo assembly (in the case of short-read sequencing). Importantly, resolution of long-range haplotype structures with minimum levels of cis/trans ambiguity (at least within each locus via short-read sequencing, and potentially along the entire HLA region via long-read sequencing) may enable the description of intergenic variation (e.g. in order to understand cis- and trans-effects of regulatory polymorphisms involving the human MHC region as well as other related gene clusters located elsewhere in the genome) [309].

Based on this remarkable NGS potential, recent studies have started showing the significant impact of NGS-based HLA genotyping data for obtaining an in-depth understanding of the biological function-significance and underlying mechanisms of this complex HLA structural diversity, which is, in turn, of great importance for many relevant clinical and research applications. Major impacts, first main reported NGS-related findings and immediate future perspectives can be summarized as follows:

**1) Clinical transplantation:**

The use of NGS for HLA typing in clinical practice provides broader coverage (not only full-length HLA sequences but also including a higher number of HLA loci) and higher resolution and phasing of HLA genes in comparison to previous legacy methods.

a) Hematopoietic stem cell transplantation (HSCT): improved analytic accuracy and efficiencies (in cost (particularly, high volume HLA typing laboratories), turnaround time and throughput) of HLA genotyping by NGS (so far, only based in short-read sequencing since its scalability can meet the high test volume demands in the HSCT and registries setting) has been a key recent advancement for donor bone marrow (BM) and umbilical cord blood (UCB) stem cell registries and the field of HSCT. The degree of matching of HLA genes at the allele-level is the most critical determinant of immunologic compatibility between donor and recipient for HSCT [112-114][117][119]. In this sense, and in comparison to previous traditional HLA typing methods, the diagnostic value of NGS-based methods for HLA genotyping is reasonably well-established as it clearly improves unambiguous high-resolution HLA genotyping (lower error rates) and facilitates the identification of rare [208][209], null [210] and novel alleles (and, importantly, respective haplotypes) [172][211] (with important potential impacts in donor selection and transplant outcome), providing a very accurate HLA matching donor-recipient information (e.g. NGS advances have led to a demonstrated 3.5-fold reduction in genotyping error rates compared to Sanger Sequencing (SBT)) and an eventual simplified donor search process in which, as an example, donor confirmatory HLA typing step could become optional but not mandatory as it is currently) [212][213]. However, the clinical utility of these improvements has not been well-defined yet (i.e. explore the extent of diversity outside of the antigen recognition domain (ARD) and to determine the impact of this diversity on transplant outcome). Most updated current donor selection criteria (based on guidelines from National Marrow Donor Program /Center for International Blood and Marrow

Transplant Research (NMDP/CIBMTR)) for standard matched unrelated allogeneic HSCT donors consider high-resolution allele-level (2-field) matching of *HLA-A*, *-B*, *-C*, *-DRB1* and also, very recently included, *-DPB1* as essential for optimal outcomes, with compatibility for *-DQB1* and *-DRB3/4/5* as significant factors for consideration [214]. Thus, while in HSCT using donors matched to “G” and “P” group alleles is currently well-accepted and clinically tolerated [215] (although with some drawbacks such as that they might also contain null alleles), it is still unclear whether more stringent matching that could be enabled by NGS methods would improve clinical outcomes. In this regard, very recent published studies have suggested that when HLA genotyping is done at very high-resolution (up to the 4-field), including exons outside of the ARD, introns, and untranslated regions, it can significantly improve outcomes for recipients in unrelated allogeneic HSCT [216][217]. Nevertheless, as indicated by some other expert groups in the HSCT field [218], larger cohorts for further clinical outcome retrospective multi-center studies are still required in order to confirm this (e.g. by comparing HSCT outcomes for patients who are matched with their donors according to very recent standard-of-care criteria to those that are more accurately matched using newer NGS methodologies). Although it may be challenging because it will require very large sample sizes and at a wide geographic scale; since the impact of amino acid sequence variation caused by substitutions in exons outside ARD regions, specifically, in donor-recipient pairs will be difficult to assess in these HSCT outcome studies because these mismatches do not occur relatively very often [219]. Some of these first larger studies are now in progress (e.g. the CIBMTR has initiated a study evaluating ultra-high-resolution matching in a cohort of nearly 6000 recent transplants (from 2000 to 2017)) [218].

At the same time, from the perspective of HSCT, better knowledge of very high-resolution (up to the 4-field, provided by NGS-based HLA genotyping) HLA allele/haplotype

frequencies in worldwide and local (within each country/geographical region) donor registries (which may be representative of regional diversity within a given population/region) can certainly allow to: better delineate donor search strategy with a more accurate estimation of the probability of finding a match in a particular population and, consequently, to better adjust priority on waiting lists; to guide also the well-representative (with an adequate size of the given registry/bank showing optimal number of units and, especially, number of different HLA haplotypes) [472] and, at the same time, targeted collection of BM/UCB from under-represented populations/regions to focus donor selection on specific donor population groups; thereby, and overall, optimizing the development of stem cell repositories such as UCB banking and in BM donor programs [214][215][220]. Moreover, although UCB is primarily used in HSCT, the use of UCB-derived products for regenerative medicine and other transplantation applications is growing. Induced pluripotent stem cell lines (iPSCs) derived from the UCB of HLA-homozygous donors (presenting homozygous major conserved extended HLA haplotypes), for instance, have been established as a very suitable alternative to autologous iPSCs [221][222][545][546]. Hence, NGS-based studies have started to characterize the extent of HLA diversity in national BM and UCB donor registries (e.g. among others, American (NMDP), British, Argentinian, Dutch, Norwegian, German, Polish and Belgian donor registry populations) [179][223-227][474][476][943] as well as local donor registries (e.g. [221]), where, typically, substantial differences in HLA allele and haplotype frequencies exist between regional recruitment centers and neighboring regions within a given country [136][220].

Furthermore, very informative high-resolution HLA genotyping data generated by NGS methods is likely to require changes [215] and improved and complex developments of the current matching donor-selection algorithms (e.g. [228]) that may increase efficiency and



accuracy of the donor search process (e.g. refined matching strategies for unrelated donor search that may be more beneficial for HSCT outcome minimizing potential complications).

Also, as previously mentioned, NGS-based HLA genotyping approaches enable not only the description of coding (exon sequences) but also of non-coding (intron and UTR sequences) and intergenic polymorphisms which modulate gene expression and its regulation at all different levels (e.g. transcriptional control, translational control, intracellular processing (including protein folding and stability), cell-membrane presentation and cellular interaction) [207]. Therefore, in the context here of HSCT, NGS-based HLA genotyping data, ultimately, is also expected to allow the comprehensive characterization of HLA cell surface expression levels of mismatched alleles and the detection of possible known and novel/deleterious mutations (e.g. key amino acid substitutions, impact in mRNA splicing sites, impact in microRNA (miRNA) binding sites and identification of unknown stop codons that may result in no expression of alleles (defining novel null alleles)), which are not detected by legacy HLA genotyping methodologies, with limited sequence coverage and phasing capacity [76]. Importantly, this may contribute to reveal additional HLA allele mismatches and, consequently, to understand their biological (their putative immunogenicity) and clinical significance (permissive or non-permissive nature) that are not well-understood currently. At the same time, defined polymorphisms associated with specific HLA expression patterns may be surrogate markers of HLA haplotypes associated with higher risks of post-transplant complications (e.g. GvHD or engraftment failure) [215]. Important examples of the role (and its potential clinical relevance) of HLA expression pattern (associated with certain SNP variation in the UTRs) have been recently described in relation to *HLA-C* [229] (where *HLA-C* expression levels define permissible mismatches and risk for GvHD) and *HLA-DPB1* [230-232] loci in the context of HSCT. Interestingly, the clinical utility that the *HLA-DPB1* locus

plays in transplantation has grown significantly in the past decade, where HLA-DPB1 matching status has been found to be critically important in the HSCT outcome [204][230-232]. In fact, *HLA-DPB1* locus represents an important challenge for obtaining a maximum degree of HLA matching during donor search, since it presents a weak LD with the rest of the HLA class II genes within the given haplotype due to existing hotspot of recombination between DQ and DP loci [92] (as a result, the frequency distributions of extended haplotypes including *HLA-DPB1* locus are quite spread in human populations). Consequently, finding *HLA-DPB1* matched (thus, 12/12 matched, considering also *HLA-A*, *-B*, *-C*, *-DQB1* and *DRB1* loci) donors for a significant proportion of patients has been assumed to be unrealistic (being very unlikely) [233]. This fact, among others, has led and required to define a model of permissiveness (as a way to assess clinical significance/impact) for a HLA-DPB1 mismatch in order to maximize the chances for finding a matched donor [230-236]. Thus, transplants with well-tolerated (permissive) HLA-DPB1 mismatches seem to carry outcomes similar to HLA-DPB1 matched transplants, whereas non-permissive HLA-DPB1 mismatches result in higher transplant-related and overall mortality [233]. Currently, the permissiveness of a HLA-DPB1 mismatch can be determined either by the expression levels of the mismatched *HLA-DPB1* alleles [230-232] (Expression model criteria); alternatively, by the relative immunogenicity of T cell epitopes (TCE) of the patient and donor *HLA-DPB1* alleles [234] (Structural TCE model criteria), or by an in-silico functional distance score between these two previous models [235]. Traditionally, determination of the permissive or non-permissive nature of mismatched *HLA-DPB1* alleles in the selection process of unrelated donors for HSCT as well as the evaluation of risk for acute GvHD [112-114] have been assessed using the Structural T cell epitope (TCE) algorithm model criteria as it is based only on the sequence content of exon 2 of *HLA-DPB1* locus, encoding the considered critical ARD and routinely

tested by legacy typing methods [234]. Nonetheless, TCE model does not account for the level of expression of the *HLA-DPB1* alleles. The SNP rs9277534, within the 3'UTR of the *HLA-DPB1* gene, is an expression level marker (although it may not be directly involved in regulation of expression but it segregates with multiple undescribed potential regulatory polymorphic elements) defining two differentiated clades (high- (G) and low-expression (A)) of lineage-specific expression in *HLA-DPB1* locus. This SNP rs9277534 has been found to serve also as a marker for permissive HLA-DPB1 mismatches (to prospectively identify DPB1-mismatched donors who generate a permissive DPB1 mismatch against low-expression patient DPB1 alleles) and it has been associated with GvHD (when donor-recipient HLA mismatches involving a high-expression DPB1 variant in the patient) (termed as Expression model criteria) [230-232]. Thus, it is expected that complete resolution of HLA-DPB1 cis/trans ambiguities by NGS long-read sequencing methods (enabling the accurate definition of *HLA-DPB1* allele combinations) will allow to accurately model HLA-DPB1 permissible mismatches and the risk for GvHD (as well as to understand the level of association between these two current Structural and Expression DPB1 models [235][236]), phasing precisely polymorphisms in exon 2 (TCE model), polymorphisms of other relevant genomic regions (such as exon 3) and polymorphisms in expression level markers such as the SNP rs9277534 in the 3'UTR [204]. In addition, data from complete phasing of haplotypes within and outside the HLA region may allow to resolve not just the associated polymorphic markers but the specific discrete genomic elements/regions directly involved in the regulation and definition of these mentioned and other HLA expression patterns [207].

Application of NGS (providing a full HLA gene panel including typically up to 11 classical HLA loci: *HLA-A*, *-B*, *-C*, *-DRB1*, *-DRB3,4,5*, *-DQA1*, *-DQB1*, *-DPA1* and *-DPB1*) has also allowed to obtain extended HLA genotyping including loci less or not tested routinely (e.g.

*HLA-DPB1*, *-DPA1*, *-DQA1* and *-DRB3/4/5*) until now in the clinical transplantation setting. Thus, the impact of the mismatches of these traditionally untested loci (which it is thought to play also an important role in the HSCT outcome [118]) can be now evaluated. For instance, this has been found to be important when:

- i) Considering the HSCT patients' level of humoral sensitization against HLA (allo-immunogenicity) prior to transplant. Application of NGS allows full evaluation of the nature of HLA allele and antigen/epitope mismatches. Where, full NGS HLA gene panel, coupled with detailed antibody testing (solid-phase antigen immunoassays, typically using single antigen beads (SABs)) assessing reactivity for HLA antigens/epitopes, can be used in the analysis of donor-specific HLA antibodies (DSAs) in the patient [211][237].
  - ii) In addition, NGS-based HLA genotyping has proven to be useful in detecting unexpected mismatches. In the related allogeneic HSCT context [238], this can be observed in those instances when, for example, crossing-over events between HLA loci would create mismatches between donor-recipient pairs of related siblings. As it has been recently described [237] for unexpected mismatches found in *HLA-DP* loci and caused, most likely in this case, by crossing-over events in the intervening region between *HLA-DQB1* and *-DPA1*, where hotspot of recombination has been identified [92].
- b) Solid organ transplantation (SOT): many of the NGS-related applications and advancements that have been just mentioned for HSCT are also found similarly in the context of SOT [161]. At the same time, the SOT setting presents certain singularities, challenges and specific requirements in which recent application of NGS-based HLA genotyping has started to show also its value and potential contributions [161]. In this regard, as previously mentioned, HLA genotyping in SOT is necessary for determining not only HLA-matching

status between donor-recipient pairs but also for assessing patients' anti-HLA antibody profiles [119].

In the case of HSCT, both donor search (in related allogeneic HSCT probability of identifying a HLA-identical sibling donor mainly depends on the number of siblings; whereas in unrelated (URD) allogeneic HSCT average probability of finding a highly matched unrelated donor mainly is based on the ethnic origin of the patients (since worldwide registry of donors has an uneven representation of different ethnic groups (e.g. more Northern Europeans than African Americans)), on the frequency of the patient's HLA haplotypes and on the matching grade required (e.g. 8/8 or 10/10 or 12/12)) and HLA matching requirements are highly strict, in order to minimize critical life-threatening post-transplant complications such as GvHD or engraftment failure [112-114][215]. In contrast, in the SOT setting, although characterization of HLA polymorphisms (both genotypes (allele-level) and phenotypes (antigen/epitope-level)) of donors and recipients is essential and HLA matching affects outcomes and long-term graft survival [239], general clinical practice has been favoring utilization of relatively effective immunosuppressive drugs (although immunosuppression enables the prevention and management of early-phase T cell-mediated rejection, still the humoral allo-response is not well-managed by current treatments and it remains as a major source of late graft loss (late-phase B cell-mediated chronic rejection)) and less of HLA matching [161]. This is mainly because of three relevant limiting factors that exist in the SOT setting:

- i) Shortage of available organs (e.g. where a worldwide search for HLA-compatible donors is not feasible for logistic reasons). Thus, long waiting lists for deceased donors and limited numbers of living donors are the current scenarios.

ii) The humoral sensitization history against HLA antigens presented by each given patient (with a variable, unique and dynamic HLA antibody reactivity history profile). Moreover, HLA sensitization status has been traditionally evaluated by a panel-reactive antibody (PRA) assay (which is most commonly based on a solid-phase HLA antibody assay that uses up to ~100 microbeads coated with purified class I or II HLA molecules of a single specificity (SABs), and, thus, it can detect anti-HLA class I or II antibodies in patient's tested serum). Where the PRA score value reflects the expected percentage of representative organ donor HLA antigens that will potentially induce an allo-antibody reaction on the given patient tested and, thus, estimating the degree of "transplantability" for that patient. Consequently, a high PRA score value indicates that the given patient is primed (allo-sensitized) to react immunologically against a large proportion of the respective organ donor population and, thus, being considered at high-risk of developing transplant rejection. However, this PRA assay is method-dependent and does not consider all the diversity found in HLA allele frequency distributions between different ethnic groups. In contrast, a more accurate calculated PRA (cPRA), or also termed as virtual PRA (vPRA), correctly estimates the percentage of not suitable organ donors for recipient candidates and it is increasingly being used in organ allocation algorithms. As this estimation computes the PRA percentage using both anti-HLA class I and class II antibody specificities tested, which assigned to each patient the unacceptable HLA antigens/alleles, and the respective frequencies values of the assigned unacceptable HLA antigens/alleles (defined at 1-field or 2-field resolution) described in a representative organ donor population cohort [240].

iii) The highly polymorphic complexity of HLA system [56].

As a result, all these facts make absolute HLA matching in SOT virtually impossible, especially in such as clinical time-constrained context (being even more critical in the case of

deceased donor organ transplants than for living donation) [161][220]. Therefore, the currently adopted clinical strategy and practice for donor-recipient HLA assessment and matching in SOT establishes a balance of HLA antigens/alleles matching grade in the donor-recipient pair that can also fit properly with the given patient's HLA antibody reactivity history profile, and always compensated with the most optimal immunosuppressive medications and regimens, thereby reducing as much as possible the risk on DSAs formation after transplantation and, thus, graft failure (where early graft failure is significantly associated with de-novo DSAs) [161]. However, this current approach in SOT still leads to a significant average loss of grafts (~40%), as it has been evaluated during the last decade [241]. In this complex and restricted scenario, there are two other important aspects (in addition to the HLA matching grade as a criteria factor, and apart from standard medical considerations such as the evaluation of the medical necessity and feasibility between recipient candidates for a given available organ donor) that are also considered and analyzed for SOT in order to maximize the chances of organ donor availability and compatibility/matching (especially for hypersensitized patients) as well as to establish the most suitable and efficient organ allocation scoring system and recipient-donor pair selection criteria:

i) Defining the immunogenicity/acceptability grade of HLA mismatches: in contrast to HSCT, where HLA matching grade at the allele-level is the key element for the decision-making process and in which this immunogenicity definition of HLA mismatches has been so far only addressed for *HLA-C* (permissible mismatches based on expression patterns) [229] and *HLA-DPB1* (TCE Structural model and Expression model) [230-236] loci. In SOT, not only the number of HLA mismatches is taking into account as a potent predictor of transplant outcome (determining thus also organ allocation and recipient-donor pair selection criteria) but, importantly, the differential allo-immunogenicity (based on the

respective identified antigen/epitope load) of each of the given HLA mismatches is also considered. Being well-accepted that not every HLA mismatch may have an equal effect on promoting graft failure via activation of the allo-antibody response (due to preexisting (memory) and/or de-novo DSAs) [242]. Consequently, an organ donor with acceptable HLA mismatches of low immunogenicity may be found the most suitable option in the most common cases of not having a perfect HLA matched organ donor [242]. Defining HLA mismatch acceptability of organ transplant donors for recipients has traditionally been based on serologically defined HLA antigens (low 1-field resolution typing), known as antigen-based HLA matching or HLA antigen analysis [242][243]. Nonetheless, high-resolution (2-field) molecular HLA typing at the allele-level has resulted in increased knowledge of amino acid sequences of HLA alleles which has facilitated the identification of potential immunogenic HLA epitopes that may constitute the specific anti-HLA core targets responsible of these post-transplant allo-antibody responses [242]. HLA epitopes have been defined by some research groups as molecular entities denominated “eplets” [244]. These eplets contain amino acids that linearly can be continuous or discontinuous but are clustered closely together (radius of  $\sim 3 \text{ \AA}$  patches) and that represent the potential functional epitope of the antibody determining specificity in the given three-dimensional structure of the HLA molecule [244]. Thus, it is well-accepted that anti-HLA allo-antibodies specifically recognize a wide range of epitopes present on HLA antigens and that molecularly defined high-resolution alleles corresponding to the same low resolution antigen can possess different epitope repertoires. Hence, determination of HLA compatibility at the allele-level (or epitope-based HLA matching) represents a more accurate approach than at the antigen-level in order to identify suitable donors for sensitized patients in the context of SOT or platelet transfusions [119][245][246]. Where, HLA epitope analysis (performed by



programs such as HLAMatchmaker (B-cell epitopes) [247] or PIRCHE-II (T cell helper epitopes) [248]) consists on the comparison of amino acid sequences of the given HLA allele mismatches of a certain donor to antibody reactivity patterns of the respective patient to detect correlations between the two, based on the supposition that there is an association between amino acid sequence similarities or differences and antibody reactivity [249][250]. Therefore, this epitope-based HLA matching (or HLA epitope analysis) can be used to identify acceptable mismatches for sensitized patients and to develop permissible mismatch strategies for non-sensitized patients. Nevertheless, still there is not enough conclusive data to support the added clinical benefit of epitope-based matching over traditional antigen-based matching, as there is also an ongoing debate regarding the utility and suitability of high-resolution epitope-based HLA matching at the allele-level in the SOT setting [243][249-251]. In this sense, large-scale multi-center studies need to be conducted to confirm the potential utility of HLA epitope-based matching, which also needs to be standardized and validated (e.g. the threshold of eplet mismatches for developing adverse allograft outcome has not been well-established yet), before it can be recommended in widespread clinical practice [243]. At least, so far, it is thought that epitope-based HLA matching can definitely contribute to avoid future allo-sensitization with the development of anti-HLA antibodies and to allow selection of a suitable allograft for highly sensitized patients through virtual crossmatch (where physical crossmatch assay results can be predicted reliably based on patient antibody profile and donor HLA antigen/epitope mismatches, allowing donors' organs, especially from deceased donors, to be allocated more efficiently and without the need for a time-consuming prospective physical crossmatch) [243][252].

ii) Defining the specificity (i.e. HLA antibody recognition patterns) and nature of detected antibodies in patient's serum: testing or screening for anti-HLA antibodies is mandatory to identify pre-sensitized patients. Distinguishing anti-HLA antibodies from those against non-HLA antigens is very relevant in order to predict accurately the solid organ transplant outcome. Thus, high-resolution HLA typing combined with detailed SAB-based solid-phase antibody screening at the allele-level provides the best level of discrimination possible for a sensitive and specific detection of anti-HLA antibodies [161]. The identity and intensity of anti-HLA antibodies is useful not only in safely finding an organ donor for a sensitized recipient, but also in deciding on which sensitized patients require treatment (i.e. desensitization via plasma exchange, intravenous immunoglobulin (IVIG) or rituximab) prior to transplantation. With the introduction of the solid-phase bead-based multiplex flow cytometry assays for HLA antibody detection, the detailed characterization of HLA antibody specificities (SABs) has become possible with a high degree of accuracy [253]. However, it still presents technical limitations, particularly false positives due to denatured HLA antigens [254]. At the same time, still there is no consensus to define an appropriate cut-off for solid-phase antibody assays that allows accurate prediction of physical crossmatch [254].

Thus, having a full understanding of the complexities, singularities and challenges of HLA matching in the SOT setting. Several clinical laboratory groups have recently reported first important and compelling evidences of some of the areas and scenarios in which adoption of NGS-based HLA genotyping can be not only feasible but clearly beneficial in the context of SOT [255-258]. In summary:

i) Regarding living donation:

- Until now clinical laboratories supporting SOT programs have mostly relied on SSO-based HLA typing methods employed for high test volumes and on SBT technology employed for obtaining high-resolution HLA data, although with the previously described workflow, throughput and ambiguity limitations [35][76][85]. On the other hand, robust, single-pass and cost-effective NGS-based HLA genotyping approaches, which can be routinely performed with a relatively adequate turnaround time for living donation (e.g. 2-3 days in the case of NGS short-read sequencing strategies), allow both high-resolution typing for all 11 major classical HLA loci (eliminating almost all ambiguities, including the cis/trans ambiguities found in SBT) and a very suitable scalability (even higher than the one offered by SSO technologies) to meet the high test volume demand in the SOT setting.

- NGS-based HLA genotyping data can provide critical information in both solid organ pre- and post-transplant settings for living donation:

- \*At the 2-field resolution level (covering all the coding region of HLA genes):

- NGS HLA genotyping data is informative for living donation evaluations. For non-sensitized patients with multiple potential donors, the rank order of preferred donors is directly dependent on HLA 2-field typing and criteria for matching. Consequently, there is an optimal matching categorization, prioritization, and selection of donors during donor searching process.

- In living donation, availability of NGS HLA genotyping data (providing full HLA gene panels) of the given organ donor and recipient candidates avoids the traditional necessity of using inaccurate statistical imputation/extrapolation methods (based on described 1-field or 2-field resolution HLA haplotype frequency tables of a given

representative population cohort (e.g. [259] or HaploStats database from NMDP)) and other computational approaches for, ultimately, clinical decision-making in regards to patient management. At the same time, cPRA/vPRA score values can be more accurately and unambiguously estimated, when this estimation is based on NGS HLA genotyping data for describing the HLA allele frequency distributions of representative organ donor population cohorts, thus significantly improving the organ donor assignment, being this especially critical for hypersensitized patients [260].

- In combination with detailed SAB-based solid-phase antibody screening at the allele-level, NGS HLA genotyping data allows the fine-mapping of epitope mismatches during epitope analysis at the allele-level and, thus, assessing and monitoring patients' anti-HLA antibody profiles and recognition patterns with high specificity and sensitivity. So, in contrast to SSO and SBT legacy typing methods, NGS-based HLA genotyping enables the characterization of complete and unambiguous protein sequences of specific alleles, identifying mismatched epitopes in both ARD and non-ARD portions of the HLA molecule. At the same time, the own NGS high-resolution and unambiguous HLA genotyping data obtained from both the given prospective organ donor and recipient candidates allows to correctly interpret the results of the antibody reactivity patterns of the given recipient candidates (being of great importance especially for highly sensitized patients with difficulty receiving transplant offers) and also to detect and to troubleshoot possible common errors found in these SAB-based solid-phase antibody screening techniques (with a more thorough and credible evaluation of bead reactivity, which, in turn, also facilitates the subsequent correlation and interpretation of corresponding physical crossmatch results). Main examples are the following [258]:

- Discernment of allele-specific (particularly when certain DSA involve HLA alleles that are not discriminated by lower resolution methods such as SSO typing), possible anti-self antibodies, or artifacts in the setting of bead reactivity through accurate characterization of anti-HLA antibodies.
- NGS unambiguous 2-field HLA genotyping data enables the attribution of positive results to anti-HLA or non anti-HLA antibodies. Thus, DSAs can be characterized appropriately if the incompatible mismatched HLA alleles of the graft are known and represented by the SAB panel used for detection.
- Retrospective post-transplant NGS-based HLA genotyping can be useful for the appropriate selection of SAB bead/s for DSAs detection and monitoring (as it is necessary to detect antibodies developed against mismatched HLA alleles on the allograft) in sensitized patients, which can influence decisively in patient management (e.g. clinical decision-making about immunosuppressive medications and regimens). In this regard, de-novo anti-DQ antibodies, predominantly in the post-transplant setting and critically implicated in late-phase antibody-mediated rejection and allograft loss, are particularly challenging to properly characterize. Due to current SAB panels do not assess with enough specificity HLA-DQ antibody reactivity at the allele-level. Furthermore, when particular antibody specificities are excluded from the SAB panel used (i.e. when donor's alleles are not represented by available SAB panels), epitope analysis with NGS HLA typing data can provide the best alternative “surrogate” specificities for monitoring DSAs on the given patient still with enough accuracy.

- NGS unambiguous 2-field HLA genotyping data and subsequent epitope fine analysis can also help to identify characteristic antibody reactivity patterns due to common epitopes that may be present on different beads, having respectively lower mean fluorescence intensity (MFI) values than expected (the so-called “epitope sharing” phenomenon, being in part as a consequence of the known patchwork nature of HLA sequence polymorphism [145]). Where precise identification of shared epitopes (e.g. broad shared epitope Bw4) between different solid-phase beads is essential for proper interpretation of antibody screening results avoiding the underrecognition of DSAs (false negatives) and, in turn, favoring also the correct interpretation of initially unexpected positive physical crossmatches [261]. Therefore, NGS-based HLA genotyping allows to better understand DSA development or absence and assessing antibody cross-reactivity to products of related alleles and their potential involvement in antibody-mediated rejection [161].

\*At the 4-field resolution level (covering expanded exons, enhancer or promoter regions, introns and the untranslated regions):

- Although using donors matched to “G” and “P” group allele designations (i.e. focusing on the ARD only) still is well-accepted in the clinical transplantation setting (for both HSCT and SOT) [215][258]. Several recent studies have shown that immunogenic molecular HLA epitope targets of DSAs can be located also in non-ARD-encoding regions (e.g. leader, alpha 3 (in class I), alpha 2 (in class II), transmembrane and cytoplasmic domains), which are unambiguously defined only by NGS-based HLA typing methods [262-264] (e.g. DSA against the  $\alpha 2$  domain of the DQ $\alpha$  chain). Thus, this is calling into question current assumptions regarding immunogenic epitopes (i.e. epitopes must be present only on the mature protein,

solvent accessible, on the ‘top’ peptide binding surface (antigen recognition domain (ARD)) of the molecule, restricted to the same class as the antibody, and in the same position on the target allele if reactive to more than one locus) [247][248]. Therefore, non-ARD-encoding regions may have an important impact when performing virtual crossmatches in the SOT setting or, at the same time, in HSCT donor selection especially for patients with complex sensitizations [258].

- Characterization not only of all HLA exons but also of the enhancer or promoter regions, introns and the untranslated regions (5’UTR and 3’UTR) by NGS-based HLA typing methods can enable to assess in-depth the impact of HLA polymorphism on HLA expression and overall immune reactivity [207]. As recently reported in a recent journal issue with a series of studies describing illustrative examples (that are later described in this section) about the biological/functional role and significance of non-ARD regions of HLA molecules [265] as well as in other previously published reviews (e.g. [266]).

- Furthermore, despite non-ARD mismatches are presumed to be rare (low frequency) events, only application of NGS-based HLA typing methods can accurately evaluate the real frequency of these events and, thus, their clinical relevance [258].

ii) Regarding SOT from deceased donors:

- Although available NGS-based HLA typing methods are presently not viable for rapid turnaround prospective high-resolution (2-field) typing of deceased donors (i.e. having HLA genotyping results in less than 7h to minimize ischemia time, where traditional SSP and RT-PCR are still the most optimal methods for that strict turnaround time requirement). The highly time-effective (of the order of minutes to hours) workflow of

novel NGS single-molecule long-read sequencing technologies allows to consider a real “STAT” prospective high-resolution HLA typing protocol able to generate high-resolution HLA genotypes in time for deceased donor organ allocation [159][161][512][516][519]. Hence, one of the key aspects in the development and refinement of this type of ultra-fast HLA typing approach is to establish a rapid but efficient and specific targeted enrichment method (probably different from the current PCR-based systems that still require several hours) for the HLA genomic regions, which can also provide full-length and unambiguous characterization of the HLA polymorphism for all major HLA genes to be tested [159][161][512][516][519].

## 2) Population genetics:

Recent application of NGS-based HLA genotyping in population studies has started to enable a full assessment of HLA diversity (at both allele and haplotype (LD patterns) levels) at the genomic 4-field resolution level in worldwide human population cohorts (comprising both unrelated individuals [267-283] or nuclear families (trios/quartets) [284-287] respectively that are representative of a given well-defined population (>100 individuals per cohort)), being this a very important breakthrough in the field of population immunogenetics.

At the same time, NGS HLA population-based studies (which are presumed to have a relevant statistical power and level of representativeness and biological significance) are providing novel and relevant insights to better identify, interpret and understand the HLA genomic diversity (at both allele and haplotype (LD patterns) levels) in relation to a variety of contexts such as:

- a) Fine-mapping of HLA-disease associations (defining associations of HLA alleles/haplotypes with certain diseases (particularly those with an autoimmune component)) and pharmacogenetics (defining associations of HLA alleles/haplotypes with certain drug-



induced hypersensitivity reactions) [120][121][126][127][142][170][207][288-294]. Also, specific and trans-ethnic epidemiology programs on HLA linked diseases can definitely benefit from NGS HLA populations studies [288].

b) Filling unambiguously (by full-length sequencing and complete phasing of heterozygous positions even over long distances without remaining HLA genotype ambiguities) the gaps [146][172][204][463][475] in the existing incomplete HLA allele sequences IPD-IMGT/HLA database [87][295] and, consequently, enhancing, among other bioinformatics features, both reference-based and de-novo assembly algorithms as well as integrative parameters related to ethnicity frequency information as part of the HLA sequence data bioinformatics analysis process for assignment of HLA genotypes [296][463]. At the same time, the rapid and massive discovery of new HLA alleles (up to the 4-field of resolution) by application of NGS is causing the vast expansion of the most updated HLA alleles list in official databases (e.g. IPD-IMGT/HLA database) [362-364]. Consequently, this latter may represent an important challenge for the NGS-based HLA genotype calling process performed by software programs, which will constantly need to be addressed in order to be still compatible with clinical turnaround time requirements [296].

c) Establishing new or improved underrepresented population-specific reference HLA databases with improved accuracy (e.g. [297]). Since current reference HLA alleles sequence databases and related worldwide human populations genomic projects (e.g. 1,000 Genomes Project [305]) have historically been based largely upon European (also known as Caucasian or Caucasoid) populations [464], and the full extent of HLA diversity in African [465][496], Middle-Eastern [466][467] or Asian [468-470][497] populations as well as certain isolated and rural ethnic tribes/groups [471] still remains poorly understood [499]. Assessment of HLA diversity across worldwide human populations may also shed light into the evolution of

HLA polymorphism [104][496][497]. Hence, a comprehensive description of the global HLA genetic variation will be also very relevant to health and disease [499].

d) Evaluating the prevalence of HLA alleles (especially for those alleles previously belonging to an allele ambiguity group), and haplotypes, within a given population and across worldwide populations more precisely. In fact, very recent NGS-based HLA studies carried out in population cohorts [267-287] and bone marrow registries [224-227] have unexpectedly found a greater number of examples of HLA alleles with relatively common frequency that were previously considered to be rare (when only using legacy methods for HLA genotyping), and actually represent the most common allele of the 1st-2nd fields allele group [300]. For instance, a study of an Argentinian registry population cohort has described that HLA alleles such as *HLA-A\*80:01:01:02* or *-B\*15:03:01:02* appear to be more common than their “:01” counterpart at the 4-field [224]. Thus, application of NGS technology enables the characterization of intronic diversity within genes sharing exon sequences (even alleles closely related), obtaining a highly improved assignment of HLA allele prevalence and, consequently, determination of LD for the description of haplotypes [224]. This refined description of HLA allele prevalence (since it is influenced by natural selection [104]) may contribute to define in great detail HLA evolutionary mechanisms such as allele diversification and convergent evolution in modern populations [300]. Furthermore, this information is invaluable for updating and refining current CWD catalogues [137][300]. In this sense, a very new report (covering the time period 2012-2018, and thanks to late advances supporting cost-effective high volume DNA sequencing as well as the efforts of registries to grow and diversify their volunteer donor pool in terms of multiple ancestries and regional HLA variation found within a given same country [136][481]) has been recently published. It has been named as “3.0.0 CIWD” and it compiles an updated catalog of common, intermediate

(as a novelty) and well-documented (CIWD) *HLA-A*, *-B*, *-C*, *-DRB1*, *-DRB3*, *-DRB4*, *-DRB5*, *-DQB1* and *-DPB1* alleles from over eight million individuals using data (both G and P level resolution assignments) from twenty worldwide (World Marrow Donor Association (WMDA)) unrelated HSCT volunteer donor registries according to IPD-IMGT/HLA version 3.31.0 released in January 2018) [479]. Importantly, this study points out, and tries to address them as well, some of the current challenges (that still need further revision) in order to be able to establish the most representative, integrative and updated HLA CIWD catalog [479]:

-Over time (especially during these last two decades), the constant evolving ability to assign and report HLA diversity due to the almost exponential increase of described HLA alleles [87] in parallel with improvements in the allele resolution level (from 1-field to the 4-field routinely tested now via NGS) based on the advancements in chemistry and technology basis for HLA genotyping methods (i.e. initially using serologic typing of HLA proteins later replaced by DNA-based methods (which identified the presence or absence of specific polymorphisms through the binding of oligonucleotide probes or primers) and later, in turn, by DNA sequencing (first, SBT-based with many ambiguities in resolution and phasing, and most recently NGS-based with phased high-resolution and minimum ambiguity)) [35][76]. All these constant changes have created a wide variety of HLA assignments found in the millions of individuals listed in worldwide donor registries.

-Even now for NGS-based HLA genotyping methods, the lack of uniformity in relation to the given coverage (exons, introns, 5' and 3' UTRs) of the HLA sequence for the tested loci between the respective method(s) and reagents of different vendors and/or in-house developments has also created an additional level of variability within the databases of unrelated hematopoietic stem cell donor registries. Thus, this makes it difficult to determine if a given assignment truly reflects any unresolved ambiguity. For example, *HLA-*

*B\*51:01:01:01* might be assigned by a given laboratory based on a sequence that includes all exons and introns but not the 3'UTR. Alleles identified later that differ in this 3'UTR (e.g., *HLA-B\*51:01:01:02*) make the assignment ambiguous. Therefore, standardization of this 3- and 4-field nomenclature accompanies differential practices among registries, both in typing methodology and standardization. Likewise, it is not clear from the HLA assignment what level of resolution was applied. For example, a two-field assignment, *HLA-A\*01:01*, is not clear as to what alleles are included (e.g., *HLA-A\*01:103* because it is included in the *HLA-A\*01:01:01G* group) or excluded (*HLA-A\*01:87N* because it is a non-expressed allele and the assignment was provided without the "P" because of registry specifications). Thus, this type of variations has presented a major challenge for determining accurately the frequency of individual alleles.

-Inclusion and accurate estimation of the worldwide distribution of null alleles (there are about 464 class I and 124 *HLA-DRB1*, *-DRB3*, *-DRB4*, *-DRB5*, *-DQB1*, *-DPB1* non-expressed alleles (IPD-IMGT/HLA version 3.31.0) described [87]) are still pending.

-Overrepresentation/underrepresentation of worldwide human population groups (represented by the different unrelated hematopoietic stem cell donor registries) as well as considering accordingly geographic/ancestral/ethnic categories with the difficulty of their broad variability.

Lastly, future studies based on novel long-read sequencing strategies may contribute to a more comprehensive and diverse analysis of HLA and, thus, improve estimation frequencies in different geographic/ancestral/ethnic population groups to keep updating and refining these CWD catalogues.

e) Assessing signatures of demographic events and selective pressures reflected in HLA population variation as well as ancestry remoteness and relatedness (to shed light on possible common origins of certain populations/ethnic groups throughout human peopling history and more recently migration waves and/or admixture episodes in modern populations) between different populations and geographical regions [104][137][301].

f) Unravelling in-depth the vast diversity (e.g. identifying routinely novel, null and rare alleles) of the HLA system [172][211][302-304][323] and, thus, having a better understanding of the evolutionary mechanisms (and the respective putative evolutionary relationships) that shaped the HLA diversity patterns observed in human populations [104][137][273][307][308] (which has been mostly described only at the ARD-encoding exons level until recently [305][306]) and, importantly, also identified more precisely the main immunobiological roles of HLA loci and other linked genes within the human MHC region [104][137][178][288][309].

g) Studying the impact of coding and non-coding variation (and also the phased variation within and between HLA loci) on HLA expression (at both transcriptional (RNA) control and translational (protein) control levels) [309], on HLA biology and function at different levels (e.g. intracellular processing (including protein folding and stability, and binding of the peptide), cell-membrane presentation and functional cellular interaction and signaling) [265][266][288][309-317], in relation to susceptibility/protection to infectious diseases [309][318-322][513], and, as previously described, on immune reactivity in the context of transplantation [204][211][229-237][255-258].

h) Implementation of NGS-based HLA typing methodologies (enabling the sequencing of previously unsequenced regions of HLA alleles) in the routine clinical practice has allowed

the rapid and massive discovery of new alleles [172][211][302-304][323]. Which, in turn, may be a potential challenge to the current four colon-delimited field HLA nomenclature system nomenclature naming conventions and, thus, it may require to be honed [74][94][178]. Therefore, a proper updated HLA nomenclature must be assigned by the WHO Nomenclature Committee for Factors of the HLA System [74][94], in order to: address the growing numbers of new alleles being discovered; to reduce naming complexity; and to provide the most informative nomenclature system possible (e.g. to better distinguish 5'- and 3'UTR variations from the intronic ones). Furthermore, several more precise and adapted HLA allelic polymorphism and ambiguity complementary reporting systems (as previously described) have been developed and are in process to be established for this new and very recent era of NGS-based genotyping in the histocompatibility and immunogenetics field [324-327]. It is noteworthy the recently implemented GFE system (which describes the distribution of polymorphism between the various features that are known for HLA alleles) that complements the current HLA allele nomenclature (which only defines silent, replacement and non-coding polymorphisms) [327][365]. At the same time, it is expected NGS-based HLA genotyping data will provide a more comprehensive serological equivalent list for HLA class I and II alleles (e.g. compiling a list of HLA class I and II alleles (especially in the case of novel and infrequent alleles) for which serological patterns and the serological assignments or equivalents (i.e. association of HLA alleles to serologically defined HLA antigens) have not been identified (e.g. [518]), may have been incorrectly categorized or not listed yet) and, thus, to update the last 2008 HLA Dictionary (whose information has been traditionally contributed from several sources such as: WHO Nomenclature Committee for Factors of the HLA System, the UCLA International Cell Exchange Program, NMDP, recent publications and individual HLA laboratory groups) [86].

Referring back to the HLA population genetics topic itself (as just previously summarized in the listed point “e”), NGS-based HLA genotyping studies of worldwide population cohorts have allowed a very comprehensive HLA assessment of population diversity, in which molecular distances between alleles and haplotypes in terms of nucleotide differences (including genetic variation found at non-coding regions) can be now taking into account [104][137][178][301]. In this regard, and as expected [104][137], it has been found how NGS-based HLA genotyping data at the 4-field significantly furthers genetic distance computation (e.g. when being evaluated by phylogenetic dendrograms) between and within populations (e.g. see [328]). Furthermore, NGS-based HLA population studies have revealed, so far, both some unique and striking as well as some other more commonly shared 4-field extended haplotype associations (e.g. *HLA-DQA1\*05:01:01* variants, *HLA-B\*18:01:01* variants, *HLA-C\*05:01:01* variants or *HLA-C\*06:02:01* variants), previously undescribed at 2-field resolution by legacy molecular HLA genotyping methodologies, among worldwide human populations/ethnic groups (e.g. see [224][268][286][287][328]). As a very specific example illustrating this latter observation; in a recent large US European American population cohort the two 4-field variants *HLA-C\*06:02:01:01* and *HLA-C\*06:02:01:03* are distinctly associated with *HLA-A\*30:01:01* and *HLA-A\*29:02:01:02* respectively [268]. Thus, a significant portion of “silent” 3-/4-field HLA polymorphism, previously undescribed in population genetics studies, may be relevant to infer geographical and/or historical relationships between human populations [137]. In this sense, the information relative to 4-field extended haplotype associations at the population level can be extremely helpful, as an example, for selecting matched unrelated donors in HSCT and as a first reference source for future studies and current database registries, such as Bone Marrow Donor Worldwide (BMDW) database (<https://wmda.info/>) (World Marrow Donor Association (WMDA) took over the activities of Bone Marrow Donors Worldwide (BMDW)

and the NetCord Foundation since 2017), that may present incomplete genotyping data for some of the HLA loci. Moreover, some recent NGS-based HLA studies have shown that comparison of the nucleotide diversity at different coding and non-coding regions of HLA genes (which are presumably driven by distinct types of selective pressures) may facilitate to disentangle demographic from natural selection effects on HLA allele/haplotype diversity patterns at a given population/geographical region (e.g. *HLA-A*, *-B* and *-DRB1* genes may underwent similar selective pressures in the Uyghur (Central Asia) and Ami (Taiwan) populations which present very distinct demographic histories) [137][178][301]. In parallel, more sophisticated statistical methods, based on high-resolution HLA haplotype frequencies, are required to resolve demographic events such as the recent admixture between populations which can mask the ancestral haplotype frequency distribution and present an important challenge, as an example, for the HLA matching process in the transplantation setting [329]. Also, a more accurate generation of HLA phased data (at both the level of SNPs within a locus and among loci describing haplotypes) can be achieved now via NGS sequencing and, consequently, it is possible to obtain a highly improved and more reliable estimation of LD values and patterns of HLA haplotypes at the population level. Being this LD parameter very useful to understand in-depth worldwide human population history and past migrations [104][137]. For instance, LD may serve for the assessment of ancestry remoteness of populations as this parameter is expected to decrease with time through recombination events [26][137]. In fact, it has been generally observed how LD values for different HLA loci pairs tend to be low in populations of remote ancestry. Thus, and in line with the “out-of-Africa” migration theory of worldwide human population dispersion outside Africa, some studies have shown how lower LD values have been detected on African and African-descent populations (more likely to be populations of remote ancestry) in comparison to other European, Chinese or Japanese population cohorts



(being these more likely to represent more modern populations) [26][137][330]. At the same time, recent NGS HLA studies have revealed that African and African-descent populations present a relatively rich repository for HLA genetic variation (especially considering the high levels of heterozygosity detected across HLA loci), which might have arisen in the African continent before human dispersion [279][328]. In fact, these more recent observations are also in line with previous studies. Which already showed that in African populations, the genetic distances (although only based on HLA allele/haplotype class I frequency distributions) between each other are greater than those observed between European (Caucasoid) populations. Thus, the remarkable current allelic and haplotypic diversity in the HLA system as well as their variable distribution in different sub-Saharan populations is probably the result of evolutionary forces and environments that have acted on each individual recent population and/or in their earliest ancestors [496]. Furthermore, in regards to the effect of the quality of the phasing of NGS HLA genotyping data (so far, mostly generated by NGS short-read or 2<sup>nd</sup> generation sequencing platforms) on the accuracy of LD estimation for 2-locus haplotypes; interestingly, a recent study [331] has described the existing overestimation of LD (higher values) for EM estimated haplotypes (using the statistical EM algorithm approach, typically carried out for the inference of haplotype frequency distributions in population cohorts comprising unrelated individuals, where the phasing between HLA loci is uncertain [101][332-336]) in comparison to LD values estimated by accurately phased HLA haplotypes (most typically built by family-based allele segregation analysis [101][337-339]). This observed singularity is even more ostensible in those 2-locus haplotype pairs known for presenting a weak LD, where a broader range, and also more rare, of combinations of these loci occur (e.g. *HLA-A* ~ *HLA-C*, *HLA-B* ~ *HLA-DRB1* or *HLA-DQB1* ~ *HLA-DPA1*) [92]. This aforementioned finding may be explained by the well-known fact that the EM algorithm approach not only estimates haplotype

frequency distributions from unphased (or phased-unknown) HLA genotyping data but also considers the assumption that all HLA loci are under Hardy-Weinberg Equilibrium Proportions (HWEP). When, in reality, HWEP deviation is commonly observed at multiple HLA loci in population studies due to, as an example, population stratification among other causative factors [268][340][341]. Thus, these aspects may be causing that EM algorithm approach underestimates the frequency of rare ( $n < 3-4$ ) haplotypes, and thereby LD values of 2-locus or other extended haplotypes are overestimated [268][286][331]. Nonetheless, EM algorithm approach is still considered a very useful tool for inferring haplotype frequency distributions (despite their estimation is not completely accurate, particularly at low frequencies), especially from HLA genotyping data of registries comprising unrelated potential HSCT donors [179][223-227][259][299][474][476][481][943] as well as in the case of population-specific cohorts of unrelated individuals [267-283] which, in both cases, generally lack information on family pedigrees. On the other hand, HLA allele segregation analysis of NGS HLA genotyping data in nuclear families generally allow to establish a very accurate phasing (intra- and interlocus) and, thus, a robust determination of HLA haplotype (via family pedigree analysis or phasing analysis with Mendelian constraints), including the reliable identification of haplotypes encompassing rare, null or novel alleles that exist in the given general population [284-287]. Moreover, HLA family-based studies [101][287] are especially invaluable for the assessment of distinctly admixed and/or not very well-characterized populations (e.g. the distinctive nature of highly admixed Argentinian [224] or Mexican Admixed [267][522][547-552] populations which include an important variety and disparity of ethnic backgrounds of Amerindian, European, African and even Asian origin within the general population). In addition, NGS HLA family-based studies facilitate (in those instances when the most common segregation pattern of alleles of HLA loci, inherited as HLA chromosomal haplotype blocks from both parents to

their offspring [100], is not observed) the detection of potential chromosomal crossing-over events between HLA loci (as a consequence of occasional existing meiotic recombination) [92]. Although there are some exceptions and limitations that make difficult to determine, in general, a very precise and absolute recombination fraction for HLA haplotypes (e.g. crossover events cannot be determined reliably among trio families; also, only in families that have more than two children, it is usually possible to identify the specific parental haplotypes that participate in a given crossover event; on the other hand, even in families with two or more children, crossover events within homozygous HLA loci may not be detected) [284-287][331]. Also, despite it presents a high robustness and accuracy, the family-based approach for phasing and determination of haplotype frequencies is not very feasible for large-scale studies at the population level, in contrast to the EM algorithm approach in large-scale study cohorts with samples collected from unrelated individuals of a given population [342]. Consequently, as a main drawback, unless significantly large numbers of families at a highly wide geographical scale can be studied, HLA allele and haplotype (although very accurately phased) distributions reported by family-based studies (generally presenting a non-random nature with some plausible biases) may not be very representative, reflective and generalizable of their respective larger populations and, thus, neither meaningful for anthropological analysis [284-287][331]. Nonetheless, large family-based studies can be used to refine segregation patterns of alleles in worldwide populations. Thus, these NGS studies may contribute to construct accurately high-resolution HLA haplotype databases that can be useful for, as some main examples: improving NGS HLA genotyping software applications and, especially, their algorithms for the assembly of consensus sequences (especially for challenging loci to be correctly genotyped such as *HLA-DPA1*, *HLA-DPBI* or *HLA-DRB3/4/5*) [284][331]; enabling and facilitating, together with the clonal property of NGS platforms, the identification of rare variant sequences as well as

detection of PCR and sequencing errors of given NGS HLA genotyping methods [343]; and downstream analyses for studies in fields such as: disease association (e.g. which may facilitate, by discarding possible hitch-hiking effects, identification of specific allele within an haplotype as the determinant factor for a certain disease association) [207]; transplantation (e.g. more confident predictions of match grade in HSCT) [211-220]; evaluation and detection of chromosomal crossover events and/or potential HLA genotyping errors (e.g. homozygous overcalling at *HLA-DPAI* locus (with only a single consensus sequence resolved by NGS instead of two) or accurate haplotype designation and determination of copy number variation of *HLA-DRB* genes) [331][344]; and validation of new HLA analysis tools, novel HLA genotyping technologies (including cross-platform validations) and statistical estimation approaches of HLA haplotype frequencies and LD values at the population level [284-287][331][345][346]. At the same time, despite these current limitations found in both HLA population family-based and studies of unrelated individuals using currently available NGS HLA typing technologies (so far, mostly based on NGS short-read sequencing platforms); the very recent application (although still in evolving development) and optimization of NGS long-read or 3<sup>rd</sup> generation sequencing approaches, it is expected to facilitate this complex analysis and phasing of the entire HLA haplotype region and hence enhance estimation of the LD parameter [159][165-173][204]. It has been extensively reported that natural selection (in the context of the relevant and complex role of HLA molecules in the immune response and the related immunoregulatory mechanisms, and in addition to populations' demography and migration contributions) seems to operate on the observed diversity of HLA alleles at different loci and the LD between neighboring and non-neighboring loci [22][23][56][103][104]. In turn, both of these may influence on the distribution of respective extended haplotypes [100]. Unlike traditional molecular HLA genotyping methods, NGS-based HLA studies (both based on

unrelated individuals or families within a given population) are also able to report maximum extended 9-locus (including all main classical HLA class I (-A, -B, -C) class II (-DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3/4/5) loci) haplotype associations at the highest 4-field allele resolution level. So far, one of the main general observations about these haplotype frequency distributions in different NGS HLA worldwide populations studies has been that many identical haplotypes across 7 loci (comprising *HLA-A~B~C~DRB3/4/5~DRB1~DQA1~DQB1*, and excluding *HLA-DPA1* and *-DPB1*) become extremely divergent in terms of the multiplicity of *HLA-DP* alleles with which they associate [268][286][287][331]. Thus, as previously mentioned, since *HLA-A~B~C~DRB1~DQB1* haplotypes carry very differing *HLA-DPB1* associations (which is even more pronounced at the 3-/4-field resolution), these associations also impact considerably (and, in fact, negatively) the likelihood of finding unrelated donors (URD) in HSCT [233]. This seems to be especially due to the weak LD between *HLA-DP* and the rest of the class II haplotype since existing hotspot of recombination is present between *HLA-DQ* and *-DP* loci [92]. At the same time, the non-coding sequence variation confers an additional higher level of diversity which in combination with weak haplotype associations (as it is observed for *HLA-DP* loci) imply a substantial widespread distribution of extended HLA haplotypes presenting very low frequency values that, in turn, require to be precisely estimated and where it is also required a large sample size of the respective given population cohorts in order to be representative and meaningful for evaluating anthropological aspects based on these 4-field extended HLA haplotype frequency distributions [268][286][287][331].

It is also noteworthy that the majority of these cited NGS HLA population studies (and their respective previously summarized findings) [267-287] (including the NGS HLA Spanish population study that is part of the present thesis work [269]) were greatly spurred by and carried out as part of the activities, components and projects of the past 17<sup>th</sup> International HLA

and Immunogenetics Workshop (17<sup>th</sup> IHIW), held in September 2017 [347][348][487]. The 17<sup>th</sup> IHIW was primarily focused on the applications of NGS technology in the histocompatibility and immunogenetics fields, establishing a preliminary multicenter quality control project [349] to validate various NGS HLA (and also KIR separately) genotyping platforms and related genotyping software analysis programs and, importantly, creating an innovative centralized database [297] for: HLA genotyping data collection (including both sequence and allele-name/genotype data according to v.3.25.0 IPD-IMGT/HLA (released July 2016)) for the different 17<sup>th</sup> IHIW components and projects; for its visualization; and for the posterior data management and analysis with different computational tools (e.g. to submit, to report and to transfer accurately NGS HLA genotyping data; to calculate allele and haplotype frequencies; to estimate LD; to validate HLA haplotypes; or using HLA DNA sequence alignment tool for reporting novel variants in the generated consensus sequences) [324-327][331][350-354]. Thus, under the auspices of the 17<sup>th</sup> IHIW, full-length description of coding and non-coding variation of classical HLA genes and also, separately, of KIR genes [355] were analyzed at high resolution through the application of various NGS platforms (mostly of 2<sup>nd</sup> generation as well as a few of 3<sup>rd</sup> generation sequencing methods) in the context of different components and topics: population genetics (worldwide unrelated- and family-based population studies) [267-287]; refined and extended molecular characterization of HLA genes of the International Histocompatibility Working Group (IHIWG) reference cell lines (representing an important resource for quality control purposes, validation and optimization of genotyping methods for HLA and KIR gene clusters as well as MHC extended haplotype studies) [356]; pharmacogenetics and disease association (for example, see [291-294]); immunogenetics of aging [357]; HSCT [309]; and mapping of serologic epitopes [262][263] among other main projects. Also, in addition to optimized and standardized short-read [179-193]; very initial (still

in development) long-read [159][165-171][358] NGS technologies; and NGS dual-combined approaches [172][173][204][359]; there was a 17<sup>th</sup> IHIW component dedicated to the full-length hemizygous Sanger sequencing approach (based on group-specific priming, and especially designed for genotyping HLA class I genes) as a valuable alternative to unambiguously identify and confirm (by resolving genotype ambiguity due to cis–trans polymorphism and allele ambiguity due to polymorphism located outside the ARD region) the genomic sequence of alleles with novel variants, alleles with unknown intron and exon sequences, or consensus sequences obtained by PCR-based short-read and long-read NGS technologies that may be difficult to interpret due to, for instance, existing homopolymer stretches or STR regions [360][361]. Overall, the main breakthrough of this past 17<sup>th</sup> IHIW [487] was the completion of a very refined and full-length characterization of extended HLA, and also KIR, genomic reference sequences defined by allele name and respective consensus sequences at the population level using different NGS platforms (enabling an extensive phasing across genes, minimal genotyping ambiguity and in-depth description of allelic diversity) together with related resources and tools available for the international histocompatibility and immunogenetics community and to continue completing and updating (by replacing or confirming incompletely defined sequences) [146][463] the IPD-IMGT/HLA database [87][362] with full-length HLA sequences as well as for KIR genes in the respective IPD-KIR database [363][364]. At the same time, there are still a number of remaining challenges to be addressed, which have been now set as goals for the upcoming 18<sup>th</sup> IHIW in 2021 in order to achieve: improved automated annotation and curation systems of sequence variation and its related nomenclature [365] (e.g. which may be able to deal with either partial-gene sequence and/or multiple consensus blocks per gene provided by many NGS platforms as well as defining SNP locations within the genes); and application of newer developments in HLA and KIR genes

(as well as for other human MHC and LRC genes and/or complete regions) sequencing platforms that: can generate more comprehensive, extensive and reliable consensus and/or directly generated sequences; that can also circumvent phasing uncertainties, especially of distal polymorphic positions, and resolve rest of remaining ambiguities, especially those associated with DNA regions presenting repetitive and extensive low-complexity and imbalanced genomic sequence composition (e.g. homopolymer stretches; STR regions; high AT- or GC-rich regions) [366]. Moreover, in relation to NGS technology and its application to HLA genotyping, it is expected that current major NGS limitations (e.g. time required to sequence and analyze data, costs, coverage, sequencing error rates and multiplexing capabilities) may be overcome by introduction of novel sequencing instruments with varied improved chemistry approaches and engineering design [367][368]. Ultimately, characterization of unambiguous and full genomic allelic and haplotypic polymorphisms of the HLA (and KIR, respectively) region with an exhaustive description of all possible existing HLA-allele and -haplotype SNPs at the population level (and also at the individual level, in the context of precision medicine and personalized treatment) may contribute to understand and to elucidate which particular variants (and at which particular genomic locations: regulatory, untranslated, coding or non-coding region) are functionally relevant, hence having a crucial effect on immune system phenotypes (regarding antigen presentation to immune effector cells and its regulatory influence in downstream humoral and cellular immune responses) [120] and being intrinsically linked to certain risk factors and risk stratification in the context of, as some of the main examples, transplantation (donor registries) and disease associations [358][369][370]. Nonetheless, due to the vast genomic diversity within the human MHC region (unambiguously, extensively and fully described at high-resolution now via NGS technologies), the evolutionary and functional significance of the HLA region may be difficult to assess and define (especially from a statistical



point of view) in most cases as a consequence of the relatively low frequency distributions observed at 3-/4-field high-resolution HLA alleles and haplotypes, being this especially the case of rare non-coding sequence variants [104][137][152][301]. Within the context of HSCT, it is highly unlikely to have a fully 10/10 matched URD identified for patients affected by haematological disease [371], and even more unlikely now as a consequence of this increased level of HLA polymorphism (at coding and non-coding sequences) described, since the HLA haplotype frequency distributions are considerably more spread out at these higher resolution levels. In this sense, the choice of haploidentical donors for HSCT may represent a more plausible and relevant alternative (in which, in spite of that level of HLA disparity, so far it has shown favorable outcomes in the engraftment, a decline in the rate of GVHD, and an improvement in the morbidity and mortality of patients, similar to those reported for unrelated HSCTs) when a fully matched donor is not available [372][373]. At the same time, the expected, and already current, huge and rapid increase of reported HLA alleles (newly discovered and/or whose genomic sequence has been now entirely characterized) and respective related haplotypes, based on the massive amount of HLA genomic phased sequence data generated at high-resolution (and hence immense HLA diversity being described) in the era of NGS, also present a series of new challenges and possible pitfalls that definitely require new analytical and statistical approaches and methods in addition to a thorough revision of, as some main examples: so far reported disease-associated alleles or haplotypes, reference cell line panels and HLA sequence databases, CWD (and CIWD recently defined) catalogues and donor-recipient matching algorithms and likelihood estimates in the transplantation field [104][137][152][153][178][265][301][479].

Current NGS-based HLA population studies enable the description at high-resolution of new common alleles as well as new rare alleles (especially at non-coding regions) although in a

context of both limited sample size and incomplete geographical and regional coverage of worldwide population cohorts studied so far (in comparison to the given vast HLA genomic diversity that has started to be found recently via NGS). Moreover, these current sampling limitations certainly avoid an entirely accurate estimation of allele frequency distributions, as it is also the case of haplotype frequency distributions. Thus, the HLA genetic profile of a given population cohort may not be still very truthful and representative at such high resolution levels, which require (as a statistical condition) very large population study cohorts at a highly wide geographical scale to be conducted [137][301]. At the same time, evaluation and definition of rare alleles/haplotypes have to be consistent with geographical origin/distribution and linkage disequilibrium [374]. Furthermore, it is also important to consider the contribution of rare alleles/haplotypes to the overall genetic diversity pool (i.e. the overall HLA allele repertoire) as well as the rapid turnover that may occur in a significant number of HLA alleles [137][375][376], since generation of novel MHC immune gene variants confers a selective advantage in host-pathogen coevolution [104][137][301][377][378]. Consequently, statistical assessment of population relatedness (e.g. considering the computation of genetic distances using phylogenetic dendrograms) and other related population data analyses (e.g. HLA evolution or disease-association studies) only based on allele/haplotype frequencies may not be suitable at this 3-/4-field allele resolution level. Since there is an inaccurate estimation of rare alleles as well as a reduced statistical power (and, thus, a lack of reproducibility) because of this rarity and higher population divergence [137][379]. In this sense, a series of new statistical methods (including the use of computer simulation models and neural network programs) [104][137][380-382] and approaches (e.g. coalescent theory; sample frequency spectrum (SFS) and polymorphism/divergence ratio; excess of identity by descent (IBD) regions; long-term shared polymorphism; or composite likelihood), most of them based on the interrogation of

HLA sequence instead of simply allele/haplotype frequency data, have been suggested and initially tested [104]. Nevertheless, these new strategies may still require further optimization and a higher level of robustness, accuracy and consistency with previous studies [104][137]. Overall, direct examination of human MHC region sequence (providing invaluable information not only regarding molecular HLA polymorphism but also in relation to, for instance, mutation/recombination rates and linkage disequilibrium) by these new analytical and statistical methods may: facilitate identification and discernment of natural selection (where several selective regimes account for the patterns of variation of HLA loci) and demographic (e.g. population size changes, migrations or admixture) effects on the evolution of different HLA polymorphisms, respectively, at various timescales; as well as to assess the effect of epistatic interactions (e.g. HLA-KIR) and epigenetic mechanisms (e.g. DNA methylation, histone modifications and non-coding RNA (such as miRNA)) in order to unveil also the complex crosstalk between genetic and environmental factors [104][137][288][521].

In relation to the remaining previously listed **a)** to **h)** points, some of them are further discussed in the next pages showing recent and significant applications of NGS-based HLA genotyping in some other related fields at the population-level.

### **3) Disease Associations (as previously listed on point “a”):**

Till date, over 160 complex diseases (including autoimmune diseases, cancer and infectious diseases) [27][120][383-390][513] and multiple hypersensitivity reactions (phenotypically distinct T cell mediated ADRs) [110][126][127][391] (see respective **Figures I-15** and **I-16**) have been relatively and/or significantly associated with a variety of numerous HLA variants (haplotypes, alleles and/or their respective SNPs structures) as well as other gene polymorphisms (generally immune-related and commonly in close linkage with the human

MHC region). Nevertheless, the current understanding and knowledge of the biological significance (and underlying molecular mechanisms) of these HLA disease- and hypersensitivity-associations have been defined only partially and still remain very limited [121][207], with only few exceptions (e.g. *HLA-B\*57:01*-associated abacavir hypersensitivity [392-394]; or the proposed “shared epitope-coding *HLA-DRB1* alleles” hypothesis in rheumatoid arthritis (RA) [395-397]). Thus, despite the human MHC region shows the highest number of disease associations across the human genome [120], fine-mapping and deconvolution of discrete causal variants in the human MHC are still uniquely challenging and remain elusive due to the main following reasons [121][207][398]:

(a) The extreme (and still not completely described [146][463]) sequence polymorphism (e.g. it has been estimated that HLA genes may contain on the order of millions of alleles per locus [306][376]) and the complex mosaic structural genetic architecture (e.g. numerous SNPs; regions showing repetitive and extensive low-complexity and imbalanced sequence composition (such as STR regions, homopolymer stretches and AT- and GC-rich regions); loci in extensive LD, with also plenty of recombination hotspots that may break the haplotype structure; as well as many structural variations including gene inversions, deletions, duplications and copy number variants (CNV) [206]) which may play an important role in the haplotype structure and, thus, function. Furthermore, HLA allele polymorphisms that are associated with diseases may include both ARD-encoding regions (defining the peptide binding and presentation but also the specificity of the entire HLA-peptide-TCR interaction [399][400]) and non-ARD-encoding regions (for protein elements like transmembrane and cytoplasmic regions, which can also affect activation of the engaged cells) [121][265][266]. Moreover, these HLA sequence and structural polymorphisms greatly vary (at the allele/haplotype frequency level and also, in the case of haplotypes, at the level of LD strength

and length) between human populations and ethnic groups/geographical regions [309][401]. At the same time, due to this extensive LD, for many reported associations between a disease phenotype and a particular variant in the human MHC region, it may not be feasible to determine whether the variant is causal or whether that association simply reflects LD with the true causal variation [402].

(b) The high density of human MHC genes, which are in extensive LD and are tightly related to innate and adaptive immune functions, show a pleiotropic nature. Since many of these clustered MHC genes are involved in a variety of both physiological and immune-mediated disease pathways (where MHC gene clusters encode a series of proteins for cellular and extracellular antigen presentation to circulating T cells, inflammatory and immune-responses, heat shock, complement cascade systems, cytokine signaling, and the regulation of various aspects of cellular development, differentiation, and apoptosis) [403]. Thus, many immune-mediated diseases (e.g. autoimmune diseases) show an important overlap between disease-associated loci (even finding in some instances discordant associations between autoimmune diseases) [207][385][404]. Also, it is noteworthy that the majority of these immune-mediated diseases are polygenic, heterogeneous and multifactorial, in which, so far, an incomplete penetrance of known HLA genes or loci has been observed [207][383][405]. Another aspect to consider is the concept of evolutionary-genetic tradeoffs, having the 8.1 ancestral haplotype (AH) (containing *HLA-A\*01*, *-C\*07*, *-B\*08*, *TNFAB\*a2b3*, *TNFN\*S*, *C2\*C*, *Bf\*s*, *C4A\*Q0*, *C4B\*I*, *DRB1\*03:01*, *DRB3\*01:01*, *DQA1\*05:01*, *DQB1\*02:01*, *-DPA1\*01*, *-DPB1\*03:01*, *-TAP1\*01:01*, *-TAP2\*02:01*) as a prototypic example. Since it has been widely reported the association of this extremely long 8.1 AH haplotype (spanning ~3-4 Mb) for conferring susceptibility to autoimmune disorders (as a deleterious effect) on one hand, and protection against infections (from selective pressures) on the other hand, which both may be also under

the influence of other genetic and environmental co-factors [406][407]. Furthermore, beyond functions related to host defense, it has been recently reported how HLA and non-HLA genes of the human MHC cluster may be also involved in neurobiological processes including both synaptogenesis and synaptic pruning [401][406]. Thus, immune dysregulation of a certain set of human MHC genes (including HLA genes) may also influence susceptibility towards neurological diseases [290-294]. In addition, small molecules and human metabolites may dramatically interfere with the peptide repertoire and, thus, immunomodulating the HLA-peptide-TCR interaction, which, in turn, may be also implicated in the etiology of certain autoimmune/inflammatory diseases [408][409].

(c) Human MHC genes (including HLA and non-HLA genes, and their respective molecules ultimately) show a coordinated “multigene” and “long-range haplotype” (both cis- and trans-) interaction (in a likely allele/haplotype-specific pattern) to both regulate and carry out their functions [207]. However, as an example, only few studies have interrogated the variation and phasing of human MHC class III clusters (such as *TNF*, *BTNL2*, and *C4*) with the classical HLA class I and II alleles [401]. Besides HLA-disease association studies, in the transplantation field there is also an increasing interest in the full characterization of the gamma block (which contains numerous inflammatory and immune regulatory genes) within the human MHC (as well as other less characterized MHC blocks so far) [27] and the evaluation of the impact of gamma block matching on clinical outcomes in HSCT [410][411]. Moreover, it is also noteworthy the regulation of expression and, consequently, biological functions of human MHC genes by epigenetic mechanisms (e.g. DNA methylation, histone modifications and non-coding RNA (such as miRNA)) acting at key regulatory sites especially located in non-coding regions (e.g. enhancers, silencers, promoter and untranslated regions (5'- and 3'-UTRs)) [121][288][521]. In addition, gene-gene epistatic interactions (e.g.

HLA-KIR) also seem to contribute significantly to disease associations involving the combined presence and interplay of variants within and outside the human MHC, showing a higher overall risk for the given disease phenotype [104][121][290].

(d) The technical limitations and inherent biases of traditional analysis approaches and sequencing/genotyping technologies in studying human MHC gene variation (and particularly HLA genes) and its association with disease phenotypes. First, assessment of the differential presence/absence of HLA alleles/haplotypes in a patient-disease cohort versus a considered healthy control cohort has been the most traditional and extended approach for establishing the HLA association of specific diseases [412]. Nonetheless, many of these candidate gene studies have also indicated that HLA variation alone may not be sufficient as a susceptible/protective genetic factor to deduce the underlying mechanism of those diseases initially HLA-associated [401]. At the same time, this traditional case-control approach presents important limitations, especially for the distinction between the effects of HLA loci that show a strong LD complicating the deconvolution of discrete causal factors [402]. Thus, in order to try to differentiate the effects of tightly linked loci under this traditional approach, this type of studies require to be conducted with very large sample sizes and by analyzing an enough number of different ethnic groups. This may facilitate the identification of recombination events and, thus, the detection of the real and distinctive causative HLA variation pattern [412]. In addition, conditional analyses can be applied to separate allelic from haplotypic association (i.e. thus, to discern hitchhiking effects of given detected associations) [402]. Another methodological challenge in this type of traditional studies is the so-called “population stratification”. In which unrecognized ethnic/regional differences between the disease and control population groups can be detected as genetic HLA-disease associations that are actually not related to the disease of interest. Thus, designing ethnicity-

matched case-control studies with a consistent sampling strategy and large sample size (that even may allow stratified analyses) can contribute to minimize this type of issue [384]. Second, current GWAS (including HLA genotype imputation; although being very informative for indicating possible regions of interest, it still presents limited density of SNP arrays and a lack of allelic high-resolution, LD and specific population/ethnicity-reference data) [413] or exome sequencing (being unable to evaluate non-coding variation) [414] approaches present limited ability to detect and interconnect the actual causative genetic variants and their role in the given disease phenotype [121][207]. In fact, GWAS data itself has shown that 90% of causal autoimmune disease variants reside within non-coding regions of the human genome, with many of these disease-association SNPs mapping to the human MHC [415]. Moreover, within these 90% non-coding variants about 60% are mapping to immune cell enhancer-like elements most of them involved in transcription of non-coding RNAs such as miRNAs, which may also play a role in the pathophysiology of certain HLA-associated diseases [121][288][416]. Therefore, not only HLA structural domains involved in peptide binding but also regulatory variants operating at the human MHC region are important for understanding both the evolution of HLA polymorphisms at the population-level and the role that they play in disease [104][309][521].

All of these previous considerations show the inherent difficulty in fully deconvoluting the causative aspects of HLA-associated diseases. In order to overcome these difficulties, more comprehensive approaches have been recently developed, which are focused on targeted deep-sequencing strategies (based on NGS) of the entire human MHC region [195] (allowing for long-range haplotype structure to be fully resolved in phase; although, so far, current sequencing methods have described only homozygous DNA samples) in conjunction with multidisciplinary analysis methods (for instance, based on pattern recognition and complexity theory [417]) to



decipher more accurately causal nucleotide changes, regulatory factors and interactions of extended genomic regions being part of complex underlying mechanisms associated with disease phenotypes [121]. In addition, recognizing the relevance of non-coding variants and their association with disease phenotypes, many efforts have been devoted to identifying DNA elements and to describing their variation as well as their interactive relationships and, in turn, their regulation in gene expression [121][309]. In this sense, it is noteworthy the contributions made so far by the ENCODE (ENCyclopedia Of DNA Elements) consortium [418] as well as by several studies that have examined expression quantitative trait locus/loci (i.e. SNP variants altering gene expression, named as eQTLs) [419][420]. These identified SNPs (operating as single SNP units or as several haplotype-specific SNPs) may influence expression over several HLA genes within an haplotype or on a certain HLA allele [309] and at different genomic locations (cis-eQTLs and trans-eQTLs) [401]. Thus, human MHC polymorphism also displays a transcript diversity, where regulatory variants of gene expression may be a key component of biological mechanisms underlying the MHC-associated phenotypes and diseases [421]. Therefore, the comprehensive global map of regulatory gene expression variation may facilitate the fine-mapping of disease causal variants within an associating locus or region contained by the human MHC [401]. Overall, complex HLA-associated with disease risks and adverse effects of drugs may be caused by a combination of variants: operating at both cis- and trans-; being both rarely and commonly frequent; located in coding and non-coding regions; and from within and outside the human MHC [207]. Recently, studies based on 4-field resolution assignments of HLA alleles/haplotypes using NGS approaches have started to elucidate more accurately the role of HLA variation in the multifactorial etiopathogenesis of complex diseases [288-294] and adverse effects of drugs [422][517].

At the same time, application of NGS-based HLA typing methods may be also useful for a better characterization of clinical phenotypes and for monitoring and predicting more accurately the given disease course and the response to respective current and novel (even individualized) therapies as well as conditioning regimens (in the case of HSCT) [288]. In this sense, NGS technology presents a high depth of sequencing coverage enabling not only a high-resolution characterization of HLA polymorphism (even from challenging, but precious, biological sample sources (such as umbilical cord blood (UCB) samples) for DNA extraction [473]) but also a reliable detection of minor and rare mutational genomic events that were unattainable by previous legacy methods when studying diseases [121]. Thus, application of NGS has allowed to evaluate more precisely the role of HLA and non-HLA genes in complex diseases, having also a significant impact in cancer biology and cancer immunotherapies [121][424]. In relation to HLA, NGS may allow to have a better understanding of the immune escape underlying mechanisms exploited by cancer [387][510][942] including HLA allele-/haplo-specific loss heterozygosity (LOH) and/or reduction in expression (impaired at different levels: during transcription or affecting some of the elements of the antigen processing machinery) of intact and functional HLA molecules loaded with the respective peptide (where ligandome originating from cancer-associated antigens is also important to be characterized [314][423]) on the cell surface [386-390]. At the same time, NGS-based HLA genotyping facilitates considerably the detection of loss of HLA function mutations in cancer patients (e.g. presenting myelodysplastic syndromes (MDS)). Being critical in those cases to confirm the presence of novelties (especially for the verification of detected null alleles) in the germline from a non-blood based source (e.g. buccal swab saliva sample), since this information determines the search for the appropriate HSCT donor [425].

It is also expected that all this knowledge (provided by NGS-based HLA allele and haplotype frequency distributions datasets at the worldwide population-level) can contribute to establish strategies for improving the efficacy of both current and novel immunotherapies (that can be also used in combination with, or alternatively to, the application of monoclonal antibodies and/or cytokines) [488-490]:

- Which are mainly based on the use of HLA restricted peptide/epitope-based vaccines to treat both cancer and infectious (viral/bacterial/parasitic) diseases.

- As well as those adoptive autologous/allogeneic cell-based therapies (e.g. infusions of TCR-engineered T cells, tumor-infiltrating lymphocyte (TIL), Chimeric Antigen Receptor (CAR) T cells or virus/bacteria/parasite-specific T cells) for treatment in the context of oncology and also to combat viral infections (for instance, caused by Epstein-Barr virus (EBV) or cytomegalovirus (CMV) and that are commonly found in the setting of HSCT in patients with primary immunodeficiency disorders (PIDDs)) in addition to bacterial and parasitic diseases.

Since a thorough and most updated depiction of HLA diversity and respective HLA allele/haplotype frequencies panel (especially of the most common ones) of a given population group may be highly informative for developing and establishing an optimal and effective (clinically as well as cost-wise) design of the following main examples as for therapeutic approaches that have been recently explored so far:

- Updated and extended population-specific panel of the most common HLA alleles (in order to reach the majority of targeted patients from the corresponding population) as potential novel targets for development of more efficient peptide/epitope-based vaccines in therapies for treating cancer and infectious diseases, and in particular for those patients presenting a highly resistant and/or refractory profile. Here, it may be also essential to take into consideration the specific HLA allele repertoire restrictions [70][89][553] of either malignant

tumor- or infection-related antigen presentation on the given diseased recipient tissue in order to induce efficient and specific anti-tumoral/anti-infectious immune responses that, at the same time, can be clinically safe for the patient.

-Updated and extended population-specific panel of the most common full (i.e. encompassing at least the 11 major classical HLA class I and class II loci) extended HLA haplotypes, in this case, may serve to define and construct the most suitable (in terms of histocompatibility barriers to be considered) and representative (considering an adequate size and comprised diversity of the given therapeutic cell registry/bank) population-specific allo-donors pool of cell-based therapy products. Where these therapeutic allogeneic genetically engineered (e.g. via lentivirus/retrovirus vector systems or via more recent and sophisticated gene editing systems such as CRISPR/Cas9) T cells are collected and stored (generally via cryopreservation) in a donor registry or bank specific for a given population, or even locally organized per region within the same country, in order to cover the majority of the given targeted patient/recipient population to be treated for cancer and infectious diseases respectively. Here, although most of the current cell-based therapy products are still of autologous origin (the so-called *per se* personalized medicine approach (which is highly costly and labor-intensive) where the cells source is directly from the own given patient); alternatively, the more recent selection and use of allogeneic cells from donors (thus called allo-donors) show clear and significant advantages especially in relation to required logistics, coverage, volume and costs associated to the development of these therapeutic genetically engineered T cells registries/banks. As a consequence, the selection criteria and prioritization (ranking) of allo-donors, that made-up this type of registries as providers of these cell-based therapy products, is primarily conditioned by considerations on histocompatibility barriers since these therapeutic allogeneic genetically engineered T cells do express HLA molecules

that most likely cause rejection (or even GvHD effect) events when there are HLA mismatches between respective donor and potential targeted recipient (similarly, as it occurs in the case of strategic donor recruitment and planning in HSCT (using in this case BM and UCB cell sources from related/unrelated allo-donors)). On the other hand, there may be cases in which given HLA mismatching situations could be used for therapeutic purposes (e.g. GvL effect in patients with cancer), however these desired therapeutic effects are still very difficult to predict and to clinically manage. Consequently, construction of registries of therapeutic allogeneic genetically engineered T cells exclusively selected from HLA homozygous allo-donors (making up the so-called haplo-banks/-registries/-repositories) has been proposed [491] to be an effective (and more adequate) way to match a maximal number of targeted patients receiving this type of cell-based therapy. Hence, this approach actually consists on achieving the matching on only one of the two HLA haplotypes (ideally, one of the most commonly found within a given population) of the recipient in the absence of allele mismatches for the other haplotype due to the haplo-homozygosity of the donor T cells' HLA [492][950].

In the regenerative medicine field, and similarly to the above mentioned application, moderate sized banks of iPSCs lines, once again, exclusively selected from HLA homozygous allo-donors can be constructed. Where a diverse set of these HLA haplo-homozygous iPSCs lines respectively carrying the most common HLA haplotypes of a given population (once again, based on respective HLA haplotype frequency distributions dataset) can make up this population-specific repository (which can be also locally organized per regions within the same country) [492][493][950]. Thus, differentiated cells from iPSCs (since these present unique properties of self-renewal and plasticity to be differentiated *in vitro* to many diverse cell lineages) can be generated depending on respective clinical need (e.g. various

conditions/injuries that can be potentially treated are hematopoietic disorders, musculoskeletal injury, spinal cord injury, cardiovascular injury, liver damage; however, there are limitations such as safe iPSCs infusion, post-treatment adverse effects and standardization of protocols to generate large amounts of pure good quality differentiated cells originally from iPSCs) [494]. Importantly, in addition to HLA matching requirements based on this HLA haplo-homozygosity of the given donor, a certain patient can only receive iPSCs-differentiated cells from a respective healthy donor, that has been previously screened and tested, and is considered free of any known disorder or condition [494]. Interestingly, as previously mentioned, in addition to its role in HSCT (e.g. to treat blood diseases and inherited metabolic disorders), there are new and emerging uses of umbilical cord blood (UCB) units in regenerative therapy and immune modulation, including the generation of iPSCs derived from UCB. Consequently, an haplo-bank of most frequent HLA haplo-homozygous iPSCs lines can be fittingly and effectively constructed starting from an existing UCB registry population which has already been characterized for HLA genotypes [221][222][495][545][546].

#### 4) HLA diversity and evolution (as previously listed on point “f”):

Characterization of DNA variability at both coding and non-coding sequence regions by using NGS has been found to be very informative for deciphering in-depth the HLA diversity and the related evolutionary processes and origins found in human populations. Based on the most updated known and described (greatly improved with application of NGS technology although still with important gaps in the related reference HLA sequence databases [87][146][463]) HLA allelic polymorphism and diversity, it is noteworthy some of the general and preliminary observations/patterns (which some appear to be common whereas some other may vary across human populations) that NGS-based HLA population studies have shown so far:

(a) In relation to classical HLA class I loci, although *HLA-B* locus most commonly presents the highest allele diversity at the 2-field resolution level (defining protein-coding alleles) in comparison to *HLA-A* and *-C* loci. Interestingly, these two latter HLA class I loci appear to show a relatively higher allelic diversity at the level of synonymous (3-field) and non-coding sequences (4-field) in comparison to *HLA-B* locus. Therefore, these differences found in the HLA allelic diversity (between the protein (2-field alleles) and nucleotide (3- and 4-field alleles) levels for these HLA class I loci) may suggest that these different HLA class I genes likely underwent dissimilar selective forces that have been shaping HLA allelic diversity and defining its mode of evolution according to specific pathogen-driven selection mechanisms and/or demographic events over time on a given population or ethnic group of a certain geographical region [104][136][137][260][301]. At the same time, introduction of NGS sequencing has allowed to assess the fine nucleotide molecular diversity between different genomic regions across HLA loci (for both class I and class II) at the population level. Where such observed distinctions between individual HLA class I and/or class II genomic regions may reflect not only their specific modes of evolution but also, consequently, their roles played in the immune process [301]. For instance, a recent study evaluated the genomic variability of the entire *HLA-A* gene by using massively parallel NGS sequencing (including all exons and introns and regulatory segments (containing also the extended promoter)) in a highly admixed population sample from Brazil [307]. In that particular study, it has been observed how *HLA-A* promoter regulatory segments present few but divergent sequences, which are in close association with the coding regions and, thus, might be presumably under the same positive selective pressure. Whereas, as possible exceptions of the high genetic diversity found at *HLA-A* locus, the 3'UTR segment seem to be highly conserved (this conservation is probably related to the maintenance of an adequate *HLA-A* expression pattern

via an important post-transcription regulation) as well as the evidence of purifying selection at exon 4 (tentatively, maintaining an invariable or similar  $\alpha 3$  domain to assure a proper  $\alpha 3$ /CD8 interaction) [307]. Another recent publication about the Mandenka population from Eastern Senegal is also among the first HLA studies that has compared the molecular diversity between different HLA regions at the population level by using NGS sequencing [273][301]. This latter study has shown, for example, how the highest genetic diversity at HLA class I loci seems to be lying at exon 3 (encoding the  $\alpha 2$  domain of HLA class I molecules, with a tentative greater involvement in peptide presentation in comparison to  $\alpha 1$  domain encoded by exon 2), with the exception of *HLA-C* locus presenting a lesser variability of both ARD- and non-ARD-encoding codons in comparison to *HLA-A* and *-B* loci [273][301]. Thus, by analyzing full-length HLA sequences, several aspects of the genetic diversity within the main classical HLA loci have been so far evaluated in these NGS population-level studies, such as: the level of association between regulatory segments with the respective coding allele regions; the definition of distinct patterns of molecular DNA diversity among certain exons and introns which, together with certain population parameters, allow to better postulate and to discriminate between the effects of, on one side, positive selection or demographic expansion and, on the other side, balancing selection or demographic contraction; [301]; moreover, the identified differences in these HLA regions at the population level may be also explained, at least until certain extent, by signatures of adaptations to peculiar environments (e.g. distinct pathogen richness or prevalence) during human evolution globally and/or regionally [137][301]; in addition, NGS HLA studies may allow to examine more in depth (at both ARD and non-ARD regions) proposed HLA evolutionary mechanisms, that seem to be even opposite to each other in certain instances, such as: the hypothesis of divergent allele advantage (which explains the high levels of heterozygosity found at HLA genes in worldwide



open populations) [426]; and, conversely, the recently proposed model of joint divergent asymmetric selection [427] (which is based on the multi-locus and pleiotropic effects that HLA genes (e.g. observed in classical HLA class I loci) may coordinately present in relation to its main biological function of antigen presentation as part of the immune response) that may explain the fact that some small isolated human populations often exhibit a reduced diversity at individual HLA genes and HLA regions in strong LD [104][137][301].

(b) Regarding classical HLA class II loci, so far, it has been generally observed how *HLA-DPA1* and *HLA-DQA1* loci, that encode alpha subunits ( $DP\alpha$  and  $DQ\alpha$  respectively), present a striking high allelic diversity at the 4-field level unlike what is observed at the 2-field level. In contrast to this previous observation, the counterpart *HLA-DPB1* and *HLA-DQB1* loci (encoding beta subunits ( $DP\beta$  and  $DQ\beta$  respectively)) display a higher allele diversity at the 2-field level instead of at the 4-field level. Similarly for HLA-DR molecules, *HLA-DRA* locus (encoding the respective alpha subunit) shows virtually no protein sequence diversification whereas *HLA-DRB1* and *HLA-DRB3/4/5* loci (which encode the respective beta subunits of HLA-DR molecules) present a high allele diversity at the 2-field level but not as elevated at the 4-field level. Hence, it can be conjectured that classical HLA class II “A” (*DPA1*, *DQA1*, *DRA*) loci, despite worldwide human population diversity, appear to be under high selective pressure in which synonymous and non-coding allele variants are predominantly generated over protein-coding allele variants. In this way, from a functional perspective, allele diversification at the protein level (defining contact positions of these alpha subunits for the binding with the respective beta subunits) of HLA class II “A” loci may be restrained and relatively conserved to facilitate and ensure the pairing with the respective wide allele range of classical HLA class II “B” loci (*DPB1*, *DQB1*, *DRB1/3/4/5*) [268]. Moreover, since HLA class II “A” (*DPA1*, *DQA1*, *DRA*) loci variants are mostly located in non-coding regions, these

5-UTR/3'UTR/intronic polymorphisms may be involved in regulatory functions related to cell surface expression of respective HLA class II molecules and the stability of their peptide binding groove [178][288][309]. In fact, this shows a very similar pattern as it is observed in the case of  $\beta 2$  microglobulin, which is the stable non-polymorphic structural component (associated non-covalently to the given heavy  $\alpha$  chain encoded by *HLA-A*, *-B*, *-C* genes respectively) of classical MHC class I molecules [27][35][268]. In contrast, and in comparison to classical HLA class II "A" loci, the classical HLA class II "B" loci present a higher allele diversity at the protein level defining an extensive peptide binding repertoire that is intimately related with their immunobiological role for antigen presentation [56][104][268]. At the same time, recent refined analysis (at a fine molecular scale of the DNA sequence variation) in certain human populations of these classical HLA class II "B" loci have shown how *HLA-DRB1* and *HLA-DQB1* loci display a high allele diversity at the ARD-encoding exon 2; whereas the *HLA-DPB1* locus exhibits much greater variability at non-ARD-encoding exons than at the ARD-encoding exon 2. This may suggest existing distinct functional roles and modes of evolution between these classical HLA class II "B" loci [273][301][428]. Furthermore, full-length HLA sequencing enabled by NGS (including both 2<sup>nd</sup> and, especially, 3<sup>rd</sup> generation platforms) has allowed the interrogation of the previously underestimated genomic variation in certain HLA genes considered to be monomorphic or highly conserved (at least at the protein level) such as *HLA-DRA* locus. In this sense, some recent studies have shown how SNPs and polymorphism clusters within the introns and 3'UTR region of *HLA-DRA* locus (in spite of being not as polymorphic as the other class II alpha genes (*HLA-DQA1* and *HLA-DPA1*)) define distinct gene lineages, which, in turn, facilitate the identification and definition of *HLA-DRA~DRB3/4/5~DRB1~DQB1* haplotype patterns [429]. A more comprehensive HLA sequence assessment by NGS has also contribute

to shed light on understanding the evolution (defining allelic/haplotypic lineages/clades) and biological functions of the DRB region, particularly the previously not well-characterized second expressed DRB genes (*HLA-DRB3/4/5*) [308][430]. For instance, a very recent and thorough study has evaluated coding and non-coding regions (including analysis of the variations of simple repeat stretches with the basic structures poly(A), poly(T) as well as microsatellite (GT)<sub>x</sub>(GA)<sub>x</sub> motif of STR repeats (especially those in the vicinity of the polymorphic exon 2 region)) of *HLA-DRB5* alleles. In this regard, SNP analysis has shown that most of the diversity is located in the segment defined between exon 1 and intron 3 in comparison to other downstream segments of *HLA-DRB5* locus. Furthermore, that study points out that plausible coevolution with proximal exons may suggest a relevant biological/regulatory role (that needs to be further evaluated) of these composite intron microsatellites (GT)<sub>x</sub>(GA)<sub>x</sub>, which define specific *DRB5* lineages and also associate distinctly with haplotypes and ethnic groups [308][431]. In relation to the relatively low diversity found in the coding region of second expressed DRB genes (as it is *HLA-DRB5*), this study also conjectured that elective pressures such as convergent evolution may be acting and, thus, the exon variability would be restricted to that of the intron microsatellites due to their physical proximity [308][431]. At the same time, based on the prominent variation identified in *HLA-DRB3/4/5~DRB1~DQA1~DQB1* haplotypes, it has been previously postulated that the different respectively encoded HLA-DQ, -DP, and -DR molecules may complement coordinately their biological functions in antigen presentation and regulation of the immune response [409][432]. Similarly, other NGS studies have been focused on *HLA-DPA1* and *HLA-DPBI* loci, describing in-depth variations in coding regions and previously undescribed variability in non-coding regions [359][433][434]. In *HLA-DPA1* locus, its intron 1 shows many SNPs and indels which contribute to the striking nucleotide diversity of this

long segment (~3.5-3.6 kb) as well as it contains length polymorphisms; whereas much shorter (~0.2-0.3 kb) intron 2 and intron 3 are not that variable in size and present low polymorphism [433]. In relation to non-coding regions of *HLA-DPBI* locus, its intron 2 appears to present the largest nucleotide variability. In turn, this *HLA-DPBI* intron 2 harbors one variable STR region (the tetranucleotide motif (AAGG<sub>(4-17)</sub>) that is adjacent to the 5' side of exon 3. Interestingly, it has been observed how this particular composite intron microsatellite repeats motif distinguishes two clades of *HLA-DPBI* alleles: one group presents only 4 repeats (denominated as “short STR”); while the other has 8-14 repeats (termed as “long STR”) [433]. Furthermore, it has been found how 3'UTR variants of *HLA-DPBI* [230-236], which correlate with high (rs9277534-G) and low (rs9277534-A) cell surface expression, also associate tightly with these mentioned short and long *HLA-DPBI* intron 2 STR (AAGG<sub>(4-17)</sub>) length variants, respectively [433]. Nevertheless, further functional studies are still required to better clarify the plausible molecular mechanism (e.g. alternative splicing (with the splicing defect involving exon 3) or, on the other hand, transcriptional repression) of the regulatory role of these short/long STR variants of intron 2 in the expression of *HLA-DPBI* gen [236]. At the same time, recent identification of unexpected extensive polymorphism in the promoter region of *HLA-DP*, may also contribute for a better understanding of the regulation of the cell surface expression of this heterodimer [434]. In relation to the 3'UTR region, the SNP rs9277534 (G/A) correlates with two clades that are fully differentiated by 174 fixed polymorphisms throughout a stretch of this 3'UTR end of *HLA-DPBI* locus, where A-clade alleles (including the including the most frequent *HLA-DPBI*\*04:01, *-DPBI*\*02:01 and *-DPBI*\*04:02 alleles) may have experienced a more recent divergence in comparison to those G-clade alleles [359].

(c) In addition, the high-coverage capability of NGS-based HLA genotyping methods has enabled to initially obtain a full description of the nucleotide diversity along the major non-classical HLA class I (HLA Ib) genes, HLA pseudogenes (e.g. *HLA-H*), class I-like genes *MICA/MICB* and non-classical HLA class II genes in large cohorts (i.e. in a population level approach) at high-resolution. Some striking and novel findings of the following recent studies are noteworthy:

*i)* Recent studies have described the full-length *HLA-G*, *-E* and *-F* gene variability (including the regulatory segments (both distal and proximal promoter regions and 3'UTRs), all exons and introns) in several different worldwide populations by using NGS technology (e.g. see [279-283]). In these studies, application of NGS technology has allowed the identification of variable sites and haplotype structures (e.g. promoter/coding/3'UTR haplotypes) along different genomic regions of these non-classical HLA class I loci. Interestingly, it has been found that sequence variability is much greater (especially at the regulatory regions) than previously thought, where known and new *HLA-G*, *-E* and *-F* coding alleles have been also described. Nevertheless, these gene coding regions remained relatively conserved, since all the coding alleles (described so far) converged to encode few molecules for each gene [279][280]. Hence, these findings may illustrate the immunomodulatory role (with well-characterized immunomodulatory activities, including downregulation of the innate and adaptive immune responses and the induction of tolerance) of these non-classical HLA class I genes as well as their role as immune checkpoint genes in tumor immunosurveillance, transplantation or protection against infection diseases [65-68]. Since low variability at their coding regions (likely under high selective pressures and particularly in relation to the non-synonymous substitutions) may sustain certain optimal protein modifications at the ligand-binding sites that critically

define the molecular binding to a corresponding wide diverse range of leukocyte receptors on target cells and, thus, to ensure this biological regulatory function (which occurs at various different contexts such as pregnancy, transplantation or disease). Whereas the higher variability found at regulatory regions (e.g. enhancers, promoters and 3'UTRs) may facilitate a plausible wide range of binding sites to transcription factors (mainly at the promoter region) and post-transcriptional factors, such as microRNAs (mainly at the 3'UTR region), that may directly influence expression both quantitatively (e.g. gene variation may account for the magnitude of protein production) and qualitatively (e.g. gene variation may account for protein modification) [279][280] in a likely allele/haplotype-specific expression pattern. At the same time, functional studies are required to define the exact mechanisms underlying the tentative correlations found between, for instance, mRNA expression levels and specific given allele/haplotype groups [279][280]. In addition to the full description of allelic polymorphism of *HLA-E*, *-F* and *-G* genes, NGS sequencing also allows to determine the poorly understood linkage disequilibrium between classical *HLA* class-Ia and non-classical *HLA* class I-b alleles, with special interest (previously described only by very few studies [435][436][498]) between *HLA-A* and *HLA-G* genes which are in relative close physical proximity within the human MHC genomic region [27][436]. In addition to obtain a more comprehensive evaluation of worldwide genetic diversity and linkage disequilibrium for non-classical *HLA* class I genes (*HLA* class-Ib genes), *HLA* class I pseudogenes can be also evaluated in-depth by NGS-based typing methods. Notably, *HLA-H* pseudogene (defined as a non-functional gene because it is deleted at different frequencies in humans and because it encodes a potentially non-functional truncated protein) remains scarcely explored. In fact, recent studies (although not via NGS yet in this case) have pointed out that this *HLA* pseudogene may have, unexpectedly, relevant functional roles (e.g. *HLA-H*

transcriptional activity and HLA-E mobilization at the cell surface by the HLA-H signal peptide) [500]. Thus, further insight provided by NGS studies into *HLA-H* genetic diversity may allow to understand how this variability potentially affects its expression as well as its allelic diversity and LD patterns displayed with other HLA class-Ia and class-Ib genes at the population-level (e.g. it has been initially described strong LD between *HLA-H* and *-A*, and between *HLA-H* and *-G*, where these three genes have shown distinct worldwide allelic distribution. Conversely, *HLA-E* and *HLA-F* both have apparently shown weak LD, displayed restricted allelic diversity and practically no difference in their global distribution) [498].

*ii*) Furthermore, long (~11-13 kb) class I-like genes *MICA/MICB* have been recently characterized in full length [437][511] using dual redundant reference sequencing approaches [172][173][204][359] by combining short-read sequencing data with long-read sequencing data. So far, unexpected high sequence variability (identifying SNPs and indels in both exons and introns) has been observed, which may tentatively influence on receptor interaction and, thus, immune regulation. Therefore, this may be of interest to be further evaluated especially in the context of transplantation (e.g. assessment of *MICA/MICB* matching status and its impact on transplant outcomes) [554]. Following this latter example, future improved NGS-based HLA approaches (especially by long-read sequencing) may open the window for the assessment of genetic variation at the population level across other genes within the human MHC region such as *HLA-DMB* [438] or *TAP1* [439] (which may be important to explain HLA-disease associations or transplant outcomes) and their specific linkage with physically closed and well-characterized classical HLA genes such as *HLA-DRB1* [440]. Likewise, application of NGS technology may also enable the characterization of other non-classical (HLA-IIb) HLA class II genes and class II pseudogenes of interest

(e.g. *HLA-DQA2* and *-DQB2* polymorphisms and their linkage disequilibrium with classical *HLA-DQB1*) [506].

(d) Moreover, the high-scalability and high-coverage of NGS (especially with the most recent long-read sequencing platforms) has also allowed moving beyond HLA genes by developing reliable high-throughput and population-scale genotyping methods and algorithmic analysis pipelines [296] for other immunogenetically relevant loci such as (ABO) blood groups [441] and the KIR cluster of genes (at first, being less heterozygous than HLA system (i.e. high level of homologous sequence) and presenting many allele combinations heterozygous positions which are more than 800 bp apart) [194][442-445].

**5) Regulatory variation on HLA expression and its effect on HLA function (as previously listed on point “g”):**

As previously mentioned, NGS-based HLA genotyping methods have contributed significantly for the (almost) unambiguous, 4-field high-resolution and phased sequencing of HLA genes, being very informative and useful for many applications [137]. Nevertheless, in order to increase our knowledge of HLA diversity and related functions, HLA expression and its regulation (at both transcriptional (RNA) control and translational (protein) control levels) are also pivotal features that have not been well-defined yet and, thus, HLA expression is still clinically poorly utilized and undervalued as well as it has been rarely incorporated into HLA-disease association studies or evolutionary analyses as phenotypic information (e.g. traditionally, only serological methods have been used to assess expression) [104][446].

As recently reviewed in [951]: Evidence from the IPD-IMGT/HLA database (<http://www.ebi.ac.uk/ipd/imgt/hla/allele.html>) [87][295][362] shows less than 30% of the full-length HLA gene is transcribed into protein (only 7% for HLA DRB1). The remaining 70%–



93% of the gene contains sites for transcription promoters, inhibitors, alternative splice sites, methylation sites, binding sites for post-translational miRNA degradation and many other functions as yet undetermined. The vast quantity of information contained within noncoding regions reinforces the biological importance of this portion of the DNA. In fact, evidence is growing to show how polymorphisms in the 5'UTR will affect subsequent RNA translation, and indeed, there are data to show how polymorphisms even further upstream can promote DNA unwinding to allow RNA polymerase access to the transcription sites. Once transcribed, polymorphisms in the 3'UTR may allow miRNA binding, rapid RNA degradation and reduced protein synthesis. HLA transcription is also subject to alternative splicing, although it is not clear whether this happens in isolated cases or is a common occurrence, or whether external stimuli such as infection are always required.

Since NGS enables the refined description of non-coding region variation within the HLA region, this has started to facilitate the discernment of the effect of regulatory variation on HLA expression and how exactly HLA diversity and polymorphisms shape allelic lineage-specific as well as haplotype-specific expression (where each pattern/associated lineage appear to have a distinctive functional unit, related evolutionary mechanism and history, serological affinities, binding repertoire and expression level) and, in turn, how this may be associated with disease susceptibility (particularly in the case of infectious diseases, since HLA expression determines an individual's ability to respond to virus/bacteria infection) [309][318-322] and transplantation outcomes (e.g. knowledge of the basal and induced expression of each HLA antigen could allow a more informed stratification of transplant risk for each patient/donor combination) [204][211][229-237][255-258][951]. To date, HLA class I expression (whose patterns seem to be relatively consistent (HLA-B similar to HLA-A and higher than HLA-C)) [447] has been better characterized than in the case of HLA class II genes (which appears to follow a more

complex expression model than just the 2-state model (basal and induced) described for HLA class I genes) [309][448]. Yet accurate quantification of relative expression levels of HLA loci is challenging due to the complex and vast polymorphism of the HLA system [448] as well as due to the technical limitations of methods such as: array-based expression tests using probes [421]; quantitative PCR (qPCR) using primers [449]; quantitative flow cytometry (antibody-based testing which allows for assessments of variations in expression levels of HLA molecules in cells over time) [450]; whole-transcriptome sequencing (also known as RNA-Seq: NGS RNA-based HLA genotyping sequencing strategy that is also able to quantify the abundance of mRNA originating from each gene or exon) [451] and even using mass spectrometry (which allows the identification of HLA-bound peptides, where the target peptides may be unique to each HLA protein and, thus, these can be used for quantitation) [448]. With the exception of RNA-Seq (in which implementation of long-read NGS technology (in some cases combined with short-read sequencing strategies) has allowed a significant improvement in the expression estimates) [451][507-509], all the HLA expression estimates provided by these other mentioned methods are still inherently biased since all of them account for a very limited range of polymorphism [104][448]. Furthermore, HLA expression appears to vary not only at the allele/haplotype/loci level but also depending on the tissue adding another layer of complexity. At the same time, diversity found in HLA peptide repertoires, modes of peptide binding, specificity mismatches between the HLA molecules and related components of the antigen processing machinery (e.g. TAP) as well as cognate T cell recognition [399] may also define HLA molecule assembly and cell surface stability (cell surface longevities (half-lives)) and, consequently, HLA expression [450]. Nevertheless, the role of these factors have not been yet fully evaluated and neither the correlation with this complex regulatory genomic polymorphism that determines HLA expression [309]. Moreover, the prediction of regulatory single nucleotide

polymorphisms (rSNPs) in proximal promoters of disease-related HLA genes could be a useful tool for personalized medicine in both patient stratification and customized therapies for transplantation and cancer immunotherapy.

Recent studies have identified regulatory non-coding variation (enhancer, promoter regions, introns and UTRs) of HLA expression that is clearly associated with disease [309][318-322]. For instance, HLA-C expression and its diversity is an important determinant in influencing disease outcome especially in the case of HIV-1 infection [448][452]. Accumulating evidence from different studies has shown how a promoter region SNP of HLA-C is linked with control of HIV infection, being also in LD with a 3'UTR variant that regulates binding of micro-RNA [319]. In neurological diseases, anomalous gene expression (e.g. soluble isoform of *HLA-B\*44:02*) [453] and increased expression of HLA-associated alleles (*HLA-DRB1\*15*) for susceptibility upon activation by associated environmental factors (vitamin D) in the disease etiology [454] have been described as in the case of multiple sclerosis (MS). Interestingly, many miRNA transcripts have been located within a LD block that also contains a disease associated SNP, suggesting that these miRNA transcripts may play a role in the etiology of the numerous diseases associated with the human MHC [416][521].

In the transplantation field, application of NGS is also having a great impact in the assessment of HLA expression in order to elucidate its effect in transplantation outcome [204][211][229-237][255-258]. The main current serological test to determine the presence and strength of DSAs is the (physical) flow cytometric crossmatch (FCXM) method, which is commonly used in combination with other antibody testing solid phase immunoassays [455]. In addition, and as previously mentioned, histocompatibility and transplantation laboratories are increasingly using virtual crossmatching (VXM) to reliably predict recipient and donor compatibility assessment, especially in clinical situations when there is not enough time to perform prospective physical

crossmatch assays such as in the case of deceased donors' organs allocation [243][252]. Information from crossmatching (XM) tests (physical or virtual) is crucial in order to establish a threshold for assessing the risk of antibody mediated rejection (e.g. patients transplanted across positive XM tests tend to have a higher incidence of early graft loss and reduced graft survival) and determining eligibility for transplantation (e.g. when performing transplants with certain HLA mismatches, the XM data supports donor selection and in determining the extent of desensitization needed for a particular DSA barrier) [455]. Furthermore, currently both FCXM (which also depends on factors such as purity of the lymphocyte population, cell-serum ratio, the detection antibody used or the presence of auto-antibody) and VXM mainly rely on donor and recipient HLA genotypes (highly phased and unambiguously resolved up to the 4-field now by NGS) and recipient's antibody profiling [456]. Nevertheless, they still fail to take donor HLA expression (including also the evaluation of plausible differences in living versus deceased donors) into consideration. Even though several studies have pointed out that donor HLA expression may have an important influence on XM tests results, as well as in the development and optimization of epitope-based matching algorithms and data-driven modeling approaches for predicting recipient and donor compatibility, yet it has not been thoroughly investigated [456]. Interestingly, a recent study [456] has shown the feasibility of introducing novel NGS-based RNA-Seq methods (enabling simultaneous determination of donor-recipient high-resolution HLA genotyping and relative donor HLA expression) as part of the recipient and donor compatibility assessment and respective histocompatibility clinical protocol. Thus, these novel RNA-Seq methods in combination with long-read single molecule sequencing strategies using third generation sequencing platforms eventually may allow a much faster crossmatching process, although still with some important technical limitations (e.g. RNA isolation and stabilization) and lack of knowledge in relation to HLA class II expression and its

correlation with FCXM results (since overrepresentation of T cells within the total lymphocyte population prevents accurate evaluation of HLA class II expression). So far, it has been observed a direct correlation between this RNA-Seq estimated donor HLA expression (at the HLA locus level, being only evaluated for class I) to which the DSA (if consistent) is against and FCXM median channel shifts (MFI) values (only at the HLA class level) [456]. Therefore, these novel RNA-Seq methods in combination with epitope analysis at the allele-level may increase the accuracy and reliability of XM tests significantly [258][457]. Moreover, evaluating in-depth the expression patterns of HLA expression variants (including alleles that show: encoded soluble proteins (S); low expression levels (L); and questionable expression status (Q)) may contribute to improve HLA allogenicity prediction algorithms since these variants appear to be also fully functional [458], reflecting once again the complex structure and evolutionary mechanisms of the human MHC system [459][460]. In addition to NGS-based RNA-Seq technology approaches, introduction of precise novel gene editing systems, such as CRISPR/Cas9, may be a useful analysis tool of HLA protein expression and the assessment of its variants [461]. Also this latter tool has the potential for considering the development of immunocompatible pluripotent stem cells via CRISPR-based human leukocyte antigen engineering for transplantation applications [462].

### **III. PREVIOUS STUDIES ON HLA DIVERSITY IN SPANISH POPULATION**

#### **10. SPANISH DEMOGRAPHIC HISTORY, GENETIC LANDSCAPE OF THE IBERIAN PENINSULA AND OVERVIEW OF HLA SPANISH POPULATION STUDIES**

Mainland Spain is located on the Iberian Peninsula (which also contains mainland Portugal) at the southwestern edge of Europe. The Iberian Peninsula is separated from the rest of Central and North Europe by a range of mountains (called the Pyrenees) in the North-East, and from North Africa by the Strait of Gibraltar in the South. Moreover, the Spanish mainland is bordered to the South and East almost entirely by the Mediterranean Sea and to the West by the Atlantic Ocean and Portugal (that borders Spain on its northern and eastern frontiers). The current Spanish territory also includes the Balearic Islands in the Mediterranean Sea, the Canary Islands off the North African Atlantic coast and two port cities, Ceuta and Melilla, located on the northern coast of Africa (thus, showing the historical Spanish influence in this area). The unique geographical location of the Iberian Peninsula (and of mainland Spain in particular), as it is presenting a wide access to the Mediterranean Sea and Atlantic Ocean and being the westernmost region of Europe and also the nearest European region to the African continent, was strategically pivotal for facilitating the Spanish kingdom (initiated by the dynastic union of the Catholic Monarchs in 1469), as well as to the previous cultures that had also settled in the Iberian Peninsula, the control on overseas trade and for its own territory defense throughout the history. In this sense, this privileged geographical location of the Iberian Peninsula also relatively facilitated the Spanish kingdom to organize transatlantic maritime expeditions that led to the final discovery of the Americas in 1492 [555]. In relation to the geographical characteristics of the Iberian Peninsula, it is noteworthy its orography [556], since mainland Spain is very mountainous in relation to some other European countries. First, there is a central big plateau called “Meseta Central”, which occupies most of the peninsula and it is split in two parts by a mountain range called Central

System. Second, this central plateau is, in turn, surrounded by mountains ranges: the Galician Massif to the Northwest, the Cantabrian Mountains to the North, the Iberian System to the East and both the Sierra Morena and the Baetic System (which is divided into two ranges: Sub-baetic and Penibaetic) to the South (see **Figure I-22** and **Figure I-23**).



**Figure I-22.** Topographic map of Spain including the Iberian Peninsula, Balearic Islands (in the Mediterranean Sea) and the Canary Islands (off the North African Atlantic coast). (From: [http://infantes-science5.blogspot.com/2013/05/the-relief-of-spain\\_7.html](http://infantes-science5.blogspot.com/2013/05/the-relief-of-spain_7.html))





**Figure I-23.** Political map of Spain. Autonomous Communities of Spain (respective administrative capital in parentheses): Galicia (Santiago de Compostela); Asturias (Oviedo); Cantabria (Santander); Pais Vasco (Basque Country) (Vitoria); La Rioja, (Logroño); Navarra (Pamplona); Aragón (Zaragoza); Cataluña (Catalonia) (Barcelona); Comunidad Valenciana (Valencia); Murcia (Murcia); Andalucía (Andalusia) (Sevilla); Extremadura (Mérida); Castilla-La-Mancha (Toledo); Castilla y León (Valladolid); Islas Baleares (Balearic Islands) (Palma de Mallorca); Islas Canarias (Canary Islands) (Santa Cruz de Tenerife jointly with Las Palmas de Gran Canaria); and the two autonomous cities Ceuta and Melilla. Figure is obtained and adapted from: <https://www.ezilon.com/maps/europe/spain-maps.html>

Importantly, as a consequence (at least to a certain extent) of this singular orographic organization across the Iberian Peninsula, Spanish territory has showed throughout the history (and also nowadays) an extensive cultural and social diversity (e.g. linguistically diverse) within its entire population shaped by a complex demographic history. In fact, this population diversity observed in Spain has been inherited from the contribution of different European (e.g. German tribes), Atlantic (e.g. Celts), African (e.g. Carthaginians first, and later North African Muslim



Arabs and Berbers), Middle-East (e.g. Phoenicians first, and later Sephardic Jews (originally from the Levant)) and Mediterranean (e.g. Greeks and Romans) populations during different past time periods. Some of these civilizations settled more in certain regions than others (in general, sequentially at different times throughout the history) of the Iberian Peninsula at times when intra-migration flows and, thus, population admixture were partially limited for many centuries favoring the settlement of some isolated population nuclei especially within these different mountainous regions (e.g. Galicia, Asturias, Cantabria and all the Pyrenees area in northern Spain or, in a lesser extent, some southeastern regions of Spain in the Sierra Morena and the Baetic System areas) [260][557-561]. This is currently reflected, for instance, by the very different languages and dialects that are spoken and officially recognized in Spain. Thus, in addition to Castilian Spanish (“*Castellano*”) as the main language spoken, Catalan in the East, Galician in the Northwest, Euskera in the Western Pyrenees area (including Basque Country and Navarra regions) and the Astur-Leonese languages (as the “*Bable*” language spoken in Asturias) are examples of other spoken languages in Spain. The Spanish population diversity resulting from different civilizations that migrated and inhabited the Iberian Peninsula throughout its history is well documented [555]. Chronologically, considering the majority of the more modern era history timeline in the Iberian Peninsula it can be included:

-During the pre-Roman Iron Age: first Iberians (concentrated in central regions of the Iberian Peninsula), Celts (northern regions), Lusitanians (western regions) and Tartessians (southwestern regions); as well as posterior settlements of Phoenicians, Greeks and Carthaginians (especially along the eastern coast).

-Then, in the majority of the territory of the Iberian Peninsula: firstly Romans (2nd century BCE - 5th century CE); followed by Germanic tribes (5<sup>th</sup>-8<sup>th</sup> century in CE); later North African Muslim Arabs and Berbers (8<sup>th</sup>-15<sup>th</sup> centuries in CE); and, through the “*Reconquista*”, Christian

Visigoths (especially starting from the 14<sup>th</sup>-15<sup>th</sup> century in CE), that finally led to the beginning of the Spanish Kingdom with the Catholic Monarchs in 1469 [555]. From there and up to now (thus, during these last five centuries), a more homogenous and stable Spanish (Iberian) population group settled in the Iberian Peninsula that mainly came from this Christian Visigoth lineage although still showing important cultural, social and genetic signatures from other past civilizations and genetic backgrounds as well.

Thus, these different past civilizations of the Iberian Peninsula had also a significant cultural and technological exchange and enrichment over the centuries. In this sense, it is noteworthy a very distinctive period (in comparison to other regions of Europe) of the history that occurred in the Iberian Peninsula denominated the “*Convivencia*” during the Middle Ages between 8<sup>th</sup>-15<sup>th</sup> century in the CE, and being especially important in the 12<sup>th</sup> and 13<sup>th</sup> centuries. In which cultures (with very distinct geographical origins as well as cultural and religious traditions between them) of Christian Visigoths, North African Muslim Arabs-Berbers and Sephardic Jews coexisted for several centuries under the Islamic rule in the Iberian Peninsula [562]. Christian Visigoths had emerged as a western branch of the nomadic German tribes or Goths; and whose kingdom (after defeating the Roman Empire) occupied what is currently southwestern France and the Iberian Peninsula from the 5<sup>th</sup> century to the 8<sup>th</sup> century CE until North African Muslim Arabs and Berbers conquered most part of the Iberian Peninsula forcing the Christian Visigoths to retreat to northern regions of the peninsula [555]. The Muslim forces who occupied the Iberian Peninsula in the 8<sup>th</sup> century CE were mainly Berbers (comprising an ethnic group indigenous to Northwest Africa or the so-called Maghreb; which also were closely related to the indigenous people of the Canary Islands [563-568]) together with Muslim Eastern Arabs (originally from Eastern and Southern Syria as well as from the Arabian Peninsula) under the suzerainty of the Arab Umayyad Caliphate of Damascus [569]. In the case of Sephardic Jews (also known as Western Sephardim), it has been

hypothesized (as it is still uncertain) a very early Jewish presence in the Iberian Peninsula due to their plausible connection with Phoenicians and even Tartessians during the pre-Roman Iron Age. However, it is thought that still most of them had arrived from (at least) the later centuries of the Roman Period, either voluntarily or originally as slaves brought from the Near East/Levant (e.g. Lebanon and Israel areas) by the Romans (around the 1st century CE). Yet, very important Jewish congregations (many of them with a pivotal economic role in the Iberian society and as a demographically non-negligible minority) were present across the entire Iberian Peninsula as well as other Spanish territories (such as the Balearic Islands, where Jewish groups were locally known as “*Chuetas*” [570][571]) throughout the history and until the present day [572]. This “*Convivencia*” period of time, which was uniquely characterized by a relative religious, cultural and social tolerance, it was terminated under the Christian rule (at the end of the 15<sup>th</sup> century in CE, once the “*Reconquista*” was completed) when both Jewish and Muslims were forced by royal decree (The 1492 Edict of Expulsion) to either religious conversion (being denominated respectively “*conversos*” and “*moriscos*” those who converted to Catholicism) or to be expelled (as it occurred to many people from these communities). Even, few centuries later (starting in 1609) most of the “*moriscos*” (especially from eastern regions of the Iberian Peninsula) were also finally expelled under the reign of Philip III of Spain [573]. As for the Sephardic Jewish groups, although many non-converted had to emigrate outside the Iberian Peninsula (mainly to other Northern-Central European countries; as well as to North Africa, especially to Morocco and Libya; also to the Near East or Levant regions (e.g. Lebanon or Syria); and even to the Americas), many others officially converted to Catholicism and could remain in the Spanish Kingdom as Crypto-Jews [572][574]. Moreover, from a historical standpoint and as a consequence of the Jewish Diaspora (i.e. dispersion of Jews out of their ancestral homeland in the current-day Israel region, and their subsequent settlement in other parts of the world over the centuries, that initially occurred

with the large exile of Jewish people due to the historic Roman persecution in the Levant (e.g. Siege of Jerusalem in the year 70 CE)), it is important to remark that different worldwide Jewish population groups can be categorized into three major lineages based on their traditional geographical areas of settlement [575-577]:

-Eastern Jews who have been historically residing in the Near-Eastern and Middle-Eastern regions. Although it is considered that this group category (also denominated as “Mizrahi” Jews) comprises more broadly the descendants of the local Jewish communities that had existed in both the Near-/Middle-East as well as along the North African coast region.

-Sephardic Jews established in the Iberian Peninsula until the 1492 Edict of Expulsion, and after that as “*conversos*” (i.e. Crypto-Jews) who were well integrated in the Christian Iberian society.

-And, in contrast to the two above considered non-Ashkenazi Jewish population groups, the diverse community of Ashkenazi Jews who belonged to Central and Eastern European regions. Even though a large group also migrated from those European regions to the Americas, Western Europe, Australia and South Africa more recently in the 19<sup>th</sup> and 20<sup>th</sup> centuries.

Interestingly, these three main groups (shaped by geographical isolation and/or religious and sociocultural constraints against intermarriage over the centuries) share a common Near-/Middle Eastern ancestry, together with variable degrees of admixture and introgression from the corresponding host Diaspora populations [577].

Based on all these abovementioned historical facts, it is remarkable the profound cultural and anthropological heritage and genetic imprint left by all these three past civilizations in the Iberian Peninsula making up the main ancestry of modern Iberians (comprising the Spanish and Portuguese) [555][556][558][578]. At the same time, from a demographic perspective it is historically evident that the predominant contribution came from the Christian Visigoths

representing the large majority of the ancestry population in the modern Iberian Peninsula and, thus, being much less clear the demographic impact in the case of the relatively fewer “*moriscos*” and “*conversos*” descendants. Nevertheless, recent genetic studies (also supported by some historical studies [572][573]) have shown how constant immigration events (promoting gene flow and integration) from the Near-/Middle-East and North Africa to the Iberian Peninsula over the last two millennia, followed by introgression driven by religious conversion and intramarriage (i.e. endogamy, thus to counteract the expected admixture), especially starting at this period in the Middle Ages, of both “*moriscos*” (as North African Arab-Berber descendants) and “*conversos*” (as Sephardic Jewish descendants) seem likely to have contributed to a relatively substantial genetic proportion of the gene pool ancestry of modern populations in the Iberian Peninsula [574][578]. At the same time, these two groups seem to have succeeded in maintaining, until certain extent, not only certain genetic influence (although that still needs to be further studied) of their own until the present time but also a cultural tradition in the Iberian Peninsula as well as those other neighboring foreign regions where these groups migrated after being expelled back in the 15<sup>th</sup> century [571][572][573][578][579].

Another important ethnic minority in the Iberian Peninsula, and in fact with a long presence (since their earliest settlement was documented in the Northwest region of Spain (Aragon) back in 1425) [580], is the Roma or Romani people (Gypsies). Roma people are known as “*gitanos*” in Spain and as “*ciganos*” in Portugal and they belong to the denominated Iberian Kale (“*Calé*”), an ethnic group native to Iberia and Southern France [580-584][757]. At the same time, this Iberian Romani subgroup is a branch of a much larger Roma (Gypsies) group (including a very diverse range of social and religious traditions between the existing Romani subgroups), that make up a founder population dispersed throughout Europe, whose origins might be traced (based on previous genetic, anthropological and linguistic studies) to Indo-Aryan ethnic groups from ancient

India back in the 10<sup>th</sup> century (both modern-day northwestern regions of India and northeastern regions of Pakistan) as the ancestral population [580-589]. Until recently, knowledge about the origin, migration history and background of Roma people has been limited mainly due to their nomadic nature, broad cultural and social diversity and illiteracy (e.g. no written history or genealogy). Thus, despite these limitations, several recent studies have inferred (and it is now generally accepted) that the migration of the first ancestors of Roma groups from modern-day Northwest India/Northeast Pakistan regions (e.g. Punjab region, Sindh or Baluchistan) to Middle-East and, later, to Europe occurred between the 10<sup>th</sup> (if not earlier) and the 14<sup>th</sup> centuries, in a number of waves [580-589]. These first Indo-Aryan ethnic groups were forced to migrate westward due to the expansion of Muslim invaders (e.g. the Ghaznavid dynasty first, and later the Seljuk Empire) as well as due to the periods of famine in those ancient Indian regions. Furthermore, Roma people's migration routes (where in most cases they were considered aliens and were commonly enslaved or persecuted as outlaws) included first regions such as Persia, Armenia and Anatolia. During the Byzantine Empire, it was here where Romani people acquired the ethnic name they bear still today: *tsigane* (in Greek *athínganos* or *atsínganos*). Later, the conquest of the Byzantine territory by the Ottoman (Turkish) Empire forced the Christians and many of the Indo-Aryan descendants (or proto-Romani people) to emigrate further to the West as well as by the time the "Black Death" reached that area. Thus, in this migration process proto-Romani peoples diverged into three different major groups between the 6<sup>th</sup> and the 11<sup>th</sup> century. Those groups that moved as far as Western Europe are identified as Rom or Roma/Romani people. While the ones who remained in Persia and Anatolia are denominated Dom (or Domi) people, who also dispersed along the North African region. A divergent third group comprises those who remained in Armenia and are known as Lom (or Bosha) people. Additionally, there were some other Indo-Aryan ethnic groups that remained in ancient India and were established in the region of the Rajasthan known

today as the Rajasthani people (who are native) and the Banjara people (also known as Laman or Lambadi, who are nomadic) who might be closely related with the common Indo-Aryan ancestors of the emigrated and divergent Rom, Dom and Lom peoples [580]. Thus, the Rom or Romani people continue migrating to the Peloponnese region, the rest of the Balkans area, and later being rapidly widespread in all East, Central and West Europe (including the Iberian Peninsula) by the 15<sup>th</sup> century. Even from Europe, some Romani groups also migrated to the Americas later around the 18<sup>th</sup>-19<sup>th</sup> centuries period. Moreover, there are also documented historical records of Romani slaves that were brought to the European colonies in the New World to forcedly work in the galleys and on plantations [522][590]. Therefore, especially with the contribution of genetic studies [582][585-589], it has been possible to confirm the origins of modern-day Romani groups distributed around the world linked to these Indo-Aryan ethnic groups from ancient India back in the 10<sup>th</sup> century and thus to discard other previous initial theories that postulated their origins in modern-day Egypt (as apparently, the Romani people had migrated to the Peloponnese region from a region along the Adriatic coast known as "Little Egypt" (in the Tzingania region near by Methoni) and the local Greek people mistakenly believed they had come from Egypt and called them "Egyptians") or in modern-day Romania (in the historical region of Wallachia, where the long period of enslavement of the Roma people is extensively documented) [590][591]. In the Iberian Peninsula, it has been suggested that Roma people arrived via two plausible paths. Firstly, the better documented trans-Pyrenees route from France at the beginning of the 15<sup>th</sup> century by the Roma people per se [580]. Secondly, although not as well-documented, a route across the Mediterranean Sea coming from the Near East, southern European regions and, even, the North African coastline. Thus, there is the possibility that not only Roma groups but also some of the Dom people arrived to the Iberian peninsula through this route as well) during the 15<sup>th</sup> century [583][584][591]. Nevertheless, genetic studies based on uniparental and/or biparental genetic

markers (e.g. single nucleotide polymorphisms (SNPs), mitochondrial markers (mtDNA), Y-haplogroup or microsatellite markers) have not found evidence yet supporting a putative North African origin of the Iberian Roma groups which present a complex demographic history [582][585-588][592-594]. Furthermore, although sharing a common origin, the different European Romani Gypsy groups are highly heterogeneous (with also significant genetic substructure) as a consequence of genetic drift and different levels of admixture with neighboring host populations [582][585-588][592-594][757]. Moreover, in the Iberian society (especially in Southern regions), and in contrast to many other European regions, Romani Gypsies (although with an important nomadic way of life) were more socially accepted, being highly involved in the development of regional folkloric culture and who generally adopted and practiced Christian traditions [580][581]. In addition, contrary to many other European Roma population groups, the Iberian Gypsies are non-Romani-speakers as they experienced a profound linguistic immersion after entering Iberia long time ago. Where Spanish Gypsies (“*gitanos*”) speak Spanish language with the *Calo* dialect just being a reminiscent reference language (i.e large amount of original Romani loan words) [580][581]. At the present time, as a demographically non-negligible minority, Romani Gypsies are estimated to amount to 650.000-800.000 individuals in Spain (representing 1.5% out of the entire population) according to recent studies and official demographic population databases [595-598].

Finally, mostly after this “*Convivencia*” period of time during the Middle Ages between 8th-15th century in the CE, it is also important to underscore that Spain did not receive any new major inward population contributions from neighboring populations, while important emigration events to the Americas and to Europe gradually took place along these last five centuries due to socioeconomic factors [555][556][578]. More recently, substantial emigration from rural to urban areas within Spain occurred during the 19<sup>th</sup> and 20<sup>th</sup> centuries. Interestingly, the trend of migration



events observed in Spain for the past centuries has been recently shifted and, indeed, reversed during the last decades of the 20<sup>th</sup> century and the beginning of the 21<sup>st</sup> century [599]. In fact, relative to its population size, Spain has become very rapidly one of the most important immigration destinations in Europe due to socioeconomic factors. Currently, immigrants account for up to 13.0% of the present Spanish population according to official demographic population databases [595]. Furthermore, out of this 13.0% of immigrant population in Spain [260][595][600]:

-Approximately 5.1% corresponds to the Central, Caribbean and Southern American group (mainly from Andean countries such as Ecuador, Bolivia, Peru and Colombia), and thus it represents the largest immigrant population group (with a common shared language, family ties and cultural proximity) in Spain.

-There is also an important European immigrant component (accounting for 4.7%), primarily coming from Eastern Europe and other Mediterranean regions (especially from Romania).

-And finally, the North African and sub-Saharan African incoming component (especially from Morocco and Senegal) accounts for 2.1% of this immigrant population in Spain.

Therefore, at the present time and in the future decades, this strikingly new demographic scenario in Spanish general population may have important implications to be especially considered in the biomedical field and particularly in the fields of pharmacogenetics, transplantation (e.g. of relevance for optimizing donor search and donor recruitment strategies for UCB and BM registries), and both regenerative and immune modulation therapies and related biobanks [600].

Referring now to the genetic landscape that has been described so far in the Iberian Peninsula, and in particular for the Spanish general population, by main previous population genetics studies. On one hand, and in comparison to those very early studies [557], Spanish population genetic structure and diversity have been delineated mainly and more thoroughly by the analysis of

uniparental genomic markers (such as mitochondrial DNA (mtDNA) and the non-recombining portion of Y chromosome, including their respective SNPs and STRs and also considering respective haplogroups) [565][578][601], whose inheritance is not altered by recombination events or driven by natural selection but only by possible mutations, being this particularly useful in anthropological studies for reconstructing population expansions and migrations based on descent relationships [542][543]. On the other hand, other major Spanish population genetics studies have been also focused on autosomal DNA markers (including high density arrays of SNP, STR and CNV variants of certain polymorphic regions of the human genome, many of them also used for genome-wide studies, thus allowing to disentangle the extensive and fine-scale population structure in the Spanish (and Iberian) population) [260][558][602-606]; which, in turn, can be useful for inferring biogeographical ancestry, population admixture and also population stratification and substructure [542][543]. In this sense, main findings from these two types of approaches have allowed to establish certain widely accepted conventions in regards of the very complex demographic history and genetic makeup in the Iberian Peninsula in addition to the also evident existence of a current remarkable regional genetic variation across the Spanish territory. In summary, the present-day Iberian gene pool has been shaped by differentiated contributions from a diverse group of population migrations throughout the history:

-As an initial point in history established here, starting with the local Paleolithic Iberian population, which already existed by 50,000 BCE, and that would have received an important African gene flow (between 8,000 and 4,000 BCE) due to the vast desertification that took place across the Sahara region.

-Secondly, it has been also well-defined genetic signatures of population migrations to the Iberian Peninsula along the history, even back from the Neolithic period (approximately 5500–3000 BCE), from both Central European regions (trans-European migrations through the

Pyrenees Mountains and from the Atlantic seaboard of the West of Europe (Bay of Biscay regions) by respective, and subsequent, Celtic, Roman, Germanic population groups among others) and a significant maritime colonization around the Mediterranean Basin including from both the Levant (presenting Arab and Jewish backgrounds) and the Northern African coastlines (Berber-Arab and even sub-Saharan backgrounds) as previously commented.

- Furthermore, these genetic studies have also revealed a relatively significant genetic structure (i.e. population stratification) among broad geographic regions of the Iberian Peninsula in the current population. Where some of the local patterns that have been detected could be tentatively explained by the complex orography of the Iberian Peninsula resulting on the historical settlement of certain isolated population nuclei for long periods of time in geographical niches such as: the Galician Massif to the Northwest; the Cantabrian Ridge to the North; the higher Ebro Valley to the Northeast; the Western Pyrenees region; and the Baetic System (a mountain range across Southern Spain). Nevertheless, despite the unique geographical landscape of the Iberian Peninsula and its tentative influence on this population stratification, it is widely accepted that these striking regional variations in the genetic makeup of the modern-day Iberian general population are better fitted with historical, political and cultural constraints that influenced migratory patterns and the relationships between populations throughout the history. Thus, singular genetic signatures of relatedness have been observed between those regions sharing a common linguistic background, illustrating a significant influence of the cultural diversity in both Spain and Portugal. As a remarkable and representative example, genetic data supports that the linguistic and geopolitical boundaries present around the end of the time of Muslim rule (8<sup>th</sup>-15<sup>th</sup> centuries in CE) in the Iberian Peninsula might have a significant and long-term impact on the genetic population structure currently observed in Spain and Portugal.

-At the same time, even in the case of a traditionally considered genetic isolate and supposed outlier in Western Europe as the Basque population (in Northern Spain) based on early genetic studies (using classical markers) [557]. Actually, it has been recently disclosed in various genetic studies (e.g. based on either a dense map of genome-wide SNPs or HLA genotyping data [260][558][559][607][608]) that people of Basque origin are not genetically differentiated from other non-Basque Iberian populations. Where a primary North African-Iberian Mediterranean genetic substrate is part of the overall Basque gene pool as it has been described for other Southern European populations [609]. In fact, it has been also widely reported how these current Basque, Iberian and North African populations cluster more closer together than to the rest of Central and Northern European populations [260][558][559][578][607][608].

-On the other hand, there is indeed a set of singular Spanish population groups which also need to be remarked. These groups are related with territories geographically isolated, cutting off from the main landmass of the Iberian Peninsula. These particular Spanish population groups are found respectively in the Balearic Islands (Majorca, Minorca, and Ibiza islands situated in the Mediterranean Sea) and in the Canary Islands (Santa Cruz de Tenerife, Fuerteventura, Gran Canaria, Lanzarote, La Palma, La Gomera, El Hierro and La Graciosa islands located off the North African Atlantic coast (specifically located off the southern coast of present-day Morocco)). Which, in turn, have experienced certain unique demographic events (and even, hypothetically, some indistinguishable natural selection events) and specific gene flows (also observed at the HLA system level) over the centuries that in some cases clearly differ (or are observed in a greater manner) from those identified within the Iberian Peninsula. In the case of the Canary Islands (which, due to their oceanic volcanic origin, they have probably never been connected to any continent), modern-day populations of this archipelago show significant North African Berber genetic signatures in addition to Iberian and sub-Saharan genetic imprints. Thus,

it is widely accepted (and supported also by archaeological, anthropological and linguistic studies) that the first Canarian indigenous or native populations (e.g. known as “*Guanches*” for Tenerife, “*Benehaoritas*” for La Palma or “*Bimbapes*” for El Hierro) were of North African Berber origin. Furthermore, initial comparisons of genetic markers between these insular populations have shown some striking dissimilarities where some populations might have experienced high genetic diversity, while others were probably affected by genetic drift and/or bottlenecks during pre- and especially post-European conquest (in the 15<sup>th</sup> century by the Spanish Kingdom) times [563-568][610-613]. As for the Balearic Islands, it has been detected a significant Near-/Middle-East Jewish genetic substratum in their modern-day local insular populations, including at the HLA level [130][464][558][571][574][614-621]. In fact, it is historically well-documented that relatively large Jewish communities (where descendants of this Balearic Jewish population were known as “*Chuetas*”), as a consequence of both the initial historical Jewish Diaspora from the Levant and Near East regions and also due to the posterior 1492 Edict of Expulsion, also settled in this Mediterranean archipelago (as similarly observed in the case of other previous cultures such as Phoenicians, Greeks, Carthaginians, Romans, Vandals, Byzantines and Moors). Since this was a pivotal location in the very concurred maritime trading routes that crisscrossed the Mediterranean Sea throughout the history [558][571][574-577][620][622][623].

Overall, the general profile of genetic variation across the Iberian Peninsula seems to be relatively continuous along the main geographic axes. Nonetheless, future data based on genetic studies with larger sample sizes and at a wider geographical scale where covering a higher degree of polymorphism (e.g. hundreds of thousands of SNPs typed individually in large samples and application of both short- and long-read NGS technologies) will allow to depict a much more comprehensive map of the genetic diversity in the Iberian Peninsula.

In relation now to the diversity of the highly complex and very polymorphic HLA system so far described in the Iberian Peninsula, and in particular for the Spanish general population. Previous HLA genotyping studies (including the present thesis work; see **Figure I-24**) are in consonance with the findings just abovementioned from population genetics studies based on other uniparental/biparental and autosomal genetic markers. Thus, as previously mentioned, HLA allele and haplotype frequency distributions data and specific LD patterns shown by a given population (despite being partially shaped by host-pathogen interactions and evolutionarily driven by natural selection in ancestral populations even at a microgeographic scale [542][543][607]) are a powerful and very informative approach for inferring genetic background and ethnical composition. In turn, this key information enables a fair assessment of relatedness between modern human populations and also detection of demographic historical events as well as tracking migration waves of modern or more ancient populations [131][132][137]. As a matter of fact, due to the vast polymorphism and LD displayed by the HLA system, characterization of the diversity of a few major HLA loci can provide an equivalent level of in-depth genetic population-level data in comparison to genome-wide analyses based on hundreds of thousands of bi-allelic SNPs distributed across the whole genome [260]. However, in the recent past and before the development and application of NGS technology for molecular HLA genotyping, only restricted very low- or low-resolution HLA data from very specific Spanish (and Iberian) population groups of generally delimited geographical areas [260][546][558-561][563][564][571][600][602][603][608][624-630][757] and from certain particular disease association studies [631-636] had been reported. Since the large majority of previously reported HLA studies in Spanish population were conducted:

- Using lower resolution legacy typing methods (e.g. either 1-field or 2-field or reporting by P, G groups), where only capturing HLA polymorphism comprised by exon 2 and 3 in HLA class

I genes and exon 2 in HLA class II genes, and, thus, with an important level of both allelic and phasing ambiguities.

- HLA typing was performed for covering certain loci but not for all major classical HLA loci. As a result, most of these previous studies did not define complete extended HLA class I and class II haplotypes, and consequently Spanish (and Iberian) HLA haplotype landscape had not been accurately and thoroughly described yet.

-Additionally, these studies (with very few exceptions [221][260][546][628]) were restricted to small sample size cohorts (or where even only considering a cohort of patients but not considering a group of healthy individuals) of some Spanish (and Iberian) population groups and regions. Therefore, this group of previous HLA studies have not been still adequately representative of the entire Spanish general population, not being able to reveal a very comprehensive map of the HLA diversity in the Iberian Peninsula.

Yet, despite the paucity of early studies describing both all major HLA class I and class II loci and respective high-resolution 3-/4-field allelic/haplotypic data at the genomic level in previous Spanish population cohorts; these past studies, indeed, have reported certain relevant findings that have significantly contributed to a first (although incomplete) depiction of the HLA genetic diversity found in the Iberian Peninsula. Thus, this first group of studies have described and identified singular and most common HLA alleles and partially extended haplotypes in Spanish (Iberian) populations (mainly summarized in [558][620]). Furthermore, these initial studies have also contributed to the discovery of novel/rare alleles (thus, being able to infer tentative genetic mechanisms involved in HLA polymorphism generation, where most HLA alleles have arisen by point mutation and gene conversion or recombination events of short fragments of DNA leading to single or short amino acid motives substitutions) (e.g. [637-639] among many other new allele

reports); as well as identification of clinically relevant null alleles in both classical [640] and non-classical [641] HLA loci, being certainly characteristic of Spanish (Iberian) populations.

On the other hand, and just very recently with the introduction of NGS technologies, a few novel NGS-based HLA typing (including the present thesis work) and high-resolution (via SBT or SSO) typing data studies in Spanish population, [221][269][297][545][546][564][624] have overcome many of these past technical HLA typing limitations. As a result, these last studies have achieved to gain a better insight of the existing genetic complexity of the Spanish general population and thus of the HLA diversity (both at the allele and extended haplotype levels for the majority of the classical HLA loci) across the Iberian Peninsula. Nevertheless, and of note, all these recent studies (with the exception of the present thesis work) still show important limitations especially in regards to incomplete coverage of HLA gene sequence for many of the major HLA class I and, especially, class II loci. Moreover, HLA loci such as *HLA-DQA1*, *-DPA1* and *-DRB3/4/5* have not been well-characterized either in this more recent group of HLA studies. Main results from these more recently published works are also later reviewed, commented and compared with the current thesis work in great detail at the **DISCUSSION** section.

It is also noteworthy the significant efforts made during these last decades by Spanish Public Health institutions (and specifically coordinated by the Spanish National Transplant Organization (ONT) and the José Carreras Foundation agencies) that have led to the development of a very robust and large national transplantation network for both clinical settings in SOT [642] and HSCT (here with the Spanish registry of BM and UCB donors named as REDMO (“Red Española de Donantes de Medula Osea”) presenting a respective very large HLA typing donor pool database (n= 423,455 registered donors; as of July 2020)) [643]. In this sense, so far there have been two main reported HLA studies [260][546] that have attempted to describe the HLA diversity found in this Spanish donor registry population:



-Firstly, Romòn et al. study [260] described (comprising the largest reported sample size up to date) the HLA regional diversity found within the Iberian Peninsula although only based on 1-field very low-resolution (generic level of resolution) HLA genotyping data of a reduced set of loci (*HLA-A*, *-B*, *-C*, *-DQB1*, *-DRB1*) obtained from a very large cohort (N=63,484) of this Spanish registry. Here, Romòn et al. [260] mapped HLA variation in the majority of Iberian Peninsula (and, indeed, existing population substructure) by combining classical population genetic analyses with geographic information approaches (i.e. evaluating the correlation between variation of HLA allele/haplotype frequency distributions and latitude/longitude as geographical parameters) as previously conducted in the context of European populations [136].

-A second very recent study (Alvarez-Palomo et al. study) [546] has demonstrated the feasibility of tentatively using in the future (as part of the IPS-PANIA project) banked cord blood units from this REDMO Spanish donor registry population in order to create a well-represented national iPSC haplobank (made up of HLA-matching iPSC lines from homozygous donors as starting material for iPSC-derived cell therapies) that will cover a significant percentage of the Spanish population for future advanced therapy replacement strategies [222][491-495][950]. In this case, after inferring via EM algorithm the most common *HLA-A~C~B~DRB1~DQB1* haplotypes at 1-/2-field allele resolution level from a N=30,000 randomized cohort of this REDMO Spanish donor registry population (where 0.62% of the cord blood units were homozygous for *HLA-A*, *HLA-B* and *HLA-DRB1*). Authors of this work have estimated that ten cord blood units from homozygous donors bearing the most common haplotypes detected in Spanish registry population could provide a *HLA-A*, *HLA-B* and *HLA-DRB1* matching for approximately 30% of the population.

Nonetheless, it is important to underscore that there is an important level of variability in the REDMO Spanish donor registry population HLA typing database in regards to the number of HLA

loci and allele resolution level consistently and historically tested and reported per donor (due to application of a variety of HLA typing methodologies during the last three decades, creating not uniform high-resolution data at various definition levels (at the 1-field, 2-field, 3-field or 4-field)). This lack of uniformity in the REDMO HLA data has been critically limiting the current characterization of Spanish donor registry population. Therefore, upcoming standardized HLA typing results of most common HLA allele and extended haplotype frequency distributions of all 11 major HLA classical genes obtained from future larger NGS studies and at a wider geographic scale of the Spanish territory will provide an invaluable refined information for the potential improvement of the current registry in terms of population coverage as well as strategies of the organization/prioritization of donors recruitment. Moreover, development of local donor registries, in addition to improvements of the most national main registry, may contribute to better cover the diverse HLA genetic background found in Spain, which actually presents remarkable geographical (regional) HLA signatures.

Finally, there have been also several significant past HLA-disease association studies conducted in Spanish population, and in particular relative to pathologies presenting a major autoimmune component such as (some main examples): Behcet's disease [633], pemphigus vulgaris [636], psoriatic arthritis and ankylosing spondylitis (among other spondyloarthropathies) [631], rheumatoid arthritis [632], celiac disease [634] or neurological diseases as multiple sclerosis (MS) [635][737]. In this context, single locus HLA genotyping has been traditionally used as an ancillary testing to assist with diagnosis of these related HLA-associated diseases. Nonetheless, there has been now substantial evidence supporting that other not as well described HLA and non-HLA genes within the MHC region may be associated with the given disease. Thus, additional HLA alleles and bearing haplotypes implicated in susceptibility/protection may play a role in determining specific features of the respective disease phenotypes. However, it has not yet been

well established whether these associations are driven by direct associations or by linkage disequilibrium (LD) mechanisms. Thus, recent and future population large-scale case-control studies (ethnically- and regionally-matched) with adequate patient groups and replication cohorts, as well as confirmation studies in family pedigrees through the use of novel NGS genotyping methods, will help for the fine-mapping of the etiological role of both HLA and non-HLA genes in diseases.

| <b>Spanish Population/Region</b> | <b>Sample Size<br/>(N)</b> | <b>HLA- loci tested/<br/>Allele Resolution Level</b> | <b>Reference</b> |
|----------------------------------|----------------------------|--|------------------|
| <u>Galicia</u>                   | 264                        | A, B, C at 1-field                                   | [625]            |
| <u>Galicia</u>                   | 125                        | A, B, DRB1 at 1-field                                | [626]            |
| <u>Basque Country</u>            | 82                         | A, B, C, DQA1, DQB1,<br>DRB1, DRB3/4/5 at 1-/2-field | [608]            |
| <u>Basque Country</u>            | 100                        | A, B, C, DQA1, DPA1, DPB1,<br>DRB1 at 1-/2-field     | [627]            |
| <u>Cantabria (Pas Valley)</u>    | 88                         | A, B, DQA1, DQB1, DRB1<br>at 1-/2-field              | [559]            |
| <u>Cantabria (Cabuérniga)</u>    | 95                         |  |                  |
| <u>Cantabria (North Spain)</u>   | 83                         |  |                  |

|                                   |       |  |       |
|-----------------------------------|-------|--|-------|
| <u>Castilla y León</u>            | 1,818 | A, B and DRB1 at 1-field                         | [628] |
| <u>Navarre</u>                    | 116   | DPA1, DPB1 at 2-field                            | [630] |
| <u>Navarre</u>                    | 112   | DQA1 at 2-field                                  | [629] |
| <u>Barcelona UCB Bank</u>         | 5,458 | A, B, C, DQB1, DRB1 at 2-field                   | [221] |
| <u>Catalonia (North Girona)</u>   | 88    | A, B, C, DQA1, DPA1, DPB1,<br>DRB1 at 1-/2-field | [627] |
| <u>Majorca</u>                    | 407   | A, B, C, DQB1, DRB1 at 1-field                   | [571] |
| <u>Majorca-Jewish (“Chuetas”)</u> | 103   | A, B, C, DQB1, DRB1 at 1-field                   |       |
| <u>Minorca</u>                    | 94    | A, B, C, DQB1, DRB1 at 1-field                   |       |
| <u>Ibiza</u>                      | 88    | A, B, C, DQB1, DRB1 at 1-field                   |       |
| <u>Murcia</u>                     | 173   | A, B, C, DQB1,<br>DRB1, DRB3/4/5 at 1-/2-field   | [560] |
| <u>Granada-Almería</u>            | 125   | A, B, DQA1, DQB1,<br>DRB1 at 1-/2-field          | [561] |

|  |          |   |                                |
|--|----------|---|--------------------------------|
| <u>Madrid</u>  | 176      | A, B, C, DQA1, DQB1, DRB1, DRB3/4/5 at 1-/2-field             | [608]                          |
| <u>Madrid</u>  | 253      | A, B, C, DQB1, DRB1, DRB3/4/5 at 3-/4-field                   | [624]                          |
| <u>Madrid</u><br><u>(Amerindian immigrants)</u>                  | 173      | A, B, DQB1, DRB1 at 1-/2-field                                | [600]                          |
| <u>Tenerife</u>  | 83       | DRB1 and DQB1 at 2-field                                      | [563]                          |
| <u>Gran Canaria</u>  | 215      | A, B, C, DQB1, DRB1 at 2-field                                | [564]                          |
| <u>REDMO</u>   | 63,484   | A, B, C, DQB1, DRB1 at 1-field                                | [260]                          |
| <u>REDMO</u>   | 30,000   | A, B, C, DQB1, DRB1 at 2-field                                | [546]                          |
| <u>IBERIA</u>  | (Review) | Variety of typing methods and loci                            | [558]                          |
| <u>17<sup>th</sup>-IHIW</u><br><u>Spanish healthy population</u> | 282      | A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1, DRB3/4/5 at 3-/4-field | [269]<br>(Present Thesis Work) |
| <u>Spanish Romani Gypsy</u><br><u>of Andalusia</u>               | 80       | DRB1, DQB1 and DPB1 at 2-field                                | [757]                          |

**Figure I-24.** Summary of main HLA Spanish population studies reported in literature.

## **IV. HLA ASSOCIATION STUDIES IN MULTIPLE SCLEROSIS (MS)**

### **11. OVERVIEW OF MS GENETICS AND HLA-MS ASSOCIATION STUDIES**

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) mediated by the adaptive and innate arms of an unregulated immune response that leads to demyelination and axonal degeneration and accompanying neurological impairment and disability [644][645]. MS affects around 2.3 million people worldwide with very different incidence and prevalence among human populations [646]. In fact, MS risk varies within the same geographical region across racial and ethnic groups. Traditionally, MS had been found to be more frequent in high latitude regions and Central-Northern European populations. Notwithstanding, it has been observed an increase of the worldwide MS prevalence over the last decades (global median prevalence reported by World Health Organization (WHO) in 2013 was 33 per 100,000 [647]), occurring primarily in women and even in populations previously considered to be at low-risk, such as those with Latin American, Asian, and African ethnic backgrounds. It is not clear yet if this increase may be explained due to the overall improvement of diagnosis and reporting systems, lifestyle and diet changes, or specific evolving environmental factors [648]. It is widely accepted that MS presents a multifactorial etiology, since it has been described that a variety of both genetic and environmental factors can influence the disease risk, severity and clinical course. Vitamin D levels, smoking and Epstein Barr Virus (EBV) infection history (both anti-Epstein Barr Nuclear Antigen-EBNA-IgG seropositivity and infectious mononucleosis) are some of the well-documented environmental elements [649]. At the same time, evidence for a significant but complex genetic component in MS pathogenesis is found in the clustering of affected individuals in families (e.g. via linkage analyses employing multiple-affected member (multiplex) MS families), high disease concordance rate in monozygotic twins (20-30%), and observed differences

in disease prevalence among different ancestral groups irrespective of geographic location [650][651].

Similarly to other autoimmune diseases, MS is characterized by moderate heritability, polygenicity, and multifaceted gene-environment interactions. This polygenic model of MS heritability has offered the rationale and drive for assembling very large DNA datasets to pursue genome-wide association studies (GWAS), which (based on novel chemistries, system miniaturization, and automation strategies that have enabled a much more comprehensive and efficient DNA microarrays panel configuration) have been highly successful in identifying associated variants for susceptibility [652]. To date, targeted genomic screens in populations, mostly of European descent, have revealed 233 independent regions that are across the human genome significantly associated with susceptibility, including 32 independent allelic and locus effects within the MHC region [651][653]. Here, association to HLA gene variants (which was first described several decades ago [654-656]) carries the strongest genetic burden for MS risk, although non-HLA genes and gene-gene interactions (i.e. epistasis) seem to also play a significant role in MS pathogenesis (in the plausible scenario of an antigen(s)-specific autoimmune disease) [657-659]. The main signal in those described MS genome-wide maps has been detected at the class II region of the HLA system and it actually explains up to 10.5% of the genetic variance underlying risk. Predominantly, class II *HLA-DRB1\*15:01* allele variant has been observed as the strongest genetic determinant on MS risk, showing an average odds ratio of 3.08. In fact, this major *HLA-DRB1\*15:01* susceptibility factor has been also associated (although still without a total consensus across studies) with phenotypic markers of disease severity, in contrast to other MS associated HLA-DRB1 alleles [650][651][660][661]. Yet, complex allelic hierarchical lineages, HLA-DRB1 allelic heterogeneity across studied populations, cis/trans haplotypic effects, and

independent both risk and protective MS signals in the HLA class I region have been additionally described [290][645][650][659][662-664]. In summary:

i) *HLA-DRB1\*15:01* as a major MS risk signal and respective haplotypic associations:

From the very initial detection of the serological specificity DR2 as a highly associated factor to MS [654-656], continuous examination of HLA polymorphism coupled with incremental technological advance during the last decades and, thus, improved resolution of molecular genotyping methods allowed to finally refine this association with MS to the *HLA-DRB1\*15:01* allele [665]. Up to date, *HLA-DRB1\*15:01* represents the strongest genetic determinant on MS risk, where its allele frequency distribution has been consistently found greater in MS patients than controls in many studies (the majority being focused on populations groups of European ancestry) [666]. At the same time, *HLA-DRB1\*15:01* allele is most often embedded in a relatively common extended HLA class II haplotype in populations of European descent (e.g. HF=12.2% in European American healthy population [268]; or HF=8.6% in the present Spanish healthy population [269]), in which it is tightly linked with *HLA-DRB5\*01:01*, *HLA-DQA1\*01:02* and *HLA-DQB1\*06:02* alleles that have been also reported for their association with MS susceptibility [667][668]. Consequently, so far it has been extremely difficult to distinguish the definitive and primary predisposing locus or allele within this extended HLA class II haplotype. In fact, previous attempts to localize the MS susceptibility gene within the DR-DQ regions of the MHC have not provided consensus. For instance, on one hand, HLA-MS studies in African American population groups have clearly attributed risk to *HLA-DRB1* locus but not to *HLA-DQB1\*06:02* allele (this latter being ruled out as the causative allele of the association signal, at least in the case of this population) [669]. Conversely, studies in other different populations (e.g. Sardinians [670], Afro Brazilians [671], and Canadian families [672]) have suggested a possible secondary role of HLA-DQ variation in MS susceptibility as



it has been also observed in other autoimmune diseases (e.g. celiac disease) (reviewed in [650]). Moreover, it has been described that this characteristic extended HLA class II haplotype may also influence disease severity (e.g. based on clinical phenotypic data such as age of onset, presence of oligoclonal bands and IgG levels in the cerebrospinal fluid of MS patients and radiological imaging data of the brain) [661]. These observations suggest that this *HLA-DRB1\*15:01* bearing haplotype (where *HLA-DRB1\*15:01* alone explains up to 10.5% of the genetic variance underlying risk [652]) plays a specific and singular role in MS disease pathogenesis, and indeed being distinct and unique from other associated HLA alleles and haplotypes (where additional *HLA-DRB1* associations appear to account for less than 2% of the remaining variance [652], where these may contribute to disease risk through alternative mechanisms) [650][659]. In addition to the increased risk for disease observed for *HLA-DRB1\*15:01* homozygous genotypes (following an additive model [673]), there appears to be an epistatic effect for the *HLA-DRB1\*15:01/DRB1\*08:01* heterozygous genotype, where increased risk in comparison to other different heterozygous *DRB1\*15:01* genotypes has been described [674]. In this sense, *HLA-DRB1\*08:01* allele does not seem to confer risk on its own although a weak independent signal was detected in an Ashkenazi Jewish population cohort study [675].

**ii) Additional HLA-DRB1 risk alleles and other described HLA class II associations:**

Besides the *HLA-DRB1\*15:01* allele and its most common bearing class II haplotype (*HLA-DRB5\*01:01~HLA-DRB1\*15:01~HLA-DQA1\*01:02~HLA-DQB1\*06:02*) as a major MS risk factor frequently found in populations of Northern-Central European descent, it has been also identified a striking allelic heterogeneity in *HLA-DRB1* locus importantly associated with MS susceptibility as well as other secondary associated HLA class II signals (conferring susceptibility or protection, respectively) across worldwide populations [650][659][676][677].

In this case, it is important to underscore that the lack of clear association for these other HLA class II alleles and loci can be most likely, and generally, due to the limited statistical power presented by them since their respective allele frequency distributions are globally low and uneven (i.e. with a relatively low representability and being rarely found at population frequencies higher than 3% worldwide) [650]. In detail, differentiating between global regions and broad ethnic groups or even in the case of peculiar populations so far described:

-Some studies have reported a relevant correlation of *HLA-DRB1\*04:05* allele with MS, in addition to the also *HLA-DRB1\*15:01* detected signal, in Japanese and Asian populations [294][678-681]. In fact, this suggested independent role of *HLA-DRB1\*04:05* with MS etiology (conferring risk) has been further replicated in results from other various studied populations (e.g. Sardinians [670], Sicilians [682] and African Americans [683]). Furthermore, within the HLA class II region, *HLA-DPBI\*03:01* allele has been also significantly, and independently, associated with MS susceptibility in a Japanese population study [678] and also in a study based on Australian *HLA-DQBI\*06:02*-negative MS patients and controls [684].

-Additional HLA-DRB1 risk alleles that have been consistently reported include *HLA-DRB1\*13:03* and *-DRB1\*03:01*, especially detected in certain regions of Western Europe and the Mediterranean basin, with some specific considerations that are noteworthy. The association of *HLA-DRB1\*13:03* allele with MS was first identified through analysis of Ashkenazi and non-Ashkenazi Jewish cohorts from Israel [675]. Where, despite *HLA-DRB1\*13:03* is rarely observed at population frequencies greater than 3% worldwide [650], its association with MS appears to be more robust with a stronger effect size as it is seen in the case of previously reported HLA imputation studies based on SNP data from very large cohorts (e.g. in populations of European descent) [652][658] and in those given studied

population groups (e.g. Israeli) that show considerably higher allele frequency distributions of this variant [675][685]. Whereas in the very singular Sardinia region of Italy, where MS prevalence is strikingly high, all characteristic *HLA-DRB1\*04:05*, *-DRB1\*03:01*, *-DRB1\*13:03* and *-DPB1\*03:01* allelic signatures (in addition to major *HLA-DRB1\*15:01* risk factor) have shown positive associations to MS susceptibility [650][662][670][686][687]. It has been suggested that these results may be indicating that the HLA genetic structure and patchwork (or mosaic) observed in the present Sardinian islander population is the result of a fixation of haplotypes, which are very rare elsewhere, and are most likely to have originated from a relatively large group of diverse founder populations [688]. Similarly to *HLA-DRB1\*08:01* (with a weak independent signal detected in Ashkenazi Jewish population) [675] and in contrast to *HLA-DRB1\*13:03*, *HLA-DRB1\*03:01* is also relatively common in populations of European Mediterranean ancestry and it also seems to contribute less significantly to risk through recessive modes or interaction effects [689-693]. Thus, *HLA-DRB1\*03:01* and *HLA-DRB1\*08:01* do not appear to impact primarily disease severity, suggesting a dissimilar role in disease in comparison to those other HLA-DRB1 alleles (e.g. *HLA-DRB1\*15:01/15:03*, *-DRB1\*13:03* and *-DRB1\*04:05*) that have consistent findings of risk and severity in MS [650]. Moreover, recent studies in which HLA alleles were assigned by imputation methods proposed a positive association of *HLA-DPB1\*03:01* allele with MS risk [658][945].

-Interestingly, HLA-MS studies in African American populations [669][683] and populations of African descent [694][695] have described that this ethnic background not only shows the prototypic association of *HLA-DRB1\*15:01* (although much more weakly since it presents a low allele frequency in these populations) on MS susceptibility. But also, within the *HLA-DRB1\*15* allele group (*HLA-DRB1\*15:01* through *HLA-DRB1\*15:06*, which is distinguished

from other common HLA-DRB1 alleles by the small and hydrophobic Alanine (Ala) residue at amino acid position 71 defining a distinctive structure on the P4 pocket of the HLA-DRβ1 molecules and thus a unique peptide binding profile), the allele *HLA-DRB1\*15:03* (which is the most common *HLA-DRB1\*15* allele in these African-related populations) confers susceptibility to MS as well, being highly specific to people of African ancestry. In contrast to this clear association to MS susceptibility of both *HLA-DRB1\*15:01* and *-DRB1\*15:03* alleles (which differ only at amino acid position 30 (Tyr → His, respectively) and have an identical peptide binding motif), the structurally similar *HLA-DRB1\*15:02* allele does not seem to be associated with MS, either in populations of European descent where the frequency of this allele is very low (AF~1%) [696], or in Asian populations (primarily in Southeast Asia and Oceania) where the frequency is much higher (AF~8%) but incidence of MS appears to be much lower [697]. In fact, this *HLA-DRB1\*15:02* allele differs from the major predisposing *HLA-DRB1\*15:01* allele by a single amino acid substitution, Val → Gly at position 86, likely enlarging the P1 pocket of the peptide binding groove. However, presence of Val at HLA-DRβ86 position (initially suggested to be pivotal for susceptibility to MS) has not been observed across all main HLA-DRB1 risk alleles for MS susceptibility that have been described on different populations (e.g. Gly at HLA-DRβ86 position is also found, as an example, in *HLA-DRB1\*13:03* allele) (see **Figures I-25** and **I-26**). Moreover, Finn et al. study [698] showed that the difference in risk association with MS of *HLA-DRB1\*15:01* versus *-DRB1\*15:02* is not due to a lack of antigen presentation by DRβ1\*15:02, at least in the context of putative myelin peptides, and suggested that other mechanisms involving *HLA-DRB1\*15:01* may account for increased susceptibility to MS. In this sense, distinctive haplotypic associations observed still need to be further evaluated (in a trans-ethnic analysis at high-resolution) to allow the fine-mapping of the different elements in tight LD [297][668]:

*HLA-DRB5\*01:01~HLA-DRB1\*15:01~HLA-DQA1\*01:02~HLA-DQB1\*06:02* (European)

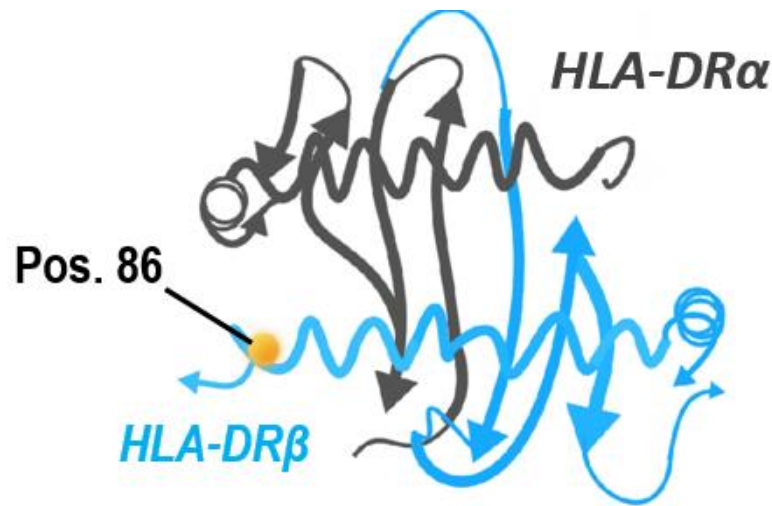
*HLA-DRB5\*01:02~HLA-DRB1\*15:02~HLA-DQA1\*01:03~HLA-DQB1\*06:01* (Asian)

*HLA-DRB5\*01:01~HLA-DRB1\*15:03~HLA-DQA1\*01:02~HLA-DQB1\*06:02* (African)

|                  |            |            |            |            |            |             |             |            |            |            |  |
|------------------|------------|------------|------------|------------|------------|-------------|-------------|------------|------------|------------|--|
| AA Pos.          | -21        | -11        | -1         | 10         | 20         | 30          | 40          | 50         | 60         | 70         |  |
| DRB1*15:01:01:01 | MVCLKLPGG  | SCMTALTIVL | MVLSSPLALS | GDTRPRFLWQ | PKRECHFFNG | TERVRFLLDRY | FYNQEEVRF   | DSDVGEFRAV | TELGREDAEY | WNSQKDILEQ |  |
| DRB1*03:01:01:01 | ---R---    | ---AV---   | -----A     | -----EY    | STS-----   | ---Y---     | -H---N---   | -----      | -----      | -----L---  |  |
| DRB1*04:05:01:01 | ---F---    | ---A---    | -----A     | -----E-    | V-H-----   | -----       | ---H---Y--- | ---Y---    | ---S---    | -----L---  |  |
| DRB1*08:01:01    | ---R---    | ---AV---   | -----A     | -----EY    | STG--Y---  | -----       | ---Y---     | ---Y---    | ---S---    | -----F--D  |  |
| DRB1*13:03:01    | ---R---    | ---AV---   | -----A     | -----EY    | STS-----   | -----       | ---Y---     | ---Y---    | ---S---    | -----D     |  |
| DRB1*15:02:01:01 | -----      | -----      | -----      | -----      | -----      | -----       | -----       | -----      | -----      | -----      |  |
| DRB1*15:03:01:01 | -----      | -----      | -----      | -----      | -----      | -----H      | -----       | -----      | -----      | -----      |  |
| AA Pos.          | 80         | 90         | 100        | 110        | 120        | 130         | 140         | 150        | 160        | 170        |  |
| DRB1*15:01:01:01 | ARAAVDTYCR | HNYGWESFT  | VQRRVQPKVT | WYPSKTQPLQ | HHNLLVCSVS | GFYPGSIEVR  | WFLNGQEEKA  | GMVSTGLIQN | GDWTFQTLWM | LETVPRSGEV |  |
| DRB1*03:01:01:01 | K-GR--N--- | -----      | ---H---    | -----      | -----      | -----       | ---R---T    | -V---H-    | -----      | -----      |  |
| DRB1*04:05:01:01 | R-----     | ---G---    | ---Y-E--   | ---A---    | -----N     | -----       | ---R---T    | -V---H-    | -----      | -----      |  |
| DRB1*08:01:01    | R--L-----  | ---G---    | ---H---    | -----      | -----      | -----       | ---R---T    | -V---H-    | -----      | -----      |  |
| DRB1*13:03:01    | K-----     | ---G---    | ---H---    | -----      | -----      | -----       | ---R---T    | -V---H-    | -----      | -----      |  |
| DRB1*15:02:01:01 | -----      | ---G---    | -----      | -----      | -----      | -----       | -----       | -----      | -----      | -----      |  |
| DRB1*15:03:01:01 | -----      | -----      | -----      | -----      | -----      | -----       | -----       | -----      | -----      | -----      |  |
| AA Pos.          | 180        | 190        | 200        | 210        | 220        | 230         |             |            |            |            |  |
| DRB1*15:01:01:01 | YTCQVEHPSV | TSPLTVEWRA | RSESAQSKML | SGVGGFVLGL | LFLGAGLFIY | FRNQKGHSGL  | QPTGFLS     |            |            |            |  |
| DRB1*03:01:01:01 | -----      | -----      | -----      | -----      | -----      | -----       | ---R---     |            |            |            |  |
| DRB1*04:05:01:01 | -----L     | -----      | -----      | -----      | -----      | -----       | -----       |            |            |            |  |
| DRB1*08:01:01    | -----      | -----S-    | -----      | -----      | -----      | -----       | -----       |            |            |            |  |
| DRB1*13:03:01    | -----      | -----      | -----      | -----      | -----      | -----       | ---R---     |            |            |            |  |
| DRB1*15:02:01:01 | -----      | -----      | -----      | -----      | -----      | -----       | -----       |            |            |            |  |
| DRB1*15:03:01:01 | -----      | -----      | -----      | -----      | -----      | -----       | -----       |            |            |            |  |

**Figure I-25.** Full Length Protein Sequence Alignment of these shown HLA-DRβ1 chains according to IPD-IMGT/HLA Release 3.41.0 (2020-07-13). The alignment above is a graphical representation to allow comparison of known sequences. Every hyphen symbol represents a given position in the sequence where amino acidic residue (aa) is identical between, in this particular case, HLA-DRβ1\*15:01 reference protein sequence aligned and compared here with sequences of other respective HLA-DRβ1 proteins of interest reported for conferring risk to

MS susceptibility. Figure and respective footnote are obtained and adapted from: <https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/align.cg> [87].



**Figure I-26.** Graphical representation of a 3D ribbon model for generic HLA-DR $\alpha$ /HLA-DR $\beta$  heterodimer molecule, indicating the location of amino acid position 86 within the P1 pocket of the peptide binding groove.

-In the case of Hispanic ethnic groups and Latin American population groups, incidence of MS appears to be much lower [699] and it is mostly due to the historical European genetic contribution (with the well-known *HLA-DRB1\*15:01* association, and even also the *HLA-DRB1\*03:01* association previously described) [700-703] from the colonization period between 15<sup>th</sup> and 19<sup>th</sup> centuries [267]. These current populations of the New World resulted from complex and ongoing admixture processes during the last five centuries involving mainly Native American (Amerindian) and European (colonizers mainly from Iberia) genetic components; and in a lesser extent (more localized depending on the region) also with the contributions of sub-Saharan African (from the transatlantic African slave trade) and South-East Asian genetic backgrounds as well as observed minorities from North Africa and Middle-East, and even including a few Romani people [522]. Thus, particularly in certain specific familial MS studies of admixed groups additional MS HLA risk signatures (i.e. ancestry-specific MS-associated HLA alleles previously mentioned) from these other ethnic

backgrounds have been also identified although being less significant [701]. Moreover, epidemiological studies show an extremely low prevalence of MS among non-mixed Amerindians [702]. Yet, the possible genetic contribution (for both risk and protection) of Native American ancestry to MS susceptibility in patients still needs to be further investigated as it has been initially described in very recent studies in MS [704] and also in a relatively similar neurological disease (of a clear autoimmune nature) denominated Neuromyelitis Optica [705].

-Although still limited and without consistent results, studies on MS patients from the Middle-East and North Africa regions have initially shown certain similar trends and allele/haplotype associations (e.g. *HLA-DRB1\*15:01*, *-DRB1\*03* and *-DRB1\*04*) as those found in European cohorts from the Mediterranean basin [706][707].

-Protective effects in HLA class II region have been also observed in some studies. *HLA-DRB1\*14:01* allele has been consistently associated with protection to MS susceptibility in European population cohorts [674][676]. Where it seems to exert a dominant counter effect on the susceptibility attributed to *HLA-DRB1\*15:01* major risk allele when both are present in the same heterozygous genotype [676]. Similar initial findings have been also identified on the tentative protective role of *DRB1\*07* allele group [690][708][709]. In African American populations, *HLA-DRB1\*11:01* and *HLA-DRB1\*04:01* have been identified as resistant alleles, while *HLA-DRB1\*14:01* and *HLA-DRB1\*09:01* are suggestive (although weakly) resistant alleles for MS among African descent populations [669][683]. Finally, similarly to the African American background, *HLA-DRB1\*09:01* has been reported to confer resistance to MS in both Japanese and Chinese populations [678][681]. While the mechanism for class II-mediated protection in MS is still unknown, engagement of MHC-promiscuous, auto-

reactive thymocytes, and resultant Treg formation, has been suggested as an explanation for these associations [710].

iii) Despite their implication in MS has been much less examined so far, some HLA class I alleles have been also reported to be associated with either risk or protection to MS susceptibility, considering also tentative epistatic interactions [711]. A possible epistatic risk interaction between *HLA-DRB1\*15:01* and *HLA-A\*03:01* was identified in a Norwegian population study [712]. Likewise, associations for risk to MS susceptibility observed for *HLA-B\*07* allele group in European descent populations are most likely secondary to LD with the canonical extended HLA class II *DRB1\*15:01~DQB1\*06:02* risk haplotype [713]. Conversely, the HLA class I protective effect has been reported to be mainly driven by *HLA-A\*02:01* allele, which does not appear to be secondary to class II associations [652]; and also by *HLA-B\*44:02* allele (presenting the motif Bw4-80T, later explained) [714][715], however this latter is in extremely tight LD with *HLA-C\*05* allele group and thus it is difficult to discriminate between their tentative individual contributions to protection [650]. In addition, HLA class I-mediated protection has also been observed for *HLA-B\*38:01* and *HLA-B\*55:01* alleles in some HLA imputation studies in European descent populations [658][677]. At the same time, these associations detected for HLA class I alleles (commonly interpreted only in the context of peptide binding and presentation driving specific adaptive immune T cell responses) have been also analyzed and explained with respect to their role as regulatory ligands for killer-cell immunoglobulin-like (KIR) receptors [5], mainly expressed on natural killer (NK) cells and being key elements in innate immunity and possible contributors to MS pathogenesis [716]. Main studies have been focused in the following well-documented HLA-KIR interactions, where NK-mediated responses are basically governed by the avidity of interaction with HLA class I  $\alpha 1$ -helix residue-80 (summarized in [717-719]):



-Even with a very high polymorphism, HLA-B alleles can be clustered based on polymorphisms at amino acid positions 77–83 that define either the Bw6 epitope, which has not been described to interact with any KIR [720], or the Bw4 epitope (that, in turn, can be divided in several different Bw4 motif groups), a ligand for KIR3DL1 receptor [721]. Based on a dimorphism (Ile(I) versus Thr(T)) at position 80 that affects interaction with KIR3DL1, the HLA-Bw4 alleles can be further segregated into Bw4-80I or Bw4-80T subtypes. In particular, Bw4 alleles characterized by I80 give an optimal inhibitory signal, whereas those characterized by T80 are responsible for weaker inhibition. Thus, KIR3DL1 interacts exclusively with HLA class I molecules that contain the Bw4 epitope, which is present within ~33% of HLA-B allotypes and also ~20% of HLA-A allotypes [537][538].

-The dimorphism among HLA-C molecules at position 80 is recognized by KIR2DL1 and KIR2DL2/3 and dictates their reactivity with C2 (characterized by V76 and K80 residues, essentially the majority of HLA-Cw2, 4, 5, and 6 and some other alleles) or C1 (displaying V76 and N80 amino acids, mainly defined by HLA-Cw1, 3, 7, 8, and some other alleles) molecules, respectively [722][723].

The specific interaction between these highly polymorphic KIR and HLA loci appears to play a relevant role in both progression and outcome of several diseases, such as viral infection, cancer and autoimmunity (summarized in [717]). In MS studies, the most consistent finding has been so far that the combined presence of KIR3DL1 and Bw4 appears to confer protection to MS susceptibility in both European and African-American cohorts [724-727].

Finally, in the following **Figure I-27**, and for the purposes of the present thesis work, it is shown the main list of HLA-B antigens that present either serological Bw6 epitope or Bw4 epitope (including also main Bw4 motif subgroups). Amino acid residues indicated for each Bw6/Bw4

epitope correspond to positions 77-83 (**Figure I-27** is obtained and adapted from:

<http://www.dorak.info/hla/bw4bw6.html> [728]):

| SLRNLRG                 |                         | DLRTLRL                   | NLRIALR                   | NLRTALR                 | VARIOUS                |
|-------------------------|-------------------------|---------------------------|---------------------------|-------------------------|------------------------|
| <b>Bw6</b>              |                         | <b>Bw4</b>                | <b>Bw4</b>                | <b>Bw4</b>              |                        |
| <a href="#">B*07021</a> | <a href="#">B*35011</a> | <a href="#">B*0727</a>    | <a href="#">B*0803</a>    | <a href="#">B*0802</a>  | <a href="#">B*0711</a> |
| <a href="#">B*07022</a> | <a href="#">B*35012</a> | <a href="#">B*1543</a>    | <a href="#">B*1513</a>    | <a href="#">B*1301</a>  | <a href="#">B*0723</a> |
| <a href="#">B*07023</a> | <a href="#">B*3502</a>  | <a href="#">B*2703</a>    | <a href="#">B*1516</a>    | <a href="#">B*1302</a>  | <a href="#">B*0806</a> |
| <a href="#">B*0703</a>  | <a href="#">B*3503</a>  | <a href="#">B*27052</a>   | <a href="#">B*1517101</a> | <a href="#">B*1303</a>  | <a href="#">B*1557</a> |
| <a href="#">B*0704</a>  | <a href="#">B*3504</a>  | <a href="#">B*27053</a>   | <a href="#">B*1517102</a> | <a href="#">B*1304</a>  | <a href="#">B*3527</a> |
| <a href="#">B*0705</a>  | <a href="#">B*3505</a>  | <a href="#">B*27054</a>   | <a href="#">B*1523</a>    | <a href="#">B*1306</a>  | <a href="#">B*3536</a> |
| <a href="#">B*0706</a>  | <a href="#">B*3506</a>  | <a href="#">B*2707</a>    | <a href="#">B*1524</a>    | <a href="#">B*1308</a>  | <a href="#">B*3705</a> |
| <a href="#">B*0707</a>  | <a href="#">B*3507</a>  | <a href="#">B*2709</a>    | <a href="#">B*1567</a>    | <a href="#">B*1536</a>  | <a href="#">B*3920</a> |
| <a href="#">B*0708</a>  | <a href="#">B*3508</a>  | <a href="#">B*2710</a>    | <a href="#">B*2702</a>    | <a href="#">B*1809</a>  | <a href="#">B*4037</a> |
| <a href="#">B*0709</a>  | <a href="#">B*35091</a> | <a href="#">B*2713</a>    | <a href="#">B*3801</a>    | <a href="#">B*2701</a>  | <a href="#">B*4105</a> |
| <a href="#">B*0710</a>  | <a href="#">B*35092</a> | <a href="#">B*2714</a>    | <a href="#">B*3805</a>    | <a href="#">B*38021</a> | <a href="#">B*4703</a> |
| <a href="#">B*0712</a>  | <a href="#">B*3510</a>  | <a href="#">B*2716</a>    | <a href="#">B*3806</a>    | <a href="#">B*38022</a> | <a href="#">B*5305</a> |
| <a href="#">B*0713</a>  | <a href="#">B*3511</a>  | <a href="#">B*2717</a>    | <a href="#">B*3807</a>    | <a href="#">B*3804</a>  | <a href="#">B*7301</a> |
| <a href="#">B*0714</a>  | <a href="#">B*3512</a>  | <a href="#">B*2719</a>    | <a href="#">B*4013</a>    | <a href="#">B*44021</a> |                        |
| <a href="#">B*0715</a>  | <a href="#">B*3513</a>  | <a href="#">B*3701</a>    | <a href="#">B*4019</a>    | <a href="#">B*44022</a> |                        |
| <a href="#">B*0716</a>  | <a href="#">B*3514</a>  | <a href="#">B*3702</a>    | <a href="#">B*4406</a>    | <a href="#">B*44023</a> |                        |
| <a href="#">B*0717</a>  | <a href="#">B*3515</a>  | <a href="#">B*3704</a>    | <a href="#">B*4418</a>    | <a href="#">B*44031</a> |                        |
| <a href="#">B*0718</a>  | <a href="#">B*3516</a>  | <a href="#">B*4701101</a> | <a href="#">B*4425</a>    | <a href="#">B*44032</a> |                        |
| <a href="#">B*0719</a>  | <a href="#">B*3517</a>  | <a href="#">B*4701102</a> | <a href="#">B*4901</a>    | <a href="#">B*4404</a>  |                        |
| <a href="#">B*0720</a>  | <a href="#">B*3518</a>  | <a href="#">B*5303</a>    | <a href="#">B*4903</a>    | <a href="#">B*4405</a>  |                        |
| <a href="#">B*0721</a>  | <a href="#">B*3519</a>  |                           | <a href="#">B*51011</a>   | <a href="#">B*4407</a>  |                        |
| <a href="#">B*0722</a>  | <a href="#">B*3520</a>  | <b>SLRTLRL</b>            | <a href="#">B*51012</a>   | <a href="#">B*4408</a>  |                        |
| <a href="#">B*0724</a>  | <a href="#">B*3521</a>  | <a href="#">B*2704</a>    | <a href="#">B*51013</a>   | <a href="#">B*4410</a>  |                        |
| <a href="#">B*0725</a>  | <a href="#">B*3522</a>  | <a href="#">B*2706</a>    | <a href="#">B*51014</a>   | <a href="#">B*4411</a>  |                        |
| <a href="#">B*0726</a>  | <a href="#">B*3523</a>  | <a href="#">B*2711</a>    | <a href="#">B*51021</a>   | <a href="#">B*4412</a>  |                        |
| <a href="#">B*0801</a>  | <a href="#">B*3524</a>  | <a href="#">B*2715</a>    | <a href="#">B*51022</a>   | <a href="#">B*4413</a>  |                        |
| <a href="#">B*0804</a>  | <a href="#">B*3525</a>  | <a href="#">B*2720</a>    | <a href="#">B*5103</a>    | <a href="#">B*4414</a>  |                        |
| <a href="#">B*0805</a>  | <a href="#">B*3526</a>  | <a href="#">B*2721</a>    | <a href="#">B*5104</a>    | <a href="#">B*4416</a>  |                        |
| <a href="#">B*0807</a>  | <a href="#">B*3528</a>  | <a href="#">B*2722</a>    | <a href="#">B*5105</a>    | <a href="#">B*4417</a>  |                        |
| <a href="#">B*0809</a>  | <a href="#">B*3529</a>  | <a href="#">B*2723</a>    | <a href="#">B*5106</a>    | <a href="#">B*4420</a>  |                        |
| <a href="#">B*0810</a>  | <a href="#">B*3530</a>  |                           | <a href="#">B*5107</a>    | <a href="#">B*4421</a>  |                        |
| <a href="#">B*0811</a>  | <a href="#">B*3531</a>  |                           | <a href="#">B*5108</a>    | <a href="#">B*4422</a>  |                        |
| <a href="#">B*0812</a>  | <a href="#">B*3532</a>  |                           | <a href="#">B*5109</a>    | <a href="#">B*4424</a>  |                        |
| <a href="#">B*0813</a>  | <a href="#">B*3533</a>  |                           | <a href="#">B*5110</a>    | <a href="#">B*4426</a>  |                        |
| <a href="#">B*0814</a>  | <a href="#">B*3534</a>  |                           | <a href="#">B*5112</a>    | <a href="#">B*4427</a>  |                        |
| <a href="#">B*1309</a>  | <a href="#">B*3535</a>  |                           | <a href="#">B*51131</a>   | <a href="#">B*4704</a>  |                        |
| <a href="#">B*1401</a>  | <a href="#">B*3537</a>  |                           | <a href="#">B*51132</a>   | <a href="#">B*4902</a>  |                        |
| <a href="#">B*1402</a>  | <a href="#">B*39011</a> |                           | <a href="#">B*5114</a>    | <a href="#">B*5607</a>  |                        |
| <a href="#">B*1403</a>  | <a href="#">B*39013</a> |                           | <a href="#">B*5115</a>    |                         |                        |
| <a href="#">B*1404</a>  | <a href="#">B*39014</a> |                           | <a href="#">B*5116</a>    | <b>SLRTALR</b>          |                        |
| <a href="#">B*1405</a>  | <a href="#">B*39021</a> |                           | <a href="#">B*5117</a>    | <a href="#">B*3803</a>  |                        |
| <a href="#">B*14061</a> | <a href="#">B*39022</a> |                           | <a href="#">B*5118</a>    | <a href="#">B*4415</a>  |                        |
| <a href="#">B*14062</a> | <a href="#">B*3903</a>  |                           | <a href="#">B*5119</a>    |                         |                        |

---

|                                  |                                  |                                |
|----------------------------------|----------------------------------|--------------------------------|
| <a href="#"><u>B*1501101</u></a> | <a href="#"><u>B*3904</u></a>    | <a href="#"><u>B*5120</u></a>  |
| <a href="#"><u>B*15012</u></a>   | <a href="#"><u>B*3905</u></a>    | <a href="#"><u>B*5121</u></a>  |
| <a href="#"><u>B*15013</u></a>   | <a href="#"><u>B*39061</u></a>   | <a href="#"><u>B*5122</u></a>  |
| <a href="#"><u>B*15014</u></a>   | <a href="#"><u>B*39062</u></a>   | <a href="#"><u>B*5123</u></a>  |
| <a href="#"><u>B*1502</u></a>    | <a href="#"><u>B*3907</u></a>    | <a href="#"><u>B*5124</u></a>  |
| <a href="#"><u>B*1503</u></a>    | <a href="#"><u>B*3908</u></a>    | <a href="#"><u>B*5126</u></a>  |
| <a href="#"><u>B*1504</u></a>    | <a href="#"><u>B*3909</u></a>    | <a href="#"><u>B*52011</u></a> |
| <a href="#"><u>B*1505</u></a>    | <a href="#"><u>B*3910</u></a>    | <a href="#"><u>B*52012</u></a> |
| <a href="#"><u>B*1506</u></a>    | <a href="#"><u>B*3911</u></a>    | <a href="#"><u>B*52013</u></a> |
| <a href="#"><u>B*1507</u></a>    | <a href="#"><u>B*3912</u></a>    | <a href="#"><u>B*5202</u></a>  |
| <a href="#"><u>B*1508</u></a>    | <a href="#"><u>B*3913</u></a>    | <a href="#"><u>B*5203</u></a>  |
| <a href="#"><u>B*1509</u></a>    | <a href="#"><u>B*3914</u></a>    | <a href="#"><u>B*5301</u></a>  |
| <a href="#"><u>B*1510</u></a>    | <a href="#"><u>B*3915</u></a>    | <a href="#"><u>B*5302</u></a>  |
| <a href="#"><u>B*15111</u></a>   | <a href="#"><u>B*3917</u></a>    | <a href="#"><u>B*5304</u></a>  |
| <a href="#"><u>B*15112</u></a>   | <a href="#"><u>B*3918</u></a>    | <a href="#"><u>B*5306</u></a>  |
| <a href="#"><u>B*1512</u></a>    | <a href="#"><u>B*3919</u></a>    | <a href="#"><u>B*5307</u></a>  |
| <a href="#"><u>B*1514</u></a>    | <a href="#"><u>B*3922</u></a>    | <a href="#"><u>B*5308</u></a>  |
| <a href="#"><u>B*1515</u></a>    | <a href="#"><u>B*3923</u></a>    | <a href="#"><u>B*5701</u></a>  |
| <a href="#"><u>B*1518</u></a>    | <a href="#"><u>B*3924</u></a>    | <a href="#"><u>B*5702</u></a>  |
| <a href="#"><u>B*1519</u></a>    | <a href="#"><u>B*40011</u></a>   | <a href="#"><u>B*57031</u></a> |
| <a href="#"><u>B*1520</u></a>    | <a href="#"><u>B*40012</u></a>   | <a href="#"><u>B*57032</u></a> |
| <a href="#"><u>B*1521</u></a>    | <a href="#"><u>B*40013</u></a>   | <a href="#"><u>B*5704</u></a>  |
| <a href="#"><u>B*1522</u></a>    | <a href="#"><u>B*4002</u></a>    | <a href="#"><u>B*5705</u></a>  |
| <a href="#"><u>B*1525</u></a>    | <a href="#"><u>B*4003</u></a>    | <a href="#"><u>B*5706</u></a>  |
| <a href="#"><u>B*1527</u></a>    | <a href="#"><u>B*4004</u></a>    | <a href="#"><u>B*5707</u></a>  |
| <a href="#"><u>B*1528</u></a>    | <a href="#"><u>B*4006101</u></a> | <a href="#"><u>B*5709</u></a>  |
| <a href="#"><u>B*1529</u></a>    | <a href="#"><u>B*4006102</u></a> | <a href="#"><u>B*5801</u></a>  |
| <a href="#"><u>B*1530</u></a>    | <a href="#"><u>B*4007</u></a>    | <a href="#"><u>B*5802</u></a>  |
| <a href="#"><u>B*1531</u></a>    | <a href="#"><u>B*4008</u></a>    | <a href="#"><u>B*5804</u></a>  |
| <a href="#"><u>B*1532</u></a>    | <a href="#"><u>B*4009</u></a>    | <a href="#"><u>B*5805</u></a>  |
| <a href="#"><u>B*1533</u></a>    | <a href="#"><u>B*4010</u></a>    | <a href="#"><u>B*5806</u></a>  |
| <a href="#"><u>B*1534</u></a>    | <a href="#"><u>B*4011</u></a>    | <a href="#"><u>B*5901</u></a>  |
| <a href="#"><u>B*1535</u></a>    | <a href="#"><u>B*4012</u></a>    |                                |
| <a href="#"><u>B*1537</u></a>    | <a href="#"><u>B*4014</u></a>    |                                |
| <a href="#"><u>B*1538</u></a>    | <a href="#"><u>B*4015</u></a>    |                                |
| <a href="#"><u>B*1539</u></a>    | <a href="#"><u>B*4016</u></a>    |                                |
| <a href="#"><u>B*1540</u></a>    | <a href="#"><u>B*4018</u></a>    |                                |
| <a href="#"><u>B*1542</u></a>    | <a href="#"><u>B*4020</u></a>    |                                |
| <a href="#"><u>B*1544</u></a>    | <a href="#"><u>B*4021</u></a>    |                                |
| <a href="#"><u>B*1545</u></a>    | <a href="#"><u>B*4023</u></a>    |                                |
| <a href="#"><u>B*1546</u></a>    | <a href="#"><u>B*4024</u></a>    |                                |
| <a href="#"><u>B*1547</u></a>    | <a href="#"><u>B*4025</u></a>    |                                |
| <a href="#"><u>B*1548</u></a>    | <a href="#"><u>B*4026</u></a>    |                                |
| <a href="#"><u>B*1549</u></a>    | <a href="#"><u>B*4028</u></a>    |                                |
| <a href="#"><u>B*1550</u></a>    | <a href="#"><u>B*4029</u></a>    |                                |
| <a href="#"><u>B*1551</u></a>    | <a href="#"><u>B*4030</u></a>    |                                |
| <a href="#"><u>B*1552</u></a>    | <a href="#"><u>B*4031</u></a>    |                                |
| <a href="#"><u>B*1553</u></a>    | <a href="#"><u>B*4032</u></a>    |                                |
| <a href="#"><u>B*1554</u></a>    | <a href="#"><u>B*4033</u></a>    |                                |
| <a href="#"><u>B*1555</u></a>    | <a href="#"><u>B*4034</u></a>    |                                |
| <a href="#"><u>B*1556</u></a>    | <a href="#"><u>B*4035</u></a>    |                                |
| <a href="#"><u>B*1558</u></a>    | <a href="#"><u>B*4036</u></a>    |                                |

---

[B\\*1559](#)            [B\\*4038](#)  
[B\\*1560](#)            [B\\*4039](#)  
[B\\*1561](#)            [B\\*4101](#)  
[B\\*1562](#)            [B\\*4102](#)  
[B\\*1563](#)            [B\\*4103](#)  
[B\\*1564](#)            [B\\*4104](#)  
[B\\*1565](#)            [B\\*4106](#)  
[B\\*1566](#)            [B\\*4201](#)  
[B\\*1568](#)            [B\\*4202](#)  
[B\\*1569](#)            [B\\*4409](#)  
[B\\*1801](#)            [B\\*4501](#)  
[B\\*1802](#)            [B\\*4502](#)  
[B\\*1803](#)            [B\\*4503](#)  
[B\\*1804](#)            [B\\*4504](#)  
[B\\*1805](#)            [B\\*4505](#)  
[B\\*1806](#)            [B\\*4601](#)  
[B\\*1807](#)            [B\\*4602](#)  
[B\\*1808](#)            [B\\*4702](#)  
[B\\*1810](#)            [B\\*4801](#)  
[B\\*1811](#)            [B\\*4802](#)  
[B\\*1812](#)            [B\\*4803](#)  
[B\\*1813](#)            [B\\*4804](#)  
[B\\*2708](#)            [B\\*4805](#)  
[B\\*2712](#)            [B\\*4806](#)  
[B\\*2718](#)            [B\\*4807](#)  
                      [B\\*5001](#)  
                      [B\\*5002](#)  
                      [B\\*5004](#)  
                      [B\\*5401](#)  
                      [B\\*5402](#)  
                      [B\\*5501](#)  
                      [B\\*5502](#)  
                      [B\\*5503](#)  
                      [B\\*5504](#)  
                      [B\\*5505](#)  
                      [B\\*5507](#)  
                      [B\\*5508](#)  
                      [B\\*5509](#)  
                      [B\\*5510](#)  
                      [B\\*5601](#)  
                      [B\\*5602](#)  
                      [B\\*5603](#)  
                      [B\\*5604](#)  
                      [B\\*5605](#)  
                      [B\\*5606](#)  
                      [B\\*67011](#)  
                      [B\\*67012](#)  
                      [B\\*6702](#)  
                      [B\\*7801](#)  
                      [B\\*78021](#)  
                      [B\\*78022](#)  
                      [B\\*7803](#)

---

B\*7804

B\*7805

B\*8101

B\*8201

B\*8202

B\*83:01

---

Several main studies of HLA-MS association have been performed in Spanish population cohorts although just from very specific regions or genetic backgrounds. For the most part, cohorts of Caucasian Spanish ancestry have been evaluated from Northern and Eastern regions (Mediterranean region) in Spain as well as from few Central and Southern regions, including also one very small study in Spanish Romani Gypsy patients from Andalusia [635][689][709][729-742][761]. Moreover, all these cited studies were based on direct DNA sequencing (i.e. direct DNA-based typing techniques such as SSO or SBT) of the HLA region. However, molecular characterization of these HLA-MS associations reported in these previous studies has been highly limited due to the only use of lower resolution (e.g. either 1-field or 2-field) HLA typing methods with several drawbacks: an important level of both allelic and phasing ambiguities; covering just a few HLA loci (where most extended haplotypes have not been well-defined and thus, for example, lacking the description of relevant class II *HLA-DRB3/4/5* or HLA class I loci [293][650][668][743]); and in most cases restricted to relatively small sample size cohorts. Yet, there have been important findings that have partially described the HLA genetic background associated to risk/protection for MS in Spanish population. In summary, from these previously reported HLA studies in Spanish population (which are relatively representative of the general MS patient and healthy population groups although not completely):

- i) The most consistent finding has been the association between MS risk and the canonical DR2 (DR15) haplotype (*HLA-DRB1\*15:01~HLA-DQA1\*01:02~HLA-DQB1\*06:02*) as well as respective risk associations from these embedded individual alleles, being in line with many

studies carried out in populations of European ancestry (reviewed in [650][664]). In this sense, this may be most likely reflecting the significant genetic imprint left by Germanic tribes/Christian Visigoths in modern-day populations of the Iberian Peninsula [555][556][558][578]. Moreover, in some of these studies, *HLA-DQB1\*06:02* allele has been particularly found to be more prevalent in MS cases and to maintain its significance after correction for univariate analysis, and also when analyzed by means of a multivariate test with a logistic regression model [741][742]. Nevertheless, similar to other European descent populations, identification of the truly causative predisposing alleles within this risk haplotype for MS susceptibility is unlikely to be unraveled by HLA Caucasian population association studies due to the extraordinarily strong linkage disequilibrium displayed between *HLA-DRB1\*15:01* and *-DQB1\*06:02* alleles (reviewed in [650]). Furthermore, the possible epistatic (both cis- and trans-) interactions involved between these HLA class II genes (and also class I loci) and their contribution (within the context of the combined effect of the two parental haplotypes) to the overall risk for developing MS also need to be taken into account and, in turn, to be further elucidated in future studies [672][709][711][744-746]. At the same time, it has been also postulated that the observed effects associated to individual classical HLA alleles may be modulated (in terms of biological functions, structural properties and expression patterns [747]) to some extent by many weak effects at many loci (SNPs) across the genome (in a network of fine-regulated interactions between HLA and non-HLA risk/protection variants), thus defining a polygenic epistasis [309][659].

ii) Among these previously reported HLA-MS Spanish population studies, some of them can be considered relatively comprehensive since most of them are based on intermediate high-resolution 2-field HLA data where, at least, main HLA class II genes of interest were generally interrogated and relatively large sample sizes were studied [635][689][709][729-

742][761]. Thus, apart from the DR15 major risk signal, additional secondary (in some cases more tentative than others) but informative HLA-MS positive (suggestive to confer susceptibility) and negative (suggestive to confer protection) association signals have been also initially described both at the allelic and haplotypic levels in some given Spanish population groups/regions. In detail:

-In Northern [742] and Northeastern Spain [734][735], haplotype *HLA-DRB1\*13:03~DQA1\*05:05~DQB1\*03:01* was observed to be increased in MS patients. This finding is in consonance with, among others, studies of MS patients in Ireland [748], Sardinia [670], U.S (European American) [749] and non-Ashkenazi patients from Israel [675].

-Other studies in Spanish MS cohorts from various geographical regions, despite being located far apart, such as the Basque region (Northern Spain) [742], Canary Islands [738] and Madrid (Central region of Spain) [732] have also detected associated risk that would be conferred by haplotype *HLA-DRB1\*04:02~DQA1\*03:01~DQB1\*03:02*. Where, in particular, this positive association with MS has been linked to some DR4 alleles (*HLA-DRB1\*04:02*, *-DRB1\*04:03*, and *-DRB1\*04:04*) that indeed present Val residue at HLA-DRβ86 position. Which has been suggested to play a role for susceptibility to MS, since this amino acid position defines the configuration of P1 pocket of the peptide-binding groove and is located at the end of the alpha helix in the given HLA-DRβ chain, being thus a key element for the peptide presentation to T cells [650][732]. However, this specific molecular motif has not been consistently found in multi-ethnic MS patients, having important exceptions such as *HLA-DRB1\*04:05* risk allele (presenting Gly at HLA-DRβ86 position instead) which appears increased, for example, in MS Asian populations [678][679] and also, strikingly, in MS Sardinian patients [670]. Moreover, De la Concha et al study in

Madrid MS patients [732] showed that *HLA-DRB1\*15:01* and DRβ86-Val DR4 alleles were associated only with approximately 50% of MS in that studied cohort and thus did not account for all the susceptibility to the disease (where other different gene associations both inside and outside the HLA region may be also relevant in the pathogenesis of MS [650]). Also, it is noteworthy that when the statistically conservative Bonferroni's method correction is applied, and probably due to the relative small sample size from these previous studies in Spanish population, none of these predisposing haplotypes for MS showed a significant difference in frequency between MS patients and controls [742]. On the other hand, and very interestingly, predisposing role to MS susceptibility for *HLA-DRB1\*04* allele group and its associated bearing haplotypes have been also initially described in MS patients from the Middle-East and North Africa regions (e.g. *HLA-DRB1\*04:04~DQA1\*03~DQB1\*03:02* haplotype associated in Turkish population; or *HLA-A\*01~B\*51~DRB1\*04* and *HLA-A\*03~B\*44~DRB1\*04* haplotypes associated in Iranian population) [706][707][750][751]. Therefore, this particular HLA-MS association signature observed here could be also reflecting, at least to certain extent, the shared demographic history and, consequently, genetic substrate between certain modern-day Middle-Eastern and North African MS/healthy cohorts and the current Spanish MS/healthy cohorts due to the original settlement for many centuries (8<sup>th</sup>-15<sup>th</sup> centuries in CE) of North African Muslim Eastern Arabs and Berbers in the Iberian Peninsula [555][556][558][578].

-In relation to the *HLA-DRB1\*03:01* risk allele association (which may contribute to risk through recessive modes or interaction effects) reported for other populations mainly of European ancestry [650][670][690-693]. This HLA allele has not been detected by itself as a clear predisposing factor to MS susceptibility in Spanish population studies so far (where both cohorts of healthy controls and MS patients generally showed similar allele frequency



distributions) [689][740]. Notwithstanding, a more recent study by De la Concha et al. [733] evaluated the role of *HLA-DRB1\*03:01* allele specifically in the context of its more conserved and common bearing extended haplotypes (such as ancestral haplotypes AH 8.1 (more common in Northern European populations) or AH 18.2 (more common in European Mediterranean populations)) in certain Spanish population cohorts (from both Madrid (Central region of Spain) and Malaga (Southern Spain)) according to MS risk and production of oligoclonal IgM against myelin lipids (OCMB), thus studying more homogeneous groups of MS patients. By following this study design and approach, it was observed that diverse *HLA-DRB1\*03:01* carrying haplotypes contribute with different risk to MS susceptibility, where, in particular, the ancestral haplotype AH 18.2 (*HLA-A\*26:01~C\*05:01~B\*18:01~DRB1\*03:01~DQA1\*02:01~DQB1\*05:01*) [486][530], which is also common in Sardinians, showed the highest risk specifically to those MS patients presenting OCMB. Whereas the more common extended *HLA-DRB1\*03:01* bearing haplotype in North European regions, the AH 8.1 (*HLA-A\*01:01~C\*07:01~B\*08:01~DRB1\*03:01~DQA1\*02:01~DQB1\*05:01*) [486][530] did not show that level of risk.

-At the same time, many studies (especially in MS patient populations of European descent) have thoroughly investigated the role of relevant associated genes (both MHC- and non-MHC-related) in disease outcome, prognosis and progression in MS. Given the fact that, in addition to influencing disease susceptibility, epidemiological and genetic evidence suggests that these genetic factors (considering also their respective epistatic interactions as well as complex interactions with environmental factors) may affect phenotypic expression of the disease [454][650][752]. However, associations of MS genetic risk variants (both HLA and non-HLA) and the clinical phenotype are still debatable although many studies have approached this aim (e.g. [753]). In this sense, some of these more comprehensive HLA-MS

studies conducted in Spanish population have also attempted to elucidate the impact of the different predisposing HLA alleles/haplotypes on the clinical phenotype with special interest in the disability progression of MS [635][689][737][739][741][742]. Nevertheless, similar to other worldwide family-based or unrelated HLA-MS studies (as reviewed in [689][752]) and even though employing large numbers of patients with MS (very well characterized in relation to all standardized clinical and demographical variables), evaluation of the association of the natural history of the disease and HLA genotype has led to controversial or even statistically no significant results and, thus, this still remains unclear. Still, it is important to consider the possibility that statistical power in these HLA-MS studies series may have not be sufficient to detect the reported effects or trends. Especially, if the given respective HLA datasets were underpowered to detect modifying genes of moderate effect following stratification by clinical categories (despite these may be more homogenous). In relation to the latter point, and as an additional issue, conflicting results may also be explained due to the different criteria used for selecting, categorizing and subdividing patients in each study. Moreover, another critical confounding aspect is the not well-defined yet epistatic interactions between HLA alleles/loci (where extended MHC haplotypes could be a fundamental unit of genetic regulation/deregulation of immune response and pathogenesis in MS) [711] that could help clarifying some of these contradictory and inconsistent results found on the correlation between the HLA genotype-phenotype in MS.

-Previous HLA-MS studies in Spanish population groups have also shed some light (although still results have not been very consistent and not with a total consensus between different studies [740]) into the identification of certain HLA alleles and respective bearing haplotypes with a given putative protective role. Where their respective relatively higher frequency distributions (or overrepresentation) observed in healthy control population

cohorts could be interpreted as that they are protective factors on their own or because of their specific epistatic interactions displayed, exerting a protective effect, with those other risk alleles that imply susceptibility to develop the disease [672][709][711][744-746]. Thus, in Northern [742] and Northeastern Spain [734][735] haplotype *HLA-DRB1\*01:01~DQAI\*01:01~DQBI\*05:01* as well as its individual embedded alleles (being especially significant for *HLA-DRB1\*01:01*) appear to exhibit a protective association with MS, all retaining statistical significance after application of respective statistical correction methods [742]. In this case, some studies have suggested that *HLA-DRB1\*01:01* carrying haplotypes may be specifically interacting in trans- with *HLA-DRB1\*15:01* bearing haplotypes [674][746][754]. In contrast, other reports have revealed protective properties of the *HLA-DRB1\*01:01* allele on its own [755][756]. Furthermore, in the HLA-MS study on Basque population from Northern Spain, *HLA-DQBI\*03:03* allele also seems to exert a protective effect on the MS outcome. Interestingly, relatively common respective *HLA-DQBI\*03:03* bearing extended haplotypes such as *HLA-DRB1\*09:01~DQAI\*03:02~DQBI\*03:03* and *HLA-DRB1\*07:01~DQAI\*02:01~DQBI\*03:03* have also been detected for their putative protective role in MS susceptibility in other studies [650][664]. Also, another HLA-MS case-control study of a Spanish population cohort from Catalonia (Northeastern Spain) [709] described that *HLA-DRB1\*07* allele could exert an epistatic effect along with the *HLA-DRB1\*15* allele in an opposite direction which neutralizes this genotype, although this hypothesis still needs to be further corroborated.

-Finally, in the peculiar, but very small (n=14), Spanish Romani Gypsy MS patient population study from the Andalusian region (Southern Spain) [729], *HLA-DRB1\*15:01*, *-DQBI\*06:02* and *-DQBI\*06:08* alleles were the only positive HLA associations with MS. Where *HLA-*

*DRB1\*15:01~DQB1\*06:02* was the most frequent haplotype in this MS group. In this sense, this could be illustrating the certain level of admixture in which this Spanish Romani Gypsy ethnic group (despite showing a relatively high level of consanguinity (i.e. genetically conserved) due to the high degree of endogamy and intramarriage that this people have practiced throughout their itinerant history) may have undergone in the Iberian Peninsula with local populations of a significant Northern-Central European genetic substrate, since they were socially and culturally well integrated [580][581][757][758]. Consequently, this would be also supporting the finding from Fernandez et al. MS study [729] that described an estimated relevant prevalence of MS in Gypsies in Andalusia (52/100,000) strikingly higher than in other previously reported European Romani populations (which were much more isolated and persecuted historically and, thus, tentatively experiencing lesser admixture with local populations) [759][760], and despite being significantly less than that the one found in Caucasians from Spain (75–79/100,000) [729]. On the other hand, *HLA-DRB1\*15:02~DQB1\*05:03*, *HLA-DRB1\*15:02~DQB1\*06:01* or *HLA-DRB1\*16:01~DQB1\*06:01* haplotypes were not detected in the Gypsy MS group, whereas they were present significantly in the Gypsy healthy controls [729].

| Spanish Population/Region                    | Sample Size Patients (MS) / Controls (HC) | Main HLA-MS positive (+) / negative (-) associations identified (generally for both allele and common bearing haplotypes)                      | Ref.       |
|--|---|--|------------|
| <u>Basque (North)</u>                        | 197 / 200                                 | (+) <i>DRB1*15:01, DRB1*04:02, DRB1*13:03</i><br>(-) <i>DRB1*01:01, DQB1*03:03</i> (with <i>DRB1*07:01</i> and <i>DRB1*09:01</i> )             | [742]      |
| <u>Malaga (Caucasian, South)</u>             | 149 / 160                                 | (+) <i>DRB1*15:01, DQA1*01:02</i> and, especially, <i>DQB1*06:02</i>   | [741]      |
| <u>Madrid (Central)</u>                      | 143 / 143                                 | (+) <i>DRw15, DQw6, Dw2</i>  | [731]      |
| <u>Madrid (Central)</u>                      | 135 / 168                                 | (+) <i>HLA-DRB1*15:01</i> and <i>IL-1Ra</i> allele 2 were significantly higher in R/R MS   | [732]      |
| <u>Madrid and Malaga (Central and South)</u> | 1068 / 624                                | (+) <i>DRB1*03:01</i> -containing AH 18.2 with OCMB  | [733]      |
| <u>Calatayud, Aragon (North-East)</u>        | 31 / 895                                  | (+) <i>A19, B5, B41, Cw7, DR15(2)</i> ( <i>DRB1*15:01</i> and <i>DRB5*01:01</i> ), <i>DR6, DR13(6), DR10, DQ1</i><br>(-) <i>Cw4, DRI</i>       | [734][735] |
| <u>Asturias (North)</u>                      | 43 / 100                                  | (+) <i>B7, B27, DR2</i><br>(-) <i>B35</i>  | [761]      |
| <u>Mediterranean Spanish Basin</u>           | 194/0                                     | (+) <i>DR2</i> (familial MS dataset)   | [635][737] |
| <u>Gran Canaria (Canary Islands)</u>         | 53/55                                     | (+) <i>DR15</i> ( <i>DQA1*01:02, DQB1*06:02, DRB1*15:01</i> and <i>DRB5*01:01</i> ) and <i>DR4</i> ( <i>DRB1*04:02</i> and <i>DRB1*04:04</i> ) | [738]      |
| <u>Asturias (North)</u>                      | 121 / 156                                 | (+) <i>MICB*004</i> and <i>HLA-DRB1*15</i> belonging to the AH 7.1<br>(-) <i>DRB1*01</i>   | [730]      |

|   |            |   |       |
|---|------------|---|-------|
| <u>Catalonia</u><br>(North-East)  | 380 / 1088 | (+) <i>DRB1*15, DRB1*03</i><br>( <i>HLA-DRB1*01</i> and <i>-DRB1*04</i> alleles were independently associated with a worse prognosis when considering the time taken to reach severe disability)<br>(-) <i>DRB1*11</i>  | [689] |
| <u>Catalonia</u><br>(North-East)  | 268 / 1088 | (+) <i>HLA-DRB1*15</i> allele is associated with oligoclonal immunoglobulin IgG bands (OCB) positive patients with MS   | [739] |
| <u>Catalonia</u><br>(North-East)  | 380 / 1088 | (+) Genotypes <i>DRB1*08/*15, DRB1*03/*03, DRB1*03/*15</i> and <i>DRB1*04/*15</i> . The <i>DRB1*01/*04</i> and the <i>DRB1*15/*15</i> genotypes were associated with a worse prognosis when considering the time taken to reach severe disability.<br>(-) <i>DRB1*03/*07, DRB1*07/*08, DRB1*07/*16, DRB1*07/*15</i> | [709] |
| <u>Malaga</u><br>(Gypsy, South)   | 14 / 80    | (+) <i>HLA-DRB1*15:01~DQB1*06:02</i> and <i>HLA-DRB1*13~DQB1*06:08</i><br>(-) <i>HLA-DRB1*15:02~DQB1*05:03, HLA-DRB1*15:02~DQB1*06:01, HLA-DRB1*16:01~DQB1*06:01</i>  | [729] |
| <u>Asturias</u> (North)   | 96 / 123   | (+) <i>DR15/DQw6</i> , general <i>DR4/DQw8</i> , in primary progressive form<br>(-) <i>DRw13/DQw5</i>   | [736] |
| <b>Figure I-28.</b> Summary of main HLA-MS Spanish population studies reported in literature. |            |   |       |

In this context, mechanistically and functionally speaking, still it has not been elucidated how all these HLA-associated alleles/haplotypes exactly contribute to MS susceptibility [658][663]. Although it is thought that, consistent with the known biology of MS, disease-associated variants in HLA-DR/-DQ could primarily influence the structural characteristics of the peptide-binding groove and, thus, presumably lead to alterations of the T cell repertoire that enhance the likelihood

of an inflammatory demyelinating process [658]. At the same time, it has been observed that the implicated HLA-associated alleles/haplotypes are neither necessary nor sufficient to cause or predict completely the development of MS, thus other pivotal factors must also contribute to the disease [662]. Moreover, despite the remarkable and comprehensive molecular dissection carried out at the HLA region in MS (mainly via GWAS with large sample sets (of predominant European descent) and controlling the best possible for any population stratification (i.e. differences in genetic structure between disease and control groups)), the role of genetic variation at the HLA loci has not been completely defined, due, in part, to the aforementioned extensive LD (i.e. a non-random statistical association of the variants due to physical linkage on the chromosomes) that exists among the HLA loci. Where this widespread LD across the entire HLA region hinders the identification of the true predisposing factor(s) within the detected disease susceptibility (and protective) haplotypes. Thus, any associated marker may not itself be the causal variant but is in linkage disequilibrium with the causal variant. To distinguish between these primary and secondary risk/protection effects due to LD, several studies have applied a practical approach by scrutinizing large number of HLA haplotypes in datasets with different ancestral histories, since LD patterns can differ between populations and, thus, enabling to narrow down the putative causative regions and improve our understanding of MS pathogenesis [664][668]. Also, as an statistical analysis strategy, conditional analyses (i.e. after excluding from the dataset a given statistically significant HLA variant found, all remaining variants are re-analyzed for association to identify the possible statistically independent effects) can be applied to separate allelic from haplotypic association and thus to discern hitchhiking effects of given detected associations [401][402][412]. In addition to the extended LD structure, structural complexity and high polymorphism of the human MHC; the observed ambiguity and the lack of replication for many of the HLA associations previously identified can be also attributed to the limited number of HLA

loci analyzed, the restricted allele resolution level offered by legacy HLA genotyping methods and the relatively small sample size of preceding studies. Therefore, subsequent fine-mapping studies of the associated genetic HLA region and related studies of the functional relevance of the respective specific variants have been and are still needed to further assess initial findings of association given mainly by GWAS [662]. Within this group of fine-mapping genetic studies, it is important to note that most of the reported HLA risk/protection associations with MS have been detected through large-scale association studies based on statistical imputation of HLA alleles from SNP data (i.e. where HLA imputation relies on patterns of LD in the tagging SNP flanking regions) [658], rather than direct DNA sequencing (e.g. HLA genotyping methods). However, accuracy of imputation results is quite limited and it varies with respect to HLA locus, machine-learning (imputation algorithm) training and testing populations, as well as reliability of confidence metrics associated with each prediction. Consequently, mechanistic assessment of these associations is certainly difficult or, even, impossible if not all alleles of a HLA locus are identified and/or some of them are imputed in error [650]. Moreover, the association testing of HLA imputation methods (mostly based on reference 2-field HLA healthy and MS population datasets (predominantly of European descent)) is restricted only to variations at the peptide-binding region of the HLA molecule, omitting examination of non-coding variants that may influence HLA expression or interaction with accessory molecules, as well as it is unable to detect novel variants, which may be relatively common for a given regional/ethnic group but are not considered in these reference HLA population datasets [650]. Furthermore, patterns of LD across the HLA region and the strength of these associations can drastically vary not only by locus or allele but particularly on specific haplotypes. For instance, due to the extensive structural variation found around *HLA-DRB1* locus, where the additional *HLA-DRB* loci (*HLA-DRB3*, *-DRB4* and *-DRB5*) are present to varying degrees depending on the specific *HLA-DRB1* lineage, imputation



accuracy can be also very inadequate [650]. To circumvent these limitations observed in HLA imputation methods (and in contrast to limited traditional low-resolution, low-coverage and low-throughput HLA genotyping methods), very recent large-scale, high-throughput and high-resolution HLA analysis of MS by NGS has allowed the almost full characterization of coding and non-coding sequence variation and it has significantly facilitated the description of the specific 3-/4-field allelic LD patterns displayed by the different HLA class I and class II loci (i.e. where alleles differing only in non-coding variations have differential associations with alleles of neighboring loci across all regions of the HLA system), being this very informative for MS fine-mapping studies [293][294][650][727]. Thus, it is possible to evaluate and to deconstruct more accurately, by direct DNA sequencing via NGS-based HLA genotyping data, the plausible synergic action established between different HLA alleles/loci to confer susceptibility or protection to MS risk [677]. In fact, this fine-mapping of MS susceptibility and protection can be more accurately performed when both HLA class I and II loci are examined simultaneously by this very high-resolution typing approach. For example, the broadly genotyped allelic groups of *HLA-B\*44* and *HLA-C\*05*, which have been described as having protective effects in MS risk and progression [714], show distinctive 3-/4-field haplotypic associations. In a NGS HLA European American population study [268], HLA alleles *-C\*05:01:01:01* and *-C\*05:01:01:02*, which differ by a single nucleotide substitution in intron 2, are found with distinctively tight associations with *HLA-B\*18:01:01* and *-B\*44:02:01:01* alleles, respectively. Furthermore, in the respective Creary et al. NGS MS study also in European American population [293] analyses of the *HLA-DRB1\*04* in the absence of *HLA-DRB1\*15:01* haplotypes revealed that the *HLA-DQB1\*03:01:01:01~DQA1\*03:03:01:01~DRB1\*04:01:01:01~DRB4\*01:03:01:01* haplotype was protective, whereas the *HLA-DQB1\*03:02:01~DQA1\*03:01:01~DRB1\*04:01:01:01~DRB4\*01:03:01:01* haplotype was associated with disease susceptibility. These recent findings

underscore the importance of evaluating variants at the highest HLA allele resolution level to identify with certainty the primary associations to MS risk/protection. Therefore, larger multi-ethnic studies using NGS-based HLA genotyping combined with genotyping of a highly dense set of SNP markers are planned to further elucidate the HLA contribution to MS pathogenesis. Also, results from these future HLA studies will also provide a more comprehensive guidance for conducting more detailed functional studies to unravel the causal variants and genes in MS as well as to establish more reliable correlations with respect to pivotal clinical phenotypic data such as disease severity or progression of MS, treatment regimens (e.g. therapeutic response to interferon-beta) and prognosis.

In summary, the lesson from the study of HLA polymorphism over the last several decades has been that each incremental technological advance that leads to higher resolution has yielded further insights into the cause or mechanisms of disease. With the advent of highest-resolution NGS technologies, there is an opportunity to more comprehensively define the role of HLA in health and disease populations. The present thesis work characterizes the nature and extent of HLA variation in established and well-characterized MS cohorts in Spanish population to evaluate HLA genetic associations as well as to provide a significant public data resource with a more complete description (for the first time in this population) of HLA variation for all exons, introns, and flanking regions in a representative Spanish population healthy control group.



## ***OBJECTIVES***



On the scope of the present thesis work, the main aim of this study was to explore and evaluate the advantages and the data offered by this novel NGS technology for obtaining very high-resolution (at the 3- to 4-field resolution) with minimum ambiguity of HLA genotypes at large-scale by a high-throughput system based on the multiplexing capability of NGS technology in combination with automated liquid-handling systems for the DNA library preparation protocol. Thus, the specific objectives of this thesis work, by which this would be achieved, were:

I) Characterization of HLA allele and haplotype diversity of all major classical HLA genes (*HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1* and *-DRB3/4/5*) by application of NGS of a first representative cohort of the Spanish population that could also serve as a healthy control reference group. This NGS-based HLA Spanish population study also includes statistical analyses for: estimating HLA allele (by count) and haplotype (via an expectation-maximization (EM) algorithm) frequencies; quality control (QC) verification of the integrity of NGS-based HLA genotype data for this given population cohort by Hardy-Weinberg Equilibrium Proportions test; measures of pairwise linkage disequilibrium (LD); measurement of selection by Ewens-Watterson homozygosity statistic; and estimation of genetics distances (Nei's  $D_A$  distance) between the main Spanish subgroups of this cohort from HLA allele frequency data by using the neighbor-joining method (NJ) as an exercise (i.e. test case) to study relatedness between these subgroups.

II) Characterization of HLA allele and haplotype diversity of all major classical HLA genes (*HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1* and *-DRB3/4/5*) by application of NGS of a respective Spanish cohort of multiple sclerosis (MS) patients (recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain). Thus, a first case-control study was carried out to examine HLA-disease associations with MS in these Spanish population

cohorts as well as to attempt a fine-mapping of these allele and haplotype associations by full gene resolution level via NGS. In addition, a second exercise (i.e. test case) of this case-control study was carried out using an alternative healthy control group dataset, in this case specifically from the Spanish northeastern region of Catalonia, to evaluate possible differences in the findings of HLA-disease association with MS due to plausible regional HLA genetic variation within Spanish population.







## ***MATERIALS AND METHODS***



## **1. STUDY POPULATION, DESIGN AND DATA COLLECTION**

The study of the current thesis work included three main cohorts in order to obtain their respective NGS high-resolution HLA genotype datasets. First, as part of the NGS HLA Population Genetics Unrelated Project at the 17<sup>th</sup> International Histocompatibility and Immunogenetics Workshop (IHIW) [269][328], a representative Spanish population cohort of healthy unrelated individuals randomly selected (collection of de-identified genomic DNA samples and their shipping, to the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory for testing, were coordinated by the Spanish Working Group in Histocompatibility and Transplant Immunology (GETHIT) of the Spanish Society for Immunology (SEI)) was studied and also used as a healthy control reference group for carrying out a subsequent case-control study. Second, under the auspices (NIH-funded) of both Immunogenetics of Neurological Diseases working GrOup (INDIGO) consortium and the HLA and KIR Region Genomics in Immune Mediated Diseases Consortium (HLARGC), a Spanish cohort of multiple sclerosis (MS) patients (collection of de-identified genomic DNA samples and their shipping, to the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory for testing, were coordinated by Dr. Albert Saiz and Dr. Pablo Villoslada from the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain) was tested for HLA genotyping via NGS as part of the aforementioned case-control study to evaluate HLA-disease association at the allele and haplotype levels. At the same time, in parallel to this first exercise (i.e. test case) of case-control study and in order to evaluate possible differences in the findings of HLA-disease association with MS due to plausible regional HLA genetic variation within mainland Spain. An alternative healthy control reference group was used, which comprised exclusively healthy unrelated individuals randomly selected from the Spanish northeastern region of Catalonia (collection of de-identified genomic DNA samples and their shipping, to the Stanford Blood Center HLA Histocompatibility and Immunogenetics

Laboratory for testing, were coordinated by Dr. José Luis Caro from the Histocompatibility and Immunogenetics Laboratory, Blood and Tissue Bank, Barcelona, Catalonia, Spain).

The entire study, including these three different main cohorts, followed ethical guidelines of the most recent revision of the Declaration of Helsinki (2013) [762]:

### **1.1 17<sup>th</sup> IHIW-Spanish Unrelated Healthy Control Reference Group**

This representative Spanish population cohort included N=282 de-identified genomic DNA samples from corresponding healthy unrelated individuals randomly selected from Spain in collaboration with the Spanish Working Group in Histocompatibility and Transplant Immunology (GETHIT) of the Spanish Society for Immunology (SEI). Collection of all de-identified genomic DNA samples consisted of 11 participating clinical histocompatibility and immunogenetics (H&I) laboratories that are located in 10 different regions in Spain (Santander, Salamanca, Madrid (which included 2 different participating clinical laboratories), Barcelona, Valencia, Murcia, Córdoba, Sevilla, Málaga and Gran Canaria) which provided a set of n=25-26 de-identified genomic DNA samples per institution (see **Figure M-1** and **Table M-1**). This HLA Spanish population study was approved by the Institutional Review Board (IRB) of the 17<sup>th</sup> International Histocompatibility and Immunogenetics Workshop (IHIW) as well as by the respective local research and ethics committee of each Spanish participating institution. De-identified genomic DNA samples were tested for HLA genotyping in parallel:

i) On one hand, all N=282 de-identified genomic DNA samples were genotyped by using a commercially available HLA genotyping method (MIA FORA NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA)) whereby 11 major classical HLA genes were typed using a high-resolution (according to version 3.25.0 (released in July 2016) IPD-IMGT/HLA database, available at the time of this study) long-range PCR amplicon-based next

generation sequencing approach and a short-read sequencing platform [187][763-766] at the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory.

ii) On the other hand, the 11 Spanish participating clinical laboratories performed HLA genotyping tests (with a variable range of allele resolution level and number of HLA genes tested) of their respective sets of n=25-26 de-identified genomic DNA samples by using other HLA molecular typing techniques (either using an in-house NGS platform or commercial/in-house SSO, SSP or SBT technologies) (see **Table M-1**).

The purpose of this dual and in parallel HLA genotype testing was to evaluate, at the end, the level of concordance (and, thus, the validity of the HLA genotyping results as well as discarding possible errors associated to sample-switching, allele dropout (for the HLA loci tested respectively) and contamination) between the HLA genotyping results obtained by these other high-/intermediate-resolution methods performed at the different participating Spanish HLA laboratory groups and the NGS-based high-resolution HLA genotyping results obtained from this replicated entire Spanish population cohort at the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory (see **RESULTS** section).

For posterior statistical analysis, particularly relative to estimation of genetic distances and construction of dendograms, several Spanish geographically related sub-groups of samples were considered in the scope of this study:

- Northern-Central Spain sub-group: including samples from Santander, Salamanca and Madrid.
- Southern-Spain sub-group: samples from Málaga, Córdoba, Sevilla and Gran Canaria.
- Eastern-Spain sub-group: samples from Barcelona, Valencia and Murcia.



**Figure M-1.** Map of the geographical location of Spain (Spanish territory colored in light yellow, where Spanish autonomous regions are delimited by black borders and, in turn, respective provinces are delimited by grey borders) which shows the location of the 11 participating Spanish local clinical laboratories (coded from 1 to 11) in the collection of samples (n=282 healthy unrelated individuals) for this study. In detail:

- [1] Immunology, Hospital Universitario Marqués de Valdecilla in Santander (n=25 samples);
- [2] Molecular Biology-Hematology, Hospital Clínico Universitario, in Salamanca (n=26 samples);
- [3] Immunogenetics and Histocompatibility, Instituto de Investigación Sanitaria Puerta de Hierro in Madrid (n=25 samples);
- [4] Histocompatibility, Centro de Transfusión de la Comunidad de Madrid in Madrid (n=26 samples);
- [5] Histocompatibility and Immunogenetics, Banc de Sang i Teixits in Barcelona (n=26 samples);
- [6] Histocompatibility, Centro de Transfusión de la Comunidad Valenciana in Valencia (n=26 samples);
- [7] Immunology, Hospital Clínico Universitario Virgen de la Arrixaca in Murcia (n=26 samples);
- [8] Immunology, Hospital Universitario Reina Sofía in Córdoba (n=26 samples);
- [9] Immunology, Hospital Universitario Virgen del Rocío in Sevilla (n=25 samples);
- [10] Histocompatibility, Centro de Transfusión de Málaga in Málaga (n=26 samples);
- [11] Immunology, Hospital Universitario de Gran Canaria Dr Negrín in Las Palmas de Gran Canaria (n=25 samples).

\*(Maps of this figure are a modified version from:

[https://en.wikipedia.org/wiki/2020\\_FIBA\\_Intercontinental\\_Cup#/media/File:España-Canarias-loc.svg](https://en.wikipedia.org/wiki/2020_FIBA_Intercontinental_Cup#/media/File:España-Canarias-loc.svg))

**Table M-1.** HLA molecular typing techniques used by 11 Spanish local participating clinical and histocompatibility laboratory institutions (de-identified and coded here **A** through **K**) for testing in parallel the respective institution's sample set out of this entire 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects). HLA genotyping results (data not shown) generated present a variable range of allele resolution level and number of HLA genes tested (see **RESULTS** section).

| HLA molecular typing technique                         | Spanish local participating laboratories |
|--|--|
| In-house Next Generation Sequencing (NGS) <sup>1</sup> | E and F                                  |
| Sequence-Specific Oligonucleotide (SSO) <sup>2,3</sup> | A, B, C, D, G, H, I, J and K             |
| Sequence-Specific Primer (SSP) <sup>3</sup>            | C  |
| Sequence Based Typing (SBT) <sup>2,3</sup>             | C and D                                  |

**Notes:**

1. In-house NGS-based HLA typing method used by laboratories “E” and “F”.
2. Laboratory “D” used a commercially available SSO method for HLA typing in 7 samples (out of 26) and rest 19 samples (out of 26) were HLA typed by SBT using local reagents for *HLA-A*, *-B*, *-C* and *-DRB1* loci (In-house SBT method).
3. Laboratory “C” used a commercially available SSO method for HLA typing in 7 samples (out of 26) and rest 19 samples (out of 26) were HLA typed by either commercially available SSP method or by SBT using local reagents for *HLA-A*, *-B*, *-C* and *-DRB1* loci (In-house SBT method).

## 1.2 Spanish Multiple Sclerosis Cohort

De-identified genomic DNA samples were collected from a Spanish cohort of N=238 multiple sclerosis (MS) patients who were recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain. This group of N=238 MS patients [age average=32.5 years; and in turn, phenotypically divided in MS patients male gender sub-group (age average=32.6; n=67(28%)), and MS patients female gender sub-group (age average=32.4; n=171(72%))] was, in turn, clinically divided in three sub-groups: n=216 relapsing-remitting (RR); n=19 secondary progressive (SP); and n=3 primary progressive (PP). Nevertheless, for the purpose of this case-control study of the present thesis work, and the respective statistical analyses that were carried out, the entire MS patients group was the only one disease group considered, especially since



the MS RR sub-group and the MS female gender sub-group were very predominant in comparison to the other respective MS clinical/phenotype sub-groups. In addition, there was not phenotypic data (e.g. age or gender) available from healthy control reference groups used in the present study. All MS subjects met established McDonald diagnostic criteria [767] and were ethnically matched (defined as European-Spanish-Mediterranean (Caucasoid) ancestry) in regards to the 17<sup>th</sup> IHIW-Spanish Unrelated Healthy Control Group as well as the Northeast Spain (Catalan) Healthy Control Reference Group. This study was approved by the Institutional Review Board (IRB) (Research and Ethics Committee) of the Hospital Clínic, Barcelona, Spain. All genomic DNA samples from this Spanish MS cohort were tested using MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA)) for characterizing the full-length and/or the most possible extended sequence of *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci, and including respective HLA genotyping software analysis package (MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0, via VNC viewer, with reference to IPD-IMGT/HLA database release 3.25.0. available at the time of the current study (Immucor, Inc. Norcross, GA, USA)).

### **1.3 Northeast Spain (Catalan) Healthy Control Reference Group**

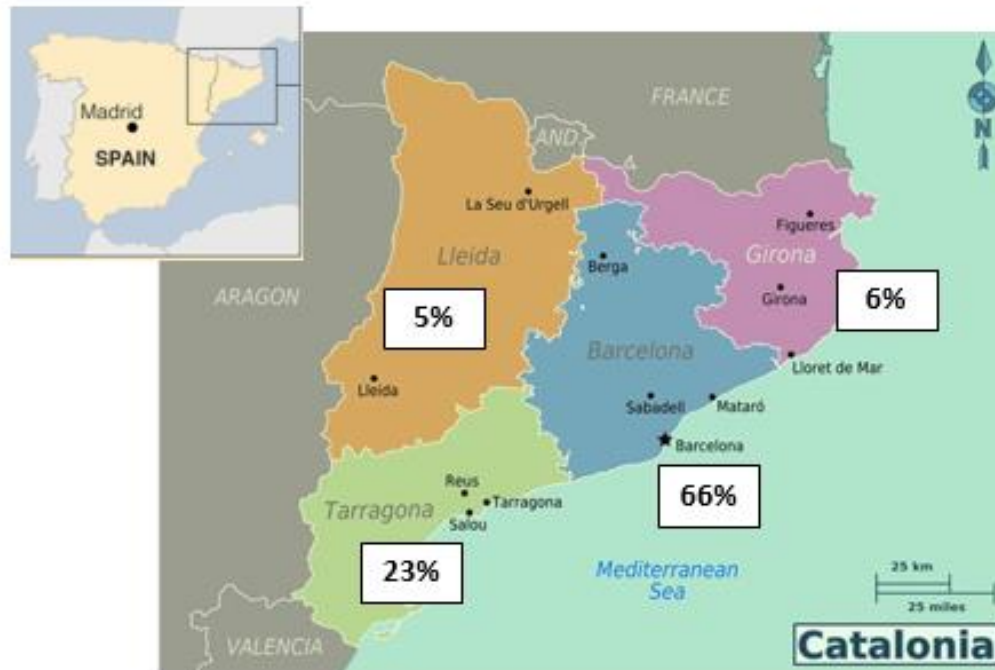
An additional HLA genotype dataset specifically representative of the Catalan population cohort from Northeast Spain (region of Catalonia) included N=196 de-identified healthy unrelated individuals randomly selected and originally recruited at the Histocompatibility and Immunogenetics Laboratory, Blood and Tissue Bank, Barcelona, Catalonia, Spain. This Northeast Spain (Catalan) healthy control reference group comprised HLA genotyping data only available in this case for: *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci, established at the 2-field allele resolution level and according to version 3.35.0 (released in January 2019) IPD-IMGT/HLA database. All subjects that were part of this Catalan healthy

control reference group ethnically matched (as previously indicated) the respective MS cohort of this study and all were individuals exclusively from the region of Catalonia, distributed in the respective Catalan provinces as follows (shown in percentages): 66% from Barcelona, 23% from Tarragona, 5% from Lleida and 6% from Girona (see **Figure M-2**). This Northeast Spain (Catalan) healthy control reference group served to verify the findings of HLA-disease associations from this initial case-control study (17th IHIW-Spanish Unrelated Healthy Control Reference Group versus Spanish Multiple Sclerosis Cohort) and to evaluate the effect of plausible regional HLA genetic variation within mainland Spain. This study was approved by the Institutional Review Board (IRB) (Research and Ethics Committees) of both the Hospital Clínic, Barcelona, Catalonia, Spain and the Histocompatibility and Immunogenetics Laboratory, Blood and Tissue Bank, Barcelona, Catalonia, Spain.

The corresponding generation of genotyping results of this HLA genotype dataset, as a healthy control reference group from Northeast Spain (region of Catalonia), was carried out originally at the Histocompatibility and Immunogenetics Laboratory, Blood and Tissue Bank, Barcelona, Catalonia, Spain. Briefly:

Blood samples (4 mL) were collected and sent to the Histocompatibility and Immunogenetics Laboratory, Blood and Tissue Bank (Banc de Sang e Teixits), Barcelona, Catalonia, Spain. The QIASymphony DNA kit (QIAGEN, Hilden, Germany) was used for genomic DNA extraction following the instructions provided by the manufacturer. Molecular HLA typing was performed using an in-house next-generation sequencing (NGS) technique in combination with the commercially available NGSgo kit (GenDx, Utrecht, Netherlands). All samples were genotyped for HLA loci, namely *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*, at a variable range of allele resolution level (between 2- to 4-field) according to version 3.35.0 (released in January 2019) IPD-IMGT/HLA database. The amplification of HLA genes was

performed using an in-house long-range PCR protocol in a single-tube tube per sample. In this NGS in-house approach, the whole gene sequence is amplified for the HLA class I loci: *HLA-A*, *-B* and *-C*. Whereas, since HLA class II genes have extremely large introns, only exons 2 and 3 of *HLA-DPB1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* were respectively amplified by multiplex PCR. The HLA amplicons and size were verified on a 1 % agarose gel. Library preparation was performed by enzymatic fragmentation of PCR amplicons and double indexing using the NGSgo kit (GenDx, Utrecht, Netherlands) according to the manufacturer's instructions. The indexed libraries were pooled, denaturized, and diluted to a final concentration of 4 nM. The pooled DNA library was sequenced on the MiSeq system (Illumina, San Diego, CA, USA). Results were analyzed with NGSEngine version 2.13.0 (GenDx, Utrecht, Netherlands) according to the manufacturer's recommendations.



**Figure M-2.** Map of the geographical location of the region of Catalonia (Spain), which shows the location of the 4 participating sub-regions (or provinces) in the collection of HLA genotyping data (n=196 healthy unrelated individuals) for this study. In detail, all subjects that were part of this Catalan healthy control reference group were individuals exclusively from the region of Catalonia, distributed in the respective Catalan provinces as follows (shown in percentages): 66% from Barcelona (Blue colored), 23% from Tarragona (Green colored), 5% from Lleida (Orange colored) and 6% from Girona (Pink colored). Respective images are obtained and adapted from:

<http://www.orangesmile.com/travelguide/catalonia/high-resolution-maps.htm>

<https://www.bbc.com/news/world-europe-20345071>

## **2. HLA CLASS I AND II NGS GENOTYPING BY A LONG-RANGE SHOT-GUN BASED SEQUENCING STRATEGY USING A SHORT-READ SEQUENCING PLATFORM**

At the time of the preparation of this thesis work (between years 2014-2019) and when this NGS HLA Spanish health and disease population cohorts study was carried out at the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory for testing both 17th IHIW-Spanish Unrelated Healthy Control Reference Group and the Spanish Multiple Sclerosis Cohort. We considered, as one of the most suitable, advanced and innovative NGS-based HLA genotyping protocol to be used, the following:

i) A kit that was commercially available and clinically validated and implemented (MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA)) for characterizing the full-length and/or spanning to the most possible extended sequence of *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci. This method also includes respective HLA genotyping software analysis package: MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0, via VNC viewer, with reference to IPD-IMGT/HLA database release 3.25.0. available at the time of the current study (Immucor, Inc. Norcross, GA, USA)) [187][763-766].

ii) Whose manufacturer's protocol was already semi-automated for liquid-handling processes, thus being very compatible for large-scale studies:

-Biomek NX<sup>P</sup> Automated Workstation, Span-8 and Gripper (A31840) (Beckman Coulter, Brea, CA, USA) for pre-PCR procedures related to initial long-range PCR reactions set-up. This automation system allows to prepare the respective nine PCR master mix reaction plates from one DNA plate of 96 samples (see more details below on **2.2 Long-range PCR of HLA Genes**).

-Biomek FX<sup>P</sup> Dual Arm System, Multichannel Pipettor and Span-8 Pipettors Workstation (A31844) (Beckman Coulter, Brea, CA, USA) for post-PCR procedures related to DNA library construction set-up. This automation system allows to run automated protocols of the **2.3 Quantification, Balancing and Pooling of PCR products** steps per 96 sample DNA set. Also, this automation system allows to run automated protocols of the **2.4 Construction of DNA Sequencing Library** steps simultaneously for two different 96 sample DNA libraries (see more details below on **2.4 Construction of DNA Sequencing Library**).

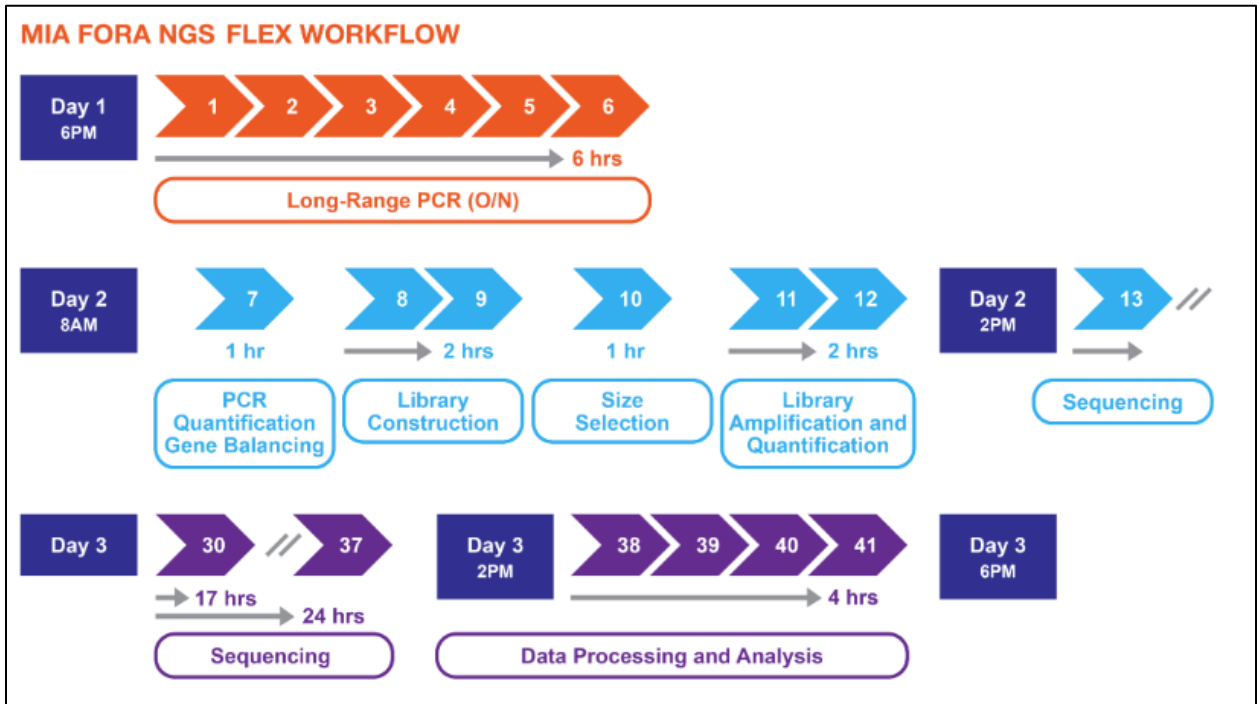
iii) And, most importantly, a protocol that combined the advantage of long-range PCR amplification (thus, defining this targeting strategy for maximum coverage of HLA genes) and

the power of high-throughput (high multiplexing capacity) and paired-end mode in a short-read sequencing platform (using both MiniSeq and 500/550 NextSeq sequencers (Illumina, San Diego, CA, USA)) with the maximum possible depth of sequencing coverage and minimum possible base-calling error rate (Q30) [152][184]. Thus, the linkage across ~400 bases from paired-end reads, together with polymorphic sites in intron regions provide important phasing information that is very useful to resolve combination ambiguities. As previously mentioned, this type of NGS-based HLA genotyping method substantially enhances the allele resolution and dramatically improves the combination resolution in comparison to the conventional SBT method [35][76].

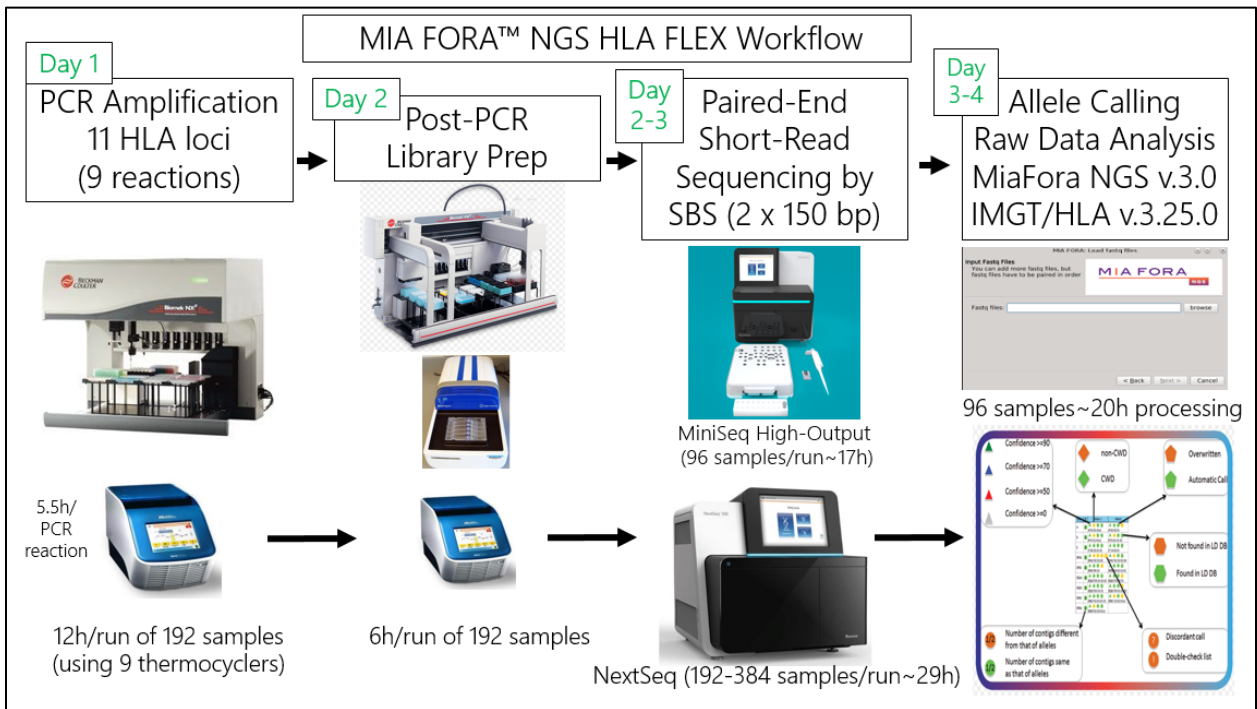
iv) Moreover, other research groups and clinical institutions have also successfully applied this same commercial HLA genotyping protocol and software analysis for large-scale HLA clinical and research population studies (see the following articles in the **BIBLIOGRAPHY** section):

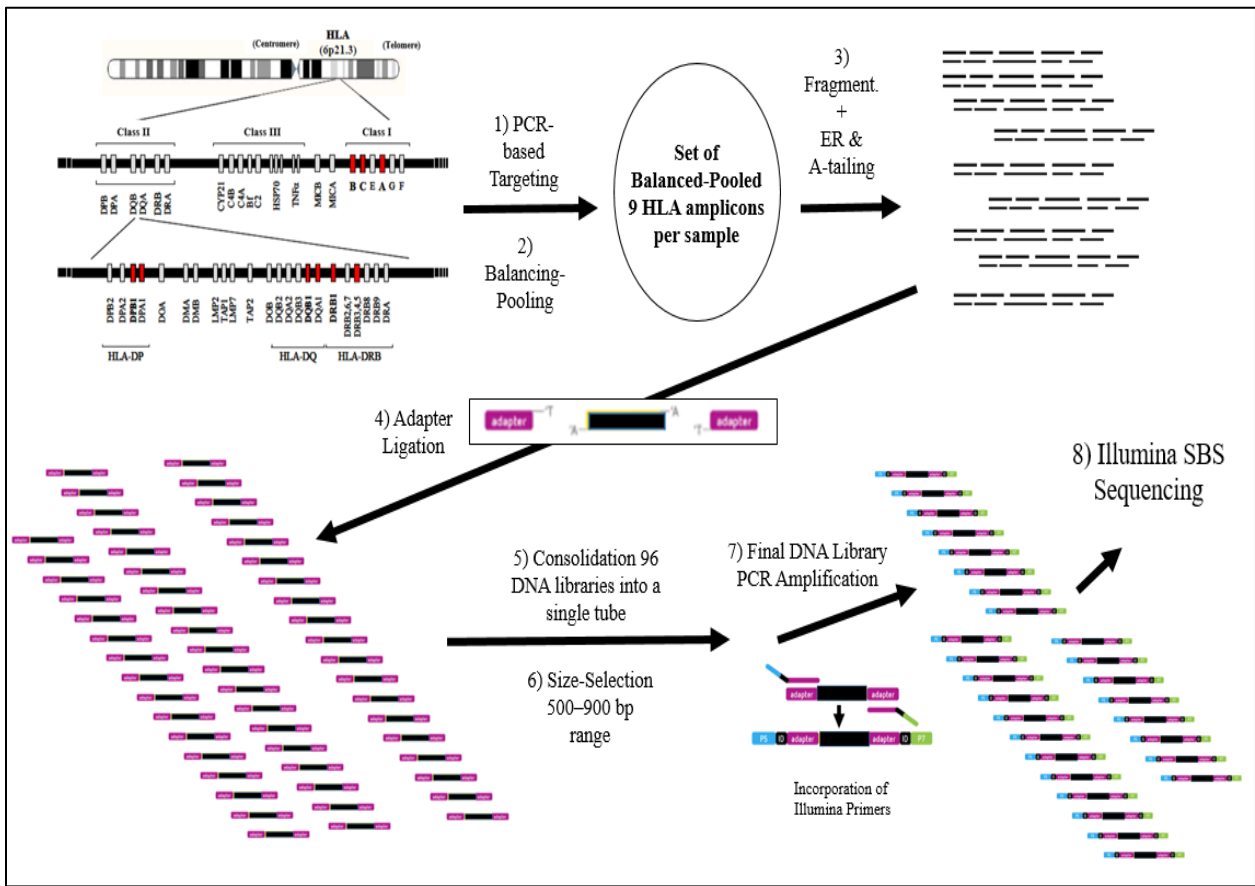
[267][268][272-274][284][286][291][293][300][308][328][356][433][473][474][476-478][482].

Therefore, MIA FORA™ NGS FLEX HLA typing protocol (see **Figure M-3** and **Figure M-4**) provides the full-length and/or extended HLA sequence and genotyping results of these clinically relevant eleven HLA loci and respective phasing information per gen to achieve high-resolution (3- up to the 4-field) HLA typing with minimum ambiguity (since the extensive coverage across HLA loci minimized the possibility for ambiguous genotypes). Furthermore, using this particular MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA), genomic DNA samples are tested here in sets of: 94 samples of interest for the given study plus NTC (negative control sample) and PTC (positive control sample) per run.



**Figure M-3.** Scheme of MIA FORA™ NGS HLA FLEX Genotyping Workflow. Courtesy: <https://www.immucor.com/en-us/Products/Pages/MIA-FORA-NGS.aspx>.





**Figure M-4.** Scheme of MIA FORA™ NGS HLA FLEX Genotyping Instrument Workflow (**Top Image**) and scheme of molecular basis of this NGS-based protocol (**Bottom Image**). Respective images are obtained and adapted from:

<https://www.thermofisher.com/order/catalog/product/4375786#/4375786>

<https://www.beckman.com/liquid-handlers/biomek-nxp>

<https://www.beckman.com/liquid-handlers/biomek-fxp>

<https://www.illumina.com/systems/sequencing-platforms/miniseq.html>

<https://sagescience.com/products/pippin-prep/>

<https://www.illumina.com/systems/sequencing-platforms/nextseq.html>

<http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf>

[531]

<https://www.gendx.com/products/ngsgo-illumina>

On the next pages, it is described a summarized version of this NGS-based HLA genotyping protocol and the subsequent HLA allele base-calling and genotyping software analysis showing



the main information and features of each step as it is fully explained in corresponding descriptions of patents [763][764], publications [187], commercial guides and protocols [765][766] (available via online), and according to International Histocompatibility and Immunogenetics Clinical and Research Laboratory Standards and Policies regulated by accreditation organizations (e.g. ASHI and EFI) [768][769]:

### **2.1 Specimen Collection and Preparation**

Human genomic double-stranded DNA (hg-dsDNA) was purified from whole blood and/or buffy coats at the respective participating Spanish clinical institutions (for both healthy control and MS disease cohorts), using any validated method (e.g. QIAamp DNA Blood Mini Kit (QIAGEN, Hilden, Germany)) that met the criteria shown here below. As a note, DNA extracted from blood and preserved in EDTA is compatible with this NGS-based HLA typing assay, whereas DNA extracted from blood preserved in heparin cannot be used in this assay. The specific criteria is the following:

- The isolated DNA should be in 10 mM Tris-HCl, pH 8.0-9.0, or in nuclease free water. If a chelating agent such as EDTA is present, the final concentration of the chelating agent should not exceed 0.5 mM.
- Final DNA concentration should be from 5 to 15 ng/μL in a volume of 115 uL per sample, located separately on each well of the DNA 96-well plate.
- Absorbance measurements of the DNA sample at 260 and 280 nm should give a ratio of 1.65 to 2.0.
- DNA can be used immediately after isolation or stored at –20°C for up to one year. Repeated freeze/thawing should be avoided since this can result in DNA degradation.

- At least 50% of genomic DNA samples must have fragments greater than 10 kb for successful long-range PCR amplification.

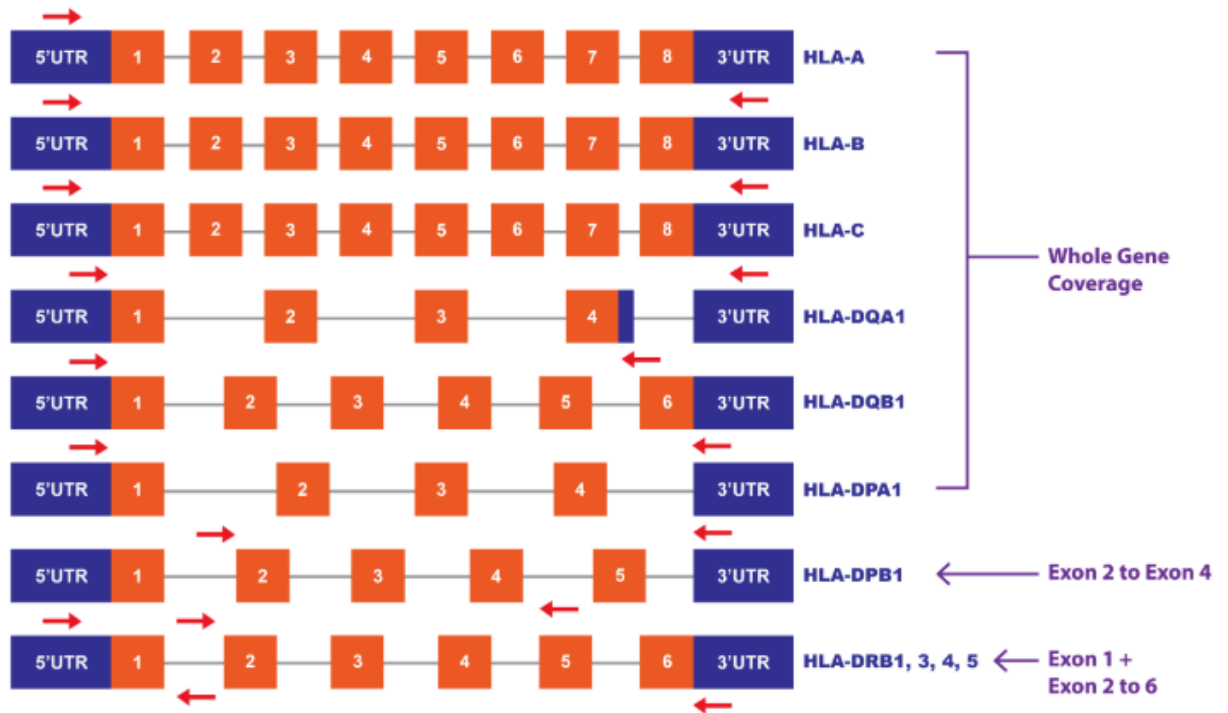
DNA 96-well plates are manually prepared, including 94 samples to be tested at a time and two controls samples: negative/blank control (NTC; made with nuclease free water on A01 well) and positive control (PTC; located in H12 well and using a DNA reference sample whose HLA genotype is already well-known and -documented). Thus, 94 samples are tested per NGS-based HLA genotyping run using this MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA).

## **2.2 Long-range PCR of HLA Genes**

The first stage of the protocol is the long-range PCR amplification for targeting the HLA genes of interest, which takes place in a Pre-PCR room to avoid any contamination (e.g. by post-PCR products found in the environment) according to International Histocompatibility and Immunogenetics Clinical and Research Laboratory Standards and Policies regulated by accreditation organizations (e.g. ASHI and EFI) [768][769]. All samples were genotyped for 11 HLA loci, namely *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*, *-DRB3*, *-DRB4* and *-DRB5*. Here, each HLA gene is amplified as one large segment in its entirety, with the exception of *HLA-DRB* loci (*HLA-DRB1* and *-DRB3/4/5* genes), due to their large size, which are co-amplified in two separated PCR amplification reactions generating one PCR product in each case denominated, respectively: DRB-S (or “DRB1”; covering exon 1 and adjacent regions at both ends); and DRB-L (or “DRB2”; covering segment from the end of intron 1 to the very first bases of 3’UTR). Thus, nine different all-in-one PCR master mixes (PCR MMs; containing a cocktail of primers unique for each gene, dNTPs, PCR buffers, and DNA polymerase enzyme) are used, denominated: *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRBS* and *-DRBL*.

In detail, the specific PCR amplification coverage (i.e. amplicon extension) in each case is (see also scheme on **Figure M-5**):

- HLA class I (*HLA-A*, *-B* and *-C* loci): for each gen, respective single amplicon encompasses from more than 200 base pairs (bp) of the 5'UTR to 100-1100 base pairs (bp) of the 3'UTR, including all exons and introns.
- HLA class II (in each tested locus):
  - HLA-DPA1*: amplicon includes all exons and introns, at least 45 bp of the 5' UTR and 25-190 bp of the 3' UTR.
  - HLA-DPB1*: coverage includes only key regions: exons 2–4 and introns 2–3.
  - HLA-DQA1*: this coverage includes all exons and introns, at least 45 bp of the 5' UTR and 25-190 bp of the 3' UTR.
  - HLA-DQB1*: amplicon includes key regions: exons 1–5 and introns 1–4.
  - HLA-DRB1*: this coverage includes all exons 1-6, introns 2–5, at least 440 bp of the 5' UTR, 12 bp of the 3' UTR, 275 bp of intron 1 adjacent to exon 1, and 210 bp of intron 1 adjacent to exon 2.
  - HLA-DRB3* and *HLA-DRB4*: respective amplicon includes all exons, introns 2–5, at least 440 bp of the 5' UTR, 12 bp of the 3' UTR, 275 bp of intron 1 adjacent to exon 1, and 210 bp of intron 1 adjacent to exon 2.
  - HLA-DRB5*: this coverage includes exons 2–6, introns 2–5, and 260 bp of intron 1 adjacent to exon 2.



**Figure M-5.** Coverage of HLA loci using MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA). Each following segment represents a different targeted HLA locus/loci. Purple boxes at each end represent 5' and 3' untranslated regions (UTR); orange boxes represent exons and black thin lines represent introns in respective HLA class I and class II loci; red arrows represent positions of primers used for long-range PCR. *HLA-DRB1, 3, 4, 5* loci are amplified together in two separate reactions (DRB-S and DRB-L). Courtesy: <https://www.immucor.com/en-us/Products/Pages/MIA-FORA-NGS.aspx>.

Therefore, nine individual long-range PCR amplifications are performed per sample. Each PCR amplification (total volume = 25 uL) contains 10 uL (containing 50-150 ng) of respective genomic dsDNA sample and 15 uL of respective solution of PCR master mix (PCR MM) (HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRBS and -DRBL). Thus, for each run of 94 samples plus NTC and PTC to be tested, there is an automated preparation of these nine different 96-well PCR plates (one plate per PCR master mix (PCR MM)) according to the manufacturer's semi-automated protocol [765].

PCR amplification reactions are performed using Veriti Thermal Cyclers (Veriti™ 96-Well Thermal Cycler, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA). The

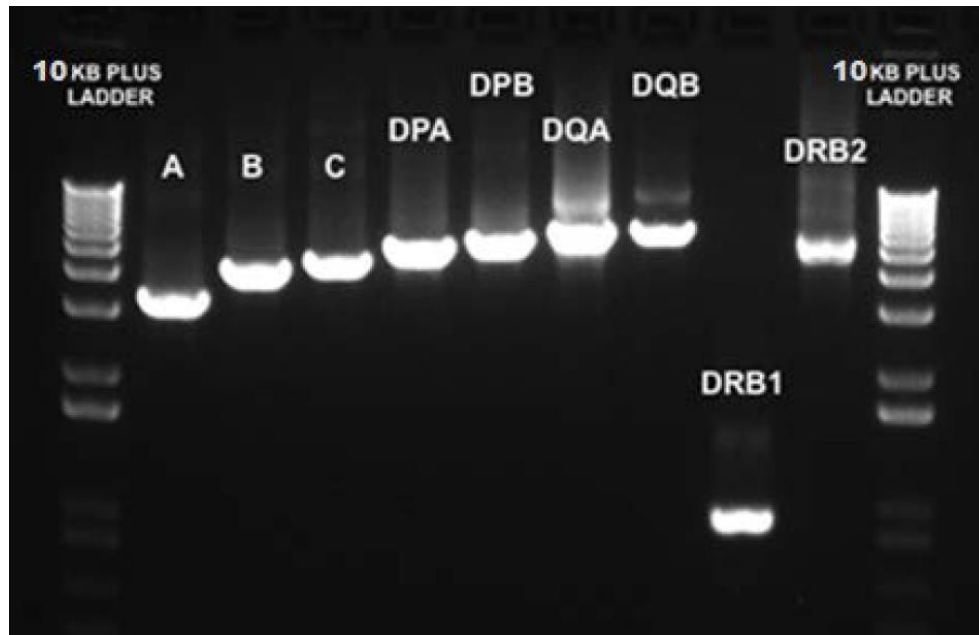
thermal cycling program (total duration t~6h) parameters and conditions for long-range PCR reactions (HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRBS and -DRBL) are the same for all eleven HLA genes tested, as follows (see **Table M-2**):

|              | Cycles (15)  |           |             | Cycles (20)  |           |             |                 |       |              |
|--------------|--------------|-----------|-------------|--------------|-----------|-------------|-----------------|-------|--------------|
| Initial Hold | Denaturation | Anneal    | Extension   | Denaturation | Anneal    | Extension   | Final Extension | Hold  | Total Volume |
| 94°C /30s    | 94°C /75s    | 60°C /30s | 66°C /7:30m | 94°C /30s    | 60°C /30s | 66°C /7:30m | 66°C /10m       | 4°C/∞ | 25 µl        |

### Amplification program for Long Range PCR: MIA\_FORA\_HLA\_PCR

**Table M-2.** Thermal cycling program (total duration t~6h) parameters and conditions for long-range PCR amplification reactions (HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRBS and -DRBL) for targeting the HLA genes (*HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4* and *-DRB5*) of interest. Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf)

Upon completion of the long-range PCR amplification reactions, a gel electrophoresis (0.8-1.0% w/v agarose gel; using Ethidium Bromide or GelRed as the fluorescent nucleic acid gel stains; 80-150 V until the dye line is approximately 75-80% of the way down the gel; a typical run time is about 1-1.5 hours, depending on the gel concentration) can be performed for the analysis of amplification products by size (see **Figure M-6** as an example).



**Figure M-6.** Example of gel electrophoresis showing robust and mostly specific amplification of targeted HLA genes (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*, *-DRB3*, *-DRB4* and *-DRB5*) of interest. Observed bands of PCR products [“A”(3.23 kb; *HLA-A* locus), “B”(4.12 kb; *HLA-B* locus), “C”(4.37 kb; *HLA-C* locus), “DPA”(4.98 kb; *HLA-DPA1* locus), “DPB1”(5.24 kb; *HLA-DPB1* locus), “DQA”(5.83 kb; *HLA-DQA1* locus), “DQB”(6.31 kb; *HLA-DQB1* locus), “DRB1” or “DRB-S”(co-amplicons of 0.94 kb; *HLA-DRB1/3/4* loci) and “DRB2” or “DRB-L”(co-amplicons of 5.56 kb; *HLA-DRB1/3/4/5* loci)] that correspond to the respective nine different long-range PCR amplification reactions according to the nine different all-in-one PCR master mixes (MMs) used. Indicated between parentheses are the average amplicon/co-amplicons lengths (in kilobases (kb)) and respective targeted HLA loci. Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525 MIA FORA NGS FLEX HLA Typing Package Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525%20MIA%20FORA%20NGS%20FLEX%20HLA%20Typing%20Package%20Insert-RUO-A.pdf)

### 2.3 Quantification, Balancing and Pooling of PCR products

After long-range PCR amplification step is performed (and from now on in the protocol, working only at the Post-PCR room designated areas), concentration of each HLA gene PCR product (corresponding to the different targeted HLA genes per sample) is quantified and then, PCR products per sample are balanced and equimolarly (equimolar amounts of the amplified gene products (amplicons)) pooled to ensure equal representation of each gene per tested sample. Since PCR yield is typically variable among different reactions, this set of steps allow to balance the final amounts of certain targeted genes with a higher PCR yield and those others with a lower PCR yield. Thus, importantly, this contributes to have, ultimately, a sufficiently

even generation of sequencing reads per HLA gen/allele per tested sample (i.e. obtaining a high evenness of coverage), considering also a high-throughput setting in which the simultaneous testing of 94 samples plus NTC and PTC per run is also evenly optimized (since it is maximized the number of PCR products that can be multiplexed per analytical sample in the final deep-sequencing process). Briefly, in detail:

- Quantification step: concentration of each PCR product is quantified by performing the PicoGreen® dsDNA quantification assay (Invitrogen/Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's semi-automated protocol and following respective particular MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA) protocol guidelines. A Victor X3 plate reader (Perkin Elmer, Waltham, MA, USA) is used to indirectly measure dsDNA concentration of PCR products (using only a minimal aliquot of the total original volume) per sample by obtaining relative fluorescent dye intensity values. These fluorescent dye intensity values shown in Victor X3 output files are then firstly converted into dsDNA concentration values within the MIA FORA™ NGS software system via the so-called Sirona Quant tool. In turn, based on this PCR products' length and calculated dsDNA concentration data, on the number of targeted HLA genes in each PCR reaction, on the number of samples to be tested and on balancing default values (recommended pmol value is 0.0035), this Sirona Quant tool generates the final instructions file for the Biomek FX<sup>P</sup> Dual Arm System, Multichannel Pipettor and Span-8 Pipettors Workstation (A31844) (Beckman Coulter, Brea, CA, USA) to be used for balancing and pooling steps.

- Balancing and Pooling steps: via an automated program, PCR products are balanced (i.e. all nine HLA PCR products per sample are made equimolar either using pre-dilutions (with 10mM Tris-HCl, pH 8.0 solution) only for certain cases; and/or pooling certain specific

volumes per HLA gen; and also by defining a final common total volume of pooling (final volume per sample = 55 uL)) and pooled in optimal equimolar amounts. Pooled volumes of each sample (where each pooled volume per sample includes the respective equimolar mixture of the nine amplicons) are then purified (also termed as bead clean-up step) using Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) via respective manufacturer's semi-automated protocol. In which always, after incubating with magnetic beads, there are several washes with newly prepared 80% EtOH and a final elution using 10mM Tris-HCl, pH 8.0 solution. Thus, starting from nine original PCR plates, after the balancing and pooling steps there is a single 96-well plate (one well per sample's balanced and pooled PCR products) prepared and ready for the next set of steps: construction of DNA sequencing library.

#### **2.4 Construction of DNA Sequencing Library**

In preparation for the final stage of sequencing for the targeted (via PCR amplification) HLA genes of interest per tested sample, the construction of DNA sequencing library consists on the following main steps:

- Equimolar mixtures of balanced and pooled HLA gene PCR products per sample are, subsequently (although this occurs during the same thermal setup reaction and cycling program), enzymatically fragmented and end-repaired (including blunt-ending and dA-tailing steps). Then, each processed sample's PCR products are purified.
- Purified and processed PCR products belonging to each sample solution are enzymatically ligated with unique index-adapters that include the so-called barcodes. Each barcode consists on a target specific sequence identifier, thus having one barcode per sample. Later in this protocol, at the end of the sequencing stage and during the posterior primary analysis of the raw



sequencing data (including steps such as demultiplexing and adapter and quality trimming of reads) these different barcodes enable to unequivocally identify and assign the source of the given genomic DNA sample linked to a particular barcode.

- After the index-adapters have been ligated, the 96 (94 testing samples plus NTC and PTC) samples are consolidated into one tube, designated as the pre-final sequencing library, whose all initial fragments still need to be (prior to sequencing):

- Size-selected (since a length of ~400-600 bp is the most suitable DNA library fragment size for Illumina sequencer systems (Illumina, Inc., San Diego, CA, USA)).

- And, then, specifically PCR amplified for selectively enriching those size-selected DNA library fragments that are also ligated to Illumina's flow cell-specific sequencer adapters (Illumina, San Diego, CA, USA). Thus, P5 and P7 adapter sequences, which are needed for binding to the Illumina flow cell, are incorporated to size-selected DNA library fragments in this step.

In detail:

#### **2.4.1 Primary DNA Library Preparation**

Pooled PCR products are enzymatically cleaved into fragments and, subsequently during the same thermal setup reaction and cycling program, end-repaired (including blunt-ending of both ends and A-tailing at the 3'-end so that, later, the respective index-adapter that has a 5'-T-overhang can be ligated). Thus, upon manual preparation of Primary Buffer Mix (Primary Enzyme Mix + Primary Buffer Mix) and following respective manufacturer's semi-automated protocol, 11 uL of Primary Buffer Mix are combined with only 14 uL (out of the original 55 uL solution) of each balanced and pooled DNA PCR HLA gene products sample respectively in a new 96-well plate denominated "Primary Prep Plate". Then, using a Veriti Thermal

Cycler (Veriti™ 96-Well Thermal Cycler, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA), the thermal cycling program (total duration t~40min) parameters and conditions for Primary DNA Library Preparation is as follows (see **Table M-3**):

| <b>MIA FORA FLEX PrimaryPrep</b> | <b>Total Volume</b> |
|----------------------------------|---------------------|
| 20°C for 10 min                  | 25 µL               |
| 75°C for 30 min                  |                     |
| 4°C ∞ hold                       |                     |

**Table M-3.** Thermal cycling program (total duration t~40min) parameters and conditions for Primary DNA Library Preparation reaction for enzymatic fragmentation and end-repaired (including blunt-ending and dA-tailing steps) of balanced and pooled HLA gene PCR products per sample. Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf)

At the conclusion of this thermal reaction, there is an immediate Post-Primary Preparation Bead Clean-Up step (as previously described) using Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) via respective manufacturer’s semi-automated protocol (final eluted volume per sample = 23 uL). After that, user directly proceeds to the index-adapter ligation step.

### **2.4.2 Index-Adapter Ligation**

Index-adapters include unique barcode sequences that are used to identify the different samples during posterior sequence analysis. The MIA FORA™ NGS HLA FLEX Typing 11 Kit (RUO) 96 Tests (Immucor, Inc. Norcross, GA, USA) contains one index-adapter plate with 96 barcodes (2.5 uM original concentration of barcode per well in a volume of 10 uL). The index-adapters (which present a 5’-T-overhang) are ligated to the A-tailed 3’-end samples’ fragments. At this step, the user must record the sample layout including the sample name and the corresponding position of the index-adapter in the plate (as well as specific the Index-Adapter 96 barcoding Plate # used). The sample layout is used later, after sequencing, for data analysis by the MIA

FORA™ NGS software system. Therefore, upon manual preparation of Index-Adapter Ligation Master Mix (Ligase Enzyme Solution + Ligase Buffer Solution) and following respective manufacturer’s semi-automated protocol: 26 uL of Index-Adapter Ligation Master Mix are combined with the original 23 uL of each balanced\_pooled\_cleaned\_fragmented\_end-repaired(A-tailed)\_cleaned DNA PCR HLA gene products sample plus 8 uL of the respective index-adapter respectively, defining thus the “Primary Prep-Index-Adapter Ligated Plate” (final volume per sample = 57 uL). Then, using a Veriti Thermal Cycler (Veriti™ 96-Well Thermal Cycler, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA), the thermal cycling program (total duration t~40min) parameters and conditions for Index-Adapter Ligation is as follows (see **Table M-4**):

| MIA FORA Ligation | Total Volume |
|-------------------|--------------|
| 25°C for 30 min   | 57 µL        |
| 65°C for 10 min   |              |
| 4°C ∞ hold        |              |

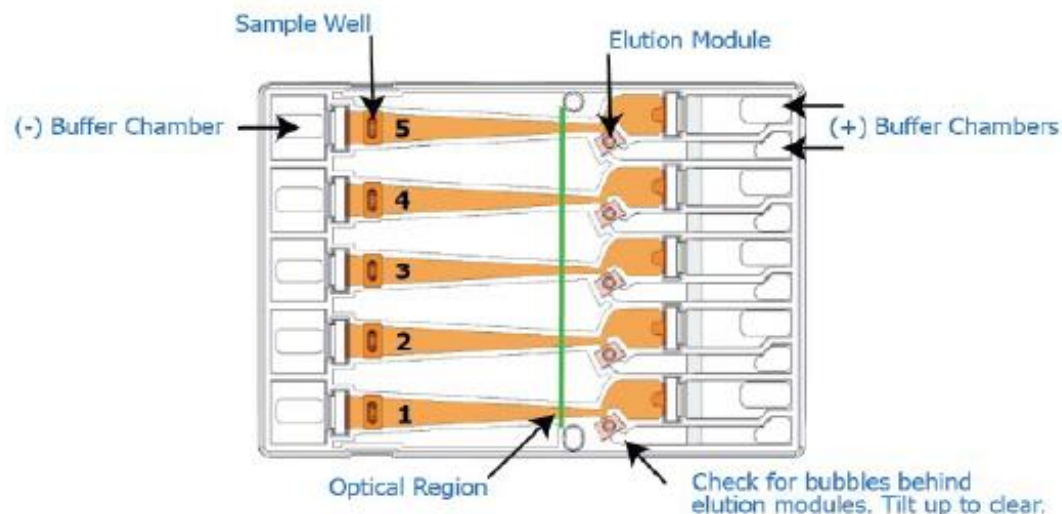
**Table M-4.** Thermal cycling program (total duration t~40min) parameters and conditions for Index-Adapter Ligation reaction. Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf)

### 2.4.3 Consolidation of Adapter Ligated Products

Adapter-ligated (balanced\_pooled\_cleaned\_fragmented\_end-repaired(A-tailed)\_cleaned DNA PCR HLA gene products sample) samples are pooled (adding 7 uL per adapter-ligated sample, making final consolidated volume of 96 samples up to = 672 uL) into a single microcentrifuge tube. Then, consolidated 96 samples solution is cleaned with Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) via respective manufacturer’s semi-automated protocol. Thus, after this Post-Adapter Ligation Bead Clean-Up step, final eluted consolidated volume of 96 samples = 60 uL.

### 2.4.4 DNA Library Size Selection

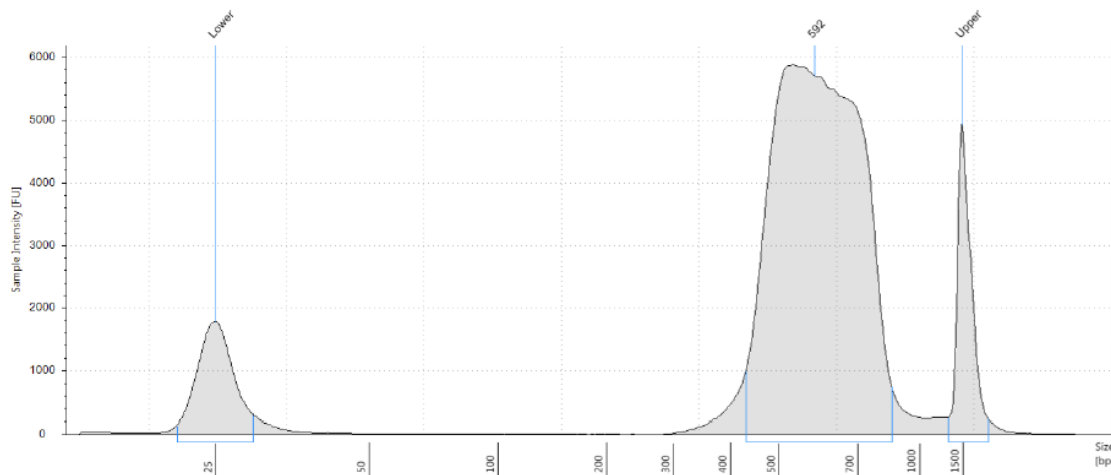
The consolidated and cleaned 96 adapter-ligated (previously balanced\_pooled\_cleaned\_fragmented\_end-repaired(A-tailed)\_cleaned DNA PCR HLA gene products sample) samples (from now also referred as DNA library) are size selected within the 500–900 bp range using a 1.5% w/v agarose gel cassette (DF Marker R2) on a PippinPrep instrument (Sage Science, Beverly, MA, USA), in a run time of size-selection of  $t \sim 45$  min, following respective manufacturer's manual protocol (see **Figure M-7**). Thus, this step allows the collection (in which final eluted volume of consolidated\_cleaned\_size-selected 96 adapter-ligated samples DNA library = 40  $\mu$ L) of specific and uniformed DNA library fragments' range size and, consequently, optimizing the posterior sequencing step in the Illumina sequencer systems (Illumina, Inc., San Diego, CA, USA).



**Figure M-7.** Schematic of PippinPrep Gel Cassette Instrument. Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf)

Then, size-selected DNA library concentration is estimated and its fragment distribution is assessed using the Agilent DNA 1000 kit (5067-1505) or the Agilent DNA High Sensitivity

kit on the 4200 TapeStation system and the D1000 ScreenTape assay (Agilent Technologies, Santa Clara, CA, USA) following respective manufacturer's manual protocol (see **Figure M-8**).



**Figure M-8.** Example of Electropherogram pattern of size-selected DNA library, showing maximum peak size (~592 bp of average) between 500 and 700 bp, separated with the D1000 ScreenTape assay on the 4200 TapeStation system.

### 2.4.5 Amplification of Size-Selected DNA Library

By following respective manufacturer's manual protocol, an aliquot (5 uL) of the size-selected 96 samples DNA library eluate from the PippinPrep is amplified, in a PCR cocktail reaction (of 12 cycles of amplification; t~30min) setup (see **Table M-5**): in combination with 25 uL of specific PCR Enzyme/Buffer Mix plus 2 uL of amplification pair of Illumina primers Fw/Rev and also adding 18 uL of nuclease-free water. These Illumina primers (P5 and P7 adapter sequences (Illumina, Inc., San Diego, CA, USA)) contain the sequences necessary for binding to the Illumina flow cell and, consequently, for optimal cluster generation.

| 1 Cycle             | 12 Cycles           |               |                  | 1 Cycle          |             | Total Volume |
|---------------------|---------------------|---------------|------------------|------------------|-------------|--------------|
| <b>Initial hold</b> | <b>Denaturation</b> | <b>Anneal</b> | <b>Extension</b> | <b>Final Ext</b> | <b>Hold</b> | 50 µL        |
| 98°C /30s           | 98°C /15s           | 65°C /30s     | 72°C /30s        | 72°C /5m         | 4°C /∞      |              |

### MIA FORA\_Library\_PCR

**Table M-5.** Thermal cycling program (total duration t~30min) parameters and conditions for PCR amplification reaction with Illumina primers of this size-selected 96 samples DNA library, using a Veriti Thermal Cycler (Veriti™ 96-Well Thermal Cycler, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA). Courtesy: [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf)

## 2.5 Preparation of Final DNA Library for Sequencing

These post-amplification steps should be performed in a sequencing room or in an area separate from library preparation (most common Post-PCR room areas), preferably in an AirClean PCR box to avoid contamination of index-adapter ligated DNA with final products that contain the Illumina cluster sequences.

Firstly, the respective amplified (and thus enriched), and previously size-selected, 96 samples DNA library product is cleaned by doing a manual Post-Amplification DNA Bead Clean-Up step using Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) following respective manufacturer's manual protocol. Final eluted volume of this final cleaned "stock" DNA library = 15 uL. The amplified DNA library should be purified within 1 hour post-amplification.

Secondly, concentration of cleaned final 96 samples DNA library is quantified by performing the Qubit® dsDNA BR (Broad-Range) Kit assay and using the respective Qubit 2.0 flurometer (ThermoFisher Scientific, Waltham, MA, USA), according to respective manufacturer's manual protocol. The Qubit® 2.0 Fluorometer provides the final DNA library

concentration (in units of ng/uL) based on the measured fluorescence intensity values of concentration references and of the sample of interest. This dsDNA library concentration reading (in ng/uL) is mathematically converted to units of nanomolar (nM), using the equation listed in **Figure M-9**:

$$\frac{(\text{concentration in ng}/\mu\text{l})}{(660 \text{ g/mol} \times \text{average library size in bp})} \times 10^6 = \text{concentration in nM}$$

**Figure M-9.** Equation for converting ng/ $\mu$ l to nM when calculating dsDNA library concentration. In this particular case and protocol established average dsDNA library size is 600 bp. The value of 660g/mol corresponds to the average molecular weight (MW) of dsDNA.

This measured and, thus, calculated concentration value is important (it should be preferably higher than 10 nM) for the preparation of the final DNA Library for sequencing as it determines how optimal the cluster generation will be.

Thirdly, preparation of the DNA library template for the sequencing run is done as follows and according to respective manufacturer's manual protocol:

- Using only an aliquot, cleaned final 96 samples DNA library is denatured (using NaOH solution) and diluted (using Tris-HCl, pH 8.0 solution) to a final concentration of 1.3 pM.
- In addition, 0.2% Illumina PhiX Control v3 (Illumina, San Diego, CA, USA) is used as spike-in to monitor sequencing quality control. PhiX is a small and well-defined genome sequence, which enables quick alignment and estimation of error rates.
- Then, final denatured and diluted DNA library spiked with 0.2 % PhiX is loaded onto a respective thawed reagent cartridge and, finally, user sets up the sequencing run.
- Sequencing is carried out on either:

-Illumina MiniSeq sequencer instrument (Illumina, San Diego, CA, USA) using respective High-Output 300 cycle paired-end (2x150bp) kit (FC-420-1003)(Illumina, San Diego, CA, USA). In this sequencing setup, a single 96 samples DNA library can be sequenced.

-Or alternatively on Illumina 500/550 NextSeq sequencer instrument (Illumina, San Diego, CA, USA) using respective Mid-Output 300 cycle paired-end (2x150bp) v2.5 kit (20024905)(Illumina, San Diego, CA, USA). In this sequencing setup, up to two different DNA library sets of 96 samples (total of 192 samples) each can be sequenced in this case (using 2 different 96 index-adaptor plates).

Overall, >80.0% of Illumina MiniSeq and/or 500/550 NextSeq (Illumina, San Diego, CA, USA) average paired-end sequencing base calls exceeded a quality score of 30 (Q30), which is equivalent to the probability of an incorrect base call 1 in 1000 times. Cluster density ranged between 180 and 260 k/mm<sup>2</sup> with 70% to 95% clusters passing the filter.

In this particular MIA FORA™ NGS HLA FLEX Typing protocol and setup, Illumina sequencers generate raw sequencing data that is stored as it is described here:

-Illumina MiniSeq (with one physical lane per dimensional surface of the flow-cell): as two major FASTQ files (R1 and R2), each of them corresponds to the respective Read 151 cycles run on each respective dimensional surface of the flow-cell.

- Illumina 500/550 NextSeq (with four physical lanes per dimensional surface of the flow-cell): as eight major FASTQ files (i.e. four R1 and four R2 FASTQ files).

## **2.6 HLA Allele Calling and Genotype Assignment Bioinformatics Analysis of Sequencing Reads**

Following sequencing, the raw sequencing data is transferred, processed and analyzed using MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0 (Immucor, Inc. Norcross,



GA, USA)), via VNC viewer, with reference to IPD-IMGT/HLA database release 3.25.0. available at the time of the current study. Thus, raw NGS reads are used as input to call genotypes for all 11 HLA loci per sample with high allele resolution (from 3- up to the 4-field). Moreover, key coverage statistics are combined in proprietary algorithms to calculate confidence scores and select the top computed alleles for each HLA gene. Basically, MIA FORA™ NGS FLEX HLA Genotyping Software performs the following main analysis steps (see **Figure M-10**):

- Firstly, this software program demultiplexes (or deconvolutes) FASTQ files according to each unique barcode (defining thus set of raw NGS reads that unequivocally belong to each respective tested sample for these 11 HLA loci).
- Secondly, this software program uses two complementary bioinformatics strategies (generally called mapping and phasing), that, in turn, are based on three orthogonal algorithms. This set of algorithms is employed to calculate a probability score and rank the genotype candidates as well as to generate consensus sequences for individual alleles. In detail (see **Figure M-11**):

-Mapping: it consists on competitive mapping of paired-end sequence reads. It is based on a Competitive Alignment Algorithm.

-Phasing: in which de-novo assembly strategies of paired-end reads enables to construct one or two phased consensus sequences (contigs). It is based on Dynamic Phasing Analysis Algorithm and Consensus Algorithms.

Thus, these two complementary bioinformatics strategies allow:

-To align, first, the reads and, later on, the built consensus sequences to HLA reference sequences. Importantly, as a unique feature of this MIA FORA™ NGS FLEX HLA

Genotyping Software, paired-end reads and consensus sequences are compared with three different sources of HLA reference sequences; (i) from the IPD-IMGT/HLA database release 3.25.0; (ii) from an internal MIA FORA™ HLA reference database generated by cloning and sequencing (using suffixes *e*, *v*, *x* in HLA allele names) as well as by an internal MIA FORA™ HLA collection (iii) of computational filled in-silico HLA sequences (using suffix *I* in HLA allele names).

-To de-novo assemble reads into each respective contig. One- or two-phased consensus sequences (contigs) per targeted HLA locus are built by de-novo assembly of mapped, paired-end reads. In the same assembly process, polymorphic sites are identified where the minor allele frequency exceeds a threshold of 0.2. Once polymorphic sites are identified, phased (where the previous Illumina paired-end sequencing process allows to generate respective fragments that can effectively anchor 2 distant polymorphisms to establish phasing), and resolved, consensus sequences (phased contigs) are built based on sequence assembly and polymorphic linkage.

-To parse the resulting alignments to provide the best match result (to keep only alignments with the, so-called, best “bit-scores”) for a genotype of each allele. Here, several bioinformatics processes are involved in order to: filter out incorrect alignments; to filter out unlikely reference candidates; to enumerate combinations of candidate alleles; to count the number of reads mapped to each combination of candidate alleles, etc. Thus, a mismatch filter eliminates alignments with mismatches or gaps and a paired-end filter increases specificity by requiring both ends to map to a single HLA reference. Then minimum coverage is computed for each allele candidate and allele pairs are computed as described in [187]. The phased resolved consensus sequences are aligned to HLA reference sequence

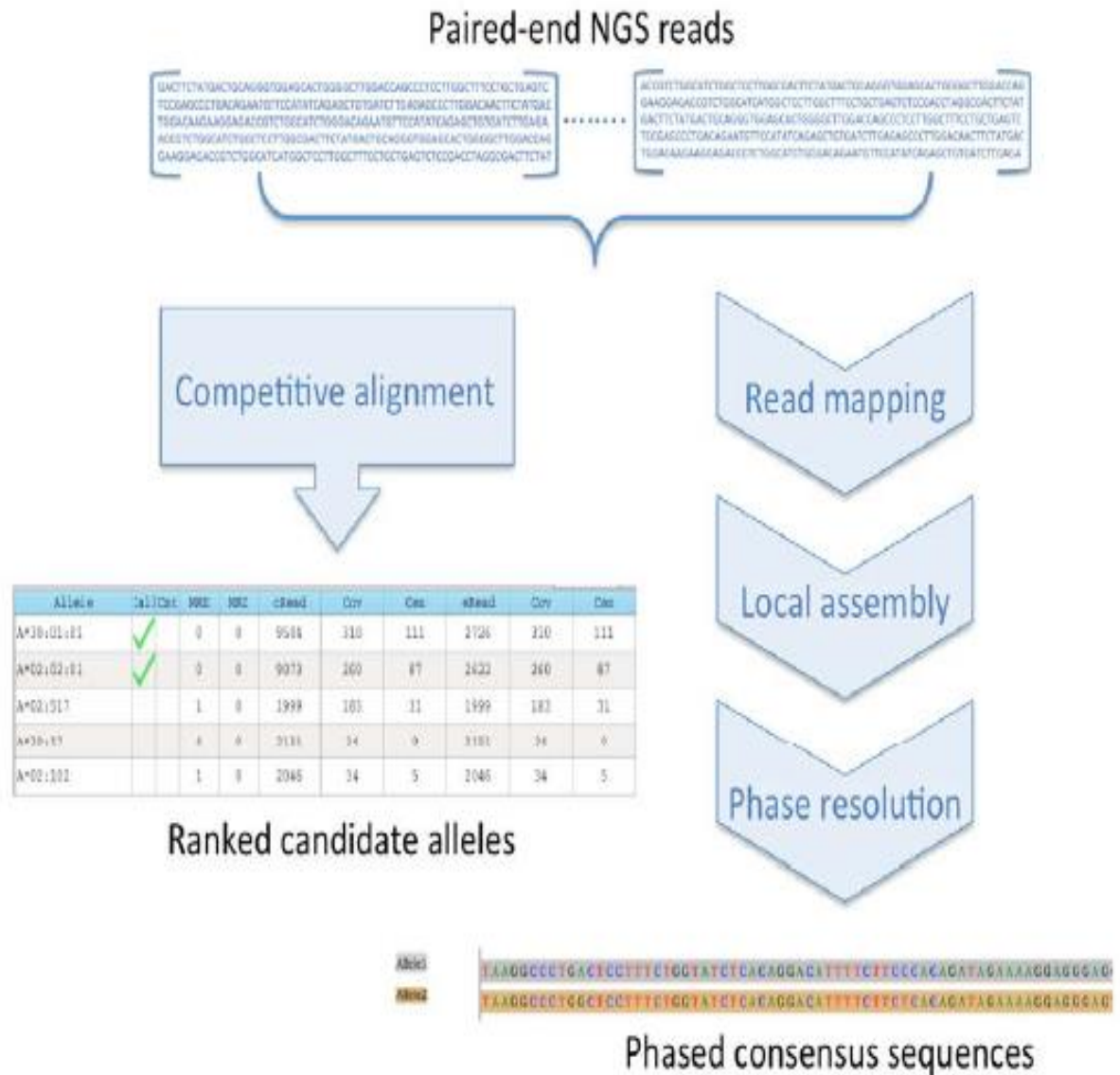
database to determine the best fit. This consensus alignment provides an independent check of the genotype call.

- Ultimately, HLA allele candidates are computed and the final HLA genotyping is called by the software. HLA genotyping results generated by the software program have to be manually reviewed by the user. To further improve the accuracy of HLA genotype algorithm, a flagging system (that the user can use for reviewing the initial automatic calls made by the software) is integrated on this software program and it is based on public information and pattern learned from results generated by these algorithms (see **Figure M-12**). In the flagging system, the linkage disequilibrium between different genes and sequencing depth are used to calibrate the reliability of genotypes (although, just as a clarification, LD data across HLA genes is not directly used within the algorithmic analyses for the HLA genotype assignment that separately takes place for each tested HLA locus). Furthermore, for a single allele, per locus, the minimum average coverage read depth for automated base call needs to be 40× at each position, and uniform coverage has to be generally observed throughout the entire region. Also, to call a heterozygous position, ratio must be at least 20%. In those instances when a given allele presents coverage less than 40×, the user can manually assign (overwrite (OWI)) genotype calls based on reviewing various quality parameters/indicators available in the MIA FORA™ NGS FLEX HLA Genotyping Software. Thus, several embedded quality metrics (e.g. read length, read quality, consensus coverage, imbalance ratio loci, and phasing) improved the accuracy and the confidence of allele calling. A particular interesting parameter is the so-called “central-reads”, which is very useful in those scenarios that require to quantify difference between the two given alignment patterns (continuous tiling pattern: when reads are mapped onto a correct reference sequence; and discontinued tiling pattern: when reads are mapped onto an incorrect reference sequence). To quantify this difference between the two

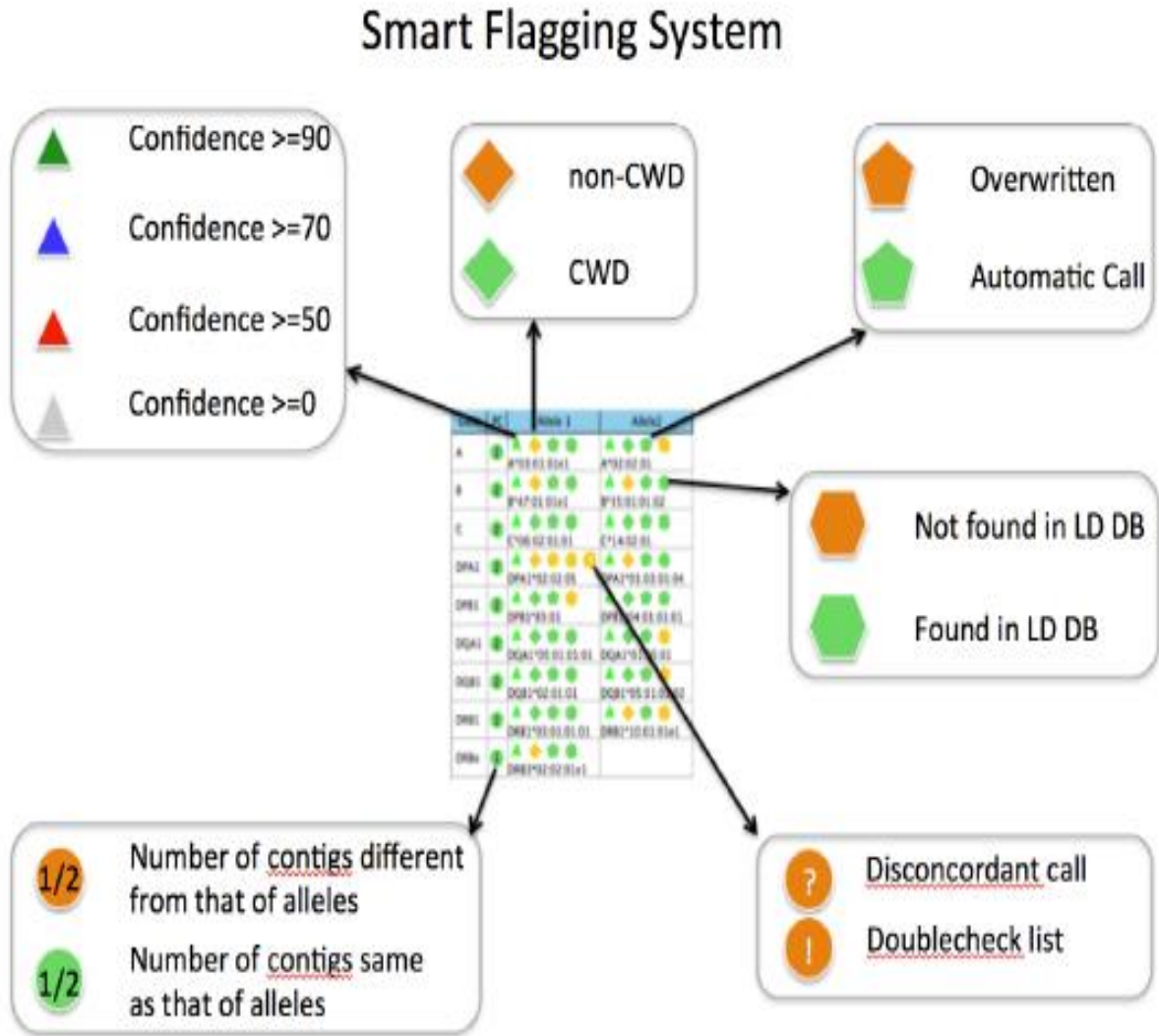
alignment patterns, the number of "central reads" for any given point is counted, where central reads are empirically defined as mapped reads for which the ratio between the length of the left arm (of aligned reads) and that of the right arm (of the aligned reads) related to a particular point is between 0.5 and 2. This central reads counting method may facilitate to distinguish true HLA alleles from sequencing artifacts and thereby improve the reliability of HLA typing.



**Figure M-10.** MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0 (Immucor, Inc. Norcross, GA, USA). Annotated detail window. Block A displays sample information; Block B displays the selected genotypes for the sample. Block C displays the Variants, Smart Guide, and LD Suggestion tables; Block D displays the table of computed allele candidates; Block E displays coverage plots and alignment browsers for mapped sequence of sample; Block F displays the alignment browser for phase resolved de novo contigs. Courtesy: <http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf>



**Figure M-11.** MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0 (Immucor, Inc. Norcross, GA, USA). MIA FORA NGS genotyping strategy. Two complementary strategies are employed to compute the best fit to HLA reference alleles and resolve consensus sequences. The left side illustrates mapping. Paired end reads are mapped using a competitive alignment algorithm to rank candidate alleles. The right side illustrates assembly and phasing. Starting with paired end reads, a multi-step process includes mapping, local assembly, and phase resolution to construct phase resolved consensus sequences. Courtesy: <http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf>



**Figure M-12.** MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0 (Immucor, Inc. Norcross, GA, USA). Flags (colored shapes) are used to depict each predicted genotype status. Indicators for confidence score (green triangle > blue > red > gray) where green is the highest confidence score and gray is the lowest confidence score, common or well-documented allele (green diamond) or not (amber diamond), whether a call has been edited (amber pentagon) or not (green pentagon), whether a call is consistent with linkage disequilibrium data (green hexagon) or not (amber hexagon), and whether special review is required (amber circle with question mark or exclamation mark), indicating a potential novel allele in the coding sequence. PC: the number of phased contigs. If the count of contigs is different from the number of alleles of the corresponding locus, it will show amber color in the circle. Otherwise, it is green color. Courtesy:

<http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf>



• Novel alleles are identified by variation in the coding sequence only. For the purpose of the present study, no attempt was made to record novel intron and 5’-/3’-UTR variation. Since, although it is mostly feasible to be evaluated using this MIA FORA™ NGS FLEX HLA Genotyping Software, in many instances it is quite challenging and complex due to low coverage and related ambiguities found in some of these intronic and 5’-/3’-UTR regions as it is explained in the following point. In addition, currently available IPD-IMGT/HLA Database resources and tools are considerably limited in terms of recorded non-coding reference sequences [87][146][463](<https://www.ebi.ac.uk/ipd/imgt/hla/blast.html>). Thus, as a criteria established by the user, found alleles with novel intron variation are consolidated to the closest and lowest 4-field allele name (see **Table M-6**).

• It is important to remark that, (as it also occurs with the majority of NGS-based HLA genotyping methods and related software HLA allele assignment analysis tools so far developed) using this MIA FORA™ NGS FLEX HLA typing method and related software analysis program, certain HLA allelic and allele combination (phase) ambiguities (see **INTRODUCTION** section for more details about HLA ambiguities) cannot be still resolved. Thus, two main ambiguity groups are noteworthy (see also **2.7 Ambiguity Groups Criteria and Standardization Assignments** for more details):

-Allelic ambiguities: as previously mentioned, there are DNA regions (mostly non-coding regions) with repetitive and extensive low-complexity and imbalanced sequence composition, and that are present along the HLA system, especially at non-coding regions, such as: homopolymer repeats poly(dA), poly(dT), poly(dG) and poly(dC) (composed of eight or more nucleotides); regions of short-tandem repeats (STRs; comprised of 1–6 bp per repeating unit); or high AT- or GC-rich regions (that often contain mononucleotide repeats of 10 or more bases). All these particular DNA regions can establish complex folded



structures in the DNA molecule that, in turn, are prone to mutation via slipped-strand mispairing (termed also as “slippage”) by the DNA polymerase during in vitro PCR-mediated DNA replication, as well as during in vivo DNA replication [203]. Thus, coverage and, consequently, reliability of the called base/s are very low at those positions. Therefore, related ambiguities found in these intronic and 5’-/3’-UTR regions make indistinguishable the 4-field allele resolution level for different HLA loci (in particular, class II under this context of MIA FORA™ NGS FLEX HLA Genotyping Software and method), being necessary to establish an ambiguity reporting criteria by the user (see **Table M-6** for more details).

-Allele combination (phase or cis/trans) ambiguities: at the same time, different described *HLA-DPBI* alleles, especially at their non-coding regions, are highly homologous (present low SNP diversity) and are yet polymorphic, although presenting very distal polymorphic positions that are difficult to be phased, especially in the currently used approach of short-read NGS sequencing. For instance in *HLA-DPBI* locus, phase is often broken in intron 2, which is approximately 4 kb in length, and can be either sparsely or densely populated with heterozygous positions, depending on the combination of alleles. Thus, *HLA-DPBI* genotypes that include alleles with identical sequences in exon-2 that differ in exon-3 generally cannot be placed in phase because of lack of informative SNPs. Consequently, different *HLA-DPBI* allele combinations may satisfy a same set of heterozygous positions but in different cis/trans combinations. These combinations of *HLA-DPBI* alleles (combination of alleles that as pairs share the same exon 2) cannot be unambiguously phased (the so-called phase ambiguities) and, thus, are equally possible. In this case, HLA genotype results output generated by the MIA FORA™ NGS FLEX HLA genotyping software includes all equally possible ambiguous *HLA-DPBI* allele pairs. Nevertheless, for posterior

statistical analysis purposes in the present study, the most likely genotype was assigned on the basis of allele distributions in unambiguous genotypes (i.e. only the *HLA-DPB1* allele pair presenting, although with some exceptions, the respective lowest 2-/3-/4-field name pair was generally considered as the “most likely one”) (see **Table M-7** for more details).

## **2.7 Ambiguity Groups Criteria and Standardization Assignments**

- First ambiguity group is according to allele ambiguities in intronic and 5’-/3’-UTR regions. Here, some HLA assignments resulted ambiguous when trying to distinguish alleles at the 4-field allele resolution level (intronic and untranslated (UTR) sequence level). In these particular cases, called allele candidates present differences only in length of either homopolymer sequences or short tandem repeats (STRs); these were not sequenced with precision by the NGS method. Due to limitations for resolving this type of ambiguities, indistinguishable alleles at the 4-field level were merged (as final genotyping results and for posterior statistical analyses) to the lowest numbered allele according to IPD-IMGT/HLA database version 3.25.0 (see second column on **Table M-6**, where selected lowest numbered HLA alleles for merging are displayed in bold letters). A complete list of indistinguishable alleles and their respective standardization criteria is shown here (see **Table M-6** on next page; obtained and adapted from [268][297]):

**Table M-6.** Characteristics of related HLA class II allele ambiguities found in intronic and 5'-/3'-UTR regions that make indistinguishable the 4-field allele resolution level according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>Ambiguity group</b>        | <b>Alleles in ambiguity group</b>   | <b>Reason for ambiguity</b> | <b>Gene region</b> | <b>Position</b>                    | <b>Motif</b> |
|-------------------------------|---|-----------------------------|--------------------|------------------------------------|--------------|
| <i>HLA-DQA1*01:01:01:02SG</i> | <i>HLA-DQA1*01:01:01:02</i><br><i>HLA-DQA1*01:01:01:03</i>                                | STR (mononucleotide)        | Intron 1           | 3107 to 3118 bp                    | A            |
| <i>HLA-DQA1*01:02:01:01SG</i> | <i>HLA-DQA1*01:02:01:01</i><br><i>HLA-DQA1*01:02:01:03</i><br><i>HLA-DQA1*01:02:01:05</i> | STR (mononucleotide)        | Intron 1           | 3100 to 3113 bp<br>3355 to 3368 bp | A<br>A       |
| <i>HLA-DQA1*01:02:01:04SG</i> | <i>HLA-DQA1*01:02:01:04</i><br><i>HLA-DQA1*01:02:01:06</i><br><i>HLA-DQA1*01:02:01:07</i> | STR (mononucleotide)        | Intron 1           | 3108 to 3121 bp                    | A            |
| <i>HLA-DQA1*01:03:01:02SG</i> | <i>HLA-DQA1*01:03:01:02</i><br><i>HLA-DQA1*01:03:01:06</i>                                | STR (mononucleotide)        | Intron 1           | 3105 to 3114 bp                    | A            |
| <i>HLA-DQA1*01:03:01:03SG</i> | <i>HLA-DQA1*01:03:01:03</i><br><i>HLA-DQA1*01:03:01:04</i>                                | STR (mononucleotide)        | Intron 1           | 3104 to 3118 bp<br>3360 to 3374 bp | A<br>A       |
| <i>HLA-DQA1*01:04:01:01SG</i> | <i>HLA-DQA1*01:04:01:01</i><br><i>HLA-DQA1*01:04:01:02</i><br><i>HLA-DQA1*01:04:01:04</i> | STR (mononucleotide)        | Intron 1           | 3353 to 3368 bp                    | A            |
| <i>HLA-DQA1*02:01:01:01SG</i> | <i>HLA-DQA1*02:01:01:01</i><br><i>HLA-DQA1*02:01:01:02</i>                                | STR (mononucleotide)        | Intron 3           | 4836 to 4848 bp                    | T            |
| <i>HLA-DQA1*05:05:01:01SG</i> | <i>HLA-DQA1*05:05:01:01</i><br><i>HLA-DQA1*05:05:01:02</i>                                | STR (tetranucleotide)       | Intron 3           | 4947 to ~5028 bp                   | TTTC         |
| <i>HLA-DQA1*05:05:01:05SG</i> | <i>HLA-DQA1*05:05:01:05</i><br><i>HLA-DQA1*05:05:01:06</i>                                | STR (tetranucleotide)       | Intron 3           | 4952 to 5033 bp                    | TTTC         |

|                               |                                    |                     |   |                                       |          |
|-------------------------------|------------------------------------|---------------------|---|---------------------------------------|----------|
| <i>HLA-DQB1*03:03:02:02</i>   | <b><i>HLA-DQB1*03:03:02:02</i></b> | Unsequenced region  | Intron 5  | 6488 bp                               | T>G SNP  |
|                               | <i>HLA-DQB1*03:03:02:03</i>        |                     |   |                                       |          |
| <i>HLA-DQB1*05:03:01:01</i>   | <b><i>HLA-DQB1*05:03:01:01</i></b> | Unsequenced region  | 5'UTR   | -157 bp                               | C>T SNP  |
|                               | <i>HLA-DQB1*05:03:01:02</i>        |                     |   |                                       |          |
| <i>HLA-DRB1*03:01:01:01SG</i> | <b><i>HLA-DRB1*03:01:01:01</i></b> | Unsequenced region  | Intron 1  | 1522 bp                               | A>T SNP  |
|                               | <i>HLA-DRB1*03:01:01:02</i>        |                     |   |                                       |          |
|                               |                                    | STR (dinucleotide)  | Intron 2  | 8412 to ~8465 bp<br>~8466 to ~8477 bp | GT<br>GA |
| <i>HLA-DRB1*04:01:01:01SG</i> | <b><i>HLA-DRB1*04:01:01:01</i></b> | STR (dinucleotide)  | Intron 2  | 8676 to ~8719 bp                      | GT       |
|                               | <i>HLA-DRB1*04:01:01:02</i>        |                     |   | ~8720 to ~8749 bp                     | GA       |
| <i>HLA-DRB1*07:01:01:01SG</i> | <b><i>HLA-DRB1*07:01:01:01</i></b> | Unsequenced region  | Intron 1  | 11734-35 bp                           | Indel CA |
|                               | <i>HLA-DRB1*07:01:01:02</i>        |                     |   | 7929 bp                               | G>A SNP  |
|                               |                                    | STR (trinucleotide) | Intron 5  | 14786 to 14890 bp                     | GAA      |
| <i>HLA-DRB1*13:01:01:01SG</i> | <b><i>HLA-DRB1*13:01:01:01</i></b> | STR (dinucleotide)  | Intron 2  | ~8417 to ~8462 bp                     | GT       |
|                               | <i>HLA-DRB1*13:01:01:02</i>        |                     |   | ~8463 to ~8504 bp                     | GA       |
| <i>HLA-DRB1*15:01:01:01SG</i> | <b><i>HLA-DRB1*15:01:01:01</i></b> | STR (dinucleotide)  | Intron 2  | 5701 to 5740 bp                       | GT       |
|                               | <i>HLA-DRB1*15:01:01:02</i>        |                     |   | 5741 to 5784 bp                       | CA       |
|                               | <i>HLA-DRB1*15:01:01:03</i>        |                     |   |                                       |          |
| <i>HLA-DRB1*15:03:01:01SG</i> | <b><i>HLA-DRB1*15:03:01:01</i></b> | STR (dinucleotide)  | Intron 2  | 5682 to 5717 bp                       | GT       |
|                               | <i>HLA-DRB1*15:03:01:02</i>        |                     |   |                                       |          |
| <i>HLA-DRB3*01:01:02:01</i>   | <b><i>HLA-DRB3*01:01:02:01</i></b> | Unsequenced region  | Intron sequence not available for<br>HLA-DRB3*01:01:02:02 |                                       |          |
|                               | <i>HLA-DRB3*01:01:02:02</i>        |                     |   |                                       |          |
| <i>HLA-DRB4*01:03:01:01</i>   | <b><i>HLA-DRB4*01:03:01:01</i></b> | Unsequenced region  | Intron 1  | 3616 bp                               | G>A SNP  |
|                               | <i>HLA-DRB4*01:03:01:03</i>        |                     | Intron 1  | 7069 bp                               | C>A SNP  |
| <i>HLA-DPB1*02:01:02</i>      | <b><i>HLA-DPB1*02:01:02</i></b>    | Unsequenced region  | Exon 5  | codon 225                             | CAA>CAG  |
|                               | <i>HLA-DPB1*02:01:19</i>           |                     |   |                                       |          |
| <i>HLA-DPB1*13:01:01</i>      | <b><i>HLA-DPB1*13:01:01</i></b>    | Unsequenced region  | Exon 1  | codon -22                             | GCG>GCA  |
|                               | <i>HLA-DPB1*107:01</i>             |                     |   | codon -14                             | ACG>ATG  |

Table adapted and originally obtained from [268] and [297].

- Second group comprises *HLA-DPB1* allele combination phase ambiguities:

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1     | HLA-DPB1_Allele2        | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|----------------------|-------------------------|----------------------------|----------------------------|
| <i>DPB1*01:01:01</i> | <i>DPB1*416:01</i>      | <i>DPB1*461:01</i>         | <i>DPB1*462:01</i>         |
| <i>DPB1*01:01:01</i> | <i>DPB1*460:01</i>      | <i>DPB1*131:01</i>         | <i>DPB1*462:01</i>         |
| <i>DPB1*01:01:01</i> | <i>DPB1*02:01:04i2</i>  | <i>DPB1*417:01</i>         | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*01:01:01</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*23:01:01</i>       | <i>DPB1*417:01</i>         |
| <i>DPB1*01:01:01</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>          | <i>DPB1*417:01</i>         |
| <i>DPB1*01:01:01</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*33:01</i>          | <i>DPB1*417:01</i>         |
| <i>DPB1*01:01:01</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*417:01</i>         | <i>DPB1*46:01:01</i>       |
| <i>DPB1*01:01:01</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*417:01</i>         | <i>DPB1*47:01</i>          |
| <i>DPB1*01:01:01</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*417:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*01:01:01</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*462:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*01:01:01</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*417:01</i>         | <i>DPB1*59:01</i>          |
| <i>DPB1*01:01:01</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*417:01</i>         | <i>DPB1*72:01i1</i>        |
| <i>DPB1*01:01:01</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*417:01</i>         | <i>DPB1*78:01</i>          |
| <i>DPB1*01:01:01</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*417:01</i>         | <i>DPB1*81:01</i>          |
| <i>DPB1*01:01:02</i> | <i>DPB1*02:01:02</i>    | <i>DPB1*162:01</i>         | <i>DPB1*461:01</i>         |
| <i>DPB1*01:01:02</i> | <i>DPB1*17:01</i>       | <i>DPB1*131:01</i>         | <i>DPB1*162:01</i>         |
| <i>DPB1*01:01:02</i> | <i>DPB1*02:01:04i2</i>  | <i>DPB1*162:01</i>         | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*01:01:02</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*162:01</i>         | <i>DPB1*46:01:01</i>       |
| <i>DPB1*01:01:02</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*162:01</i>         | <i>DPB1*47:01</i>          |
| <i>DPB1*01:01:02</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*162:01</i>         | <i>DPB1*72:01i1</i>        |
| <i>DPB1*01:01:02</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*162:01</i>         | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02</i> | <i>DPB1*03:01:08</i>    | <i>DPB1*124:01</i>         | <i>DPB1*352:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*04:02:01:01</i> | <i>DPB1*105:01</i>         | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*04:02:01:01</i> | <i>DPB1*416:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*02:01:02</i> | <i>DPB1*04:02:01:02</i> | <i>DPB1*105:01</i>         | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*04:02:01:02</i> | <i>DPB1*416:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*02:01:02</i> | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*126:01</i>      | <i>DPB1*04:01:01:01</i>    | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*126:01</i>      | <i>DPB1*04:01:01:02</i>    | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*131:01</i>      | <i>DPB1*17:01</i>          | <i>DPB1*461:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*138:01</i>      | <i>DPB1*23:01:01</i>       | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*138:01</i>      | <i>DPB1*416:01</i>         | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*02:01:02</i> | <i>DPB1*19:01</i>       | <i>DPB1*106:01</i>         | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*296:01</i>      | <i>DPB1*28:01</i>          | <i>DPB1*352:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*351:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*414:01e1</i>    | <i>DPB1*02:01:02e1</i>     | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*414:01e1</i>    | <i>DPB1*02:01:02e2</i>     | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*414:01e1</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*414:01e1</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*414:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*460:01</i>      | <i>DPB1*17:01</i>          | <i>DPB1*416:01</i>         |
| <i>DPB1*02:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*141:01</i>         | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*352:01</i>         | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*414:01</i>         | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*416:01</i>         | <i>DPB1*46:01:01i1</i>     |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*02:01:02</i>    | <i>DPB1*46:01:01</i>    | <i>DPB1*461:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02</i>    | <i>DPB1*47:01</i>       | <i>DPB1*141:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*47:01</i>       | <i>DPB1*352:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*47:01</i>       | <i>DPB1*414:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*47:01</i>       | <i>DPB1*416:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*47:01</i>       | <i>DPB1*461:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*81:01</i>       | <i>DPB1*141:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*81:01</i>       | <i>DPB1*352:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*81:01</i>       | <i>DPB1*414:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*81:01</i>       | <i>DPB1*416:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*81:01</i>       | <i>DPB1*461:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*141:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*352:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*414:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*416:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*461:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02</i>    | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*414:01</i>                |
| <i>DPB1*02:01:02</i>    | <i>DPB1*51:01i1</i>     | <i>DPB1*416:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:01:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:01:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:01:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:01:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:01:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:01:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:01e1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:01e1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:01e1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:01:01e1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:01:01e1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:01:01e1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:02:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:02:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:02:01:02</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e2</i>            | <i>DPB1*04:02:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e3</i>            | <i>DPB1*04:02:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e4</i>            | <i>DPB1*04:02:01:01</i>           |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e2</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e3</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e4</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e2</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e3</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e4</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e2</i>            | <i>DPB1*14:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e3</i>            | <i>DPB1*14:01:01i1</i>            |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*14:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*19:01</i>       | <i>DPB1*106:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e2</i>            | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e3</i>            | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e2</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e3</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*46:01:01</i>    | <i>DPB1*414:01e1</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*47:01</i>       | <i>DPB1*414:01e1</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*51:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*51:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*51:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*59:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*59:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*59:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*78:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*78:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*78:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e2</i>            | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e3</i>            | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e4</i>            | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*81:01</i>       | <i>DPB1*414:01e1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e2</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e3</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*414:01e1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e2</i>            | <i>DPB1*104:01</i>                |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e3</i>            | <i>DPB1*104:01</i>                |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*104:01</i>                |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e2</i>            | <i>DPB1*105:01</i>                |
| <i>DPB1*02:01:02e1</i>  | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e3</i>            | <i>DPB1*105:01</i>                |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1       | HLA-DPB1_Allele2        | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|------------------------|-------------------------|----------------------------|----------------------------|
| <i>DPB1*02:01:02e1</i> | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*105:01</i>         |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e2</i>     | <i>DPB1*14:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*14:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*14:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e2</i>     | <i>DPB1*23:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*23:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*23:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*29:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*29:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*29:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*33:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*33:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*33:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*39:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*39:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*39:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e2</i>     | <i>DPB1*46:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*46:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*46:01:01</i>       |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*47:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*47:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*47:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*51:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*51:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*51:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*59:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*59:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*59:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*72:01i1</i>        |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*72:01i1</i>        |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*72:01i1</i>        |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*78:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*78:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*78:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e2</i>     | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02e1</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:01:01:02</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01:02</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:01:01:01</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01:01</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:01e1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01e1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:01:01e1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01e1</i>     |



**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1       | HLA-DPB1_Allele2        | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|------------------------|-------------------------|----------------------------|----------------------------|
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:02:01:02</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:02:01:02</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e3</i>     | <i>DPB1*04:02:01:01</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:02:01:01</i>    |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e3</i>     | <i>DPB1*104:01i1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e4</i>     | <i>DPB1*104:01i1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*414:01e1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e3</i>     | <i>DPB1*105:01i1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e4</i>     | <i>DPB1*105:01i1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*19:01</i>       | <i>DPB1*106:01</i>         | <i>DPB1*414:01e1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*29:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*29:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*33:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*33:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*39:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*39:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*414:01e1</i>       | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*47:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*47:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*47:01</i>       | <i>DPB1*414:01e1</i>       | <i>DPB1*47:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e3</i>     | <i>DPB1*81:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*81:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*81:01</i>       | <i>DPB1*414:01e1</i>       | <i>DPB1*81:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*72:01i2</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*72:01i2</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*414:01e1</i>       | <i>DPB1*72:01i2</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*104:01</i>         |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*104:01</i>         |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>         | <i>DPB1*414:01e1</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e3</i>     | <i>DPB1*105:01</i>         |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*105:01</i>         |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*14:01:01</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*14:01:01</i>       |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1       | HLA-DPB1_Allele2        | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|------------------------|-------------------------|----------------------------|----------------------------|
| <i>DPB1*02:01:02e2</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*23:01:01</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*23:01:01</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*29:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*29:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*33:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*33:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*39:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*39:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e3</i>     | <i>DPB1*46:01:01</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*46:01:01</i>       |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*47:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*47:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*51:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*51:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*59:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*59:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*72:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*72:01i1</i>        |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*78:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*78:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e3</i>     | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02e2</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*81:01</i>          |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01:02</i>    |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01:01</i>    |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01e1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:01e1</i>     | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:01:01e1</i>     |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:02:01:02</i>    |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:02e4</i>     | <i>DPB1*04:02:01:01</i>    |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*104:01</i>      | <i>DPB1*02:01:02e4</i>     | <i>DPB1*104:01i1</i>       |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*414:01e1</i>       |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*105:01</i>      | <i>DPB1*02:01:02e4</i>     | <i>DPB1*105:01i1</i>       |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*14:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*19:01</i>       | <i>DPB1*106:01</i>         | <i>DPB1*414:01e1</i>       |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*29:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*29:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*33:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*33:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*39:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*39:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:02e4</i>     | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*414:01e1</i>       | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*47:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*47:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*47:01</i>       | <i>DPB1*414:01e1</i>       | <i>DPB1*47:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*51:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*59:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*78:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*02:01:02e3</i> | <i>DPB1*81:01</i>       | <i>DPB1*02:01:02e4</i>     | <i>DPB1*81:01i1</i>        |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*81:01</i>       | <i>DPB1*414:01e1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*414:01e1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*104:01</i>                |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*105:01i1</i>    | <i>DPB1*02:01:02e4</i>            | <i>DPB1*105:01</i>                |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*14:01:01i1</i>  | <i>DPB1*02:01:02e4</i>            | <i>DPB1*14:01:01</i>              |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*23:01:01i1</i>  | <i>DPB1*02:01:02e4</i>            | <i>DPB1*23:01:01</i>              |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*29:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*29:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*33:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*33:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*39:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*39:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*46:01:01i1</i>  | <i>DPB1*02:01:02e4</i>            | <i>DPB1*46:01:01</i>              |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*47:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*47:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*51:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*51:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*59:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*59:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*72:01i2</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*72:01i1</i>               |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*78:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*78:01</i>                 |
| <i>DPB1*02:01:02e3</i>  | <i>DPB1*81:01i1</i>     | <i>DPB1*02:01:02e4</i>            | <i>DPB1*81:01</i>                 |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*19:01</i>       | <i>DPB1*106:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*46:01:01</i>    | <i>DPB1*414:01e1</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*47:01</i>       | <i>DPB1*414:01e1</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*81:01</i>       | <i>DPB1*414:01e1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*72:01i1</i>     | <i>DPB1*414:01e1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*02:01:02e4</i>  | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*414:01e1</i>              |
| <i>DPB1*03:01:01</i>    | <i>DPB1*04:01:01:01</i> | <i>DPB1*124:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*04:01:01:02</i> | <i>DPB1*124:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*04:02:01:01</i> | <i>DPB1*351:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*04:02:01:02</i> | <i>DPB1*351:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*05:01:01</i>    | <i>DPB1*104:01</i>                | <i>DPB1*135:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*05:01:01</i>    | <i>DPB1*135:01</i>                | <i>DPB1*104:01i1</i>              |
| <i>DPB1*03:01:01</i>    | <i>DPB1*05:01:02</i>    | <i>DPB1*03:01:08</i>              | <i>DPB1*05:01:01</i>              |
| <i>DPB1*03:01:01</i>    | <i>DPB1*05:01:02</i>    | <i>DPB1*03:01:08</i>              | <i>DPB1*135:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*105:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*126:01</i>      | <i>DPB1*350:01</i>                | <i>DPB1*351:01</i>                |
| <i>DPB1*03:01:01</i>    | <i>DPB1*133:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*13:01:01</i>              |
| <i>DPB1*03:01:01</i>    | <i>DPB1*23:01:01</i>    | <i>DPB1*104:01</i>                | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*03:01:01</i>    | <i>DPB1*23:01:01</i>    | <i>DPB1*104:01i1</i>              | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*03:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*03:01:08</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*104:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*124:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*351:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*33:01</i>       | <i>DPB1*104:01</i>                | <i>DPB1*33:01i1</i>               |
| <i>DPB1*03:01:01</i>    | <i>DPB1*33:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*33:01i1</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1        | HLA-DPB1_Allele2       | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|-------------------------|------------------------|----------------------------|----------------------------|
| <i>DPB1*03:01:01</i>    | <i>DPB1*462:01</i>     | <i>DPB1*351:01</i>         | <i>DPB1*417:01</i>         |
| <i>DPB1*03:01:01</i>    | <i>DPB1*46:01:01</i>   | <i>DPB1*104:01</i>         | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*03:01:01</i>    | <i>DPB1*46:01:01</i>   | <i>DPB1*104:01i1</i>       | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*03:01:01</i>    | <i>DPB1*47:01</i>      | <i>DPB1*104:01</i>         | <i>DPB1*47:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*47:01</i>      | <i>DPB1*104:01i1</i>       | <i>DPB1*47:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*51:01</i>      | <i>DPB1*104:01</i>         | <i>DPB1*51:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*51:01</i>      | <i>DPB1*104:01i1</i>       | <i>DPB1*51:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*59:01</i>      | <i>DPB1*104:01</i>         | <i>DPB1*59:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*59:01</i>      | <i>DPB1*104:01i1</i>       | <i>DPB1*59:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*78:01</i>      | <i>DPB1*03:01:08</i>       | <i>DPB1*78:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*78:01</i>      | <i>DPB1*104:01</i>         | <i>DPB1*78:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*78:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*78:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*78:01</i>      | <i>DPB1*351:01</i>         | <i>DPB1*78:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*78:01</i>      | <i>DPB1*104:01i1</i>       | <i>DPB1*78:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*81:01</i>      | <i>DPB1*104:01</i>         | <i>DPB1*81:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*81:01</i>      | <i>DPB1*104:01i1</i>       | <i>DPB1*81:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*02:01:04i1</i> | <i>DPB1*104:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*03:01:01</i>    | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*104:01i1</i>       |
| <i>DPB1*03:01:01</i>    | <i>DPB1*72:01i1</i>    | <i>DPB1*104:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*72:01i1</i>    | <i>DPB1*104:01i1</i>       | <i>DPB1*72:01i2</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*02:01:04i2</i> | <i>DPB1*124:01</i>         | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*03:01:01</i>    | <i>DPB1*105:01i1</i>   | <i>DPB1*124:01</i>         | <i>DPB1*463:01</i>         |
| <i>DPB1*03:01:01</i>    | <i>DPB1*46:01:01i1</i> | <i>DPB1*124:01</i>         | <i>DPB1*46:01:01</i>       |
| <i>DPB1*03:01:01</i>    | <i>DPB1*47:01i1</i>    | <i>DPB1*124:01</i>         | <i>DPB1*47:01</i>          |
| <i>DPB1*03:01:01</i>    | <i>DPB1*51:01i1</i>    | <i>DPB1*351:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*03:01:01</i>    | <i>DPB1*72:01i2</i>    | <i>DPB1*124:01</i>         | <i>DPB1*72:01i1</i>        |
| <i>DPB1*03:01:01</i>    | <i>DPB1*81:01i1</i>    | <i>DPB1*124:01</i>         | <i>DPB1*81:01</i>          |
| <i>DPB1*03:01:08</i>    | <i>DPB1*05:01:01</i>   | <i>DPB1*05:01:02</i>       | <i>DPB1*104:01</i>         |
| <i>DPB1*03:01:08</i>    | <i>DPB1*05:01:01</i>   | <i>DPB1*05:01:02</i>       | <i>DPB1*104:01i1</i>       |
| <i>DPB1*03:01:08</i>    | <i>DPB1*28:01</i>      | <i>DPB1*124:01</i>         | <i>DPB1*296:01</i>         |
| <i>DPB1*03:01:08</i>    | <i>DPB1*414:01</i>     | <i>DPB1*104:01</i>         | <i>DPB1*352:01</i>         |
| <i>DPB1*03:01:08</i>    | <i>DPB1*414:01</i>     | <i>DPB1*352:01</i>         | <i>DPB1*104:01i1</i>       |
| <i>DPB1*03:01:08</i>    | <i>DPB1*416:01</i>     | <i>DPB1*351:01</i>         | <i>DPB1*352:01</i>         |
| <i>DPB1*03:01:08</i>    | <i>DPB1*02:01:04i2</i> | <i>DPB1*124:01</i>         | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*03:01:08</i>    | <i>DPB1*46:01:01i1</i> | <i>DPB1*124:01</i>         | <i>DPB1*46:01:01</i>       |
| <i>DPB1*03:01:08</i>    | <i>DPB1*47:01i1</i>    | <i>DPB1*124:01</i>         | <i>DPB1*47:01</i>          |
| <i>DPB1*03:01:08</i>    | <i>DPB1*51:01i1</i>    | <i>DPB1*351:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*03:01:08</i>    | <i>DPB1*72:01i2</i>    | <i>DPB1*124:01</i>         | <i>DPB1*72:01i1</i>        |
| <i>DPB1*03:01:08</i>    | <i>DPB1*81:01i1</i>    | <i>DPB1*124:01</i>         | <i>DPB1*81:01</i>          |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*104:01</i>     | <i>DPB1*04:01:01:02</i>    | <i>DPB1*104:01i1</i>       |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*104:01</i>     | <i>DPB1*124:01</i>         | <i>DPB1*350:01</i>         |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*105:01</i>     | <i>DPB1*04:01:01:02</i>    | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*138:01</i>     | <i>DPB1*126:01</i>         | <i>DPB1*23:01:01</i>       |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*138:01</i>     | <i>DPB1*126:01</i>         | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*13:01:01</i>   | <i>DPB1*133:01</i>         | <i>DPB1*350:01</i>         |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*04:01:01:01</i> | <i>DPB1*14:01:01</i>    | <i>DPB1*04:01:01:02</i>           | <i>DPB1*14:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*04:01:01:02</i>           | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*29:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*29:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*33:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*33:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*351:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*126:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*39:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*39:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*460:01</i>      | <i>DPB1*126:01</i>                | <i>DPB1*17:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*463:01</i>      | <i>DPB1*105:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*463:01</i>      | <i>DPB1*350:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*04:01:01:02</i>           | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*04:01:31</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*126:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*350:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*415:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*459:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*464:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*04:01:31</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*126:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*350:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*459:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01</i>       | <i>DPB1*464:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*51:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*51:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*59:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*59:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*78:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*78:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*04:01:01:02</i>           | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*04:01:31</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*126:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*350:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*459:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01</i>       | <i>DPB1*464:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*04:01:01:02</i>           | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*04:01:31</i>              | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*126:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*350:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*415:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*459:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i1</i>  | <i>DPB1*464:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*04:01:31</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*126:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*350:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*415:01</i>                | <i>DPB1*72:01i2</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*459:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i1</i>     | <i>DPB1*464:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*02:01:04i2</i>  | <i>DPB1*04:01:01:02</i>           | <i>DPB1*02:01:04i1</i>            |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*104:01i1</i>    | <i>DPB1*04:01:01:02</i>           | <i>DPB1*104:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*105:01i1</i>    | <i>DPB1*04:01:01:02</i>           | <i>DPB1*105:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*14:01:01i1</i>  | <i>DPB1*04:01:01:02</i>           | <i>DPB1*14:01:01</i>              |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*04:01:01:02</i>           | <i>DPB1*23:01:01</i>              |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*29:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*33:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*33:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*39:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*39:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*04:01:01:02</i>           | <i>DPB1*46:01:01</i>              |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*47:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*51:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*126:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*59:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*72:01i1</i>               |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*78:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*350:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*04:01:01:01</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*04:01:01:02</i>           | <i>DPB1*81:01</i>                 |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*104:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*138:01</i>      | <i>DPB1*126:01</i>                | <i>DPB1*23:01:01</i>              |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*138:01</i>      | <i>DPB1*126:01</i>                | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*13:01:01</i>    | <i>DPB1*133:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*351:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*126:01</i>                |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*460:01</i>      | <i>DPB1*126:01</i>                | <i>DPB1*17:01</i>                 |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*463:01</i>      | <i>DPB1*105:01</i>                | <i>DPB1*350:01</i>                |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*463:01</i>      | <i>DPB1*350:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*04:01:31</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*126:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*350:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*415:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*459:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*46:01:01</i>    | <i>DPB1*464:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*04:01:31</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*126:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*350:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*459:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*47:01</i>       | <i>DPB1*464:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>       | <i>DPB1*04:01:31</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>       | <i>DPB1*126:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>       | <i>DPB1*350:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*81:01i1</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1        | HLA-DPB1_Allele2       | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|-------------------------|------------------------|----------------------------|----------------------------|
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>      | <i>DPB1*459:01</i>         | <i>DPB1*81:01i1</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*81:01</i>      | <i>DPB1*464:01</i>         | <i>DPB1*81:01i1</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*04:01:31</i>       | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*126:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*350:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*415:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*459:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*464:01</i>         | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*04:01:31</i>       | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*126:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*350:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*415:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*459:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*464:01</i>         | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*104:01i1</i>   | <i>DPB1*124:01</i>         | <i>DPB1*350:01</i>         |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*350:01</i>         |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*51:01i1</i>    | <i>DPB1*126:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*04:01:01:02</i> | <i>DPB1*78:01i1</i>    | <i>DPB1*350:01</i>         | <i>DPB1*78:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*104:01</i>     | <i>DPB1*04:01e1</i>        | <i>DPB1*104:01i1</i>       |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*105:01</i>     | <i>DPB1*04:01e1</i>        | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*14:01:01</i>   | <i>DPB1*04:01e1</i>        | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*23:01:01</i>   | <i>DPB1*04:01e1</i>        | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*29:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*29:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*33:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*33:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*39:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*39:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*46:01:01</i>   | <i>DPB1*04:01e1</i>        | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*47:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*47:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*51:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*51:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*59:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*59:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*78:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*78:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*81:01</i>      | <i>DPB1*04:01e1</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:04i1</i> | <i>DPB1*04:01e1</i>        | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*72:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*02:01:04i2</i> | <i>DPB1*04:01e1</i>        | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*104:01i1</i>   | <i>DPB1*04:01e1</i>        | <i>DPB1*104:01</i>         |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*105:01i1</i>   | <i>DPB1*04:01e1</i>        | <i>DPB1*105:01</i>         |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*14:01:01i1</i> | <i>DPB1*04:01e1</i>        | <i>DPB1*14:01:01</i>       |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*23:01:01i1</i> | <i>DPB1*04:01e1</i>        | <i>DPB1*23:01:01</i>       |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*29:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*29:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*33:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*33:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*39:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*39:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*46:01:01i1</i> | <i>DPB1*04:01e1</i>        | <i>DPB1*46:01:01</i>       |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*47:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*47:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*51:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*51:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*59:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*59:01</i>          |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1        | HLA-DPB1_Allele2       | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|-------------------------|------------------------|----------------------------|----------------------------|
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*72:01i2</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*72:01i1</i>        |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*78:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*78:01</i>          |
| <i>DPB1*04:01:01e1</i>  | <i>DPB1*81:01i1</i>    | <i>DPB1*04:01e1</i>        | <i>DPB1*81:01</i>          |
| <i>DPB1*04:01:31</i>    | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*350:01</i>         |
| <i>DPB1*04:01:31</i>    | <i>DPB1*51:01i1</i>    | <i>DPB1*126:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*04:01:31</i>    | <i>DPB1*78:01i1</i>    | <i>DPB1*350:01</i>         | <i>DPB1*78:01</i>          |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*104:01</i>     | <i>DPB1*04:02:01:02</i>    | <i>DPB1*104:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*105:01</i>     | <i>DPB1*04:02:01:02</i>    | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*124:01</i>     | <i>DPB1*105:01</i>         | <i>DPB1*351:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*124:01</i>     | <i>DPB1*351:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*14:01:01</i>   | <i>DPB1*04:02:01:02</i>    | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*17:01</i>      | <i>DPB1*105:01</i>         | <i>DPB1*460:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*17:01</i>      | <i>DPB1*460:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01</i>   | <i>DPB1*04:02:01:02</i>    | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01</i>   | <i>DPB1*105:01</i>         | <i>DPB1*138:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01</i>   | <i>DPB1*138:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*29:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*29:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*33:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*33:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*350:01</i>     | <i>DPB1*126:01</i>         | <i>DPB1*463:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*39:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*39:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*417:01</i>     | <i>DPB1*462:01</i>         | <i>DPB1*463:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*46:01:01</i>   | <i>DPB1*04:02:01:02</i>    | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*47:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*47:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*51:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*51:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*51:01</i>      | <i>DPB1*105:01</i>         | <i>DPB1*51:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*51:01</i>      | <i>DPB1*463:01</i>         | <i>DPB1*51:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*51:01</i>      | <i>DPB1*105:01i1</i>       | <i>DPB1*51:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*59:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*59:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*78:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*78:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*81:01</i>      | <i>DPB1*04:02:01:02</i>    | <i>DPB1*81:01i1</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*04:02:01:02</i>    | <i>DPB1*02:01:04i2</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*04:02:01:02</i>    | <i>DPB1*72:01i2</i>        |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:04i2</i> | <i>DPB1*04:02:01:02</i>    | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:04i2</i> | <i>DPB1*105:01</i>         | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*104:01i1</i>   | <i>DPB1*04:02:01:02</i>    | <i>DPB1*104:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*105:01i1</i>   | <i>DPB1*04:02:01:02</i>    | <i>DPB1*105:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*14:01:01i1</i> | <i>DPB1*04:02:01:02</i>    | <i>DPB1*14:01:01</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01i1</i> | <i>DPB1*04:02:01:02</i>    | <i>DPB1*23:01:01</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01i1</i> | <i>DPB1*105:01</i>         | <i>DPB1*138:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*23:01:01i1</i> | <i>DPB1*138:01</i>         | <i>DPB1*105:01i1</i>       |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*29:01i1</i>    | <i>DPB1*04:02:01:02</i>    | <i>DPB1*29:01</i>          |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*463:01</i>         |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*33:01i1</i>    | <i>DPB1*04:02:01:02</i>    | <i>DPB1*33:01</i>          |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*39:01i1</i>    | <i>DPB1*04:02:01:02</i>    | <i>DPB1*39:01</i>          |



**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*04:02:01:01</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*04:02:01:02</i>           | <i>DPB1*46:01:01</i>              |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*105:01</i>                | <i>DPB1*46:01:01</i>              |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*47:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*105:01</i>                | <i>DPB1*47:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*51:01i1</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*51:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*59:01i1</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*59:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*72:01i1</i>               |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*105:01</i>                | <i>DPB1*72:01i1</i>               |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*78:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*463:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*04:02:01:02</i>           | <i>DPB1*81:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*105:01</i>                | <i>DPB1*81:01</i>                 |
| <i>DPB1*04:02:01:01</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*124:01</i>      | <i>DPB1*105:01</i>                | <i>DPB1*351:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*124:01</i>      | <i>DPB1*351:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*17:01</i>       | <i>DPB1*105:01</i>                | <i>DPB1*460:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*17:01</i>       | <i>DPB1*460:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*105:01</i>                | <i>DPB1*138:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*23:01:01</i>    | <i>DPB1*138:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*350:01</i>      | <i>DPB1*126:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*417:01</i>      | <i>DPB1*462:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*51:01</i>       | <i>DPB1*105:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*51:01</i>       | <i>DPB1*463:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*51:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*51:01i1</i>               |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:04i2</i>  | <i>DPB1*105:01</i>                | <i>DPB1*02:01:04i1</i>            |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*105:01</i>                | <i>DPB1*138:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*23:01:01i1</i>  | <i>DPB1*138:01</i>                | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*463:01</i>                |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*105:01</i>                | <i>DPB1*46:01:01</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*105:01</i>                | <i>DPB1*47:01</i>                 |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*105:01</i>                | <i>DPB1*72:01i1</i>               |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*105:01i1</i>              |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*78:01i1</i>     | <i>DPB1*463:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*105:01</i>                | <i>DPB1*81:01</i>                 |
| <i>DPB1*04:02:01:02</i> | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*05:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*05:01:02</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*05:01:01</i>    | <i>DPB1*352:01</i>      | <i>DPB1*05:01:02</i>              | <i>DPB1*414:01</i>                |
| <i>DPB1*05:01:01</i>    | <i>DPB1*78:01</i>       | <i>DPB1*05:01:02</i>              | <i>DPB1*78:01i1</i>               |
| <i>DPB1*05:01:01</i>    | <i>DPB1*02:01:04i2</i>  | <i>DPB1*135:01</i>                | <i>DPB1*02:01:04i1</i>            |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*05:01:01</i>    | <i>DPB1*23:01:01i1</i>  | <i>DPB1*135:01</i>                | <i>DPB1*23:01:01</i>              |
| <i>DPB1*05:01:01</i>    | <i>DPB1*29:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*29:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*33:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*33:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*46:01:01i1</i>  | <i>DPB1*135:01</i>                | <i>DPB1*46:01:01</i>              |
| <i>DPB1*05:01:01</i>    | <i>DPB1*47:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*47:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*51:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*59:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*59:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*72:01i2</i>     | <i>DPB1*135:01</i>                | <i>DPB1*72:01i1</i>               |
| <i>DPB1*05:01:01</i>    | <i>DPB1*78:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*05:01:01</i>    | <i>DPB1*81:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*81:01</i>                 |
| <i>DPB1*05:01:02</i>    | <i>DPB1*29:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*29:01</i>                 |
| <i>DPB1*05:01:02</i>    | <i>DPB1*78:01i1</i>     | <i>DPB1*135:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*104:01</i>      | <i>DPB1*105:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*104:01</i>      | <i>DPB1*105:01</i>      | <i>DPB1*104:01i1</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*106:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*19:01</i>                 |
| <i>DPB1*104:01</i>      | <i>DPB1*133:01</i>      | <i>DPB1*124:01</i>                | <i>DPB1*13:01:01</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*14:01:01</i>    | <i>DPB1*104:01i1</i>              | <i>DPB1*14:01:01i1</i>            |
| <i>DPB1*104:01</i>      | <i>DPB1*23:01:01</i>    | <i>DPB1*104:01i1</i>              | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*104:01</i>      | <i>DPB1*29:01</i>       | <i>DPB1*124:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*29:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*33:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*33:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*39:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*39:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*416:01</i>      | <i>DPB1*351:01</i>                | <i>DPB1*414:01</i>                |
| <i>DPB1*104:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*104:01i1</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*104:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*51:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*59:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*59:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*124:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*78:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*104:01i1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*02:01:04i2</i>  | <i>DPB1*124:01</i>                | <i>DPB1*02:01:04i1</i>            |
| <i>DPB1*104:01</i>      | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*105:01i1</i>    | <i>DPB1*105:01</i>                | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*105:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*104:01</i>      | <i>DPB1*14:01:01i1</i>  | <i>DPB1*14:01:01</i>              | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*23:01:01i1</i>  | <i>DPB1*23:01:01</i>              | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*33:01i1</i>     | <i>DPB1*33:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*39:01i1</i>     | <i>DPB1*39:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*124:01</i>                | <i>DPB1*46:01:01</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*124:01</i>                | <i>DPB1*47:01</i>                 |
| <i>DPB1*104:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*104:01i1</i>              |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*104:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*351:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*104:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*124:01</i>                | <i>DPB1*72:01i1</i>               |
| <i>DPB1*104:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*104:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*124:01</i>                | <i>DPB1*81:01</i>                 |
| <i>DPB1*104:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*13:01:01</i>    | <i>DPB1*133:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*105:01</i>      | <i>DPB1*14:01:01</i>    | <i>DPB1*105:01i1</i>              | <i>DPB1*14:01:01i1</i>            |
| <i>DPB1*105:01</i>      | <i>DPB1*23:01:01</i>    | <i>DPB1*105:01i1</i>              | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*105:01</i>      | <i>DPB1*29:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*33:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*33:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*39:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*39:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*463:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*105:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*105:01i1</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*105:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*463:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*51:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*59:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*59:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*78:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*463:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*463:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*105:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*463:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*105:01i1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*105:01</i>      | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*463:01</i>                |
| <i>DPB1*105:01</i>      | <i>DPB1*14:01:01i1</i>  | <i>DPB1*14:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*23:01:01i1</i>  | <i>DPB1*23:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*463:01</i>                |
| <i>DPB1*105:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*33:01i1</i>     | <i>DPB1*33:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*39:01i1</i>     | <i>DPB1*39:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*78:01i1</i>     | <i>DPB1*463:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*105:01</i>      | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*105:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*106:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*19:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*106:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*19:01</i>                 | <i>DPB1*47:01i1</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*106:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*19:01</i>                 | <i>DPB1*81:01i1</i>               |
| <i>DPB1*106:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*19:01</i>                 | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*106:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*19:01</i>                 | <i>DPB1*72:01i2</i>               |
| <i>DPB1*106:01</i>      | <i>DPB1*104:01i1</i>    | <i>DPB1*124:01</i>                | <i>DPB1*19:01</i>                 |
| <i>DPB1*124:01</i>      | <i>DPB1*138:01</i>      | <i>DPB1*23:01:01</i>              | <i>DPB1*351:01</i>                |
| <i>DPB1*124:01</i>      | <i>DPB1*138:01</i>      | <i>DPB1*351:01</i>                | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*124:01</i>      | <i>DPB1*13:01:01</i>    | <i>DPB1*133:01</i>                | <i>DPB1*104:01i1</i>              |
| <i>DPB1*124:01</i>      | <i>DPB1*460:01</i>      | <i>DPB1*17:01</i>                 | <i>DPB1*351:01</i>                |
| <i>DPB1*124:01</i>      | <i>DPB1*463:01</i>      | <i>DPB1*104:01i1</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*124:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*351:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*124:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*104:01i1</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*124:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*351:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*351:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*351:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*124:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*104:01i1</i>              |
| <i>DPB1*124:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*351:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*104:01i1</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*124:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*124:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*351:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*124:01</i>      | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*104:01i1</i>              |
| <i>DPB1*126:01</i>      | <i>DPB1*417:01</i>      | <i>DPB1*350:01</i>                | <i>DPB1*462:01</i>                |
| <i>DPB1*126:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*350:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*126:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*126:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*459:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*126:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*464:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*126:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*350:01</i>                |
| <i>DPB1*126:01</i>      | <i>DPB1*78:01i1</i>     | <i>DPB1*350:01</i>                | <i>DPB1*78:01</i>                 |
| <i>DPB1*131:01</i>      | <i>DPB1*29:01</i>       | <i>DPB1*168:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*131:01</i>      | <i>DPB1*416:01</i>      | <i>DPB1*460:01</i>                | <i>DPB1*461:01</i>                |
| <i>DPB1*131:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*168:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*131:01</i>      | <i>DPB1*02:01:04i2</i>  | <i>DPB1*17:01</i>                 | <i>DPB1*02:01:04i1</i>            |
| <i>DPB1*131:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*17:01</i>                 | <i>DPB1*46:01:01</i>              |
| <i>DPB1*131:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*47:01</i>                 |
| <i>DPB1*131:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*460:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*131:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*72:01i1</i>               |
| <i>DPB1*131:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*81:01</i>                 |
| <i>DPB1*133:01</i>      | <i>DPB1*463:01</i>      | <i>DPB1*13:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*133:01</i>      | <i>DPB1*46:01:01</i>    | <i>DPB1*13:01:01</i>              | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*133:01</i>      | <i>DPB1*47:01</i>       | <i>DPB1*13:01:01</i>              | <i>DPB1*47:01i1</i>               |
| <i>DPB1*133:01</i>      | <i>DPB1*81:01</i>       | <i>DPB1*13:01:01</i>              | <i>DPB1*81:01i1</i>               |
| <i>DPB1*133:01</i>      | <i>DPB1*02:01:04i1</i>  | <i>DPB1*13:01:01</i>              | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*133:01</i>      | <i>DPB1*72:01i1</i>     | <i>DPB1*13:01:01</i>              | <i>DPB1*72:01i2</i>               |
| <i>DPB1*133:01</i>      | <i>DPB1*29:01i1</i>     | <i>DPB1*13:01:01</i>              | <i>DPB1*29:01</i>                 |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1     | HLA-DPB1_Allele2       | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|----------------------|------------------------|----------------------------|----------------------------|
| <i>DPB1*133:01</i>   | <i>DPB1*78:01i1</i>    | <i>DPB1*13:01:01</i>       | <i>DPB1*78:01</i>          |
| <i>DPB1*138:01</i>   | <i>DPB1*17:01</i>      | <i>DPB1*23:01:01</i>       | <i>DPB1*460:01</i>         |
| <i>DPB1*138:01</i>   | <i>DPB1*17:01</i>      | <i>DPB1*460:01</i>         | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*51:01</i>      | <i>DPB1*23:01:01</i>       | <i>DPB1*51:01i1</i>        |
| <i>DPB1*138:01</i>   | <i>DPB1*51:01</i>      | <i>DPB1*23:01:01i1</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*138:01</i>   | <i>DPB1*02:01:04i2</i> | <i>DPB1*23:01:01</i>       | <i>DPB1*02:01:04i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*46:01:01i1</i> | <i>DPB1*23:01:01</i>       | <i>DPB1*46:01:01</i>       |
| <i>DPB1*138:01</i>   | <i>DPB1*46:01:01i1</i> | <i>DPB1*46:01:01</i>       | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*47:01i1</i>    | <i>DPB1*23:01:01</i>       | <i>DPB1*47:01</i>          |
| <i>DPB1*138:01</i>   | <i>DPB1*47:01i1</i>    | <i>DPB1*47:01</i>          | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*72:01i2</i>    | <i>DPB1*23:01:01</i>       | <i>DPB1*72:01i1</i>        |
| <i>DPB1*138:01</i>   | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*138:01</i>   | <i>DPB1*81:01i1</i>    | <i>DPB1*23:01:01</i>       | <i>DPB1*81:01</i>          |
| <i>DPB1*138:01</i>   | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*141:01</i>   | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*414:01</i>         |
| <i>DPB1*141:01</i>   | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*461:01</i>         |
| <i>DPB1*141:01</i>   | <i>DPB1*51:01i1</i>    | <i>DPB1*416:01</i>         | <i>DPB1*51:01</i>          |
| <i>DPB1*141:01</i>   | <i>DPB1*78:01i1</i>    | <i>DPB1*414:01</i>         | <i>DPB1*78:01</i>          |
| <i>DPB1*141:01</i>   | <i>DPB1*78:01i1</i>    | <i>DPB1*461:01</i>         | <i>DPB1*78:01</i>          |
| <i>DPB1*14:01:01</i> | <i>DPB1*23:01:01</i>   | <i>DPB1*14:01:01i1</i>     | <i>DPB1*23:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*29:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*29:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*33:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*33:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*39:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*39:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*46:01:01</i>   | <i>DPB1*14:01:01i1</i>     | <i>DPB1*46:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*47:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*47:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*51:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*59:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*78:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*81:01</i>      | <i>DPB1*14:01:01i1</i>     | <i>DPB1*81:01i1</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*14:01:01i1</i>     | <i>DPB1*72:01i2</i>        |
| <i>DPB1*14:01:01</i> | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*23:01:01i1</i> | <i>DPB1*23:01:01</i>       | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*29:01i1</i>    | <i>DPB1*29:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*33:01i1</i>    | <i>DPB1*33:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*39:01i1</i>    | <i>DPB1*39:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*46:01:01i1</i> | <i>DPB1*46:01:01</i>       | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*47:01i1</i>    | <i>DPB1*47:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*51:01i1</i>    | <i>DPB1*51:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*59:01i1</i>    | <i>DPB1*59:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*78:01i1</i>    | <i>DPB1*78:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*14:01:01</i> | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*14:01:01i1</i>     |
| <i>DPB1*168:01</i>   | <i>DPB1*02:01:04i2</i> | <i>DPB1*17:01</i>          | <i>DPB1*02:01:04i1</i>     |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*168:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*17:01</i>                 | <i>DPB1*46:01:01</i>              |
| <i>DPB1*168:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*47:01</i>                 |
| <i>DPB1*168:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*460:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*168:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*72:01i1</i>               |
| <i>DPB1*168:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*17:01</i>                 | <i>DPB1*81:01</i>                 |
| <i>DPB1*17:01</i>       | <i>DPB1*46:01:01</i>    | <i>DPB1*460:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*17:01</i>       | <i>DPB1*47:01</i>       | <i>DPB1*460:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*17:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*460:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*17:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*460:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*17:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*460:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*17:01</i>       | <i>DPB1*51:01i1</i>     | <i>DPB1*460:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*23:01:01</i>    | <i>DPB1*29:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*33:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*39:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*46:01:01</i>    | <i>DPB1*23:01:01i1</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*47:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*51:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*51:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*59:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*59:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*78:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*78:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*81:01</i>       | <i>DPB1*23:01:01i1</i>            | <i>DPB1*81:01i1</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*23:01:01i1</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*23:01:01</i>    | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*29:01i1</i>     | <i>DPB1*29:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*33:01i1</i>     | <i>DPB1*33:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*39:01i1</i>     | <i>DPB1*39:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*23:01:01</i>    | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*23:01:01i1</i>            |
| <i>DPB1*28:01</i>       | <i>DPB1*46:01:01</i>    | <i>DPB1*296:01</i>                | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*28:01</i>       | <i>DPB1*47:01</i>       | <i>DPB1*296:01</i>                | <i>DPB1*47:01i1</i>               |
| <i>DPB1*28:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*296:01</i>                | <i>DPB1*81:01i1</i>               |
| <i>DPB1*28:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*296:01</i>                | <i>DPB1*02:01:04i2</i>            |
| <i>DPB1*28:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*296:01</i>                | <i>DPB1*72:01i2</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*33:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*33:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*350:01</i>      | <i>DPB1*415:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*350:01</i>      | <i>DPB1*459:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*350:01</i>      | <i>DPB1*464:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*39:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*39:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*417:01</i>      | <i>DPB1*462:01</i>                | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*463:01</i>      | <i>DPB1*105:01i1</i>              | <i>DPB1*29:01i1</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*29:01</i>       | <i>DPB1*46:01:01</i>    | <i>DPB1*29:01i1</i>               | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*29:01</i>       | <i>DPB1*47:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*47:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*51:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*51:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*59:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*59:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*78:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*78:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*29:01i1</i>               | <i>DPB1*81:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*29:01i1</i>               | <i>DPB1*72:01i2</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*33:01i1</i>     | <i>DPB1*33:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*39:01i1</i>     | <i>DPB1*39:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*29:01</i>       | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*29:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*39:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*39:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*46:01:01</i>    | <i>DPB1*33:01i1</i>               | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*33:01</i>       | <i>DPB1*47:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*47:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*51:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*51:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*59:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*59:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*78:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*78:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*33:01i1</i>               | <i>DPB1*81:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*33:01i1</i>               | <i>DPB1*72:01i2</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*39:01i1</i>     | <i>DPB1*39:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*33:01</i>       | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*33:01i1</i>               |
| <i>DPB1*350:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*415:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*350:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*459:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*350:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*464:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*351:01</i>      | <i>DPB1*414:01</i>      | <i>DPB1*416:01</i>                | <i>DPB1*104:01i1</i>              |
| <i>DPB1*351:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*104:01i1</i>              | <i>DPB1*51:01i1</i>               |
| <i>DPB1*352:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*416:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*39:01</i>       | <i>DPB1*46:01:01</i>    | <i>DPB1*39:01i1</i>               | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*39:01</i>       | <i>DPB1*47:01</i>       | <i>DPB1*39:01i1</i>               | <i>DPB1*47:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*51:01</i>       | <i>DPB1*39:01i1</i>               | <i>DPB1*51:01i1</i>               |

**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| <b>HLA-DPB1_Allele1</b> | <b>HLA-DPB1_Allele2</b> | <b>HLA-DPB1_Alternate Allele1</b> | <b>HLA-DPB1_Alternate Allele2</b> |
|-------------------------|-------------------------|-----------------------------------|-----------------------------------|
| <i>DPB1*39:01</i>       | <i>DPB1*59:01</i>       | <i>DPB1*39:01i1</i>               | <i>DPB1*59:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*78:01</i>       | <i>DPB1*39:01i1</i>               | <i>DPB1*78:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*39:01i1</i>               | <i>DPB1*81:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*39:01i1</i>               | <i>DPB1*72:01i2</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*39:01i1</i>               |
| <i>DPB1*39:01</i>       | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*39:01i1</i>               |
| <i>DPB1*414:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*416:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*416:01</i>      | <i>DPB1*51:01</i>       | <i>DPB1*461:01</i>                | <i>DPB1*51:01i1</i>               |
| <i>DPB1*417:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*462:01</i>                | <i>DPB1*78:01i1</i>               |
| <i>DPB1*417:01</i>      | <i>DPB1*51:01i1</i>     | <i>DPB1*462:01</i>                | <i>DPB1*51:01</i>                 |
| <i>DPB1*463:01</i>      | <i>DPB1*78:01</i>       | <i>DPB1*105:01i1</i>              | <i>DPB1*78:01i1</i>               |
| <i>DPB1*463:01</i>      | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*105:01i1</i>              |
| <i>DPB1*463:01</i>      | <i>DPB1*46:01:01i1</i>  | <i>DPB1*46:01:01</i>              | <i>DPB1*105:01i1</i>              |
| <i>DPB1*463:01</i>      | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*463:01</i>      | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*105:01i1</i>              |
| <i>DPB1*463:01</i>      | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*105:01i1</i>              |
| <i>DPB1*46:01:01</i>    | <i>DPB1*47:01</i>       | <i>DPB1*46:01:01i1</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*51:01</i>       | <i>DPB1*46:01:01i1</i>            | <i>DPB1*51:01i1</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*59:01</i>       | <i>DPB1*46:01:01i1</i>            | <i>DPB1*59:01i1</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*78:01</i>       | <i>DPB1*46:01:01i1</i>            | <i>DPB1*78:01i1</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*81:01</i>       | <i>DPB1*46:01:01i1</i>            | <i>DPB1*81:01i1</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*72:01i1</i>     | <i>DPB1*46:01:01i1</i>            | <i>DPB1*72:01i2</i>               |
| <i>DPB1*46:01:01</i>    | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*47:01i1</i>     | <i>DPB1*47:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*59:01i1</i>     | <i>DPB1*59:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*72:01i2</i>     | <i>DPB1*72:01i1</i>               | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*78:01i1</i>     | <i>DPB1*78:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*46:01:01</i>    | <i>DPB1*81:01i1</i>     | <i>DPB1*81:01</i>                 | <i>DPB1*46:01:01i1</i>            |
| <i>DPB1*47:01</i>       | <i>DPB1*51:01</i>       | <i>DPB1*47:01i1</i>               | <i>DPB1*51:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*59:01</i>       | <i>DPB1*47:01i1</i>               | <i>DPB1*59:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*78:01</i>       | <i>DPB1*47:01i1</i>               | <i>DPB1*78:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*81:01</i>       | <i>DPB1*47:01i1</i>               | <i>DPB1*81:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*02:01:04i1</i>  | <i>DPB1*02:01:04i2</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*72:01i1</i>     | <i>DPB1*47:01i1</i>               | <i>DPB1*72:01i2</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*02:01:04i2</i>  | <i>DPB1*02:01:04i1</i>            | <i>DPB1*47:01i1</i>               |
| <i>DPB1*47:01</i>       | <i>DPB1*51:01i1</i>     | <i>DPB1*51:01</i>                 | <i>DPB1*47:01i1</i>               |



**Table M-7.** List of *HLA-DPB1* allele combination phase ambiguities according to v.3.25.0 IPD-IMGT/HLA database (released July 2016).

| HLA-DPB1_Allele1       | HLA-DPB1_Allele2       | HLA-DPB1_Alternate Allele1 | HLA-DPB1_Alternate Allele2 |
|------------------------|------------------------|----------------------------|----------------------------|
| <i>DPB1*47:01</i>      | <i>DPB1*59:01i1</i>    | <i>DPB1*59:01</i>          | <i>DPB1*47:01i1</i>        |
| <i>DPB1*47:01</i>      | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*47:01i1</i>        |
| <i>DPB1*47:01</i>      | <i>DPB1*78:01i1</i>    | <i>DPB1*78:01</i>          | <i>DPB1*47:01i1</i>        |
| <i>DPB1*47:01</i>      | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*47:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*59:01</i>      | <i>DPB1*51:01i1</i>        | <i>DPB1*59:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*78:01</i>      | <i>DPB1*51:01i1</i>        | <i>DPB1*78:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*81:01</i>      | <i>DPB1*51:01i1</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*72:01i1</i>    | <i>DPB1*51:01i1</i>        | <i>DPB1*72:01i2</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*51:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*59:01i1</i>    | <i>DPB1*59:01</i>          | <i>DPB1*51:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*51:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*78:01i1</i>    | <i>DPB1*78:01</i>          | <i>DPB1*51:01i1</i>        |
| <i>DPB1*51:01</i>      | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*51:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*78:01</i>      | <i>DPB1*59:01i1</i>        | <i>DPB1*78:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*81:01</i>      | <i>DPB1*59:01i1</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*72:01i1</i>    | <i>DPB1*59:01i1</i>        | <i>DPB1*72:01i2</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*59:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*59:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*78:01i1</i>    | <i>DPB1*78:01</i>          | <i>DPB1*59:01i1</i>        |
| <i>DPB1*59:01</i>      | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*59:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*81:01</i>      | <i>DPB1*78:01i1</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*72:01i1</i>    | <i>DPB1*72:01i2</i>        | <i>DPB1*78:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*78:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*78:01i1</i>        |
| <i>DPB1*78:01</i>      | <i>DPB1*81:01i1</i>    | <i>DPB1*81:01</i>          | <i>DPB1*78:01i1</i>        |
| <i>DPB1*81:01</i>      | <i>DPB1*02:01:04i1</i> | <i>DPB1*02:01:04i2</i>     | <i>DPB1*81:01i1</i>        |
| <i>DPB1*81:01</i>      | <i>DPB1*72:01i1</i>    | <i>DPB1*72:01i2</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*81:01</i>      | <i>DPB1*02:01:04i2</i> | <i>DPB1*02:01:04i1</i>     | <i>DPB1*81:01i1</i>        |
| <i>DPB1*81:01</i>      | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*81:01i1</i>        |
| <i>DPB1*02:01:04i1</i> | <i>DPB1*72:01i1</i>    | <i>DPB1*02:01:04i2</i>     | <i>DPB1*72:01i2</i>        |
| <i>DPB1*02:01:04i1</i> | <i>DPB1*72:01i2</i>    | <i>DPB1*72:01i1</i>        | <i>DPB1*02:01:04i2</i>     |

**Notes:** Table originally prepared and courtesy by: Y. Thorstenson (Senior Data Analyst) – Immucor, Inc.

Suffixes shown here in some of the HLA allele names correspond to:

a) MIA FORA™ NGS FLEX HLA Genotyping Software cloned and sequenced alleles:

e: Sequences with the suffix e (e1, e2, etc.) are those with new intronic sequence not represented in the IMGT database. The vast majority of cloned sequences are of this type.

v: Sequences with the suffix v (v1, v2, etc.) are a small subset of cloned alleles that contain new intron variants relative to existing genomic sequences in the IMGT database.

x: Sequences with the suffix x (x1, x2, etc.) are a small subset of cloned alleles that contain new exon variants. The suffix is added to the closest known reference sequence but if confirmed by IMGT the allele name will change.

b) MIA FORA™ NGS FLEX HLA Genotyping Software in-silico sequences:

Many IMGT reference sequences contain partial exon sequences. To facilitate data analysis, the closest complete exon was copied to fill in the gaps in IMGT reference sequences with an incomplete exon. The suffix i (i1, i2, etc.) is used to identify those computationally filled sequences.

Courtesy: <http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf>

- Finally, although not as an ambiguity but as a retrospective correction comment, it is also noteworthy the update of a deleted *HLA-DPA1* allele on the IPD-IMGT/HLA database:

Also, according to IPD-IMGT/HLA database version 3.25.0 there was originally an allele named as *HLA-DPA1\*02:02:01* (as it is described and left in the current study). However, later studies determined that sequence has now been shown to be in error and is identical to *HLA-DPA1\*02:07:01* (defined as the correct allele name in March 2017 by IPD-IMGT/HLA; see in <https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/deleted.cgi>).

### **3. STATISTICAL ANALYSES**

Upon completion of generation of 11 HLA loci genotyping results per sample by the MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0 (Immucor, Inc. Norcross, GA, USA)) program and, then, manual review (including annotation of these ambiguities, applying respective above mentioned standardization criteria and managing HLA genotyping data according to guidelines from [351] and [770]) by the user; a series of statistical analyses were performed for all NGS-based HLA genotyping data generated and related HLA datasets received from collaborators for this study

(see 1. STUDY POPULATION, DESIGN AND DATA COLLECTION). Performed statistical analyses can be categorized in two main groups:

- i) Population genetics statistical analyses of this immunogenetic as well as analysis methods for data primary quality control (QC) and verification of the integrity of genotyping data (as references used here, see [771][772]).
- ii) Analytical methods for HLA-disease association study with this immunogenetic data, where case-control studies were carried out (as references/examples used here, see [773][774]).

In detail:

### **3.1 Hardy-Weinberg Equilibrium Proportions (HWEP) Test**

Calculated allele frequencies (see next point **3.2 HLA Allele Frequencies Calculation**) at each HLA locus were evaluated for deviations from Hardy-Weinberg Equilibrium Proportions (HWEP) using the exact test of Guo and Thompson [775]. HWEP test allows to measure the degree to which observed genotype frequencies differ from those expected based on the allele frequencies for a given population study cohort, assuming that this studied population is suitably large (e.g. N~100 study subjects) and experiences random mating [776]. In a Hardy–Weinberg test, observed genotype counts are compared to those expected under HWEP, as calculated by generating a table of all possible genotypes, using an appropriate statistical method. The relationship between the allele and genotype frequencies under HWEP is given as (as it is described in [771]):

$$f(A_iA_i) = p_i^2$$

and

$$f(A_iA_j) = 2p_i p_j$$

Where  $p_i$  is the allele frequency of  $A_i$ . and  $p_j$  is the allele frequency of  $A_j$ . When a population is in HWE, there will not be a significant departure from these allele and genotype frequencies and there will be no change in allele frequencies between generations.

Thus, Hardy–Weinberg (HW) principle provides a useful model for primary quality control (QC) verification of the integrity of genotype data, as genotyping errors may result in both individual genotype deviations and overall deviations from HW equilibrium (HWE) [771]. Generally, departure from HWEP may occur most frequently due to genotyping errors (e.g., failure to detect a specific allele, resulting in an excess of homozygotes). Another common reason is the presence of an unknown allele which is not considered in the genotyping scheme (null allele). This happens when a variant is considered to be bi-allelic while it is actually multiallelic. Secondly, operation of selection, admixture, or nonrandom mating could also cause deviation from HWEP. In a case-control association study, it is of paramount importance that the control group does not show deviations from HWEP in order to rule out any technical errors and to avoid false-positive associations [777]. PyPop (Python for Population genomics) version 0.7.0 software was used for carrying out these HWEP tests [778]. PyPop is a framework for performing large-scale population genetic analyses on multilocus genotype data. It contains several programs and an Application Programming Interface (API) implemented in the programming language Python [778].

### **3.2 HLA Allele Frequencies Calculation**

HLA allele frequencies (or distributions) per locus for HLA genotyping data were obtained via direct counting, where the number of observations for a given allele is divided by the number

of chromosomes ( $2N$ , where  $N$  = sample size) under study [771]. PyPop version 0.7.0 software was also used to calculate allele frequencies in this manner [778].

### **3.3 Pairwise Linkage Disequilibrium Estimation**

As it is described in [771] and also very well applied as an example in [772]:

Measurement of linkage disequilibrium (LD) provides a means to assess the degree to which pairs of alleles are likely to be observed on the same haplotype and has important implications in analyzing immunogenetic data for population and disease association studies. Two different categories of LD estimations were performed in the present study:

- Haplotype-Level LD statistics:

The normalized (because in order to account for differing allele frequencies at the loci, a normalized disequilibrium value is commonly used) allele-pair-level LD measure,  $D'_{ij}$ , is the disequilibrium coefficient ( $D$ ) divided by the upper and lower bounds of  $D$  for the particular alleles at each locus (as described in [779-782]), and ranges from +1 to -1. A  $D'_{ij}$  value of 0 indicates linkage equilibrium, whereas a value of +1 indicates the complete association of a given pair of alleles in a single haplotype, and, thus, a value of exactly -1 indicates the complete absence of an haplotype comprised by those alleles. As a note, the complete absence of a particular haplotype can be inferred from a  $D'_{ij}$  value of -1 when none of the reported alleles has a frequency greater than 0.5. In the present study, pairwise LD estimate or the locus-pair-level measure, termed as  $D'$  was calculated according to Hedrick's  $D'$  statistic [98][780][787], based on the products of the allele frequencies at each locus to weight the LD contribution of specific allele pairs, using PyPop version 0.7.0 software [778].

- Global LD statistics:

For loci with more than two alleles, global LD statistics extend the haplotype-level statistics to account for all possible combinations of alleles at each locus [783]. Two different types of parameters were calculated here:

(a)  $W_n$  (Cramer's V statistic): is a multiallelic extension of the correlation measure  $r$  (correlation coefficient ( $r$ ) between the alleles at the  $p$  and  $q$  loci) [784]. The chi-square value for testing the significance of LD can be written as  $W/(2 N)$  where:

$$W = (\sum \sum D_{ij} / p_i \times q_j)^{1/2}$$

Where  $p_i$  and  $q_j$  are the observed allele frequencies at each of the two loci having  $k$  and  $l$  alleles, respectively.  $W_n$ , or Cramer's V statistic, is a normalized value that addresses differing numbers of alleles at the two loci [785][786].

$$W_n = W / \sqrt{(\min(k, l) - 1)}$$

The values of  $W_n$  fall between 0 and 1, and the significance of the overall disequilibrium is assessed using the above mentioned chi-square test. It should be noted that the  $W_n$  measure is always symmetric with respect to two loci, whereas the number of alleles reported at each locus can differ considerably. It is therefore important not to overinterpret values of  $W_n$  for locus pairs with highly asymmetric numbers of alleles. Finally, for biallelic loci,  $W_n$  is equivalent to  $r$  parameter.

(b)  $D'$ : this parameter is a second global disequilibrium statistic, which sums the absolute value of normalized  $D_{ij}$  values over all haplotypes, weighted by the frequencies of the alleles in each haplotype [787]. As with  $W_n$ ,  $D'$  values fall between 0 (equilibrium) and 1 (linkage). This is defined as:

$$D' = \frac{\sum p_i \times q_j |D'_{ij}|}{\sum p_i \times q_j}$$

In the present study, PyPop version 0.7.0 software [778] was also used to calculate all these parameters' ( $D_{ij}$ ,  $D'_{ij}$ ,  $D'$  and  $W_n$ ) values.

More recently, although out of the scope of the present thesis work, Thomson and Single described a new asymmetric pair of LD measures (ALD) that give a more complete description of LD [788]. The ALD measures are symmetric and equivalent to the correlation coefficient  $r$  when both loci are bi-allelic. When the numbers of alleles at the two loci differ, the ALD measures capture this asymmetry and provide additional detail about the LD structure. In disease association studies the ALD measures are useful for identifying additional disease genes in a genetic region, by conditioning on known effects. In evolutionary genetic studies ALD measures provide insight into selection acting on individual amino acids of specific genes, or other loci in high LD [788][789].

### **3.4 Ewens-Watterson Homozygosity (EWH) Test**

In the present study, using PyPop version 0.7.0 software [778], the Ewens-Watterson Homozygosity (EWH) test of neutrality was carried out with Slatkin's implementation of the Monte-Carlo approximation using a two-tailed test ( $p < 0.05$ ) of the null hypothesis of neutrality. This EWH test allows the measurement of selection operating on studied HLA loci. In detail, as it is described in [771] and also very well applied in [772]:

- Firstly, in this test the expected proportion of homozygotes under HWEP, for an observed value of  $k$  (number of distinct alleles ( $k$ ) for each given HLA locus) and a given sample size ( $N$ ), is used as a measure of the allele-frequency distribution and compared to the distribution expected under the neutral model for the same values of  $k$  and  $N$  [790]. Allele-frequency

distributions are used to calculate Watterson's homozygosity  $F$  statistic [791]. This is given by:

$$F = \sum p_i^2$$

Where  $p_i$  is the frequency of the  $i$ -th allele at a locus. The homozygosity test can be accomplished using the exact test described by Slatkin's Monte-Carlo implementation [792][793]. For given values of  $N$  (sample size) and  $k$  (number of distinct alleles for a given HLA locus), all possible configurations of alleles are listed (each configuration is a distinct way of distributing the  $N$  sampled genes into  $k$  allelic categories). The probability of obtaining a particular configuration can be computed under the null hypothesis of neutrality using the Ewens sampling formula [790]. The homozygosity value of each configuration along with its probability gives the sampling distribution for  $F$  under neutrality. This distribution is used to find the probability of obtaining homozygosity values equal to or larger than that observed, for a test of positive selection, by examining how many configurations result in homozygosities greater than this observed value [792]. Similarly, a test of balancing selection is based on the probability of obtaining a homozygosity value as small as or smaller than the observed value. Significant  $p$ -values of  $F$  reject the null hypothesis that the sample came from a population that is undergoing neutral evolution.

- At the same time, homozygosity values calculated for different values of  $N$  (sample size) and  $k$  (number of distinct alleles for a given HLA locus), can be directly compared by calculating the normalized deviate of homozygosity ( $F_{nd}$ ) [794]. This is given by (the difference between the observed and expected values of  $F$ , divided by the square root of the variance of the expected  $F$ ):



$$F_{nd} = (F_{obs} - F_{exp}) / \sqrt{\text{var}(F_{exp})}$$

Where  $F_{obs}$  (or  $F_o$ ) is the homozygosity value calculated for an observed frequency distribution,  $F_{exp}$  (or  $F_e$ ) is the mean homozygosity expected under the neutral model. While  $F_{nd}$  is a normalized deviate, the sampling distribution for  $F_{nd}$  is not normally distributed, so that p-values of  $F$  cannot be inferred from a given  $F_{nd}$  value using traditional parametric methods. Statistical significance for an  $F_{nd}$  value is given by the significance of the corresponding  $F_{obs}$  value. The normalized deviate of homozygosity ( $F_{nd}$ ) can also be used to characterize homozygosity values that deviate significantly from the null hypothesis in terms of modes of evolution.  $F_{nd}$  values significantly lower than 0 result from allele-frequency distributions that are more “even” than expected and are consistent with the action of balancing selection (commonly found in the HLA system).  $F_{nd}$  values significantly higher than 0 result from allele-frequency distributions that are more skewed than expected toward specific alleles and are consistent with either directional selection and/or an extreme demographic effect. Because the null-hypothesis of the EWH test is neutral evolution ( $F_{nd} = 0$ ), we used a paired sign test [795] to compare the signs of the  $F_{nd}$  values for each locus against the expectation of neutrality. To correct for the number of comparisons, the results of these tests were considered significant if the associated p-values were lower than 0.05 (indicating a statistical significance at the 5% level) or 0.01 (indicating a statistical significance at the 1% level).

### 3.5 Extended HLA Haplotype Frequencies Inference via Expectation-Maximization (EM)

#### Algorithm

Due to the nature and design of the population cohorts (healthy controls and MS disease group) of the present study, HLA haplotypes were phased-unknown since were coming from unrelated individuals. Thus, extended haplotype (encompassing different number of HLA loci (n-locus)

in each case) frequencies were estimated using the iterative expectation-maximization (EM) algorithm [333][334]. In detail:

- Here, population-level haplotype frequencies are estimated via EM using simultaneous maximum-likelihood estimation of n-locus haplotype frequencies. The expectation step determines the expected number of copies for each haplotype contributing to a given genotype. As an example, for a three locus (3-locus) haplotype, this is calculated as:

$$E[n_{abc}|P_i] = 2f_{abc}Sf_{abc}/Pr(P_i).$$

Where  $S$  is the number of ambiguous haplotypes in  $P_i$ ,  $E[n_{abc}|P_i]$  is the expected number of copies of haplotype  $H_{abc}$  within  $P_i$ , and  $f_{abc}$  is the frequency of each other possible haplotype  $H_{abc}$  to form the genotype of frequency  $P_i$ . The maximization step determines new estimates for  $f_{abc}$  for the next iteration of the algorithm. Thus, at each new iteration the estimations globally improve.

- The open-source software Hapl-o-Mat version 1.1 [342] was used to estimate these extended haplotype frequencies from the HLA genotyping datasets of the present study. This software uses a maximum likelihood estimation via an EM algorithm. Its key features are the processing of different HLA genotyping dataset resolutions and n-locus haplotypes within a given population sample and the handling of ambiguities recorded via multiple allele codes or genotype list strings. Implemented in C++, Hapl-o-Mat facilitates efficient haplotype frequency estimation from large amounts of genotype data. Its accuracy and performance has been previously tested and reported [342].

• As previously mentioned, estimated haplotype frequencies from the HLA genotyping datasets of unrelated population cohorts via an EM algorithm present inaccuracies since the performance of haplotype frequency estimation algorithms is sensitive to various aspects of the population under study, which are important to be considered when analyzing the respective output results [771][772]:

(a) Estimated frequencies for rare haplotypes (e.g.  $n = 1$  or  $2$  in a dataset), which incorporate low-frequency alleles, are often incorrect, even when the EM algorithm finds the global maximum likelihood. Thus, analytical inferences should not be made on the basis of these rare haplotypes.

(b) Diversity and complexity of HLA genotyping data poses additional challenges for haplotype estimation, such as: heterogeneity of typing resolution, heterogeneity of typing techniques, heterogeneity of allele nomenclatures, incessant discovery of new alleles, large numbers of allele per locus and high allele/haplotype diversity among human populations and regional groups (geographical diversity in HLA allele/haplotype frequency distributions). Furthermore, population substructure and/or regional variation are commonly found in HLA population-level studies owing to the fact that HLA allele/haplotype frequency distributions (and, thus, corresponding LD patterns) can reflect both the selective pressures and demographic events (which, in turn, cannot be always and totally distinguished) over human populations (even within sub-regions/sub-groups of the same considered population/region/ethnic group). Taking this into account is of critical importance, as some examples, in HLA case-control or anthropological studies. Since these HLA population singularities may be confounding in the investigation of HLA-disease association or of populations relatedness (based on these HLA allele/haplotype frequency distributions), leading erroneously to HLA susceptible/protective allele/haplotypes for a

given disease-phenotype [383][402] or mistakenly to close relations between populations respectively [542][543].

### 3.6 Genetic Distances and Dendograms

In order to study Spanish regional relatedness by comparison of allele frequencies (computing the *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* loci) between the 3 different geographical Spanish regions established for this study (Northern-Central, Eastern- and Southern-Spain) as well as between the 10 Spanish locations studied in the present work, a population dendrogram (or phylogenetic trees) was constructed using POPTREEW (web version of POPTREE software) [796]. A total of 1000 dendrogram replicates based on the matrices of Nei genetic distances ( $D_A$ ) [797] were generated using the neighbor-joining (NJ) method [798]. The root of the NJ method was calculated by the mid-point rooting method, in which the root is placed in the mid-point of the longest path of two taxa.

Nei genetic distance ( $D_A$ ) [797] is defined by:

$$D_A = 1 - \frac{1}{r} \sum_j \sum_i^{m_j} \sqrt{x_{ij} y_{ij}}$$

Where  $x_{ij}$  and  $y_{ij}$  are the frequencies of the  $i$ -th allele at the  $j$ -th locus in populations  $X$  and  $Y$ , respectively,  $m_j$  is the number of alleles at the  $j$ -th locus, and  $r$  is the number of loci used.

### 3.7 Case-Control Analyses for HLA-disease Association Study

Statistical analyses for the case-control studies were performed using R language for statistical computing with the *BIGDAWG* v.2.1 R package [799][800]. Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG), developed by Pappas et al. [799], is an automated software pipeline that performs a suite of common case-control analyses of multi-locus highly

polymorphic genetic data (e.g. HLA genes and KIR genes). BIGDAWG is integrated in the framework of the R statistical environment (<http://www.r-project.org>). As described in [799], BIGDAWG is able to accept unambiguous genotype data for case-control groups as input, and to estimate allele/haplotype frequencies to be used for chi-square ( $\chi^2$ ) testing. Thus, it also calculates odds ratios (ORs), confidence intervals (typically, 95%CI) and p-values (associated probability (p) values derived from a two-tailed Fischer's exact test; where  $p \leq 0.05$  (specific criteria applied in this thesis work) or  $p \leq 0.01$  are considered in general statistically significant) for each allele/haplotype, whose effect sizes are measured here to evaluate HLA-disease association (both susceptibility and protection). At the end of the analysis run, BIGDAWG generates respective output tables for each of these comparisons. More in detail, BIGDAWG performs overall locus-level ( $k \times 2$ ) tests of significance, followed by a series of allele-level ( $2 \times 2$ ) tests of significance for each locus. In addition, the designated healthy control group is tested for deviations from expected Hardy–Weinberg Equilibrium Proportions (HWEP) at the allele level. When multi-locus genotype data is available, BIGDAWG can estimate user-designated haplotypes and performs the same statistical calculations for each haplotype [ $k \times 2$  tests at the multi-locus level (e.g. *HLA-A–HLA-B* or *HLA-DRB1–HLA-DQA1–HLA-DQB1*) followed by  $2 \times 2$  tests at the haplotype level].

It is noteworthy that an important consideration has to be applied for rare HLA alleles/haplotypes found within the HLA genotyping data of the study cohort of interest. Since the  $\chi^2$  test can lead to false acceptance or rejection of the null hypothesis when the expected genotype counts in a contingency table are small (sparse cells) as Hollenbach et al. described [773]. Thus, the  $\chi^2$  test is inappropriate if any of the expected counts are less than one or if the expected counts are less than five in more than 20% of all cells in a contingency table [774].

Consequently, BIGDAWG combines (as the statistical strategy employed in this case) rare

alleles/haplotypes into a common class (a process called “binning”) being thus included as well for testing.

In relation to the statistical methods (and formulas) and the R statistical environment specifics used by this BIGDAWG program, it is of note the following points:

- The  $\chi^2$  statistic for a contingency table analysis of case-control data for a given genetic association is calculated as (being fully described in [773]):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where

$O_i$  = the observed count of allele/haplotype  $i$

$E_i$  = the expected count of allele/haplotype  $i$

And the derived values are summed over all cells in the tables. The expected count for each cell in the  $r \times c$  table is calculated as

$$\frac{(\text{row total allele } i) \times (\text{column total})}{2n}$$

Where, in this formula shown right above,

Column total = sum of the counts in the column

Row total = number observations of allele  $i$  in all subjects

$n$  = number individuals (cases + controls)

$2n$  = number chromosomes (cases + controls)

The degrees of freedom (*df*) for the goodness of fit  $\chi^2$  analysis are calculated from the number of alleles/haplotypes with expected values in cases and controls of five or greater, plus the combined category,  $-1$  (i.e.,  $k - 1$ ). A p-value is obtained by comparing the test statistic to the  $\chi^2$  distribution for the appropriate degrees of freedom.

Overall, the main and basic steps for contingency table analysis in case-control studies with immunogenetic data are:

- a) Construct a  $2 \times k$  table of allele (or genotype or haplotype) counts for cases and controls.
- b) Combine all alleles/haplotypes with expected values of four or less in cases or controls into a common “binned” category.
- c) Calculate the  $\chi^2$  test statistic for the table and assess significance.
- d) If results are significant at the overall locus/haplotype level, perform additional testing using  $2 \times 2$  contingency tables for each allele/haplotype (expected counts of five or more only) against all other alleles/haplotypes.

Analysis based on a series of  $2 \times 2$  tests commonly requires corrections for multiple comparisons (usually utilizing the Bonferroni’s method for correction for multiple comparisons (multiplying the value of p obtained in the statistical test by the total number of alleles tested or clinical characteristics)), with a correction factor minimally equivalent to the number of alleles tested. However, when significance of the association of all individual alleles is assessed with the a priori knowledge of overall heterogeneity at the locus, it is not necessary to correct for multiple comparisons in subsequent  $2 \times 2$  tests intended to identify the allele(s)/haplotype(s) with significant contributions to the overall

deviation at the locus/haplotype (as it was considered for the case of the present thesis work and related case-control analyses).

- All HWEP and phenotype association (haplotype and locus) analyses are currently based on a traditional  $\chi^2$  test. For HWEP deviation testing, BIGDAWG combines rare genotypes into a single common class (binning) for analysis and performs a goodness-of-fit test. The degrees of freedom (*dof*) are calculated as:

$$\text{dof} = g - (a - 1)$$

Where *g* is the number of unique non-binned genotypes and *a* is the number of unique non-binned alleles.

- For testing phenotype associations, BIGDAWG runs a test-of-independence, automatically tabulating the  $k \times 2$  contingency tables, where *k* is the number of unique haplotypes or alleles. For either testing scenario, rare cells (with expected counts less than five) are combined into a common class (binned) prior to computing the  $\chi^2$  statistic, except in cases of the test-of-independence where all cells of a given  $k \times 2$  contingency table are  $\geq 1$  and fewer than 20% of the cells have expected counts less than five. BIGDAWG's haplotype frequency estimation function requires the R *haplo.stats* package, whereas calculation of the individual haplotype/allele confidence intervals (95%CI), odds ratios (ORs), and p-values requires the R *epicalc* package. For the HLA loci, BIGDAWG performs the same analyses at the individual amino-acid/residue level. However, this residue analysis was beyond the scope of the present thesis work.





## ***RESULTS***



---

## I. NGS-BASED HLA STUDY IN 17TH-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)

### 1. EVALUATION OF CONCORDANCE OF HLA GENOTYPING RESULTS IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)

Samples of this cohort were tested in parallel with several available methods (SBT, SSO, SSP or NGS) to confirm the respective genotyping results and to identify possible causes of discrepancies. HLA genotyping results obtained for all 282 samples by using this aforementioned commercially available NGS-based HLA genotyping method [187] at Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory were 100% concordant with those available HLA typing results (e.g. *HLA-DPA1* locus was not tested locally in the Spanish institutions) obtained by using other HLA molecular typing techniques (either using an in-house NGS platform or commercial/in-house SSO, SSP or SBT technologies) respectively at the 11 local participating clinical laboratories from Spain (see **Table R-1**). Therefore, it was confirmed that all samples were tested correctly by all the participating laboratories without any sample-switching error, allele dropout (for the HLA loci tested respectively) and neither contamination.

**Table R-1.** Concordance rates\* of HLA genotyping results of this 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects) when comparing HLA genotyping results obtained by using this commercially available NGS-based typing method [187] performed at the Stanford Blood Center HLA Histocompatibility and Immunogenetics Laboratory with those available HLA typing results (with a variable range of allele resolution level and number of HLA genes tested) obtained by using other HLA molecular typing techniques (either an in-house NGS platform or commercial/in-house SSO, SSP or SBT technologies) respectively at the 11 local participating clinical laboratory groups from Spain (de-identified and coded here **A** through **K**).

| <b>Locus/<br/>Group</b> | <b>HLA-<br/>A</b> | <b>HLA-<br/>B</b> | <b>HLA-<br/>C</b> | <b>HLA-<br/>DPA1</b> | <b>HLA-<br/>DPB1</b> | <b>HLA-<br/>DQA1</b> | <b>HLA-<br/>DQB1</b> | <b>HLA-<br/>DRB1</b> | <b>HLA-<br/>DRB3</b> | <b>HLA-<br/>DRB4</b> | <b>HLA-<br/>DRB5</b> |
|-------------------------|-------------------|-------------------|-------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| <b>A</b>                | 100               | 100               | 100               | NT                   | NT                   | 100                  | 100                  | 100                  | NT                   | NT                   | NT                   |
| <b>B</b>                | 100               | 100               | 100               | NT                   | NT                   | 100                  | 100                  | 100                  | NT                   | NT                   | NT                   |
| <b>C</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | NT                   | 100                  | 100                  | 100                  | 100                  |
| <b>D</b>                | 100               | 100               | 100               | NT                   | NT                   | 100                  | 100                  | 100                  | 100                  | 100                  | 100                  |
| <b>E</b>                | 100               | 100               | 100               | NT                   | 100                  | NT                   | 100                  | 100                  | 100                  | 100                  | 100                  |
| <b>F</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | 100                  | 100                  | NT                   | NT                   | NT                   |
| <b>G</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | 100                  | 100                  | NT                   | NT                   | NT                   |
| <b>H</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | 100                  | 100                  | NT                   | NT                   | NT                   |
| <b>I</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | NT                   | 100                  | NT                   | NT                   | NT                   |
| <b>J</b>                | 100               | 100               | 100               | NT                   | NT                   | NT                   | NT                   | 100                  | NT                   | NT                   | NT                   |
| <b>K</b>                | 100               | 100               | 100               | NT                   | NT                   | 100                  | 100                  | 100                  | NT                   | NT                   | NT                   |

NT = Not tested at the Spanish local participating clinical laboratory group

100 = 100% concordance

\*Concordance rates criteria is based on how identical the HLA genotyping results are in this comparison and it is based on the maximum available allele resolution level (either 2-field, 3-field or 4-field) obtained by the respective HLA molecular typing method performed at each participating Spanish local institution group.

## **2. EVALUATION OF DEVIATIONS FROM EXPECTED HARDY-WEINBERG EQUILIBRIUM PROPORTIONS (HWE) IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

At the 3-/4-field allele resolution level, no overall deviations from expected Hardy-Weinberg Equilibrium Proportions (HWE) are observed in any of the *HLA* loci analyzed with the exception of a minor but significant departure at the *HLA-DPA1* locus (p-value = 0.0104) (**Table R-2**). To further investigate this *HLA-DPA1* departure, “collapsed” or “trimmed” (i.e. observed alleles in the present study were reduced to 2- and 3-field allele resolution level respectively for comparative analysis purposes) 2-field and 3-field *HLA* genotyping datasets of this same Spanish population cohort (n=282) were evaluated (*HLA-DPA1* locus 2-field p value of HWE = 0.1513; *HLA-DPA1* locus 3-field p value of HWE = 0.1141) and no HWE deviation was observed at any of the *HLA* loci. Furthermore, estimated homozygosity (Ewens-Watson’s homozygosity *F* statistic (*F*)) in *HLA-DPA1* locus at the 4-field allele resolution level shows a much lower value (*F*=0.164) in comparison to collapsed 2-field (*F*=0.649) and collapsed 3-field (*F*=0.635) *HLA* genotyping datasets. Altogether, this can be interpreted as estimated deviations from HWE may not be corrected properly for multiple comparisons including low number counts of alleles or genotypes when they are present at the 4-field allele resolution level. Thus, this observed deviation may be explained by the fact that *HLA* alleles presenting low frequencies (e.g. *HLA-DPA1* alleles) would not be considered properly when evaluating HWE proportions and their contribution to HWE deviation would be being estimated higher than it should be at this 4-field allele resolution level. Overall, the *HLA* dataset of the present study was considered valid for proceeding with the rest of statistical analyses.

**Table R-2.** Hardy-Weinberg Equilibrium (HWE) p-values of each *HLA* locus based on 4-field allele resolution level *HLA* genotyping data of this 17<sup>th</sup>-IHIW Spanish population cohort study (n=282 subjects).

| Locus           | p-value |
|-----------------|---------|
| <i>HLA-A</i>    | 0.5230  |
| <i>HLA-B</i>    | 0.1151  |
| <i>HLA-C</i>    | 0.9906  |
| <i>HLA-DPA1</i> | 0.0104* |
| <i>HLA-DPB1</i> | 0.9133  |
| <i>HLA-DQA1</i> | 0.3015  |
| <i>HLA-DQB1</i> | 0.7955  |
| <i>HLA-DRB1</i> | 0.8920  |

(\*) Guo and Thompson test p-values (overall) lower than 0.05 ( $p < 0.05$ ) indicate a significant (at the 5% level) deviation from expected Hardy-Weinberg Equilibrium proportions (HWEP).

When performing HWEP test, *HLA-DRB3/4/5* loci were not included as they represent a particular virtual single “locus”. Where these *HLA-DRB3/4/5* genes characteristically behave as alleles of a single locus as the presence of one of these genes at the haplotype level excludes the presence of the other two genes. This is based on the linkage constraints that exist between the *HLA-DRB3/4/5* loci and the *HLA-DRB1* locus, in which several *HLA-DRB1* allele families can be differentiated. Thus, the haplotype organization of the region encompassing *HLA-DRB1-DRB3/4/5* has been shown to be correlated with the allele present on the *HLA-DRB1* locus [344].

### **3. HLA ALLELE FREQUENCIES IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

The frequency distribution of *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*, and *-DRB3/4/5* alleles at the 4-field allele resolution level are summarized in **Table R-3**. Respectively, 36 *HLA-A*, 53 *HLA-B*, 40 *HLA-C*, 14 *HLA-DPA1*, 29 *HLA-DPB1*, 23 *HLA-DQA1*, 24 *HLA-DQB1*, 37 *HLA-DRB1*, 5 *HLA-DRB3*, 5 *HLA-DRB4* and 3 *HLA-DRB5* distinct alleles (k) were identified.

It can be observed how the most predominant HLA alleles (frequency higher than 5%) represent in the case of each locus: 6 *HLA-A* alleles (66%), 6 *HLA-B* alleles (42%), 7 *HLA-C* alleles (55%), 7 *HLA-DPA1* alleles (91%), 4 *HLA-DPB1* alleles (66%), 9 *HLA-DQA1* alleles (81%), 7 *HLA-DQB1* alleles (73%), 6 *HLA-DRB1* alleles (58%). In the case of *HLA-DRB3/4/5* alleles: *HLA-DRB3\*02:02:01:02* (15%), *-DRB4\*01:03:01:01* (16%) and *-DRB5\*01:01:01* (10%) are the most common.

**Table R-3.** *HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4, -DRB5* allele frequencies (at the 3-/4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) in this 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects). HLA alleles per locus are sorted by frequency (AF expressed in decimals) in descending order.

| Locus                | Allele counts (2n) | Allele Frequency (AF) |
|----------------------|--------------------|-----------------------|
| <b><i>HLA-A</i></b>  |                    |                       |
| <i>A*02:01:01:01</i> | 129                | 0.23370               |
| <i>A*01:01:01:01</i> | 52                 | 0.09420               |
| <i>A*03:01:01:01</i> | 48                 | 0.08696               |
| <i>A*11:01:01:01</i> | 47                 | 0.08514               |
| <i>A*29:02:01:01</i> | 46                 | 0.08333               |
| <i>A*24:02:01:01</i> | 42                 | 0.07609               |
| <i>A*32:01:01</i>    | 23                 | 0.04167               |
| <i>A*30:02:01:01</i> | 20                 | 0.03623               |
| <i>A*23:01:01</i>    | 17                 | 0.03080               |
| <i>A*26:01:01:01</i> | 13                 | 0.02355               |
| <i>A*25:01:01</i>    | 12                 | 0.02174               |
| <i>A*33:01:01</i>    | 12                 | 0.02174               |
| <i>A*31:01:02:01</i> | 11                 | 0.01993               |
| <i>A*68:02:01:01</i> | 9                  | 0.01630               |
| <i>A*02:05:01</i>    | 8                  | 0.01449               |
| <i>A*30:01:01</i>    | 8                  | 0.01449               |
| <i>A*68:01:01:02</i> | 7                  | 0.01268               |
| <i>A*68:01:02:02</i> | 6                  | 0.01087               |
| <i>A*03:01:01:05</i> | 5                  | 0.00906               |
| <i>A*03:01:01:03</i> | 4                  | 0.00725               |
| <i>A*29:02:01:02</i> | 4                  | 0.00725               |
| <i>A*66:01:01</i>    | 4                  | 0.00725               |
| <i>A*03:02:01</i>    | 3                  | 0.00543               |



---

|                         |    |         |
|-------------------------|----|---------|
| <i>A*69:01</i>          | 3  | 0.00543 |
| <i>A*01:02</i>          | 2  | 0.00362 |
| <i>A*02:17:02</i>       | 2  | 0.00362 |
| <i>A*24:02:01:04</i>    | 2  | 0.00362 |
| <i>A*24:02:01:05</i>    | 2  | 0.00362 |
| <i>A*30:02:01:02</i>    | 2  | 0.00362 |
| <i>A*33:03:01</i>       | 2  | 0.00362 |
| <i>A*68:01:02:01</i>    | 2  | 0.00362 |
| <i>A*29:01:01:01</i>    | 1  | 0.00181 |
| <i>A*30:04:01</i>       | 1  | 0.00181 |
| <i>A*30:10</i>          | 1  | 0.00181 |
| <i>A*34:02:01</i>       | 1  | 0.00181 |
| <i>A*68:17</i>          | 1  | 0.00181 |
| <br><b><i>HLA-B</i></b> |    |         |
| <i>B*07:02:01</i>       | 53 | 0.09601 |
| <i>B*44:03:01:01</i>    | 47 | 0.08514 |
| <i>B*08:01:01:01</i>    | 36 | 0.06522 |
| <i>B*51:01:01:01</i>    | 34 | 0.06159 |
| <i>B*44:02:01:01</i>    | 33 | 0.05978 |
| <i>B*35:01:01:02</i>    | 27 | 0.04891 |
| <i>B*18:01:01:01</i>    | 22 | 0.03986 |
| <i>B*49:01:01</i>       | 22 | 0.03986 |
| <i>B*14:02:01:01</i>    | 21 | 0.03804 |
| <i>B*15:01:01:01</i>    | 21 | 0.03804 |
| <i>B*35:03:01</i>       | 18 | 0.03261 |
| <i>B*27:05:02</i>       | 17 | 0.0308  |
| <i>B*38:01:01</i>       | 17 | 0.0308  |
| <i>B*18:01:01:02</i>    | 16 | 0.02899 |
| <i>B*37:01:01</i>       | 12 | 0.02174 |
| <i>B*50:01:01</i>       | 12 | 0.02174 |
| <i>B*40:01:02</i>       | 11 | 0.01993 |
| <i>B*58:01:01:01</i>    | 11 | 0.01993 |
| <i>B*40:02:01</i>       | 10 | 0.01812 |
| <i>B*35:02:01</i>       | 9  | 0.01630 |
| <i>B*14:01:01</i>       | 8  | 0.01449 |
| <i>B*35:08:01</i>       | 8  | 0.01449 |
| <i>B*55:01:01</i>       | 8  | 0.01449 |
| <i>B*13:02:01</i>       | 7  | 0.01268 |
| <i>B*50:02</i>          | 7  | 0.01268 |
| <i>B*39:01:01:03</i>    | 6  | 0.01087 |
| <i>B*52:01:01:02</i>    | 6  | 0.01087 |
| <i>B*53:01:01</i>       | 6  | 0.01087 |
| <i>B*57:01:01</i>       | 6  | 0.01087 |

---

|                      |    |         |
|----------------------|----|---------|
| <i>B*40:06:01:02</i> | 4  | 0.00725 |
| <i>B*45:01:01</i>    | 4  | 0.00725 |
| <i>B*15:03:01:02</i> | 3  | 0.00543 |
| <i>B*41:01:01</i>    | 3  | 0.00543 |
| <i>B*47:01:01:03</i> | 3  | 0.00543 |
| <i>B*07:05:01:01</i> | 2  | 0.00362 |
| <i>B*07:06:01</i>    | 2  | 0.00362 |
| <i>B*15:01:01:04</i> | 2  | 0.00362 |
| <i>B*15:220</i>      | 2  | 0.00362 |
| <i>B*27:02:01</i>    | 2  | 0.00362 |
| <i>B*15:09</i>       | 1  | 0.00181 |
| <i>B*18:01:01:03</i> | 1  | 0.00181 |
| <i>B*27:03</i>       | 1  | 0.00181 |
| <i>B*35:01:01:01</i> | 1  | 0.00181 |
| <i>B*38:20</i>       | 1  | 0.00181 |
| <i>B*39:06:02</i>    | 1  | 0.00181 |
| <i>B*40:12</i>       | 1  | 0.00181 |
| <i>B*41:02:01</i>    | 1  | 0.00181 |
| <i>B*44:04</i>       | 1  | 0.00181 |
| <i>B*44:05:01</i>    | 1  | 0.00181 |
| <i>B*51:08:01</i>    | 1  | 0.00181 |
| <i>B*52:01:02</i>    | 1  | 0.00181 |
| <i>B*57:03:01:02</i> | 1  | 0.00181 |
| <i>B*73:01</i>       | 1  | 0.00181 |
| <br><i>HLA-C</i>     |    |         |
| <i>C*07:01:01:01</i> | 64 | 0.11594 |
| <i>C*04:01:01:01</i> | 59 | 0.10688 |
| <i>C*07:02:01:03</i> | 54 | 0.09783 |
| <i>C*12:03:01:01</i> | 44 | 0.07971 |
| <i>C*16:01:01:01</i> | 30 | 0.05435 |
| <i>C*02:02:02:01</i> | 28 | 0.05072 |
| <i>C*05:01:01:02</i> | 27 | 0.04891 |
| <i>C*06:02:01:01</i> | 26 | 0.04710 |
| <i>C*03:03:01:01</i> | 24 | 0.04348 |
| <i>C*04:01:01:06</i> | 23 | 0.04167 |
| <i>C*05:01:01:01</i> | 22 | 0.03986 |
| <i>C*08:02:01:01</i> | 22 | 0.03986 |
| <i>C*01:02:01</i>    | 16 | 0.02899 |
| <i>C*03:04:01:01</i> | 16 | 0.02899 |
| <i>C*06:02:01:02</i> | 14 | 0.02536 |
| <i>C*14:02:01:01</i> | 11 | 0.01993 |
| <i>C*15:02:01:01</i> | 11 | 0.01993 |
| <i>C*07:18</i>       | 8  | 0.01449 |

|                          |     |         |
|--------------------------|-----|---------|
| <i>C*08:02:01:02</i>     | 8   | 0.01449 |
| <i>C*06:02:01:03</i>     | 5   | 0.00906 |
| <i>C*07:02:01:01</i>     | 5   | 0.00906 |
| <i>C*12:02:02</i>        | 5   | 0.00906 |
| <i>C*15:05:02</i>        | 4   | 0.00725 |
| <i>C*07:04:01:01</i>     | 3   | 0.00543 |
| <i>C*16:02:01</i>        | 3   | 0.00543 |
| <i>C*02:10:01:01</i>     | 2   | 0.00362 |
| <i>C*03:02:02:01</i>     | 2   | 0.00362 |
| <i>C*04:01:01:05</i>     | 2   | 0.00362 |
| <i>C*04:09N</i>          | 2   | 0.00362 |
| <i>C*17:01:01:05</i>     | 2   | 0.00362 |
| <i>C*02:10:01:02</i>     | 1   | 0.00181 |
| <i>C*03:04:01:02</i>     | 1   | 0.00181 |
| <i>C*03:07</i>           | 1   | 0.00181 |
| <i>C*05:09:01</i>        | 1   | 0.00181 |
| <i>C*07:01:02</i>        | 1   | 0.00181 |
| <i>C*08:25</i>           | 1   | 0.00181 |
| <i>C*12:166</i>          | 1   | 0.00181 |
| <i>C*15:05:01</i>        | 1   | 0.00181 |
| <i>C*15:05:03</i>        | 1   | 0.00181 |
| <i>C*17:03</i>           | 1   | 0.00181 |
| <b>HLA-DPA1</b>          |     |         |
| <i>DPA1*01:03:01:02</i>  | 161 | 0.29273 |
| <i>DPA1*01:03:01:04</i>  | 98  | 0.17818 |
| <i>DPA1*01:03:01:01</i>  | 76  | 0.13818 |
| <i>DPA1*01:03:01:05</i>  | 62  | 0.11273 |
| <i>DPA1*02:01:01:01</i>  | 43  | 0.07818 |
| <i>DPA1*01:03:01:03</i>  | 35  | 0.06364 |
| <i>DPA1*02:01:01:02</i>  | 25  | 0.04545 |
| <i>DPA1*02:01:02</i>     | 20  | 0.03636 |
| <i>DPA1*02:02:02</i>     | 17  | 0.03091 |
| <i>DPA1*02:02:01 (*)</i> | 5   | 0.00909 |
| <i>DPA1*02:01:08</i>     | 4   | 0.00727 |
| <i>DPA1*03:01</i>        | 2   | 0.00364 |
| <i>DPA1*01:03:05</i>     | 1   | 0.00182 |
| <i>DPA1*04:01</i>        | 1   | 0.00182 |
| <b>HLA-DPBI</b>          |     |         |
| <i>DPBI*04:01:01:01</i>  | 198 | 0.36131 |
| <i>DPBI*02:01:02</i>     | 94  | 0.17153 |
| <i>DPBI*04:02:01:02</i>  | 37  | 0.06752 |
| <i>DPBI*01:01:01</i>     | 30  | 0.05474 |

---

|                         |    |         |
|-------------------------|----|---------|
| <i>DPB1*03:01:01</i>    | 24 | 0.04380 |
| <i>DPB1*04:02:01:01</i> | 22 | 0.04015 |
| <i>DPB1*104:01</i>      | 21 | 0.03832 |
| <i>DPB1*11:01:01</i>    | 19 | 0.03467 |
| <i>DPB1*02:02</i>       | 14 | 0.02555 |
| <i>DPB1*10:01:01</i>    | 13 | 0.02372 |
| <i>DPB1*17:01</i>       | 13 | 0.02372 |
| <i>DPB1*06:01:01</i>    | 12 | 0.02190 |
| <i>DPB1*05:01:01</i>    | 10 | 0.01825 |
| <i>DPB1*13:01:01</i>    | 8  | 0.0146  |
| <i>DPB1*14:01:01</i>    | 5  | 0.00912 |
| <i>DPB1*09:01:01</i>    | 4  | 0.00730 |
| <i>DPB1*105:01</i>      | 4  | 0.00730 |
| <i>DPB1*19:01</i>       | 4  | 0.00730 |
| <i>DPB1*16:01:01</i>    | 3  | 0.00547 |
| <i>DPB1*15:01:01</i>    | 2  | 0.00365 |
| <i>DPB1*20:01:01</i>    | 2  | 0.00365 |
| <i>DPB1*23:01:01</i>    | 2  | 0.00365 |
| <i>DPB1*01:01:02</i>    | 1  | 0.00182 |
| <i>DPB1*04:01:01:02</i> | 1  | 0.00182 |
| <i>DPB1*131:01</i>      | 1  | 0.00182 |
| <i>DPB1*138:01</i>      | 1  | 0.00182 |
| <i>DPB1*26:01:02</i>    | 1  | 0.00182 |
| <i>DPB1*296:01</i>      | 1  | 0.00182 |
| <i>DPB1*59:01</i>       | 1  | 0.00182 |

***HLA-DQA1***

|                         |    |         |
|-------------------------|----|---------|
| <i>DQA1*02:01:01:01</i> | 88 | 0.16176 |
| <i>DQA1*05:05:01:01</i> | 73 | 0.13419 |
| <i>DQA1*01:02:01:01</i> | 56 | 0.10294 |
| <i>DQA1*01:01:01:02</i> | 46 | 0.08456 |
| <i>DQA1*01:03:01:02</i> | 40 | 0.07353 |
| <i>DQA1*03:01:01</i>    | 39 | 0.07169 |
| <i>DQA1*05:01:01:02</i> | 39 | 0.07169 |
| <i>DQA1*03:03:01:01</i> | 31 | 0.05699 |
| <i>DQA1*05:01:01:01</i> | 29 | 0.05331 |
| <i>DQA1*01:02:02</i>    | 14 | 0.02574 |
| <i>DQA1*01:01:02</i>    | 13 | 0.02390 |
| <i>DQA1*01:02:01:04</i> | 12 | 0.02206 |
| <i>DQA1*04:01:01</i>    | 12 | 0.02206 |
| <i>DQA1*01:04:01:01</i> | 10 | 0.01838 |
| <i>DQA1*01:05:01</i>    | 10 | 0.01838 |
| <i>DQA1*01:03:01:01</i> | 8  | 0.01471 |
| <i>DQA1*05:05:01:03</i> | 6  | 0.01103 |

---

|                         |   |         |
|-------------------------|---|---------|
| <i>DQA1*03:02</i>       | 5 | 0.00919 |
| <i>DQA1*05:01:01:03</i> | 5 | 0.00919 |
| <i>DQA1*01:04:01:03</i> | 4 | 0.00735 |
| <i>DQA1*01:05:02</i>    | 2 | 0.00368 |
| <i>DQA1*01:04:02</i>    | 1 | 0.00184 |
| <i>DQA1*05:03</i>       | 1 | 0.00184 |

**HLA-DQB1**

|                         |    |         |
|-------------------------|----|---------|
| <i>DQB1*02:02:01:01</i> | 78 | 0.14498 |
| <i>DQB1*02:01:01</i>    | 73 | 0.13569 |
| <i>DQB1*03:01:01:03</i> | 59 | 0.10967 |
| <i>DQB1*03:02:01</i>    | 50 | 0.09294 |
| <i>DQB1*06:02:01</i>    | 47 | 0.08736 |
| <i>DQB1*05:01:01:03</i> | 46 | 0.08550 |
| <i>DQB1*06:03:01</i>    | 42 | 0.07807 |
| <i>DQB1*03:01:01:01</i> | 16 | 0.02974 |
| <i>DQB1*04:02:01</i>    | 16 | 0.02974 |
| <i>DQB1*05:02:01</i>    | 16 | 0.02974 |
| <i>DQB1*05:03:01:01</i> | 15 | 0.02788 |
| <i>DQB1*05:01:01:01</i> | 13 | 0.02416 |
| <i>DQB1*03:01:01:02</i> | 10 | 0.01859 |
| <i>DQB1*05:01:01:02</i> | 10 | 0.01859 |
| <i>DQB1*06:01:01</i>    | 9  | 0.01673 |
| <i>DQB1*03:03:02:01</i> | 8  | 0.01487 |
| <i>DQB1*06:04:01</i>    | 8  | 0.01487 |
| <i>DQB1*03:03:02:02</i> | 5  | 0.00929 |
| <i>DQB1*06:09:01</i>    | 5  | 0.00929 |
| <i>DQB1*02:02:01:02</i> | 4  | 0.00743 |
| <i>DQB1*03:19:01</i>    | 4  | 0.00743 |
| <i>DQB1*03:04:01</i>    | 2  | 0.00372 |
| <i>DQB1*03:02:03</i>    | 1  | 0.00186 |
| <i>DQB1*06:39</i>       | 1  | 0.00186 |

**HLA-DRB1**

|                         |    |         |
|-------------------------|----|---------|
| <i>DRB1*07:01:01:01</i> | 87 | 0.15993 |
| <i>DRB1*03:01:01:01</i> | 73 | 0.13419 |
| <i>DRB1*15:01:01:01</i> | 50 | 0.09191 |
| <i>DRB1*13:01:01:01</i> | 42 | 0.07721 |
| <i>DRB1*01:01:01</i>    | 35 | 0.06434 |
| <i>DRB1*11:04:01</i>    | 26 | 0.04779 |
| <i>DRB1*11:01:01:01</i> | 21 | 0.0386  |
| <i>DRB1*04:05:01</i>    | 16 | 0.02941 |
| <i>DRB1*04:01:01:01</i> | 15 | 0.02757 |

|                         |    |         |
|-------------------------|----|---------|
| <i>DRB1*14:54:01</i>    | 14 | 0.02574 |
| <i>DRB1*01:02:01</i>    | 13 | 0.02390 |
| <i>DRB1*04:04:01</i>    | 13 | 0.02390 |
| <i>DRB1*13:02:01</i>    | 13 | 0.02390 |
| <i>DRB1*04:02:01</i>    | 12 | 0.02206 |
| <i>DRB1*08:01:01</i>    | 12 | 0.02206 |
| <i>DRB1*01:03</i>       | 11 | 0.02022 |
| <i>DRB1*13:03:01</i>    | 11 | 0.02022 |
| <i>DRB1*10:01:01:01</i> | 9  | 0.01654 |
| <i>DRB1*15:02:01:02</i> | 8  | 0.01471 |
| <i>DRB1*16:01:01</i>    | 8  | 0.01471 |
| <i>DRB1*04:03:01</i>    | 7  | 0.01287 |
| <i>DRB1*11:03:01</i>    | 7  | 0.01287 |
| <i>DRB1*11:02:01</i>    | 6  | 0.01103 |
| <i>DRB1*16:02:01:02</i> | 6  | 0.01103 |
| <i>DRB1*12:01:01:03</i> | 5  | 0.00919 |
| <i>DRB1*04:07:01</i>    | 4  | 0.00735 |
| <i>DRB1*09:01:02</i>    | 4  | 0.00735 |
| <i>DRB1*04:08:01</i>    | 3  | 0.00551 |
| <i>DRB1*13:05:01</i>    | 3  | 0.00551 |
| <i>DRB1*04:06:02</i>    | 2  | 0.00368 |
| <i>DRB1*14:01:01</i>    | 2  | 0.00368 |
| <i>DRB1*08:06</i>       | 1  | 0.00184 |
| <i>DRB1*10:01:01:02</i> | 1  | 0.00184 |
| <i>DRB1*11:01:02</i>    | 1  | 0.00184 |
| <i>DRB1*13:04</i>       | 1  | 0.00184 |
| <i>DRB1*14:04:01</i>    | 1  | 0.00184 |
| <i>DRB1*15:06:01</i>    | 1  | 0.00184 |

**HLA-DRB3**

|                         |     |         |
|-------------------------|-----|---------|
| <i>DRB3*00:00 (**)</i>  | 319 | 0.58425 |
| <i>DRB3*02:02:01:02</i> | 79  | 0.14469 |
| <i>DRB3*01:01:02:01</i> | 73  | 0.13370 |
| <i>DRB3*02:02:01:01</i> | 57  | 0.10440 |
| <i>DRB3*03:01:01</i>    | 15  | 0.02747 |
| <i>DRB3*02:24</i>       | 3   | 0.00549 |

**HLA-DRB4**

|                          |     |         |
|--------------------------|-----|---------|
| <i>DRB4*00:00 (***)</i>  | 383 | 0.70147 |
| <i>DRB4*01:03:01:01</i>  | 88  | 0.16117 |
| <i>DRB4*01:01:01:01</i>  | 59  | 0.10806 |
| <i>DRB4*01:03:01:02N</i> | 10  | 0.01832 |
| <i>DRB4*01:03:02</i>     | 5   | 0.00916 |
| <i>DRB4*01:03:03</i>     | 1   | 0.00183 |

**HLA-DRB5**

|                          |     |         |
|--------------------------|-----|---------|
| <i>DRB5*00:00</i> (****) | 472 | 0.86447 |
| <i>DRB5*01:01:01</i>     | 52  | 0.09524 |
| <i>DRB5*02:02</i>        | 14  | 0.02564 |
| <i>DRB5*01:02</i>        | 8   | 0.01465 |

Notes:

(\*) In March 2017, posterior to release of IPD-IMGT/HLA database version 3.25.0, the *HLA-DPA1\*02:02:01* allele was deleted from the official WHO HLA Nomenclature as its sequence has now been shown to be in error and is identical to *HLA-DPA1\*02:07:01* allele (<https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/deleted.cgi>).

(\*\*) *HLA-DRB3\*00:00* (indicating DRB3 absence) frequency of 58.4%.

(\*\*\*) *HLA-DRB4\*00:00* (indicating DRB4 absence) frequency of 70.2%.

(\*\*\*\*) *HLA-DRB5\*00:00* (indicating DRB5 absence) frequency of 86.5%.

#### **4. IDENTIFICATION OF TWO NOVEL HLA ALLELES IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

Two novel HLA alleles were identified (according to IPD-IMGT/HLA database version 3.25.0, used at the time of this study) during this Spanish population study using this aforementioned NGS-based HLA genotyping method [187] (**Figure R-1a-1f**). One individual (17<sup>th</sup> IHIW sample ID no. H00035F6, from Barcelona, Spain) presents a single base mismatch with *HLA-B\*38:20* reference allele sequence in exon 3 (codon 99), which leads to a synonymous substitution (Tyr or Y (TAC) to Tyr or Y (TAT)) (**Figure R-1a-1c**). Complete HLA genotyping result of this individual including the novel allele is:

*HLA-A\*29:02:01:01*, *HLA-A\*25:01:01*; *HLA-C\*12:03:01:01*, *HLA-C\*03:03:01:01*; *HLA-B\*38:20:02*, *HLA-B\*15:01:01:01*; *HLA-DRB4\*01:03:01:01*, *HLA-DRB3\*02:02:01:02*; *HLA-DRB1\*07:01:01:01*, *HLA-DRB1\*13:01:01:01*; *HLA-DQA1\*02:01:01:01*, *HLA-DQA1\*01:03:01:02*; *HLA-DQB1\*02:02:01:01*, *HLA-DQB1\*06:03:01*; *HLA-DPA1\*02:02:01*, *HLA-DPA1\*01:03:01:01*; *HLA-DPB1\*19:01*, *HLA-DPB1\*02:01:02*.

Also in another different subject (17<sup>th</sup> IHIW sample ID no. H00036D1, from Málaga, Spain), a single base mismatch with *HLA-DRB3\*02:02:01:01* reference allele sequence is detected in exon 3 (codon 166), which leads in this case to a non-synonymous substitution and, therefore, to an amino acid (aa) change (Arg or R (CGG) to Gln or Q (CAG)) (**Figure R-1d-1f**). Complete HLA genotyping result of this other subject including the novel allele is:

*HLA-A\*11:01:01:01*, *HLA-A\*11:01:01:01*; *HLA-C\*05:01:01:02*, *HLA-C\*15:02:01:01*; *HLA-B\*44:02:01:01*, *HLA-B\*51:01:01:01*; *HLA-DRB5\*01:01:01*, *HLA-DRB3\*02:71*; *HLA-DRB1\*15:01:01:01*, *HLA-DRB1\*03:01:01:01*; *HLA-DQA1\*01:02:01:01*, *HLA-DQA1\*05:01:01:01*; *HLA-DQB1\*06:02:01*, *HLA-DQB1\*02:01:01*; *HLA-DPA1\*01:03:01:01*, *HLA-DPA1\*01:03:01:02*; *HLA-DPB1\*04:01:01:01*, *HLA-DPB1\*02:01:02*.

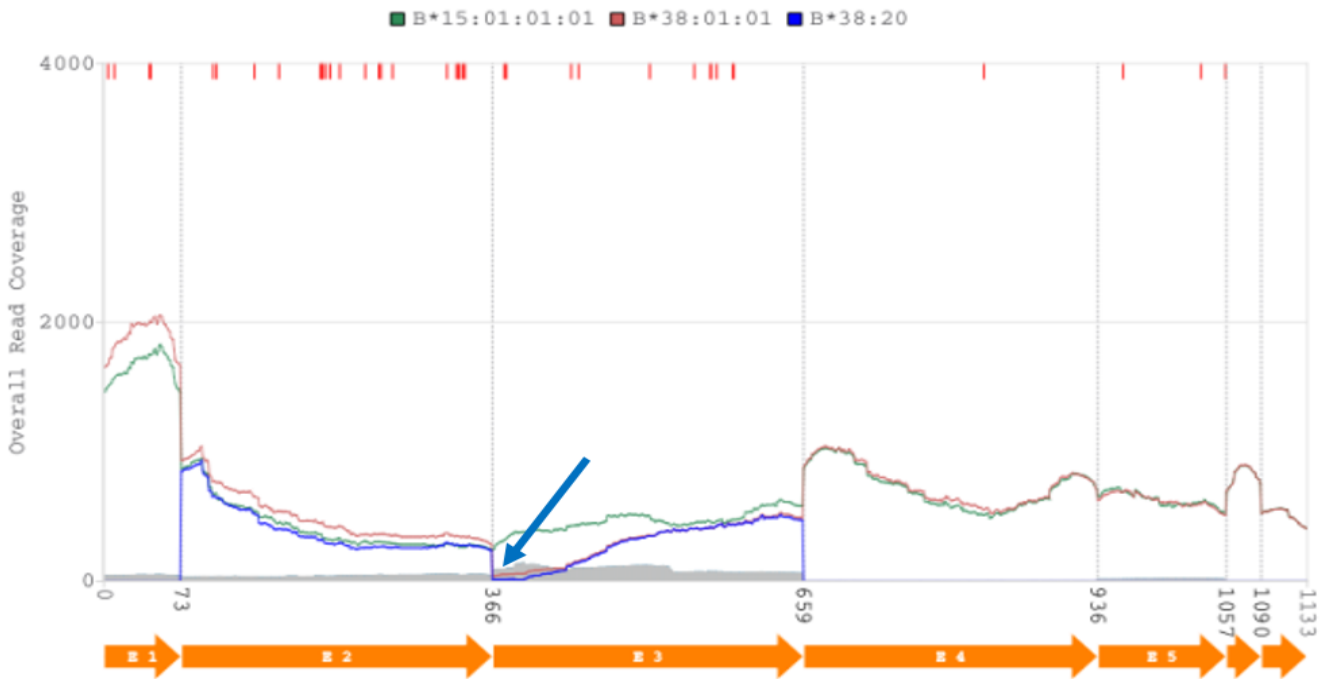
To confirm these findings, as a second HLA sequencing test performed in parallel, sequence-based typing (SBT) was performed using respective SBTexcellerator kits (GenDx, Utrecht, The Netherlands) on a 3130xL Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) and SBTengine HLA typing software version 3.14.0.2783 (GenDx, Utrecht, The Netherlands) at the corresponding local Spanish clinical laboratories of origin for these two samples.

Reported sequences of both identified exon variants were submitted to GenBank [801] and to the IPD-IMGT/HLA Database [295]. These two new alleles have been officially assigned by the WHO Nomenclature Committee for Factors of the HLA system [74]. In the case of the new *HLA-B\*38:20* allele, the official name given is *HLA-B\*38:20:02* (GenBank accession no. **MG76848** and IPD-IMGT/HLA submission no. HWS10051845). Regarding the new *HLA-DRB3\*02:02:01:01* allele, the official name given is *HLA-DRB3\*02:71* (GenBank accession no. **MG922498** and IPD-IMGT/HLA submission no. HWS10051607).



**Figure R-1a-c.** Novel allele with one base mismatch in exon 3 of *HLA-B\*38:20* allele reference allele sequence.

**a)** Coverage cDNA plots of *HLA-B\*38:20* (Blue colored coverage graph), as well as of *HLA-B\*38:01:01* (Red colored coverage graph), and *HLA-B\*15:01:01:01* (Green colored coverage graph).



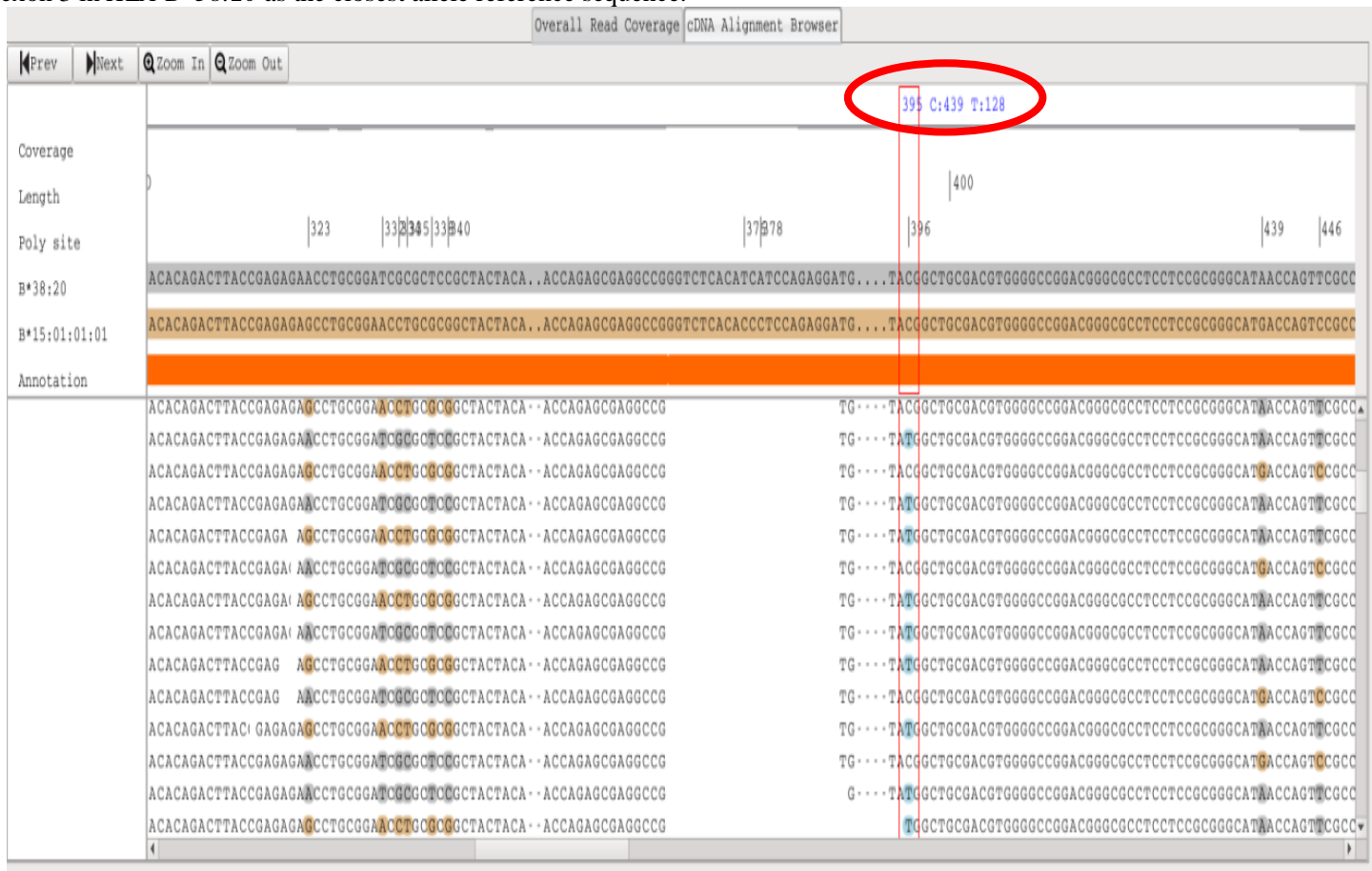
**b)** Consensus cDNA alignment in exon 3 of both contig 2 sequence versus reference sequence *HLA-B\*38:20* (as well as reference sequence *HLA-B\*38:01:01*) (above image) and contig 1 sequence versus reference sequence *HLA-B\*15:01:01:01* (bottom image).

```

>AA Codon
Contig2
AAC CTG CGG ATC GCG CTC CGC TAC TAC AAC CAG AGC GAG GCC G|GG TCT CAC ATC ATC CAG AGG ATG TAT GGC TGC
B*38:01:01
-----
B*38:20
-----

>AA Codon
Contig1
AGC CTG CGG AAC CTG CGC GGC TAC TAC AAC CAG AGC GAG GCC G|GG TCT CAC ACC CTC CAG AGG ATG TAC GGC TGC
B*15:01:01:01
-----
    
```

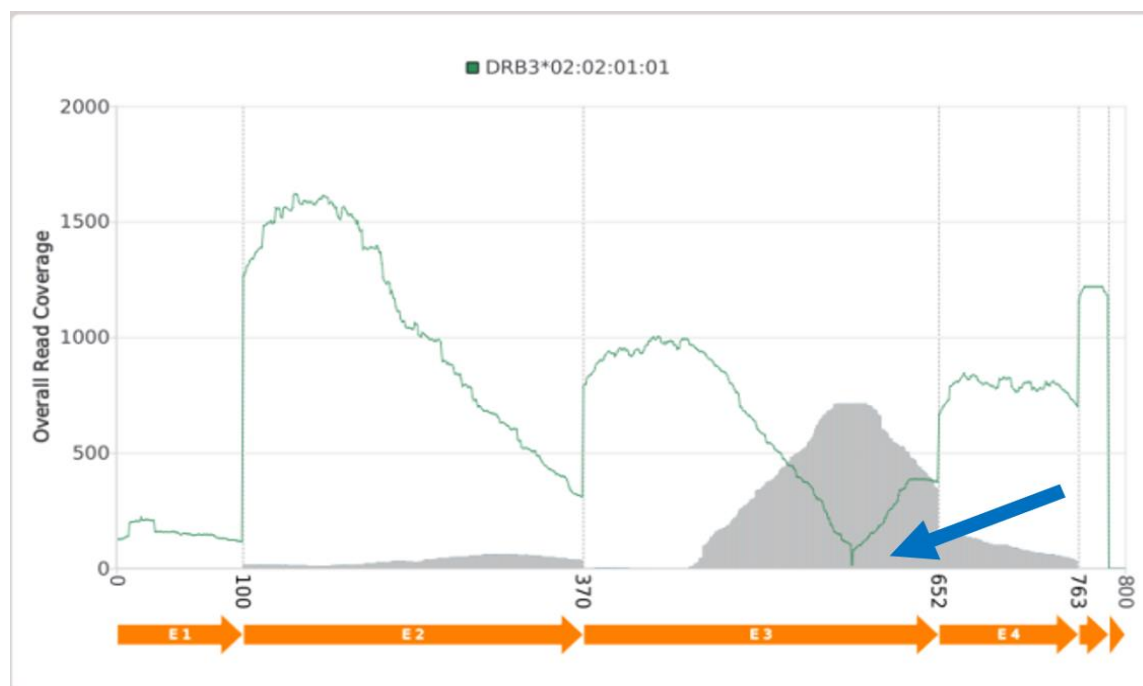
c) Consensus cDNA contig alignment browser indicates the presence of nucleotide “C” (Cytosine) and nucleotide “T” (Thymine) at position 395. Where “C” predominantly belongs to contig 1 sequence (colored in orange) for *HLA-B\*15:01:01:01* allele and “T” predominantly belongs to contig 2 sequence (colored in grey), which is a novel allele with one base mismatch in exon 3 in *HLA-B\*38:20* as the closest allele reference sequence.



17<sup>th</sup> IHIW Sample H00035F6 has one base mismatch at codon 99 in exon 3 of *HLA-B\*38:20* reference allele sequence that results in a synonymous substitution of Tyr or Y (TAC) to Tyr or Y (TAT). (a) The coverage cDNA plot of *HLA-B\*38:20* (represented here together with *HLA-B\*38:01:01*) shows a dip to the baseline (indicated by a blue arrow) in exon 3. (b) Firstly, consensus cDNA alignment allows us to confirm that the best assignment of contig 2 sequence is with *HLA-B\*38:20* reference sequence instead of *HLA-B\*38:01:01* reference sequence (shown by a green circle in the above image of (b)). Also this alignment tool detects one base mismatch (“T” instead of “C”) at codon 99 in exon 3 between this contig 2 sequence and the *HLA-B\*38:20* reference sequence (shown by the colored blue highlight box indicated by a blue circle in the above image of (b)). (c) When looking at the same position but in the consensus cDNA browser it can be detected (position 395) the presence of both nucleotide “C” (78%, predominantly belonging to contig 1 assigned to *HLA-B\*15:01:01:01*) and nucleotide “T” (22%, identifying this novel allele in contig 2 assigned to the closest reference allele *HLA-B\*38:20*) (indicated by a red circle). Therefore, all evidences show this is a novel allele closest to *HLA-B\*38:20* reference allele sequence with a single base mismatch in exon 3. In terms of plausible functional implications (i.e. antigen-presentation, related peptide specificity, protein stability/integration into the cell membrane or interaction with respective T cell co-receptor) of this novel *HLA-B* allele (named as *HLA-B\*38:20:02*): Since this detected single base mismatch with *HLA-B\*38:20* reference allele sequence in codon 99 within exon 3 {which encodes extracellular  $\alpha 2$  domain, that together with  $\alpha 1$  domain make up the peptide-binding cleft consisting of two  $\alpha$ -helices overlying a floor comprised of eight antiparallel  $\beta$ -stranded sheets of the respective expressed HLA class-I molecule [553]} leads only into a synonymous substitution [Tyr or Y (TAC) to Tyr or Y (TAT), thus remaining the identical aa or residue]. Therefore, it should not mean any alteration from the original antigen-presentation features/peptide specificity shown by *HLA-B\*38:20* allele reference.

**Figure R-1d-f.** Novel allele with one base mismatch in exon 3 of *HLA-DRB3\*02:02:01:01* reference allele sequence.

**d)** Coverage cDNA plot of *HLA-DRB3\*02:02:01:01* (Green colored coverage graph).



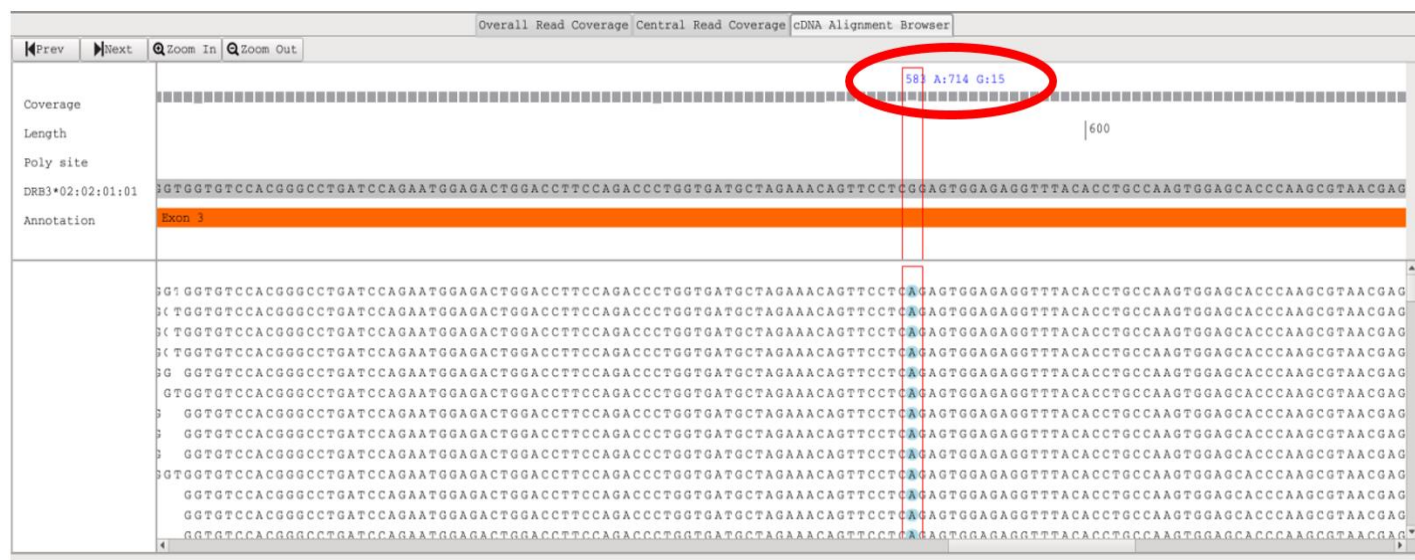
**e)** Consensus cDNA alignment in exon 3 of both contig 1 sequence (for *HLA-DRB3*) of one allele versus reference sequence *HLA-DRB3\*02:02:01:01* (above image); and on the other allele contig 1 sequence (for *HLA-DRB5*) versus reference sequence *HLA-DRB5\*01:01:01v1* (cloned and sequenced allele from internal database MIA FORA™ NGS FLEX HLA Genotyping Software) (bottom image).

```

>AA Codon          150          155          160          165
Contig1            CTG ATC CAG AAT GGA GAC TGG ACC TTC CAG ACC CTG GTG ATG CTA GAA ACA GTT CCT CAG AC
DRB3*02:02:01:01  -----
>AA Codon          150          155          160          165
Contig1            CTG ATT CAG AAT GGA GAC TGG ACC TTC CAG ACC CTG GTG ATG CTG GAA ACA GTT CCT CGA
DRB5*01:01:01v1  -----

```

**f)** Consensus cDNA contig alignment browser indicates a major presence of nucleotide “A” (Adenine) and a very low proportion of nucleotide “G” (Guanine) at position 583. Where “A” belongs to contig 1 sequence in *HLA-DRB3*, which defines a novel allele with one base mismatch in exon 3 with *HLA-DRB3\*02:02:01:01* reference allele sequence. On the other hand, nucleotide “G” most probably comes from either contig 1 of *HLA-DRB5* locus (*HLA-DRB5\*01:01:01v1*) or from contig 1 (*HLA-DRB1\*03:01:01:01*) or contig 2 (*HLA-DRB1\*15:01:01:01*) of *HLA-DRB1* locus, in an event known as “cross-mapping reads” between homologous sequences. Since there is a high similarity of certain DNA sequence regions found between *HLA-DRB1* alleles and the respective *HLA-DRB3/4/5* alleles. Thus, it is not possible to establish a clear distinction in the alignment and mapping of NGS raw reads, when attempting to define sequence and allele calling of *HLA-DRB1* gen and their respective association with *HLA-DRB3/4/5* genes [202].



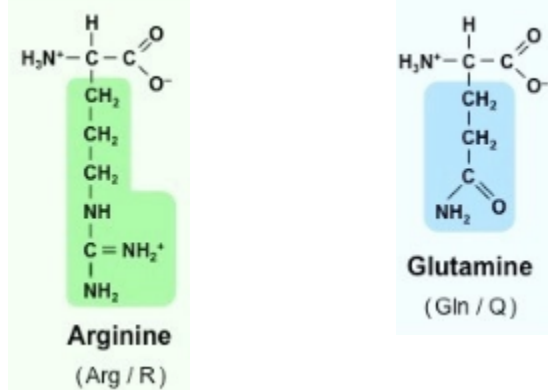
17<sup>th</sup> IHIW Sample H00036D1 shows one base mismatch in exon 3 of *HLA-DRB3\*02:02:01:01* reference allele sequence that results in a non-synonymous substitution of residue and, therefore, in an amino acid change (Arg (CGG) to Gln (CAG)). (d) The coverage cDNA plot of *HLA-DRB3\*02:02:01:01* shows a dip to the baseline (indicated by a blue arrow) in exon 3. (e) Consensus cDNA alignment shows one base mismatch (“A” instead of “G”) at codon 166 in exon 3 (indicated by a colored yellow highlight box in the above image of (e)). (f) When looking at the same position but in the consensus cDNA browser (position 583) it can be detected a dominant presence of nucleotide “A” (98%, belonging to contig 1 in *HLA-DRB3* and identifying this novel allele) and nucleotide “G” (2%, coming probably either from contig 1 of *HLA-DRB5* locus or from contig 1 or contig 2 of *HLA-DRB1* locus) (indicated by a red circle).

Therefore, all evidences show this is a novel allele closest to *HLA-DRB3\*02:02:01:01* reference allele sequence with a single base mismatch in exon 3.

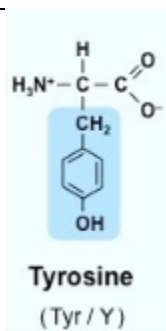
In terms of plausible functional implications (i.e. antigen-presentation, related peptide specificity, protein stability/integration into the cell membrane or interaction with respective T cell co-receptor) of this novel *HLA-DRB3* allele (named as *HLA-DRB3\*02:71*):

In this instance, the detected single base mismatch with *HLA-DRB3\*02:02:01:01* reference allele sequence in codon 166 within exon 3 {which encodes the Ig-like  $\beta 2$  domain proximal to the cell membrane of APCs, and that is part of the extracellular portion of  $\beta$  chain of the respective expressed HLA class-II molecule [553]} leads to a non-synonymous substitution and, consequently, to an amino acid (aa) change. In which, an aa with a positively charged side chain (Arg or R (CGG), that is able to specifically form ionic bonds) is replaced, in this new HLA allele and respective encoded HLA protein, by an aa with a polar but uncharged side chain (Gln or Q (CAG) that, differently, is capable of forming hydrogen bonds). As it is well-documented, the chemistry of amino acid side chains is critical to protein structure because these side chains can differently and specifically bond with one another to hold a length of protein in a certain shape or conformation [802]. Thus, this observed amino acid change (with a plausible associated alteration of aa side chains interactions and corresponding bonds) could potentially mean, in turn, a certain level of variation of protein folding and configuration within this particular above mentioned  $\beta 2$  domain. Moreover, both the  $\alpha 2$  (encoded by exon 3 of the respective given *HLA-DRA* allele pair) and  $\beta 2$  domains of HLA class II heterodimer molecules, proximal to the cell membrane of APCs, contribute to a concavity which accommodates a protrusion from the CD4 co-receptor in the T cell [803]. Consequently, the conformation (and spatial geometry) of respective HLA class II  $\alpha 2/\beta 2$  domain for binding this CD4 T cell co-receptor may be partially or minimally conditioned by this amino acid change and, thus, ultimately it may have its effect on  $\alpha\beta$ TCR T cell development and respective effector function [804]. Therefore, a clear functional relevance (and also even allogeneic from the perspective of the transplantation setting) could be potentially attributed to this new *HLA-DRB3* allele. Nevertheless, further functional studies will be necessary in order to confirm this speculative conclusion.

**Figure R-2.** Molecular structure of amino acid side chains discussed above.



**a)** Molecular structure and charge differences observed between amino acid side chains of Arg (presented by *HLA-DRB3\*02:01:01* reference allele sequence) and Gln (presented by novel *HLA-DRB3\*02:71* described here) encoded respectively by codon 166 within exon 3 of *HLA-DRB3* locus.



**b)** Molecular structure of amino acid side chain of Tyr (presented by *HLA-B\*38:20* reference allele sequence and also by *HLA-B\*38:20:02* described here) encoded by codon 99 within exon 3 of *HLA-B* locus.

## **5. EWENS-WATERSON HOMOZYGOSITY (EWH) TEST OF NEUTRALITY IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

Just as an initial and tentative analysis, Ewens-Waterson Homozygosity (EWH) test of neutrality was used for analysis of selective processes based on HLA allelic diversity at the 3-/4-field allele resolution level of this Spanish population cohort. All *HLA* loci analyzed show levels of observed homozygosity ( $F_o$ ) that are below the expected homozygosity under neutrality ( $F_e$ ) with the exception of *HLA-DPBI* locus (**Table R-4**). Furthermore, *HLA-B*, *-DQA1* and *-DQB1* are the only loci that show statistically significant deviation from neutrality and, therefore, are consistent with a more pronounced balancing selection ( $F_{nd} \ll 0$ ). As previously described across human populations [134][805] we also observed for this Spanish population cohort (in spite of presenting a relatively small sample size) an overall direction towards balancing selection for most of the classical *HLA* class I and II loci with the striking exception of *HLA-DP* genes. These latter (especially *HLA-DPBI* locus, based on our results at the 4-field allele resolution level) seem to be more under directional selection, in which only a set of few alleles become selected (e.g. *HLA-DPBI\*04:01:01:01*), as similarly observed in other previous studies [806][807]. These interpretations, however, need to be further confirmed on a larger Spanish population cohort, considering also the diverse nature of the regional subpopulations included in this study.

**Table R-4.** Ewens-Watterson Homozygosity (EWH) test of neutrality at the *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* loci based on the 3-/4-field allele resolution level HLA genotyping data (and according to IPD-IMGT/HLA database version 3.25.0) of this 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects).

| Locus           | Number of subjects typed (N) | k  | F <sub>o</sub> | F <sub>e</sub> | F <sub>nd</sub> | p-value of F |
|-----------------|------------------------------|----|----------------|----------------|-----------------|--------------|
| <i>HLA-A</i>    | 276                          | 36 | 0.0983         | 0.1048         | -0.1924         | 0.5208       |
| <i>HLA-B</i>    | 276                          | 53 | 0.0444         | 0.0661         | -1.2197         | 0.0331*      |
| <i>HLA-C</i>    | 276                          | 40 | 0.0617         | 0.0934         | -1.0858         | 0.0577       |
| <i>HLA-DPA1</i> | 275                          | 14 | 0.1639         | 0.2766         | -1.0498         | 0.0806       |
| <i>HLA-DPB1</i> | 274                          | 29 | 0.1769         | 0.1336         | 0.9376          | 0.8572       |
| <i>HLA-DQA1</i> | 272                          | 23 | 0.0872         | 0.1708         | -1.3322         | 0.0094**     |
| <i>HLA-DQB1</i> | 269                          | 24 | 0.0869         | 0.1630         | -1.2857         | 0.0120*      |
| <i>HLA-DRB1</i> | 272                          | 37 | 0.0733         | 0.1015         | -0.8761         | 0.1443       |

The normalized deviate of the Ewens-Watterson homozygosity statistic ( $F_{nd}$ ) was calculated based on the observed allele frequencies at each *HLA* locus and it is used to infer the action of balancing ( $F_{nd} \ll 0$ ) or directional selection/extreme demographic effect ( $F_{nd} \gg 0$ ) at each *HLA* locus. The results of the Ewens-Watterson Homozygosity Test are shown above. Number of unique alleles ( $k$ ); Observed F ( $F_o$ ); Expected F ( $F_e$ ); Normalized deviate of F ( $F_{nd}$ ).

(\*) p-value of F lower than 0.05 ( $p < 0.05$ ) indicates a statistical significance at the 5% level.

(\*\*) p-value of F lower than 0.01 ( $p < 0.01$ ) indicates a statistical significance at the 1% level.

Also when performing EWH test, *HLA-DRB3/4/5* loci were not included as they represent a particular virtual single “locus”. Where these *HLA-DRB3/4/5* genes characteristically behave as alleles of a single locus as the presence of one of these genes at the haplotype level excludes the presence of the other two genes. This is based on the linkage constraints that exist between the *HLA-DRB3/4/5* loci and the *HLA-DRB1* locus, in which several *HLA-DRB1* allele families can be differentiated [344].



## **6. 2-LOCUS HAPLOTYPE LINKAGE DISEQUILIBRIUM (LD) ANALYSIS IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

Estimated 2-locus haplotype frequencies and measure of overall LD (Hedrick's D' statistic) of pairs of neighboring genetic *HLA* loci (*B~C*, *DPA1~DPB1*, *DQA1~DQB1*, *DQB1~DRB1* and even (although more distant) *B~DRB1*) at the 3-/4-field allele resolution level are shown in **Table R-5**. Interestingly, it can be observed unique 2-locus haplotype associations in non-coding regions at the 4-field allele resolution level that are not apparent when testing at the 2-field level. For instance, alleles of the *HLA-B\*35* allele group show very distinctive associations with *HLA-C* alleles at the 4-field level. On one hand, at the 2-field level, *HLA-B\*35:01*, *HLA-B\*35:02*, *HLA-B\*35:03* and *HLA-B\*35:08* alleles show a strong and common association with *HLA-C\*04:01* allele. Nevertheless, at the 4-field level it can be observed that in the case of the non-coding variant *HLA-B\*35:01:01:01*, it displays a specific and conserved association with *HLA-C\*04:01:01:01*. Whereas, *HLA-B\*35:01:01:02* non-coding variant presents associations with not only *HLA-C\*04:01:01:01* allele but also with *HLA-C\*04:01:01:05* and *HLA-C\*04:01:01:06* alleles. In the case of *HLA-B\*35:02:01*, it may seem to display a specific and conserved association with *HLA-C\*04:01:01:06* in Spanish population. As for *HLA-B\*35:03:01*, it presents association with *HLA-C\*04:01:01:01*. Finally, *HLA-B\*35:08:01* shows association with *HLA-C\*04:01:01:06*. Furthermore, we also found distinctive haplotypic associations at the non-coding level in several other *HLA* class I and class II loci pairs (e.g. non-coding *HLA-DQA1\*05:01:01* variants, *HLA-B\*18:01:01* variants, *HLA-C\*05:01:01* variants or *HLA-C\*06:02:01* variants). In contrast, *HLA* loci pairs as *B\*07:02:01~C\*07:02:01:03*, *DQA1\*01:01:01:02~DQB1\*05:01:01:03* and *DQB1\*02:02:01:01~DRB1\*07:01:01:01* are some examples of 4-field highly conserved associations found in this Spanish population cohort.



**Table R-5.** HLA loci pair *B~C*, *DPA1~DPB1*, *DQA1~DQB1*, *DQB1~DRB1* and *B~DRB1* haplotypes estimated frequencies (at the 3-/4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) sorted from the lowest to the highest numbered first allele of the pair are shown for this 17th-IHIW Spanish population cohort (n=282 subjects).

| Haplotype                          | Frequency (in decimals) | LD (Hedrick's D' statistic) |
|------------------------------------|-------------------------|-----------------------------|
| <b>Locus Pair B~C</b>              |                         |                             |
| <i>B*07:02:01~C*07:02:01:03</i>    | 0.09601                 | 1                           |
| <i>B*07:05:01:01~C*15:05:02</i>    | 0.00362                 | 1                           |
| <i>B*07:06:01~C*15:05:02</i>       | 0.00362                 | 1                           |
| <i>B*08:01:01:01~C*06:02:01:03</i> | 0.00181                 | 0.14419                     |
| <i>B*08:01:01:01~C*07:01:01:01</i> | 0.05978                 | 0.90574                     |
| <i>B*08:01:01:01~C*07:02:01:01</i> | 0.00362                 | 0.35814                     |
| <i>B*13:02:01~C*06:02:01:01</i>    | 0.01268                 | 1                           |
| <i>B*14:01:01~C*08:02:01:02</i>    | 0.01449                 | 1                           |
| <i>B*14:02:01:01~C*08:02:01:01</i> | 0.03804                 | 1                           |
| <i>B*15:01:01:01~C*03:03:01:01</i> | 0.0308                  | 0.80087                     |
| <i>B*15:01:01:01~C*03:04:01:01</i> | 0.00543                 | 0.15537                     |
| <i>B*15:01:01:01~C*03:07</i>       | 0.00181                 | 1                           |
| <i>B*15:01:01:04~C*04:01:01:06</i> | 0.00362                 | 1                           |
| <i>B*15:03:01:02~C*02:10:01:01</i> | 0.00362                 | 1                           |
| <i>B*15:03:01:02~C*16:01:01:01</i> | 0.00181                 | 0.29502                     |
| <i>B*15:09~C*07:04:01:01</i>       | 0.00181                 | 1                           |
| <i>B*15:220~C*12:03:01:01</i>      | 0.00362                 | 1                           |
| <i>B*18:01:01:01~C*03:03:01:01</i> | 0.00181                 | 0.00207                     |
| <i>B*18:01:01:01~C*05:01:01:01</i> | 0.03804                 | 0.95266                     |
| <i>B*18:01:01:02~C*07:01:01:01</i> | 0.01087                 | 0.29303                     |
| <i>B*18:01:01:02~C*12:03:01:01</i> | 0.01812                 | 0.59252                     |
| <i>B*18:01:01:03~C*02:10:01:02</i> | 0.00181                 | 1                           |
| <i>B*27:02:01~C*02:02:02:01</i>    | 0.00362                 | 1                           |
| <i>B*27:03~C*07:01:02</i>          | 0.00181                 | 1                           |
| <i>B*27:05:02~C*01:02:01</i>       | 0.0163                  | 0.5486                      |
| <i>B*27:05:02~C*02:02:02:01</i>    | 0.01449                 | 0.4423                      |
| <i>B*35:01:01:01~C*04:01:01:01</i> | 0.00181                 | 1                           |
| <i>B*35:01:01:02~C*04:01:01:01</i> | 0.03623                 | 0.70971                     |
| <i>B*35:01:01:02~C*04:01:01:05</i> | 0.00362                 | 1                           |
| <i>B*35:01:01:02~C*04:01:01:06</i> | 0.00906                 | 0.17714                     |
| <i>B*35:02:01~C*04:01:01:06</i>    | 0.0163                  | 1                           |
| <i>B*35:03:01~C*04:01:01:01</i>    | 0.01993                 | 0.56457                     |
| <i>B*35:03:01~C*08:25</i>          | 0.00181                 | 1                           |
| <i>B*35:03:01~C*12:03:01:01</i>    | 0.01087                 | 0.27559                     |

|                                    |         |          |
|------------------------------------|---------|----------|
| <i>B*35:08:01~C*04:01:01:06</i>    | 0.01087 | 0.73913  |
| <i>B*35:08:01~C*12:03:01:01</i>    | 0.00362 | 0.18504  |
| <i>B*37:01:01~C*06:02:01:01</i>    | 0.02174 | 1        |
| <i>B*38:01:01~C*07:02:01:03</i>    | 0.00181 | -0.39869 |
| <i>B*38:01:01~C*12:03:01:01</i>    | 0.02899 | 0.93608  |
| <i>B*38:20~C*12:03:01:01</i>       | 0.00181 | 1        |
| <i>B*39:01:01:03~C*07:02:01:01</i> | 0.00181 | 0.19121  |
| <i>B*39:01:01:03~C*12:03:01:01</i> | 0.00906 | 0.8189   |
| <i>B*39:06:02~C*07:02:01:01</i>    | 0.00181 | 1        |
| <i>B*40:01:02~C*03:04:01:01</i>    | 0.01993 | 1        |
| <i>B*40:02:01~C*02:02:02:01</i>    | 0.01449 | 0.78931  |
| <i>B*40:02:01~C*03:04:01:02</i>    | 0.00181 | 1        |
| <i>B*40:02:01~C*07:01:01:01</i>    | 0.00181 | -0.1375  |
| <i>B*40:06:01:02~C*15:02:01:01</i> | 0.00725 | 1        |
| <i>B*40:12~C*15:05:03</i>          | 0.00181 | 1        |
| <i>B*41:01:01~C*07:01:01:01</i>    | 0.00181 | 0.2459   |
| <i>B*41:01:01~C*17:01:01:05</i>    | 0.00362 | 1        |
| <i>B*41:02:01~C*17:03</i>          | 0.00181 | 1        |
| <i>B*44:02:01:01~C*02:02:02:01</i> | 0.00362 | 0.01239  |
| <i>B*44:02:01:01~C*05:01:01:02</i> | 0.0471  | 0.96061  |
| <i>B*44:02:01:01~C*05:09:01</i>    | 0.00181 | 1        |
| <i>B*44:02:01:01~C*07:04:01:01</i> | 0.00362 | 0.64547  |
| <i>B*44:02:01:01~C*12:03:01:01</i> | 0.00362 | -0.23967 |
| <i>B*44:03:01:01~C*04:01:01:01</i> | 0.02355 | 0.19002  |
| <i>B*44:03:01:01~C*04:09N</i>      | 0.00362 | 1        |
| <i>B*44:03:01:01~C*05:01:01:02</i> | 0.00181 | -0.56501 |
| <i>B*44:03:01:01~C*15:02:01:01</i> | 0.00181 | 0.0063   |
| <i>B*44:03:01:01~C*16:01:01:01</i> | 0.05072 | 0.92713  |
| <i>B*44:03:01:01~C*16:02:01</i>    | 0.00362 | 0.63564  |
| <i>B*44:04~C*16:01:01:01</i>       | 0.00181 | 1        |
| <i>B*44:05:01~C*02:02:02:01</i>    | 0.00181 | 1        |
| <i>B*45:01:01~C*06:02:01:03</i>    | 0.00725 | 1        |
| <i>B*47:01:01:03~C*06:02:01:01</i> | 0.00362 | 0.65019  |
| <i>B*47:01:01:03~C*07:18</i>       | 0.00181 | 0.32353  |
| <i>B*49:01:01~C*02:02:02:01</i>    | 0.00181 | -0.1039  |
| <i>B*49:01:01~C*07:01:01:01</i>    | 0.03804 | 0.94858  |
| <i>B*50:01:01~C*04:01:01:01</i>    | 0.00543 | 0.16024  |
| <i>B*50:01:01~C*06:02:01:02</i>    | 0.0163  | 0.74349  |
| <i>B*50:02~C*04:01:01:01</i>       | 0.00362 | 0.20023  |
| <i>B*50:02~C*06:02:01:02</i>       | 0.00906 | 0.70685  |
| <i>B*51:01:01:01~C*01:02:01</i>    | 0.00906 | 0.26737  |
| <i>B*51:01:01:01~C*02:02:02:01</i> | 0.01087 | 0.16271  |
| <i>B*51:01:01:01~C*03:03:01:01</i> | 0.00362 | 0.02317  |
| <i>B*51:01:01:01~C*04:01:01:01</i> | 0.00543 | -0.17448 |

|                                    |         |          |
|------------------------------------|---------|----------|
| <i>B*51:01:01:01~C*04:01:01:06</i> | 0.00181 | -0.29412 |
| <i>B*51:01:01:01~C*14:02:01:01</i> | 0.01993 | 1        |
| <i>B*51:01:01:01~C*15:02:01:01</i> | 0.01087 | 0.51562  |
| <i>B*51:08:01~C*16:02:01</i>       | 0.00181 | 1        |
| <i>B*52:01:01:02~C*12:02:02</i>    | 0.00906 | 1        |
| <i>B*52:01:01:02~C*12:166</i>      | 0.00181 | 1        |
| <i>B*52:01:02~C*07:02:01:01</i>    | 0.00181 | 1        |
| <i>B*53:01:01~C*04:01:01:01</i>    | 0.01087 | 1        |
| <i>B*55:01:01~C*01:02:01</i>       | 0.00362 | 0.22761  |
| <i>B*55:01:01~C*03:03:01:01</i>    | 0.00725 | 0.47727  |
| <i>B*55:01:01~C*03:04:01:01</i>    | 0.00362 | 0.22761  |
| <i>B*57:01:01~C*06:02:01:01</i>    | 0.00725 | 0.65019  |
| <i>B*57:01:01~C*07:01:01:01</i>    | 0.00362 | 0.2459   |
| <i>B*57:03:01:02~C*08:02:01:01</i> | 0.00181 | 1        |
| <i>B*58:01:01:01~C*03:02:02:01</i> | 0.00362 | 1        |
| <i>B*58:01:01:01~C*05:01:01:01</i> | 0.00181 | 0.05317  |
| <i>B*58:01:01:01~C*06:02:01:01</i> | 0.00181 | 0.04597  |
| <i>B*58:01:01:01~C*07:18</i>       | 0.01268 | 0.87246  |
| <i>B*73:01~C*15:05:01</i>          | 0.00181 | 1        |

**Locus Pair DPA1~DPB1 (\*)**

|  |         |          |
|--|---------|----------|
| <i>DPA1*01:03:01:01~DPB1*02:01:02</i>    | 0.11182 | 0.76617  |
| <i>DPA1*01:03:01:01~DPB1*02:02</i>       | 0.01457 | 0.50095  |
| <i>DPA1*01:03:01:01~DPB1*03:01:01</i>    | 0.0018  | -0.70321 |
| <i>DPA1*01:03:01:01~DPB1*04:01:01:01</i> | 0.00244 | -0.9514  |
| <i>DPA1*01:03:01:01~DPB1*104:01</i>      | 0.00149 | -0.7196  |
| <i>DPA1*01:03:01:01~DPB1*15:01:01</i>    | 0.00292 | 0.76886  |
| <i>DPA1*01:03:01:01~DPB1*16:01:01</i>    | 0.00365 | 0.61299  |
| <i>DPA1*01:03:01:02~DPB1*02:01:02</i>    | 0.04254 | -0.15065 |
| <i>DPA1*01:03:01:02~DPB1*02:02</i>       | 0.00721 | -0.03317 |
| <i>DPA1*01:03:01:02~DPB1*03:01:01</i>    | 0.00185 | -0.85556 |
| <i>DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.20282 | 0.52191  |
| <i>DPA1*01:03:01:02~DPB1*104:01</i>      | 0.03683 | 0.94508  |
| <i>DPA1*01:03:01:02~DPB1*15:01:01</i>    | 0.00073 | -0.31812 |
| <i>DPA1*01:03:01:03~DPB1*03:01:01</i>    | 0.03832 | 0.86647  |
| <i>DPA1*01:03:01:03~DPB1*06:01:01</i>    | 0.0219  | 1        |
| <i>DPA1*01:03:01:03~DPB1*20:01:01</i>    | 0.00365 | 1        |
| <i>DPA1*01:03:01:04~DPB1*02:01:02</i>    | 0.01353 | -0.5545  |
| <i>DPA1*01:03:01:04~DPB1*02:02</i>       | 0.00377 | -0.16641 |
| <i>DPA1*01:03:01:04~DPB1*04:01:01:01</i> | 0.15059 | 0.7663   |
| <i>DPA1*01:03:01:04~DPB1*04:01:01:02</i> | 0.00182 | 1        |
| <i>DPA1*01:03:01:04~DPB1*105:01</i>      | 0.00182 | 0.08869  |

|  |         |          |
|--|---------|----------|
| <i>DPA1*01:03:01:04~DPB1*16:01:01</i>    | 0.00182 | 0.18995  |
| <i>DPA1*01:03:01:04~DPB1*23:01:01</i>    | 0.00365 | 1        |
| <i>DPA1*01:03:01:05~DPB1*04:02:01:01</i> | 0.04015 | 1        |
| <i>DPA1*01:03:01:05~DPB1*04:02:01:02</i> | 0.06569 | 0.96953  |
| <i>DPA1*01:03:01:05~DPB1*105:01</i>      | 0.00182 | 0.15432  |
| <i>DPA1*01:03:01:05~DPB1*138:01</i>      | 0.00182 | 1        |
| <i>DPA1*01:03:01:05~DPB1*14:01:01</i>    | 0.00182 | 0.09794  |
| <i>DPA1*01:03:01:05~DPB1*59:01</i>       | 0.00182 | 1        |
| <i>DPA1*01:03:05~DPB1*02:01:02</i>       | 0.00182 | 1        |
| <i>DPA1*02:01:01:01~DPB1*04:01:01:01</i> | 0.00182 | -0.93564 |
| <i>DPA1*02:01:01:01~DPB1*04:02:01:02</i> | 0.00182 | -0.65556 |
| <i>DPA1*02:01:01:01~DPB1*10:01:01</i>    | 0.00182 | -0.01968 |
| <i>DPA1*02:01:01:01~DPB1*11:01:01</i>    | 0.03285 | 0.94289  |
| <i>DPA1*02:01:01:01~DPB1*13:01:01</i>    | 0.01277 | 0.86436  |
| <i>DPA1*02:01:01:01~DPB1*131:01</i>      | 0.00182 | 1        |
| <i>DPA1*02:01:01:01~DPB1*14:01:01</i>    | 0.00182 | 0.13188  |
| <i>DPA1*02:01:01:01~DPB1*17:01</i>       | 0.02372 | 1        |
| <i>DPA1*02:01:01:02~DPB1*01:01:02</i>    | 0.00182 | 1        |
| <i>DPA1*02:01:01:02~DPB1*02:01:02</i>    | 0.00182 | -0.76681 |
| <i>DPA1*02:01:01:02~DPB1*03:01:01</i>    | 0.00182 | -0.08667 |
| <i>DPA1*02:01:01:02~DPB1*05:01:01</i>    | 0.00547 | 0.26654  |
| <i>DPA1*02:01:01:02~DPB1*09:01:01</i>    | 0.0073  | 1        |
| <i>DPA1*02:01:01:02~DPB1*10:01:01</i>    | 0.02007 | 0.8388   |
| <i>DPA1*02:01:01:02~DPB1*14:01:01</i>    | 0.00547 | 0.58088  |
| <i>DPA1*02:01:01:02~DPB1*26:01:02</i>    | 0.00182 | 1        |
| <i>DPA1*02:01:02~DPB1*01:01:01</i>       | 0.0365  | 1        |
| <i>DPA1*02:01:08~DPB1*01:01:01</i>       | 0.0073  | 1        |
| <i>DPA1*02:02:01~DPB1*04:01:01:01</i>    | 0.00182 | -0.44646 |
| <i>DPA1*02:02:01~DPB1*19:01</i>          | 0.0073  | 1        |
| <i>DPA1*02:02:02~DPB1*01:01:01</i>       | 0.01095 | 0.31547  |
| <i>DPA1*02:02:02~DPB1*04:01:01:01</i>    | 0.00182 | -0.8372  |
| <i>DPA1*02:02:02~DPB1*05:01:01</i>       | 0.01277 | 0.6904   |
| <i>DPA1*02:02:02~DPB1*10:01:01</i>       | 0.00182 | 0.04737  |
| <i>DPA1*02:02:02~DPB1*11:01:01</i>       | 0.00182 | 0.02502  |
| <i>DPA1*02:02:02~DPB1*296:01</i>         | 0.00182 | 1        |
| <i>DPA1*03:01~DPB1*105:01</i>            | 0.00365 | 1        |
| <i>DPA1*04:01~DPB1*13:01:01</i>          | 0.00182 | 1        |

## Locus Pair DQA1~DQB1

|  |         |          |
|--|---------|----------|
| <i>DQA1*01:01:01:02~DQB1*05:01:01:03</i> | 0.08396 | 1        |
| <i>DQA1*01:01:02~DQB1*05:01:01:01</i>    | 0.02425 | 1        |
| <i>DQA1*01:02:01:01~DQB1*05:02:01</i>    | 0.00373 | 0.02697  |
| <i>DQA1*01:02:01:01~DQB1*06:01:01</i>    | 0.00187 | 0.01153  |
| <i>DQA1*01:02:01:01~DQB1*06:02:01</i>    | 0.08582 | 0.97634  |
| <i>DQA1*01:02:01:01~DQB1*06:03:01</i>    | 0.0056  | -0.29101 |
| <i>DQA1*01:02:01:01~DQB1*06:04:01</i>    | 0.00187 | 0.02697  |
| <i>DQA1*01:02:01:01~DQB1*06:09:01</i>    | 0.00187 | 0.11037  |
| <i>DQA1*01:02:01:04~DQB1*05:01:01:03</i> | 0.00187 | -0.02899 |
| <i>DQA1*01:02:01:04~DQB1*06:04:01</i>    | 0.01306 | 0.87214  |
| <i>DQA1*01:02:01:04~DQB1*06:09:01</i>    | 0.00746 | 0.79542  |
| <i>DQA1*01:02:02~DQB1*05:02:01</i>       | 0.02612 | 1        |
| <i>DQA1*01:03:01:01~DQB1*06:01:01</i>    | 0.01493 | 1        |
| <i>DQA1*01:03:01:02~DQB1*06:03:01</i>    | 0.07276 | 0.97287  |
| <i>DQA1*01:03:01:02~DQB1*06:39</i>       | 0.00187 | 1        |
| <i>DQA1*01:04:01:01~DQB1*05:03:01:01</i> | 0.01866 | 1        |
| <i>DQA1*01:04:01:03~DQB1*05:03:01:01</i> | 0.0056  | 1        |
| <i>DQA1*01:04:02~DQB1*05:03:01:01</i>    | 0.00187 | 1        |
| <i>DQA1*01:05:01~DQB1*05:01:01:02</i>    | 0.01866 | 1        |
| <i>DQA1*01:05:02~DQB1*05:03:01:01</i>    | 0.00187 | 0.4856   |
| <i>DQA1*01:05:02~DQB1*06:02:01</i>       | 0.00187 | 0.45194  |
| <i>DQA1*02:01:01:01~DQB1*02:02:01:01</i> | 0.14552 | 1        |
| <i>DQA1*02:01:01:01~DQB1*03:03:02:01</i> | 0.01493 | 1        |
| <i>DQA1*03:01:01~DQB1*03:02:01</i>       | 0.0709  | 0.97178  |
| <i>DQA1*03:01:01~DQB1*03:04:01</i>       | 0.00187 | 0.46076  |
| <i>DQA1*03:02~DQB1*03:03:02:02</i>       | 0.00933 | 1        |
| <i>DQA1*03:03:01:01~DQB1*02:02:01:02</i> | 0.00746 | 1        |
| <i>DQA1*03:03:01:01~DQB1*03:01:01:01</i> | 0.01866 | 0.60277  |
| <i>DQA1*03:03:01:01~DQB1*03:02:01</i>    | 0.02052 | 0.30294  |
| <i>DQA1*03:03:01:01~DQB1*03:02:03</i>    | 0.00187 | 1        |
| <i>DQA1*03:03:01:01~DQB1*03:04:01</i>    | 0.00187 | 0.47036  |
| <i>DQA1*03:03:01:01~DQB1*04:02:01</i>    | 0.0056  | 0.15257  |
| <i>DQA1*04:01:01~DQB1*04:02:01</i>       | 0.02239 | 1        |
| <i>DQA1*05:01:01:01~DQB1*02:01:01</i>    | 0.0541  | 1        |
| <i>DQA1*05:01:01:02~DQB1*02:01:01</i>    | 0.07276 | 1        |
| <i>DQA1*05:01:01:03~DQB1*02:01:01</i>    | 0.00933 | 1        |
| <i>DQA1*05:03~DQB1*03:01:01:01</i>       | 0.00187 | 1        |
| <i>DQA1*05:05:01:01~DQB1*03:01:01:02</i> | 0.01866 | 1        |
| <i>DQA1*05:05:01:01~DQB1*03:01:01:03</i> | 0.11007 | 1        |
| <i>DQA1*05:05:01:01~DQB1*03:19:01</i>    | 0.0056  | 0.71121  |
| <i>DQA1*05:05:01:03~DQB1*03:01:01:01</i> | 0.00933 | 0.82821  |

|  |         |          |
|--|---------|----------|
| <i>DQA1*05:05:01:03~DQB1*03:19:01</i>    | 0.00187 | 0.24151  |
| <b>Locus Pair DQB1~DRB1</b>              |         |          |
| <i>DQB1*02:01:01~DRB1*03:01:01:01</i>    | 0.13534 | 1        |
| <i>DQB1*02:02:01:01~DRB1*07:01:01:01</i> | 0.14474 | 1        |
| <i>DQB1*02:02:01:02~DRB1*04:05:01</i>    | 0.00752 | 1        |
| <i>DQB1*03:01:01:01~DRB1*04:01:01:01</i> | 0.00752 | 0.26357  |
| <i>DQB1*03:01:01:01~DRB1*04:07:01</i>    | 0.00752 | 1        |
| <i>DQB1*03:01:01:01~DRB1*04:08:01</i>    | 0.00376 | 0.65633  |
| <i>DQB1*03:01:01:01~DRB1*11:01:01:01</i> | 0.00188 | 0.02397  |
| <i>DQB1*03:01:01:01~DRB1*12:01:01:03</i> | 0.0094  | 1        |
| <i>DQB1*03:01:01:02~DRB1*11:04:01</i>    | 0.0188  | 1        |
| <i>DQB1*03:01:01:03~DRB1*04:05:01</i>    | 0.00188 | -0.41667 |
| <i>DQB1*03:01:01:03~DRB1*11:01:01:01</i> | 0.03759 | 0.94667  |
| <i>DQB1*03:01:01:03~DRB1*11:02:01</i>    | 0.00564 | 0.44     |
| <i>DQB1*03:01:01:03~DRB1*11:03:01</i>    | 0.0094  | 1        |
| <i>DQB1*03:01:01:03~DRB1*11:04:01</i>    | 0.0282  | 0.52615  |
| <i>DQB1*03:01:01:03~DRB1*13:03:01</i>    | 0.0188  | 0.89818  |
| <i>DQB1*03:01:01:03~DRB1*13:05:01</i>    | 0.00564 | 1        |
| <i>DQB1*03:02:01~DRB1*04:01:01:01</i>    | 0.0188  | 0.68465  |
| <i>DQB1*03:02:01~DRB1*04:02:01</i>       | 0.02256 | 1        |
| <i>DQB1*03:02:01~DRB1*04:03:01</i>       | 0.0094  | 0.68465  |
| <i>DQB1*03:02:01~DRB1*04:04:01</i>       | 0.02444 | 1        |
| <i>DQB1*03:02:01~DRB1*04:05:01</i>       | 0.0188  | 0.5861   |
| <i>DQB1*03:02:03~DRB1*04:05:01</i>       | 0.00188 | 1        |
| <i>DQB1*03:03:02:01~DRB1*07:01:01:01</i> | 0.01504 | 1        |
| <i>DQB1*03:03:02:02~DRB1*09:01:02</i>    | 0.00752 | 1        |
| <i>DQB1*03:03:02:02~DRB1*13:01:01:01</i> | 0.00188 | 0.13496  |
| <i>DQB1*03:04:01~DRB1*04:03:01</i>       | 0.00188 | 0.49333  |
| <i>DQB1*03:04:01~DRB1*04:08:01</i>       | 0.00188 | 0.49716  |
| <i>DQB1*03:19:01~DRB1*11:02:01</i>       | 0.00564 | 0.74715  |
| <i>DQB1*03:19:01~DRB1*13:04</i>          | 0.00188 | 1        |
| <i>DQB1*04:02:01~DRB1*04:03:01</i>       | 0.00188 | 0.11628  |
| <i>DQB1*04:02:01~DRB1*04:06:02</i>       | 0.00376 | 1        |
| <i>DQB1*04:02:01~DRB1*08:01:01</i>       | 0.02256 | 1        |
| <i>DQB1*04:02:01~DRB1*13:03:01</i>       | 0.00188 | 0.06272  |
| <i>DQB1*05:01:01:01~DRB1*01:02:01</i>    | 0.02444 | 1        |
| <i>DQB1*05:01:01:02~DRB1*10:01:01:01</i> | 0.01692 | 1        |
| <i>DQB1*05:01:01:02~DRB1*10:01:01:02</i> | 0.00188 | 1        |
| <i>DQB1*05:01:01:03~DRB1*01:01:01</i>    | 0.06391 | 1        |
| <i>DQB1*05:01:01:03~DRB1*01:03</i>       | 0.02068 | 1        |
| <i>DQB1*05:01:01:03~DRB1*13:02:01</i>    | 0.00188 | -0.11037 |

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>DQB1*05:02:01~DRB1*11:01:02</i>    | 0.00188 | 1        |
| <i>DQB1*05:02:01~DRB1*15:06:01</i>    | 0.00188 | 1        |
| <i>DQB1*05:02:01~DRB1*16:01:01</i>    | 0.01504 | 1        |
| <i>DQB1*05:02:01~DRB1*16:02:01:02</i> | 0.01128 | 1        |
| <i>DQB1*05:03:01:01~DRB1*14:01:01</i> | 0.00376 | 1        |
| <i>DQB1*05:03:01:01~DRB1*14:04:01</i> | 0.00188 | 1        |
| <i>DQB1*05:03:01:01~DRB1*14:54:01</i> | 0.02256 | 0.92085  |
| <i>DQB1*06:01:01~DRB1*15:01:01:01</i> | 0.00188 | 0.02296  |
| <i>DQB1*06:01:01~DRB1*15:02:01:02</i> | 0.01504 | 1        |
| <i>DQB1*06:02:01~DRB1*08:06</i>       | 0.00188 | 1        |
| <i>DQB1*06:02:01~DRB1*14:54:01</i>    | 0.00188 | -0.0906  |
| <i>DQB1*06:02:01~DRB1*15:01:01:01</i> | 0.08083 | 0.95115  |
| <i>DQB1*06:03:01~DRB1*11:04:01</i>    | 0.00188 | -0.51282 |
| <i>DQB1*06:03:01~DRB1*13:01:01:01</i> | 0.07143 | 0.94571  |
| <i>DQB1*06:03:01~DRB1*15:01:01:01</i> | 0.00564 | -0.20833 |
| <i>DQB1*06:04:01~DRB1*13:02:01</i>    | 0.01504 | 1        |
| <i>DQB1*06:09:01~DRB1*13:02:01</i>    | 0.00752 | 0.79499  |
| <i>DQB1*06:09:01~DRB1*15:01:01:01</i> | 0.00188 | 0.12066  |
| <i>DQB1*06:39~DRB1*13:01:01:01</i>    | 0.00188 | 1        |

**Locus Pair B~DRB1**

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*07:02:01~DRB1*01:01:01</i>       | 0.00702 | 0.0149   |
| <i>B*07:02:01~DRB1*01:03</i>          | 0.01454 | 0.68915  |
| <i>B*07:02:01~DRB1*03:01:01:01</i>    | 0.00356 | -0.72234 |
| <i>B*07:02:01~DRB1*04:02:01</i>       | 0.00368 | 0.07859  |
| <i>B*07:02:01~DRB1*04:04:01</i>       | 0.00184 | -0.19527 |
| <i>B*07:02:01~DRB1*04:05:01</i>       | 0.00197 | -0.29879 |
| <i>B*07:02:01~DRB1*07:01:01:01</i>    | 0.00747 | -0.5115  |
| <i>B*07:02:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.01716  |
| <i>B*07:02:01~DRB1*11:04:01</i>       | 0.00187 | -0.59012 |
| <i>B*07:02:01~DRB1*13:01:01:01</i>    | 0.00223 | -0.69838 |
| <i>B*07:02:01~DRB1*15:01:01:01</i>    | 0.04958 | 0.49079  |
| <i>B*07:05:01:01~DRB1*04:01:01:01</i> | 0.00184 | 0.48582  |
| <i>B*07:05:01:01~DRB1*04:05:01</i>    | 0.00184 | 0.48485  |
| <i>B*07:06:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.49159  |
| <i>B*07:06:01~DRB1*11:01:01:01</i>    | 0.00184 | 0.47992  |
| <i>B*08:01:01:01~DRB1*03:01:01:01</i> | 0.04503 | 0.65346  |
| <i>B*08:01:01:01~DRB1*04:02:01</i>    | 0.00184 | 0.0203   |
| <i>B*08:01:01:01~DRB1*08:01:01</i>    | 0.00184 | 0.0203   |
| <i>B*08:01:01:01~DRB1*13:01:01:01</i> | 0.00527 | 0.00505  |
| <i>B*08:01:01:01~DRB1*13:03:01</i>    | 0.00184 | 0.0284   |
| <i>B*08:01:01:01~DRB1*14:54:01</i>    | 0.00651 | 0.20145  |

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*08:01:01:01~DRB1*15:01:01:01</i> | 0.00202 | -0.65914 |
| <i>B*13:02:01~DRB1*03:01:01:01</i>    | 0.00184 | 0.01001  |
| <i>B*13:02:01~DRB1*07:01:01:01</i>    | 0.00919 | 0.65989  |
| <i>B*13:02:01~DRB1*08:01:01</i>       | 0.00184 | 0.12352  |
| <i>B*14:01:01~DRB1*04:07:01</i>       | 0.00368 | 0.49254  |
| <i>B*14:01:01~DRB1*07:01:01:01</i>    | 0.00919 | 0.55361  |
| <i>B*14:01:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.11028  |
| <i>B*14:02:01:01~DRB1*01:02:01</i>    | 0.01287 | 0.52085  |
| <i>B*14:02:01:01~DRB1*03:01:01:01</i> | 0.00686 | 0.06067  |
| <i>B*14:02:01:01~DRB1*04:05:01</i>    | 0.00184 | 0.02672  |
| <i>B*14:02:01:01~DRB1*07:01:01:01</i> | 0.00551 | -0.06207 |
| <i>B*14:02:01:01~DRB1*11:04:01</i>    | 0.00368 | 0.05483  |
| <i>B*14:02:01:01~DRB1*13:01:01:01</i> | 0.00332 | 0.01415  |
| <i>B*14:02:01:01~DRB1*13:03:01</i>    | 0.00184 | 0.05621  |
| <i>B*14:02:01:01~DRB1*14:54:01</i>    | 0.00085 | -0.10538 |
| <i>B*15:01:01:01~DRB1*01:01:01</i>    | 0.00194 | -0.18152 |
| <i>B*15:01:01:01~DRB1*04:01:01:01</i> | 0.00735 | 0.23868  |
| <i>B*15:01:01:01~DRB1*04:05:01</i>    | 0.00184 | 0.02672  |
| <i>B*15:01:01:01~DRB1*07:01:01:01</i> | 0.00368 | -0.37471 |
| <i>B*15:01:01:01~DRB1*08:01:01</i>    | 0.00184 | 0.04835  |
| <i>B*15:01:01:01~DRB1*11:03:01</i>    | 0.00551 | 0.40676  |
| <i>B*15:01:01:01~DRB1*13:01:01:01</i> | 0.01093 | 0.23855  |
| <i>B*15:01:01:01~DRB1*15:06:01</i>    | 0.00184 | 1        |
| <i>B*15:01:01:01~DRB1*16:01:01</i>    | 0.00184 | 0.0916   |
| <i>B*15:01:01:04~DRB1*08:01:01</i>    | 0.00184 | 0.48872  |
| <i>B*15:01:01:04~DRB1*13:01:01:01</i> | 0.00184 | 0.45817  |
| <i>B*15:03:01:02~DRB1*07:01:01:01</i> | 0.00184 | 0.20642  |
| <i>B*15:03:01:02~DRB1*13:01:01:01</i> | 0.00184 | 0.27756  |
| <i>B*15:03:01:02~DRB1*14:04:01</i>    | 0.00184 | 1        |
| <i>B*15:09~DRB1*13:02:01</i>          | 0.00184 | 1        |
| <i>B*15:220~DRB1*07:01:01:01</i>      | 0.00184 | 0.40481  |
| <i>B*15:220~DRB1*11:01:01:01</i>      | 0.00184 | 0.47992  |
| <i>B*18:01:01:01~DRB1*03:01:01:01</i> | 0.03421 | 0.82204  |
| <i>B*18:01:01:01~DRB1*07:01:01:01</i> | 0.00184 | -0.71578 |
| <i>B*18:01:01:01~DRB1*13:01:01:01</i> | 0.00255 | -0.18184 |
| <i>B*18:01:01:01~DRB1*16:02:01:02</i> | 0.00184 | 0.13155  |
| <i>B*18:01:01:02~DRB1*08:01:01</i>    | 0.00184 | 0.05556  |
| <i>B*18:01:01:02~DRB1*11:04:01</i>    | 0.01467 | 0.47368  |
| <i>B*18:01:01:02~DRB1*12:01:01:03</i> | 0.00184 | 0.17576  |
| <i>B*18:01:01:02~DRB1*14:54:01</i>    | 0.00184 | 0.04329  |
| <i>B*18:01:01:02~DRB1*15:01:01:01</i> | 0.00739 | 0.17537  |
| <i>B*18:01:01:02~DRB1*15:02:01:02</i> | 0.00184 | 0.09848  |
| <i>B*18:01:01:03~DRB1*11:01:02</i>    | 0.00184 | 1        |
| <i>B*27:02:01~DRB1*01:01:01</i>       | 0.00184 | 0.46562  |



|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*27:02:01~DRB1*07:01:01:01</i>    | 0.00184 | 0.40481  |
| <i>B*27:03~DRB1*04:03:01</i>          | 0.00184 | 1        |
| <i>B*27:05:02~DRB1*01:01:01</i>       | 0.01027 | 0.30448  |
| <i>B*27:05:02~DRB1*04:01:01:01</i>    | 0.00225 | 0.05369  |
| <i>B*27:05:02~DRB1*04:03:01</i>       | 0.00184 | 0.11688  |
| <i>B*27:05:02~DRB1*04:04:01</i>       | 0.00184 | 0.04895  |
| <i>B*27:05:02~DRB1*04:08:01</i>       | 0.00184 | 0.31313  |
| <i>B*27:05:02~DRB1*11:02:01</i>       | 0.00551 | 0.48485  |
| <i>B*27:05:02~DRB1*13:01:01:01</i>    | 0.00184 | -0.19048 |
| <i>B*27:05:02~DRB1*14:54:01</i>       | 0.00219 | 0.05724  |
| <i>B*27:05:02~DRB1*16:02:01:02</i>    | 0.00184 | 0.14141  |
| <i>B*35:01:01:01~DRB1*13:01:01:01</i> | 0.00184 | 1        |
| <i>B*35:01:01:02~DRB1*01:01:01</i>    | 0.01821 | 0.35479  |
| <i>B*35:01:01:02~DRB1*01:02:01</i>    | 0.00184 | 0.03246  |
| <i>B*35:01:01:02~DRB1*01:03</i>       | 0.00385 | 0.15122  |
| <i>B*35:01:01:02~DRB1*03:01:01:01</i> | 0.00396 | -0.35713 |
| <i>B*35:01:01:02~DRB1*04:01:01:01</i> | 0.00184 | 0.02171  |
| <i>B*35:01:01:02~DRB1*11:01:01:01</i> | 0.00184 | 0.00174  |
| <i>B*35:01:01:02~DRB1*13:01:01:01</i> | 0.00724 | 0.08711  |
| <i>B*35:01:01:02~DRB1*14:01:01</i>    | 0.00368 | 1        |
| <i>B*35:01:01:02~DRB1*15:01:01:01</i> | 0.00166 | -0.60681 |
| <i>B*35:01:01:02~DRB1*15:02:01:02</i> | 0.00184 | 0.08285  |
| <i>B*35:02:01~DRB1*11:01:01:01</i>    | 0.00224 | 0.10076  |
| <i>B*35:02:01~DRB1*11:04:01</i>       | 0.01246 | 0.74104  |
| <i>B*35:02:01~DRB1*13:01:01:01</i>    | 0.00184 | 0.03674  |
| <i>B*35:03:01~DRB1*01:01:01</i>       | 0.0077  | 0.1798   |
| <i>B*35:03:01~DRB1*03:01:01:01</i>    | 0.0029  | -0.3461  |
| <i>B*35:03:01~DRB1*04:03:01</i>       | 0.00184 | 0.11353  |
| <i>B*35:03:01~DRB1*04:08:01</i>       | 0.00184 | 0.31052  |
| <i>B*35:03:01~DRB1*07:01:01:01</i>    | 0.00078 | -0.85278 |
| <i>B*35:03:01~DRB1*09:01:02</i>       | 0.00368 | 0.48289  |
| <i>B*35:03:01~DRB1*11:01:01:01</i>    | 0.00184 | 0.01763  |
| <i>B*35:03:01~DRB1*12:01:01:03</i>    | 0.00184 | 0.17262  |
| <i>B*35:03:01~DRB1*14:54:01</i>       | 0.007   | 0.24727  |
| <i>B*35:03:01~DRB1*15:01:01:01</i>    | 0.00368 | 0.02114  |
| <i>B*35:08:01~DRB1*03:01:01:01</i>    | 0.0022  | 0.01784  |
| <i>B*35:08:01~DRB1*04:01:01:01</i>    | 0.00184 | 0.10019  |
| <i>B*35:08:01~DRB1*04:02:01</i>       | 0.00184 | 0.10526  |
| <i>B*35:08:01~DRB1*04:03:01</i>       | 0.00184 | 0.13006  |
| <i>B*35:08:01~DRB1*07:01:01:01</i>    | 0.00331 | 0.07789  |
| <i>B*35:08:01~DRB1*13:02:01</i>       | 0.00184 | 0.10358  |
| <i>B*35:08:01~DRB1*15:01:01:01</i>    | 0.00184 | 0.03644  |
| <i>B*37:01:01~DRB1*01:01:01</i>       | 0.00184 | 0.0203   |
| <i>B*37:01:01~DRB1*01:02:01</i>       | 0.00184 | 0.06089  |

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*37:01:01~DRB1*11:01:01:01</i>    | 0.00184 | 0.04653  |
| <i>B*37:01:01~DRB1*11:03:01</i>       | 0.00368 | 0.2696   |
| <i>B*37:01:01~DRB1*13:01:01:01</i>    | 0.00285 | 0.05656  |
| <i>B*37:01:01~DRB1*13:02:01</i>       | 0.00368 | 0.14626  |
| <i>B*37:01:01~DRB1*15:01:01:01</i>    | 0.0045  | 0.12336  |
| <i>B*37:01:01~DRB1*15:02:01:02</i>    | 0.00184 | 0.10526  |
| <i>B*38:01:01~DRB1*04:02:01</i>       | 0.00551 | 0.22581  |
| <i>B*38:01:01~DRB1*04:04:01</i>       | 0.00184 | 0.04715  |
| <i>B*38:01:01~DRB1*07:01:01:01</i>    | 0.00196 | -0.60824 |
| <i>B*38:01:01~DRB1*08:06</i>          | 0.00184 | 1        |
| <i>B*38:01:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.08244  |
| <i>B*38:01:01~DRB1*11:01:01:01</i>    | 0.00172 | 0.01705  |
| <i>B*38:01:01~DRB1*11:04:01</i>       | 0.00368 | 0.07336  |
| <i>B*38:01:01~DRB1*13:01:01:01</i>    | 0.00919 | 0.23506  |
| <i>B*38:01:01~DRB1*13:03:01</i>       | 0.00368 | 0.15543  |
| <i>B*38:20~DRB1*07:01:01:01</i>       | 0.00184 | 1        |
| <i>B*39:01:01:03~DRB1*03:01:01:01</i> | 0.00368 | 0.23001  |
| <i>B*39:01:01:03~DRB1*13:01:01:01</i> | 0.00184 | 0.09695  |
| <i>B*39:01:01:03~DRB1*16:01:01</i>    | 0.00368 | 0.32338  |
| <i>B*39:01:01:03~DRB1*16:02:01:02</i> | 0.00184 | 0.15737  |
| <i>B*39:06:02~DRB1*08:01:01</i>       | 0.00184 | 1        |
| <i>B*40:01:02~DRB1*01:02:01</i>       | 0.00184 | 0.06865  |
| <i>B*40:01:02~DRB1*04:01:01:01</i>    | 0.00184 | 0.06513  |
| <i>B*40:01:02~DRB1*04:04:01</i>       | 0.01103 | 0.53433  |
| <i>B*40:01:02~DRB1*07:01:01:01</i>    | 0.00184 | -0.43156 |
| <i>B*40:01:02~DRB1*15:01:01:01</i>    | 0.00368 | 0.09901  |
| <i>B*40:02:01~DRB1*04:05:01</i>       | 0.00184 | 0.07273  |
| <i>B*40:02:01~DRB1*04:07:01</i>       | 0.00368 | 0.49064  |
| <i>B*40:02:01~DRB1*11:04:01</i>       | 0.00368 | 0.15985  |
| <i>B*40:02:01~DRB1*13:01:01:01</i>    | 0.00551 | 0.24143  |
| <i>B*40:02:01~DRB1*16:01:01</i>       | 0.00368 | 0.23596  |
| <i>B*40:06:01:02~DRB1*03:01:01:01</i> | 0.00184 | 0.13376  |
| <i>B*40:06:01:02~DRB1*15:01:01:01</i> | 0.00184 | 0.17409  |
| <i>B*40:06:01:02~DRB1*16:01:01</i>    | 0.00184 | 0.23881  |
| <i>B*40:06:01:02~DRB1*16:02:01:02</i> | 0.00184 | 0.24164  |
| <i>B*40:12~DRB1*07:01:01:01</i>       | 0.00184 | 1        |
| <i>B*41:01:01~DRB1*03:01:01:01</i>    | 0.00184 | 0.23001  |
| <i>B*41:01:01~DRB1*07:01:01:01</i>    | 0.00184 | 0.20642  |
| <i>B*41:01:01~DRB1*13:05:01</i>       | 0.00184 | 0.32964  |
| <i>B*41:02:01~DRB1*13:03:01</i>       | 0.00184 | 1        |
| <i>B*44:02:01:01~DRB1*01:01:01</i>    | 0.00669 | 0.04909  |
| <i>B*44:02:01:01~DRB1*03:01:01:01</i> | 0.00924 | 0.02103  |
| <i>B*44:02:01:01~DRB1*04:01:01:01</i> | 0.00511 | 0.13252  |
| <i>B*44:02:01:01~DRB1*04:02:01</i>    | 0.00358 | 0.10798  |

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*44:02:01:01~DRB1*04:03:01</i>    | 0.00184 | 0.0875   |
| <i>B*44:02:01:01~DRB1*07:01:01:01</i> | 0.00342 | -0.64769 |
| <i>B*44:02:01:01~DRB1*08:01:01</i>    | 0.00368 | 0.11285  |
| <i>B*44:02:01:01~DRB1*09:01:02</i>    | 0.00184 | 0.20157  |
| <i>B*44:02:01:01~DRB1*10:01:01:01</i> | 0.00184 | 0.05371  |
| <i>B*44:02:01:01~DRB1*11:01:01:01</i> | 0.00184 | -0.21501 |
| <i>B*44:02:01:01~DRB1*11:03:01</i>    | 0.00184 | 0.0875   |
| <i>B*44:02:01:01~DRB1*12:01:01:03</i> | 0.00368 | 0.36125  |
| <i>B*44:02:01:01~DRB1*13:01:01:01</i> | 0.01206 | 0.13174  |
| <i>B*44:02:01:01~DRB1*13:02:01</i>    | 0.00184 | 0.01731  |
| <i>B*44:02:01:01~DRB1*15:01:01:01</i> | 0.00219 | -0.60738 |
| <i>B*44:03:01:01~DRB1*01:01:01</i>    | 0.00201 | -0.63104 |
| <i>B*44:03:01:01~DRB1*03:01:01:01</i> | 0.00197 | -0.82599 |
| <i>B*44:03:01:01~DRB1*04:02:01</i>    | 0.00378 | 0.09469  |
| <i>B*44:03:01:01~DRB1*04:05:01</i>    | 0.00722 | 0.1758   |
| <i>B*44:03:01:01~DRB1*07:01:01:01</i> | 0.06211 | 0.68394  |
| <i>B*44:03:01:01~DRB1*11:01:01:01</i> | 0.0038  | 0.01503  |
| <i>B*44:03:01:01~DRB1*15:01:01:01</i> | 0.00368 | -0.52696 |
| <i>B*44:04~DRB1*11:01:01:01</i>       | 0.00184 | 1        |
| <i>B*44:05:01~DRB1*16:01:01</i>       | 0.00184 | 1        |
| <i>B*45:01:01~DRB1*03:01:01:01</i>    | 0.00184 | 0.13376  |
| <i>B*45:01:01~DRB1*11:01:01:01</i>    | 0.00144 | 0.16288  |
| <i>B*45:01:01~DRB1*11:04:01</i>       | 0.00224 | 0.26991  |
| <i>B*45:01:01~DRB1*16:02:01:02</i>    | 0.00184 | 0.24164  |
| <i>B*47:01:01:03~DRB1*04:05:01</i>    | 0.00368 | 0.65644  |
| <i>B*47:01:01:03~DRB1*12:01:01:03</i> | 0.00184 | 0.32715  |
| <i>B*49:01:01~DRB1*01:01:01</i>       | 0.00236 | -0.09423 |
| <i>B*49:01:01~DRB1*01:03</i>          | 0.00184 | 0.05259  |
| <i>B*49:01:01~DRB1*04:03:01</i>       | 0.00368 | 0.25561  |
| <i>B*49:01:01~DRB1*04:04:01</i>       | 0.00184 | 0.03802  |
| <i>B*49:01:01~DRB1*04:05:01</i>       | 0.00316 | 0.06975  |
| <i>B*49:01:01~DRB1*07:01:01:01</i>    | 0.00551 | -0.14734 |
| <i>B*49:01:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.07365  |
| <i>B*49:01:01~DRB1*10:01:01:02</i>    | 0.00184 | 1        |
| <i>B*49:01:01~DRB1*11:01:01:01</i>    | 0.00368 | 0.05711  |
| <i>B*49:01:01~DRB1*11:02:01</i>       | 0.00551 | 0.47893  |
| <i>B*49:01:01~DRB1*13:02:01</i>       | 0.00551 | 0.19835  |
| <i>B*49:01:01~DRB1*13:05:01</i>       | 0.00368 | 0.65262  |
| <i>B*50:01:01~DRB1*03:01:01:01</i>    | 0.00551 | 0.13376  |
| <i>B*50:01:01~DRB1*07:01:01:01</i>    | 0.00735 | 0.20642  |
| <i>B*50:01:01~DRB1*10:01:01:01</i>    | 0.00184 | 0.09106  |
| <i>B*50:01:01~DRB1*11:03:01</i>       | 0.00184 | 0.12352  |
| <i>B*50:01:01~DRB1*13:03:01</i>       | 0.00551 | 0.25632  |
| <i>B*50:02~DRB1*01:02:01</i>          | 0.00184 | 0.12187  |

|                                       |         |          |
|---------------------------------------|---------|----------|
| <i>B*50:02~DRB1*04:06:02</i>          | 0.00368 | 1        |
| <i>B*50:02~DRB1*08:01:01</i>          | 0.00184 | 0.12352  |
| <i>B*50:02~DRB1*13:01:01:01</i>       | 0.00184 | 0.07114  |
| <i>B*50:02~DRB1*13:03:01</i>          | 0.00368 | 0.27097  |
| <i>B*51:01:01:01~DRB1*01:01:01</i>    | 0.00448 | 0.00779  |
| <i>B*51:01:01:01~DRB1*03:01:01:01</i> | 0.00585 | -0.30241 |
| <i>B*51:01:01:01~DRB1*04:01:01:01</i> | 0.00231 | 0.02287  |
| <i>B*51:01:01:01~DRB1*04:04:01</i>    | 0.00368 | 0.09744  |
| <i>B*51:01:01:01~DRB1*04:05:01</i>    | 0.00236 | 0.0188   |
| <i>B*51:01:01:01~DRB1*04:08:01</i>    | 0.00184 | 0.28889  |
| <i>B*51:01:01:01~DRB1*07:01:01:01</i> | 0.01239 | 0.04563  |
| <i>B*51:01:01:01~DRB1*08:01:01</i>    | 0.00551 | 0.2      |
| <i>B*51:01:01:01~DRB1*09:01:02</i>    | 0.00184 | 0.2      |
| <i>B*51:01:01:01~DRB1*11:01:01:01</i> | 0.00919 | 0.1873   |
| <i>B*51:01:01:01~DRB1*11:04:01</i>    | 0.00184 | -0.38462 |
| <i>B*51:01:01:01~DRB1*13:01:01:01</i> | 0.00318 | -0.341   |
| <i>B*51:01:01:01~DRB1*14:54:01</i>    | 0.00184 | 0.00952  |
| <i>B*51:01:01:01~DRB1*15:01:01:01</i> | 0.00619 | 0.00794  |
| <i>B*51:08:01~DRB1*16:02:01:02</i>    | 0.00184 | 1        |
| <i>B*52:01:01:02~DRB1*11:04:01</i>    | 0.00184 | 0.12484  |
| <i>B*52:01:01:02~DRB1*15:02:01:02</i> | 0.00919 | 0.83085  |
| <i>B*52:01:02~DRB1*15:01:01:01</i>    | 0.00184 | 1        |
| <i>B*53:01:01~DRB1*01:02:01</i>       | 0.00368 | 0.31701  |
| <i>B*53:01:01~DRB1*04:05:01</i>       | 0.00184 | 0.14141  |
| <i>B*53:01:01~DRB1*13:02:01</i>       | 0.00551 | 0.48776  |
| <i>B*55:01:01~DRB1*04:04:01</i>       | 0.00184 | 0.10358  |
| <i>B*55:01:01~DRB1*07:01:01:01</i>    | 0.00184 | -0.21839 |
| <i>B*55:01:01~DRB1*11:01:01:01</i>    | 0.00184 | 0.08987  |
| <i>B*55:01:01~DRB1*13:02:01</i>       | 0.00184 | 0.10358  |
| <i>B*55:01:01~DRB1*13:03:01</i>       | 0.00184 | 0.10694  |
| <i>B*55:01:01~DRB1*14:54:01</i>       | 0.00184 | 0.10189  |
| <i>B*55:01:01~DRB1*15:01:01:01</i>    | 0.00184 | 0.03644  |
| <i>B*55:01:01~DRB1*16:01:01</i>       | 0.00184 | 0.11194  |
| <i>B*57:01:01~DRB1*04:01:01:01</i>    | 0.0032  | 0.27003  |
| <i>B*57:01:01~DRB1*07:01:01:01</i>    | 0.00599 | 0.45621  |
| <i>B*57:01:01~DRB1*11:04:01</i>       | 0.00184 | 0.12484  |
| <i>B*57:03:01:02~DRB1*14:54:01</i>    | 0.00184 | 1        |
| <i>B*58:01:01:01~DRB1*03:01:01:01</i> | 0.00184 | -0.32254 |
| <i>B*58:01:01:01~DRB1*04:02:01</i>    | 0.00184 | 0.0704   |
| <i>B*58:01:01:01~DRB1*07:01:01:01</i> | 0.00551 | 0.13427  |
| <i>B*58:01:01:01~DRB1*10:01:01:01</i> | 0.00368 | 0.20617  |
| <i>B*58:01:01:01~DRB1*11:01:01:01</i> | 0.00184 | 0.05441  |
| <i>B*58:01:01:01~DRB1*13:02:01</i>    | 0.00184 | 0.06865  |
| <i>B*58:01:01:01~DRB1*13:04</i>       | 0.00184 | 1        |

|                                    |         |        |
|------------------------------------|---------|--------|
| <i>B*58:01:01:01~DRB1*14:54:01</i> | 0.00184 | 0.0669 |
| <i>B*73:01~DRB1*04:05:01</i>       | 0.00184 | 1      |

## Notes:

(\*) In March 2017, posterior to release of IPD-IMGT/HLA database version 3.25.0, the *HLA-DPA1\*02:02:01* allele was deleted from the official WHO HLA Nomenclature as its sequence has now been shown to be in error and is identical to *HLA-DPA1\*02:07:01* allele (<https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/deleted.cgi>).

### **7. GLOBAL MEASURES OF PAIRWISE LINKAGE DISEQUILIBRIUM (LD) FOR HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DOB1, -DRB1 AND -DRB3/4/5 LOCI IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

To evaluate the overall linkage disequilibrium (LD) we considered (**Table R-6**) two different locus-pair level measures. The D' (normalized Hedrick's D' statistic) parameter, expressed as the normalization of the product of allele frequencies at each locus, weights the LD contribution of specific allele pairs [98][780][787]. Whereas the second parameter, Wn (Cramer's V statistic), calculates also a normalization in this case of the chi-square statistic for deviations between observed and expected haplotype frequencies [785]. As expected, the strongest associations are observed for the contiguous and/or physically close HLA loci pairs including *DRB1~DRB5/4/3*, *DRB1~DQA1*, *DQA1~DQB1* and *DRB1~DQB1* followed by *B~C*. *HLA-DPA1~DPB1* pair appears associated with less strength. Interestingly, in spite of *HLA-A~C* pair being physically closer than *HLA-A~B* the strength of the LD in the latter loci pair is higher, suggesting that differences in diversity between *HLA-B* and *-C* loci may play a role in determining this measurement. Associations between *HLA-A~B* and *HLA-B~DRB1* appear in similar ranges. *HLA-DP* loci show weaker LD associations than any of the other pairwise comparisons. As previously reported [92][806][807], LD patterns of *HLA-DP* loci seem to be driven primarily in a different manner

compared to the other *HLA* loci (e.g. relatively higher rate of recombination and combined *DPA1~DPB1* amino acid epitope have been suggested to contribute on this distinctive selection).

**Table R-6.** Global measures of pairwise linkage disequilibrium (LD) for *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci at the 3-/4-field resolution level (and according to IPD-IMGT/HLA database version 3.25.0) in this 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects).

| Locus Pair<br>HLA- | D'      | Wn      |
|--------------------|---------|---------|
| <i>B~C</i>         | 0.93630 | 0.77226 |
| <i>A~C</i>         | 0.59490 | 0.43926 |
| <i>A~B</i>         | 0.64088 | 0.45530 |
| <i>DPA1~DPB1</i>   | 0.82896 | 0.71883 |
| <i>DQA1~DQB1</i>   | 0.97854 | 0.78901 |
| <i>DQA1~DRB1</i>   | 0.98990 | 0.86147 |
| <i>DQB1~DRB1</i>   | 0.97446 | 0.80953 |
| <i>DPB1~DRB1</i>   | 0.47923 | 0.37854 |
| <i>DPB1~DQB1</i>   | 0.43446 | 0.38512 |
| <i>B~DRB1</i>      | 0.70620 | 0.46705 |
| <i>B~DQA1</i>      | 0.66583 | 0.49954 |
| <i>B~DQB1</i>      | 0.65365 | 0.49127 |
| <i>DRB1~DRB3</i>   | 0.96724 | 0.86672 |
| <i>DRB1~DRB4</i>   | 0.97158 | 0.69772 |
| <i>DRB1~DRB5</i>   | 1       | 1       |

## **8. ESTIMATION OF EXTENDED HLA HAPLOTYPE FREQUENCIES IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

Maximum likelihood estimation via an expectation-maximization (EM) algorithm is a statistical method commonly used for HLA haplotype inference and estimation of haplotype frequency distributions in unrelated individuals from a population-specific genotype data as in the present study. Moreover, this statistical method serves as an alternate approach when it is not possible to rely on family segregation studies [342]. Inferred extended HLA haplotypes (encompassing 6-locus, 7-locus and 9-locus respectively) were evaluated for the estimation of haplotype frequencies in this Spanish population cohort (see also data on 17<sup>th</sup>-IHIWS database [297] for further details):

*HLA~A~C~B~DRB3/4/5~DRB1~DQB1* (**Table R-7**);

*HLA~A~C~B~DRB3/4/5~DRB1~DQA1~DQB1* (**Table R-8**); and

*HLA~A~C~B~DRB3/4/5~DRB1~DQA1~DQB1~DPA1~DPB1* (**Table R-9**).

Similarly to what it was found in 2-locus haplotypes, it can be observed very distinctive extended haplotype associations in non-coding regions at the 4-field level that are not apparent, and indeed unattainable, at lower allele resolution level (2-field or 3-field) results that are obtained when using legacy methodologies (e.g. SSP or SSO, or even SBT depending on the given HLA gene sequence coverage) with important limitations in sequence coverage and phasing in comparison to NGS-based HLA genotyping [137][178]. Just to be noted, in the present Spanish population cohort, 3-/4-field HLA data of most common extended haplotype frequency distributions is generally shown in the present thesis work document (see **RESULTS** section (**Tables R-5, R-7, R-8, R-9**)). Whereas all respective collapsed 2-field HLA data of most common extended haplotype frequency distributions is mostly not shown in the present thesis work document with some exceptions that are later highlighted in some parts of the **DISCUSSION** section.

**Table R-7.** HLA-A~C~B~DRB3/4/5~DRB1~DQB1 extended haplotypes with estimated frequencies (HF, in decimals) of 0.020 or more (at the 3-/4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) of the 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects). HLA haplotypes are sorted by frequency in descending order.

| Haplotype HLA-  | HF    | Rank# |
|---|-------|-------|
| <i>A*01:01:01:01~C*07:01:01:01~B*08:01:01:01~DRB3*01:01:02:01~DRB1*03:01:01:01~DQB1*02:01:01</i>    | 0.080 | 1     |
| <i>A*29:02:01:01~C*16:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQB1*02:02:01:01</i> | 0.075 | 2     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01~DRB5*01:01:01~DRB1*15:01:01:01~DQB1*06:02:01</i>          | 0.050 | 3     |
| <i>A*30:02:01:01~C*05:01:01:01~B*18:01:01:01~DRB3*02:02:01:01~DRB1*03:01:01:01~DQB1*02:01:01</i>    | 0.050 | 3     |
| <i>A*03:01:01:01~C*07:02:01:03~B*07:02:01~DRB5*01:01:01~DRB1*15:01:01:01~DQB1*06:02:01</i>          | 0.045 | 4     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01~DRB3/4/5*Absent~DRB1*01:03~DQB1*05:01:01:03</i>           | 0.030 | 5     |
| <i>A*11:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB3/4/5*Absent~DRB1*01:01:01~DQB1*05:01:01:03</i>     | 0.025 | 6     |
| <i>A*03:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB3/4/5*Absent~DRB1*01:01:01~DQB1*05:01:01:03</i>     | 0.020 | 7     |
| <i>A*33:01:01~C*08:02:01:01~B*14:02:01:01~DRB3/4/5*Absent~DRB1*01:02:01~DQB1*05:01:01:01</i>        | 0.020 | 7     |
| <i>A*25:01:01~C*12:03:01:01~B*18:01:01:02~DRB5*01:01:01~DRB1*15:01:01:01~DQB1*06:02:01</i>          | 0.020 | 7     |
| <i>A*23:01:01~C*04:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQB1*02:02:01:01</i>    | 0.020 | 7     |
| <i>A*24:02:01:01~C*04:01:01:06~B*35:02:01~DRB3*02:02:01:02~DRB1*11:04:01~DQB1*03:01:01:02</i>       | 0.020 | 7     |



**Table R-8.** HLA-A~C~B~DRB3/4/5~DRB1~DQA1~DQB1 extended haplotypes with estimated frequencies (HF, in decimals) of 0.020 or more (at the 3-/4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) of the 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects). HLA haplotypes are sorted by frequency in descending order.

| Haplotype HLA-   | HF    | Rank# |
|--|-------|-------|
| <i>A*01:01:01:01~C*07:01:01:01~B*08:01:01:01~DRB3*01:01:02:01~DRB1*03:01:01:01~DQA1*05:01:01:02~DQB1*02:01:01</i>    | 0.078 | 1     |
| <i>A*29:02:01:01~C*16:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQA1*02:01:01:01~DQB1*02:02:01:01</i> | 0.078 | 1     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01</i>    | 0.052 | 2     |
| <i>A*03:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01</i>    | 0.047 | 3     |
| <i>A*30:02:01:01~C*05:01:01:01~B*18:01:01:01~DRB3*02:02:01:01~DRB1*03:01:01:01~DQA1*05:01:01:01~DQB1*02:01:01</i>    | 0.047 | 3     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB3/4/5*Absent~DRB1*01:03~DQA1*01:01:01:02~DQB1*05:01:01:03</i>        | 0.031 | 4     |
| <i>A*11:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB3/4/5*Absent~DRB1*01:01:01:01~DQA1*01:01:01:02~DQB1*05:01:01:03</i>  | 0.026 | 5     |
| <i>A*03:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB3/4/5*Absent~DRB1*01:01:01:01~DQA1*01:01:01:02~DQB1*05:01:01:03</i>  | 0.021 | 6     |
| <i>A*33:01:01:01~C*08:02:01:01~B*14:02:01:01~DRB3/4/5*Absent~DRB1*01:02:01~DQA1*01:01:02~DQB1*05:01:01:01</i>        | 0.021 | 6     |
| <i>A*25:01:01:01~C*12:03:01:01~B*18:01:01:02~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01</i>    | 0.021 | 6     |
| <i>A*23:01:01:01~C*04:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQA1*02:01:01:01~DQB1*02:02:01:01</i> | 0.021 | 6     |
| <i>A*24:02:01:01~C*04:01:01:06~B*35:02:01:01~DRB3*02:02:01:02~DRB1*11:04:01~DQA1*05:05:01:01~DQB1*03:01:01:02</i>    | 0.021 | 6     |
| <i>A*02:01:01:01~C*03:04:01:01~B*40:01:01:02~DRB4*01:03:01:01~DRB1*04:04:01~DQA1*03:01:01~DQB1*03:02:01</i>          | 0.021 | 6     |

**Table R-9.** HLA-A~C~B~DRB3/4/5~DRB1~DQA1~DQB1~DPA1~DPB1 extended haplotypes with estimated frequencies (HF, in decimals) of 0.020 or more (at the 3-/4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) of the 17<sup>th</sup>-IHIW Spanish population cohort (n=282 subjects). HLA haplotypes are sorted by frequency in descending order.

| Haplotype HLA-   | HF    | Rank# |
|--|-------|-------|
| <i>A*01:01:01:01~C*07:01:01:01~B*08:01:01:01~DRB3*01:01:02:01~DRB1*03:01:01:01~DQA1*05:01:01:02~DQB1*02:01:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.061 | 1     |
| <i>A*03:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.061 | 1     |
| <i>A*01:01:01:01~C*07:01:01:01~B*08:01:01:01~DRB3*01:01:02:01~DRB1*03:01:01:01~DQA1*05:01:01:02~DQB1*02:01:01:01~DPA1*02:01:02:01~DPB1*01:01:01:01</i> | 0.061 | 1     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.061 | 1     |
| <i>A*29:02:01:01~C*16:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQA1*02:01:01:01~DQB1*02:02:01:01~DPA1*02:01:01:01~DPB1*11:01:01:01</i> | 0.061 | 1     |
| <i>A*29:02:01:01~C*16:01:01:01~B*44:03:01:01~DRB4*01:01:01:01~DRB1*07:01:01:01~DQA1*02:01:01:01~DQB1*02:02:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.051 | 2     |
| <i>A*30:02:01:01~C*05:01:01:01~B*18:01:01:01~DRB3*02:02:01:01~DRB1*03:01:01:01~DQA1*05:01:01:01~DQB1*02:01:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.051 | 2     |
| <i>A*25:01:01:01~C*12:03:01:01~B*18:01:01:02~DRB5*01:01:01:01~DRB1*15:01:01:01~DQA1*01:02:01:01~DQB1*06:02:01:01~DPA1*01:03:01:02~DPB1*04:01:01:01</i> | 0.041 | 3     |
| <i>A*02:01:01:01~C*07:02:01:03~B*07:02:01:01~DRB3/4/5*Absent~DRB1*01:03~DQA1*01:01:01:02~DQB1*05:01:01:03~DPA1*01:03:01:02~DPB1*04:01:01:01</i>        | 0.041 | 3     |
| <i>A*02:01:01:01~C*03:04:01:01~B*40:01:02:01~DRB4*01:03:01:01~DRB1*04:04:01~DQA1*03:01:01:01~DQB1*03:02:01:01~DPA1*01:03:01:03~DPB1*06:01:01:01</i>    | 0.031 | 4     |
| <i>A*24:02:01:01~C*04:01:01:06~B*35:02:01:01~DRB3*02:02:01:02~DRB1*11:04:01~DQA1*05:05:01:01~DQB1*03:01:01:02~DPA1*01:03:01:04~DPB1*04:01:01:01</i>    | 0.031 | 4     |
| <i>A*03:01:01:01~C*04:01:01:01~B*35:01:01:02~DRB3/4/5*Absent~DRB1*01:01:01:01~DQA1*01:01:01:02~DQB1*05:01:01:03~DPA1*01:03:01:02~DPB1*04:01:01:01</i>  | 0.020 | 5     |
| <i>A*11:01:01:01~C*05:01:01:02~B*44:02:01:01~DRB3/4/5*Absent~DRB1*01:01:01:01~DQA1*01:01:01:02~DQB1*05:01:01:03~DPA1*01:03:01:02~DPB1*04:01:01:01</i>  | 0.020 | 5     |

**HF:** Haplotype frequencies

Haplotype frequencies were estimated using a maximum likelihood estimation via an expectation-maximization (EM) algorithm with Hapl-o-Mat version 1.1 software [342]. Only haplotypes with frequencies higher than 0.020 are shown and haplotypes are sorted from the highest to the lowest frequency.

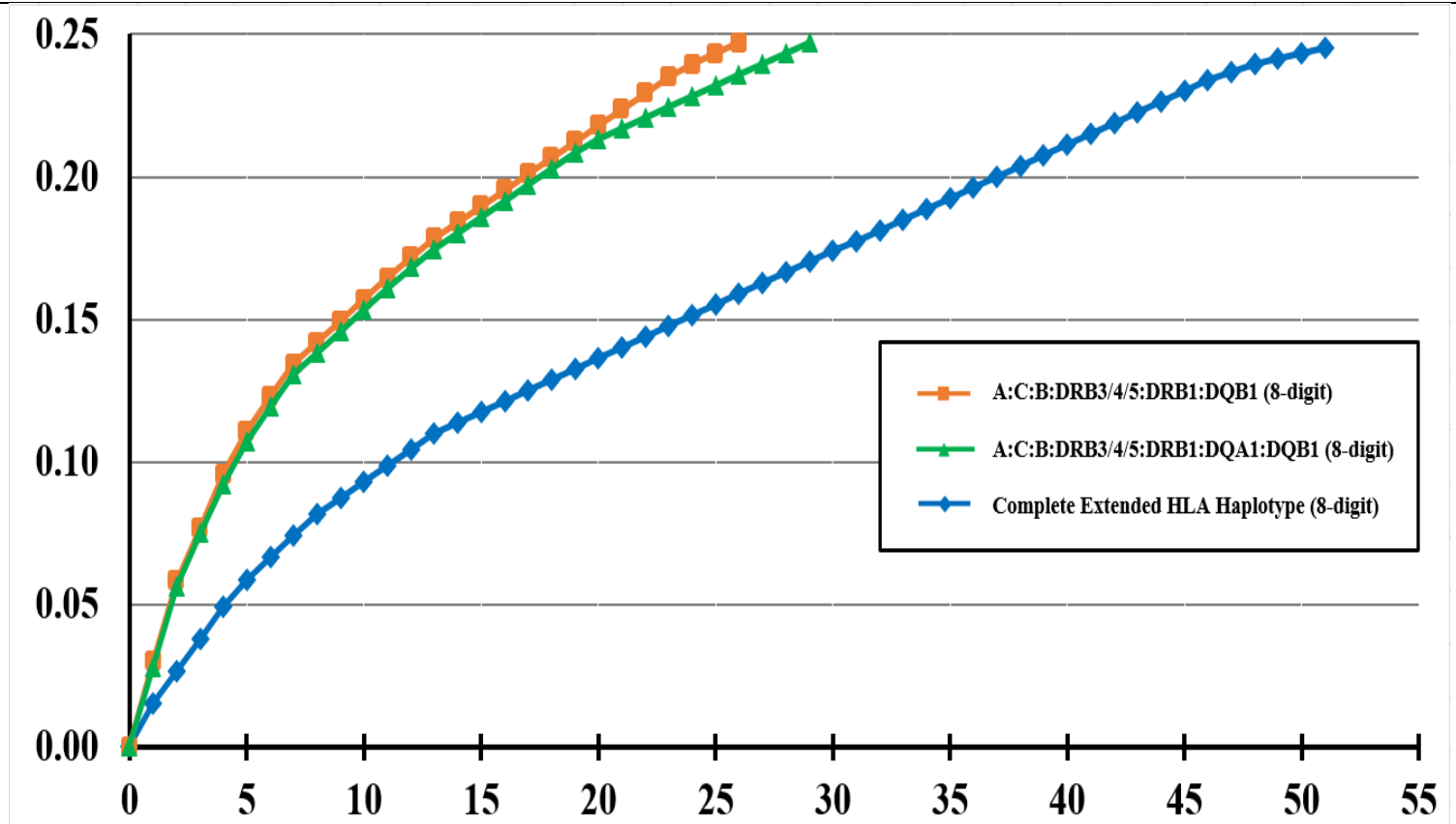
For estimation of extended haplotype frequencies, *HLA-DRB3*, *-DRB4*, and *-DRB5* genes are considered alleles of a single locus as the presence of one of these genes excludes the presence of the other two genes at the haplotype level [344]. Thus, extended haplotype frequencies results are in accordance with linkage constraints that exist respectively between the *HLA-DRB3/4/5* loci and the *HLA-DRB1* locus, in which several *HLA-DRB1* allele families are defined [344].

### **9. EVALUATION OF 3-/4-FIELD EXTENDED HAPLOTYPE DIVERSITY GIVEN BY INCLUSION OF *HLA-DPA1* AND *-DPB1* LOCI, IN 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

The patterns of population distribution of HLA haplotypes estimated at three different HLA haplotype extension degrees ((1, orange colored graph) *HLA-A~C~B~DRB3/4/5~DRB1~DQB1*; (2, green colored graph) *HLA-A~C~B ~DRB3/4/5~DRB1~DQA1~DQB1*; and (3, blue colored graph) complete extended haplotype *HLA-A~C~B ~DRB3/4/5~DRB1~DQA1~DQB1~DPA1~DPB1* of this Spanish population reference cohort were evaluated by comparing their respective cumulative haplotypes frequencies (sorted from the most frequent to the least frequent) as it is shown through **Figure R-3**. It can be observed how the two less extended types of HLA haplotypes [(1) *HLA-A~C~B~DRB3/4/5~DRB1~DQB1* and (2) *HLA-A~C~B ~DRB3/4/5~DRB1~DQA1~DQB1*] present a similar distribution and they virtually overlap. However, in the case of the complete extended 9-locus haplotype (when including *HLA-DPA1* and *HLA-DPB1* loci) (3) its cumulative frequencies distribution is shifted to the right in comparison to the other two haplotype distributions (1) and (2). This points out that more haplotypes (number of distinct haplotypes~50) are required, in the case of this complete extended haplotype distribution (3), to cover the same combined cumulative frequency presented by the other two type of haplotypes distributions (number of distinct haplotypes~25-30) (1)(2). Therefore,

the haplotype diversity is dramatically increased when including *HLA-DPA1* and *-DPB1* loci and, consequently, the linkage disequilibrium (LD) decreases at this maximum level of haplotype extension (9-locus). As previously mentioned and reported in other recent studies, many identical haplotypes across 7 loci (comprising *HLA-A~B~C~DRB3/4/5~DRB1~DQA1~DQB1*, and excluding *HLA-DPA1* and *-DPB1*) become extremely divergent in terms of the multiplicity of *HLA-DP* alleles with which they associate [268][286][287][331]. This seems to be especially due to the weak LD between *HLA-DP* and the rest of the class II haplotype since existing hotspot of recombination is present between *HLA-DQ* and *-DP* loci [92]. Therefore, this effect has direct implications and consequences, for example, in relation to the lesser likelihood of finding unrelated donors (URD) in HSCT when evaluating 3-/4-field resolution in addition to include *HLA-DP* loci in the given URD search process [233]; or, also, in the very high sample size that may be required for study cohorts in order to be representative and meaningful for evaluating anthropological aspects based on these 3-/4-field extended HLA 9-locus haplotype frequency distributions.

**Figure R-3.** Graph of cumulative frequencies (in decimals) of several different extended 4-field (or 8-digits) HLA haplotypes plotted against respective number of haplotypes defined in each given distribution.



(ORANGE graph) Cumulative frequencies of the *HLA-A~C~B~DRB3/4/5~DRB1~DQB1* extended haplotypes (8-digits or 4-field allele resolution level) sorted from the most frequent to the least frequent were plotted against the respective number of haplotypes.

(GREEN graph) Cumulative frequencies of the *HLA-A~C~B~DRB3/4/5~DRB1~DQA1~DQB1* extended haplotypes (8-digits or 4-field allele resolution level) sorted from the most frequent to the least frequent were plotted against the respective number of haplotypes.

(BLUE graph) Cumulative frequencies of the *HLA-A~C~B~DRB3/4/5~DRB1~DQA1~DQB1~DPA1~DPB1* extended haplotypes (8-digits or 4-field allele resolution level) sorted from the most frequent to the least frequent were plotted against the respective number of haplotypes.

---

**10. HLA ALLELE FREQUENCY DISTRIBUTIONS AND RELATEDNESS WITHIN SPANISH REGIONAL GROUPS FROM 17<sup>TH</sup>-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

Disparity/similarity of allelic distributions were evaluated within this Spanish population cohort based on the results (at the 3-/4-field allele resolution level) of the current study. In this sense, a comparison of HLA allele distributions was carried out (specifically based on allele frequencies found at *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* loci) between the 3 different geographical Spanish regions established (Northern-Central, Eastern and Southern Spain) as well as between the 10 Spanish locations included in the present study (see **Table R-10.a-d**).

**Table R-10. Common HLA alleles and level of allele sharing found between Spanish regions**

**Table R-10.a)** Spanish regions and locations established for this study (all Spanish region/location HLA datasets at the 3- up to the 4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0 are from this same present study)

| ID number | Spanish region/location                  | Number of subjects ( <i>n</i> ) | Spanish geographical region group assigned | HLA dataset Reference |
|-----------|--|---------------------------------|--|-----------------------|
| 1         | Spain (ESP)                              | 282                             | -  | Present Work          |
| 2         | Eastern Spain (EastESP)                  | 78                              | -  | Present Work          |
| 3         | Northern-Central Spain (NorthCentralESP) | 102                             | -  | Present Work          |
| 4         | Southern Spain (SouthESP)                | 102                             | -  | Present Work          |
| 5         | Barcelona                                | 26                              | Eastern Spain (EastESP)                    | Present Work          |
| 6         | Valencia                                 | 26                              | Eastern Spain (EastESP)                    | Present Work          |
| 7         | Santander                                | 25                              | Northern-Central Spain (NorthCentralESP)   | Present Work          |
| 8         | Salamanca                                | 26                              | Northern-Central Spain (NorthCentralESP)   | Present Work          |
| 9         | Murcia                                   | 26                              | Eastern Spain (EastESP)                    | Present Work          |
| 10        | Madrid                                   | 51                              | Northern-Central Spain (NorthCentralESP)   | Present Work          |
| 11        | Sevilla                                  | 25                              | Southern Spain (SouthESP)                  | Present Work          |
| 12        | Cordoba                                  | 26                              | Southern Spain (SouthESP)                  | Present Work          |
| 13        | Malaga                                   | 26                              | Southern Spain (SouthESP)                  | Present Work          |
| 14        | Gran Canaria (GranCanaria)               | 25                              | Southern Spain (SouthESP)                  | Present Work          |

**Table R-10.b)** Common HLA alleles. Relative comparison of the top 10 *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* allele frequencies (AF, in decimals) (at the 3- up to the 4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0) of the present entire 17th-IHIW Spanish population cohort (n=282 subjects) with respective defined Spanish region/site population groups.

| HLA-A                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|----------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>A*02:01:01:01</i> | 0.23 | 0.19  | 0.26   | 0.24  | 0.19 | 0.19 | 0.24  | 0.29 | 0.19 | 0.25 | 0.26 | 0.21  | 0.25 | 0.25     |
| <i>A*01:01:01:01</i> | 0.09 | 0.09  | 0.07   | 0.12  | 0.02 | 0.08 | 0.08  | 0.04 | 0.17 | 0.08 | 0.12 | 0.13  | 0.12 | 0.13     |
| <i>A*03:01:01:01</i> | 0.09 | 0.10  | 0.09   | 0.07  | 0.08 | 0.06 | 0.06  | 0.04 | 0.17 | 0.13 | 0.08 | 0.10  | 0.10 | 0.02     |
| <i>A*11:01:01:01</i> | 0.09 | 0.04  | 0.12   | 0.08  | 0.06 | 0.08 | 0.12  | 0.10 | 0.00 | 0.13 | 0.10 | 0.08  | 0.12 | 0.04     |
| <i>A*29:02:01:01</i> | 0.08 | 0.10  | 0.09   | 0.06  | 0.17 | 0.08 | 0.08  | 0.12 | 0.04 | 0.09 | 0.06 | 0.08  | 0.08 | 0.04     |
| HLA-A                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
| <i>A*24:02:01:01</i> | 0.08 | 0.10  | 0.08   | 0.05  | 0.13 | 0.13 | 0.04  | 0.12 | 0.04 | 0.08 | 0.04 | 0.08  | 0.06 | 0.04     |
| <i>A*32:01:01</i>    | 0.04 | 0.07  | 0.04   | 0.02  | 0.10 | 0.08 | 0.06  | 0.00 | 0.04 | 0.04 | 0.04 | 0.02  | 0.00 | 0.04     |
| <i>A*30:02:01:01</i> | 0.04 | 0.04  | 0.03   | 0.04  | 0.00 | 0.06 | 0.04  | 0.00 | 0.06 | 0.04 | 0.00 | 0.04  | 0.02 | 0.10     |
| <i>A*23:01:01</i>    | 0.03 | 0.02  | 0.03   | 0.04  | 0.04 | 0.00 | 0.02  | 0.06 | 0.02 | 0.02 | 0.04 | 0.06  | 0.02 | 0.04     |
| <i>A*26:01:01:01</i> | 0.02 | 0.01  | 0.02   | 0.03  | 0.00 | 0.00 | 0.00  | 0.02 | 0.04 | 0.03 | 0.08 | 0.04  | 0.02 | 0.00     |

| HLA-B                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|----------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>B*07:02:01</i>    | 0.10 | 0.09  | 0.10   | 0.09  | 0.15 | 0.08 | 0.12  | 0.06 | 0.04 | 0.12 | 0.10 | 0.12  | 0.10 | 0.06     |
| <i>B*44:03:01:01</i> | 0.09 | 0.08  | 0.12   | 0.05  | 0.12 | 0.08 | 0.12  | 0.13 | 0.06 | 0.12 | 0.02 | 0.06  | 0.08 | 0.04     |
| <i>B*08:01:01:01</i> | 0.07 | 0.06  | 0.06   | 0.07  | 0.02 | 0.06 | 0.08  | 0.06 | 0.12 | 0.04 | 0.08 | 0.04  | 0.10 | 0.08     |
| <i>B*51:01:01:01</i> | 0.06 | 0.04  | 0.08   | 0.06  | 0.02 | 0.04 | 0.10  | 0.04 | 0.06 | 0.09 | 0.06 | 0.08  | 0.08 | 0.04     |
| <i>B*44:02:01:01</i> | 0.06 | 0.05  | 0.08   | 0.04  | 0.04 | 0.08 | 0.12  | 0.10 | 0.04 | 0.05 | 0.08 | 0.06  | 0.04 | 0.00     |
| HLA-B                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
| <i>B*35:01:01:02</i> | 0.05 | 0.04  | 0.07   | 0.03  | 0.06 | 0.02 | 0.02  | 0.06 | 0.06 | 0.10 | 0.04 | 0.00  | 0.06 | 0.04     |
| <i>B*18:01:01:01</i> | 0.04 | 0.05  | 0.04   | 0.03  | 0.02 | 0.06 | 0.06  | 0.02 | 0.08 | 0.04 | 0.02 | 0.04  | 0.02 | 0.04     |
| <i>B*49:01:01</i>    | 0.04 | 0.05  | 0.02   | 0.05  | 0.08 | 0.04 | 0.00  | 0.02 | 0.04 | 0.02 | 0.06 | 0.10  | 0.04 | 0.02     |
| <i>B*14:02:01:01</i> | 0.04 | 0.03  | 0.04   | 0.05  | 0.04 | 0.02 | 0.04  | 0.02 | 0.02 | 0.04 | 0.08 | 0.06  | 0.04 | 0.02     |
| <i>B*15:01:01:01</i> | 0.04 | 0.04  | 0.04   | 0.03  | 0.08 | 0.04 | 0.02  | 0.08 | 0.00 | 0.03 | 0.02 | 0.04  | 0.02 | 0.06     |

| HLA-C                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|----------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>C*07:01:01:01</i> | 0.12 | 0.13  | 0.09   | 0.13  | 0.10 | 0.12 | 0.06  | 0.13 | 0.17 | 0.09 | 0.14 | 0.12  | 0.15 | 0.10     |
| <i>C*04:01:01:01</i> | 0.11 | 0.10  | 0.12   | 0.09  | 0.10 | 0.08 | 0.06  | 0.17 | 0.13 | 0.13 | 0.10 | 0.08  | 0.12 | 0.08     |
| <i>C*07:02:01:03</i> | 0.10 | 0.09  | 0.11   | 0.09  | 0.15 | 0.08 | 0.12  | 0.08 | 0.04 | 0.12 | 0.10 | 0.12  | 0.10 | 0.06     |
| <i>C*12:03:01:01</i> | 0.08 | 0.12  | 0.07   | 0.06  | 0.15 | 0.12 | 0.08  | 0.10 | 0.08 | 0.04 | 0.06 | 0.08  | 0.06 | 0.06     |
| <i>C*16:01:01:01</i> | 0.05 | 0.06  | 0.07   | 0.03  | 0.10 | 0.04 | 0.08  | 0.08 | 0.04 | 0.07 | 0.04 | 0.04  | 0.04 | 0.02     |



| HLA-C                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|----------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>C*02:02:02:01</i> | 0.05 | 0.06  | 0.04   | 0.05  | 0.08 | 0.08 | 0.02  | 0.02 | 0.04 | 0.05 | 0.06 | 0.08  | 0.04 | 0.04     |
| <i>C*05:01:01:02</i> | 0.05 | 0.03  | 0.07   | 0.04  | 0.00 | 0.06 | 0.10  | 0.10 | 0.04 | 0.04 | 0.06 | 0.06  | 0.02 | 0.02     |
| <i>C*06:02:01:01</i> | 0.05 | 0.04  | 0.03   | 0.06  | 0.00 | 0.08 | 0.02  | 0.00 | 0.06 | 0.05 | 0.04 | 0.08  | 0.06 | 0.08     |
| <i>C*03:03:01:01</i> | 0.04 | 0.03  | 0.05   | 0.04  | 0.06 | 0.04 | 0.06  | 0.08 | 0.00 | 0.03 | 0.04 | 0.04  | 0.02 | 0.08     |
| <i>C*04:01:01:06</i> | 0.04 | 0.04  | 0.05   | 0.03  | 0.06 | 0.02 | 0.02  | 0.04 | 0.06 | 0.07 | 0.08 | 0.00  | 0.04 | 0.02     |

| HLA-DQB1                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|-------------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>DQB1*02:02:01:01</i> | 0.14 | 0.19  | 0.16   | 0.10  | 0.29 | 0.12 | 0.20  | 0.14 | 0.15 | 0.14 | 0.04 | 0.12  | 0.10 | 0.16     |
| <i>DQB1*02:01:01</i>    | 0.14 | 0.19  | 0.16   | 0.11  | 0.29 | 0.19 | 0.20  | 0.14 | 0.23 | 0.18 | 0.08 | 0.12  | 0.13 | 0.16     |
| <i>DQB1*03:01:01:03</i> | 0.11 | 0.04  | 0.11   | 0.16  | 0.04 | 0.02 | 0.09  | 0.10 | 0.08 | 0.13 | 0.20 | 0.18  | 0.15 | 0.09     |
| <i>DQB1*03:02:01</i>    | 0.09 | 0.09  | 0.09   | 0.10  | 0.10 | 0.04 | 0.07  | 0.16 | 0.13 | 0.07 | 0.08 | 0.12  | 0.04 | 0.16     |
| <i>DQB1*06:02:01</i>    | 0.09 | 0.10  | 0.08   | 0.09  | 0.13 | 0.10 | 0.13  | 0.06 | 0.06 | 0.07 | 0.10 | 0.02  | 0.12 | 0.11     |
| HLA-DQB1                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
| <i>DQB1*05:01:01:03</i> | 0.09 | 0.07  | 0.09   | 0.10  | 0.08 | 0.08 | 0.04  | 0.06 | 0.06 | 0.12 | 0.06 | 0.16  | 0.12 | 0.05     |
| <i>DQB1*06:03:01</i>    | 0.08 | 0.08  | 0.09   | 0.07  | 0.13 | 0.06 | 0.11  | 0.14 | 0.04 | 0.04 | 0.08 | 0.08  | 0.08 | 0.05     |
| <i>DQB1*03:01:01:01</i> | 0.03 | 0.04  | 0.04   | 0.01  | 0.00 | 0.10 | 0.02  | 0.04 | 0.02 | 0.06 | 0.02 | 0.02  | 0.00 | 0.00     |
| <i>DQB1*04:02:01</i>    | 0.03 | 0.01  | 0.03   | 0.05  | 0.00 | 0.02 | 0.09  | 0.00 | 0.02 | 0.01 | 0.04 | 0.10  | 0.00 | 0.05     |
| <i>DQB1*05:02:01</i>    | 0.03 | 0.01  | 0.03   | 0.05  | 0.02 | 0.02 | 0.02  | 0.04 | 0.00 | 0.02 | 0.02 | 0.06  | 0.04 | 0.07     |

| HLA-DRB1                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
|-------------------------|------|-------|--------|-------|------|------|-------|------|------|------|------|-------|------|----------|
| <i>DRB1*07:01:01:01</i> | 0.16 | 0.19  | 0.16   | 0.13  | 0.29 | 0.12 | 0.20  | 0.13 | 0.15 | 0.16 | 0.04 | 0.12  | 0.13 | 0.25     |
| <i>DRB1*03:01:01:01</i> | 0.13 | 0.16  | 0.14   | 0.11  | 0.08 | 0.18 | 0.06  | 0.13 | 0.23 | 0.19 | 0.08 | 0.12  | 0.13 | 0.10     |
| <i>DRB1*15:01:01:01</i> | 0.09 | 0.10  | 0.10   | 0.08  | 0.13 | 0.08 | 0.18  | 0.08 | 0.08 | 0.06 | 0.08 | 0.04  | 0.12 | 0.10     |
| <i>DRB1*13:01:01:01</i> | 0.08 | 0.08  | 0.10   | 0.06  | 0.13 | 0.06 | 0.10  | 0.15 | 0.04 | 0.06 | 0.08 | 0.06  | 0.08 | 0.02     |
| <i>DRB1*01:01:01</i>    | 0.06 | 0.19  | 0.16   | 0.13  | 0.29 | 0.18 | 0.20  | 0.15 | 0.23 | 0.19 | 0.08 | 0.12  | 0.13 | 0.25     |
| HLA-DRB1                | ESP  | E-ESP | NC-ESP | S-ESP | Bar. | Val. | Sant. | Sal. | Mur. | Mad. | Sev. | Cord. | Mal. | GranCan. |
| <i>DRB1*11:04:01</i>    | 0.05 | 0.05  | 0.05   | 0.04  | 0.06 | 0.04 | 0.00  | 0.08 | 0.06 | 0.07 | 0.04 | 0.02  | 0.08 | 0.02     |
| <i>DRB1*11:01:01:01</i> | 0.04 | 0.00  | 0.05   | 0.05  | 0.00 | 0.00 | 0.10  | 0.02 | 0.00 | 0.05 | 0.06 | 0.06  | 0.06 | 0.04     |
| <i>DRB1*04:05:01</i>    | 0.03 | 0.04  | 0.01   | 0.04  | 0.02 | 0.04 | 0.00  | 0.02 | 0.06 | 0.01 | 0.04 | 0.02  | 0.04 | 0.06     |
| <i>DRB1*04:01:01:01</i> | 0.03 | 0.03  | 0.04   | 0.01  | 0.02 | 0.06 | 0.06  | 0.06 | 0.00 | 0.02 | 0.00 | 0.02  | 0.00 | 0.04     |
| <i>DRB1*14:54:01</i>    | 0.03 | 0.03  | 0.02   | 0.02  | 0.00 | 0.04 | 0.02  | 0.06 | 0.06 | 0.00 | 0.02 | 0.00  | 0.02 | 0.06     |

## Abbreviations:

ESP = Spain (ESP); E-ESP = Eastern Spain (EastESP); NC-ESP= Northern-Central Spain (NorthCentralESP); S-ESP= Southern Spain (SouthESP); Bar. = Barcelona; Val. = Valencia;  
 Sant.= Santander; Sal.= Salamanca; Mur.= Murcia; Mad.= Madrid; Sev.= Sevilla; Cord. = Cordoba; Mal. = Malaga; GranCan. = Gran Canaria

**Table R-10.c)** Distance matrix of respective Nei genetic distances ( $D_A$ ) between present entire 17th-IHIW Spanish population cohort (n=282 subjects) and the different region/site population groups that were estimated by using the distribution of *HLA-A*, -*B*, -*C*, -*DQB1* and -*DRB1* alleles (based on Spanish HLA dataset at the 3- up to the 4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0 from the present study).

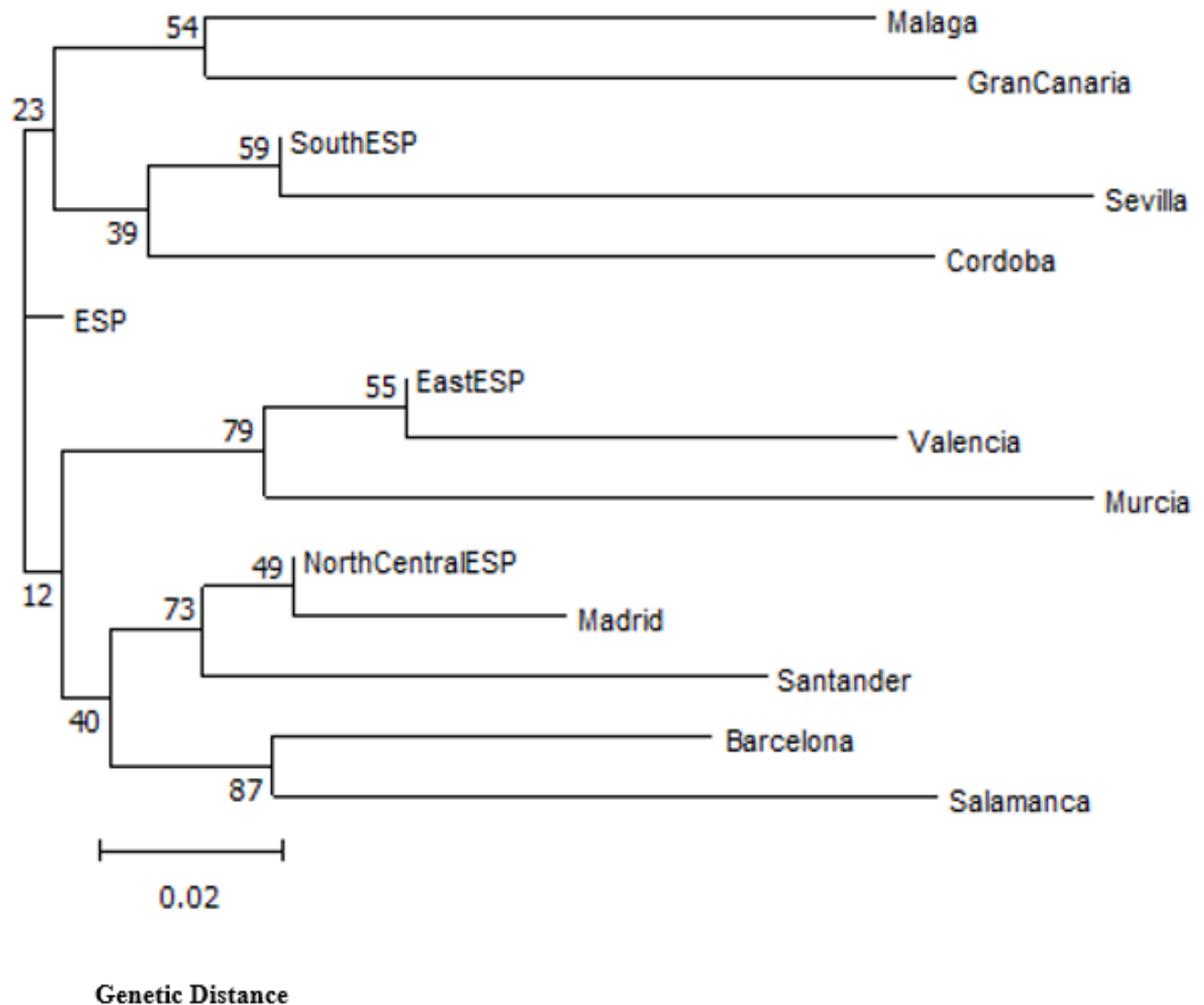
## Distance matrix

|        | ESP | E-Esp | NC-ESP | S-ESP | Bar.  | Val.  | Sant. | Sal.  | Mur.  | Mad.  | Sev.  | Cord. | Mal.  | GranCan. |
|--------|-----|-------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| ESP    |     | 0.023 | 0.018  | 0.018 | 0.090 | 0.090 | 0.088 | 0.109 | 0.118 | 0.059 | 0.111 | 0.108 | 0.098 | 0.107    |
| E-ESP  |     |       | 0.045  | 0.062 | 0.041 | 0.036 | 0.106 | 0.120 | 0.058 | 0.086 | 0.151 | 0.135 | 0.134 | 0.138    |
| NC-ESP |     |       |        | 0.052 | 0.080 | 0.108 | 0.042 | 0.066 | 0.139 | 0.018 | 0.137 | 0.116 | 0.111 | 0.130    |
| S-ESP  |     |       |        |       | 0.122 | 0.129 | 0.115 | 0.141 | 0.145 | 0.085 | 0.081 | 0.086 | 0.072 | 0.069    |
| Bar.   |     |       |        |       |       | 0.158 | 0.127 | 0.119 | 0.179 | 0.103 | 0.201 | 0.175 | 0.153 | 0.175    |
| Val.   |     |       |        |       |       |       | 0.157 | 0.175 | 0.188 | 0.155 | 0.221 | 0.171 | 0.200 | 0.191    |
| Sant.  |     |       |        |       |       |       |       | 0.181 | 0.195 | 0.117 | 0.216 | 0.147 | 0.175 | 0.170    |
| Sal.   |     |       |        |       |       |       |       |       | 0.210 | 0.153 | 0.201 | 0.201 | 0.178 | 0.184    |
| Mur.   |     |       |        |       |       |       |       |       |       | 0.168 | 0.198 | 0.224 | 0.204 | 0.209    |
| Mad.   |     |       |        |       |       |       |       |       |       |       | 0.154 | 0.139 | 0.132 | 0.179    |
| Sev.   |     |       |        |       |       |       |       |       |       |       |       | 0.194 | 0.198 | 0.247    |
| Cord.  |     |       |        |       |       |       |       |       |       |       |       |       | 0.207 | 0.192    |
| Mal.   |     |       |        |       |       |       |       |       |       |       |       |       |       | 0.154    |

## Abbreviations:

ESP = Spain (ESP); E-ESP = Eastern Spain (EastESP); NC-ESP= Northern-Central Spain (NorthCentralESP); S-ESP= Southern Spain (SouthESP); Bar. = Barcelona; Val. = Valencia;  
Sant.= Santander; Sal.= Salamanca; Mur.= Murcia; Mad.= Madrid; Sev.= Sevilla; Cord. = Cordoba; Mal. = Malaga; GranCan. = Gran Canaria

**Table R-10.d)** Neighbor-joining (NJ) dendrogram illustrates relatedness between the present entire 17th-IHIW Spanish population cohort (n=282 subjects) and the 3 different geographical regions as well as the 10 different specific sites of origin of samples of the present Spanish population study. Nei genetic distances ( $D_A$ ) between these population sub-groups were estimated by using *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* allele frequencies (based on Spanish HLA dataset at the 3- up to the 4-field allele resolution level and according to IPD-IMGT/HLA database version 3.25.0 from the present study). Bootstrap values from 1000 replicates are depicted. The root of the NJ method is calculated by the mid-point rooting method, in which the root is placed in the mid-point of the longest path of two taxa.



In general, these major Spanish population regions and different individual local sub-groups, which were compared according to these HLA-A, -B, -C, -DQB1 and -DRB1 allele distributions in this NJ relatedness analysis, are also clustered according to their geographical location, thus illustrating the existing HLA regional variation within Spanish general population [260].

Despite of limitations in the sample size shown by these different Spanish population sub-groups in the present study. At the HLA allele level, it can be observed that most frequent alleles at a national level (considering entire Spanish population, termed as “ESP”, n=282) are fairly evenly distributed and well represented among the different Spanish regions and locations evaluated here, with some minor exceptions (specifically found at the different 10 Spanish locations level presenting a limited and small sample size comparatively) that need to be further analyze by future larger-scale population studies (see **Table R-10.b**). Taking into account genetic distances evaluated here (see **Tables R-10.c** and **R-10.d**), the present entire Spanish population cohort shows a Mediterranean genetic substrate that seems to be represented more predominantly by Eastern and Central regions/locations situated within the Central Plateau (i.e. Meseta Central or Central Castilian Plateau) as previously described [221][260][545][546][558-561][563][564][571][600][608][624-630][757]. Whereas the most Northern and Southern regions/locations (which are mountainous areas that are more isolated geographically unlike this Central Castilian Plateau region in mainland Spain; or even being very unique island areas such as Canary Islands) diverge from this aforementioned Mediterranean Spanish HLA genetic background as reported in previous works [221][260][545][546][558-561][563][564][571][600][608][624-630][757]. For instance, although we considered Barcelona location as part of the Eastern region of Spain (representing the Mediterranean Spanish Basin) for this study, we clearly observed how this Catalan location seems to be more related to other Northern locations than to Mediterranean sites such as Valencia or Murcia. Interestingly,

Salamanca location population group (situated very close to the frontier that separates Spain from Portugal) describes a pronounced distinctive HLA distribution in comparison to other Northern-Central locations in Spain as previously described and it also exemplifies the extensive HLA diversity found within the Iberian Peninsula [221][260][628]. Furthermore, the striking divergence observed in Malaga and Gran Canaria locations (see **R-10.d**) may be explained by the reported historical genetic contribution from North African Berber and Muslim Arab population ancestries [563-568][578][612][613][808].

We also attempted to do this regional study at the extended HLA haplotype level (data not shown). However, due to these limited small sample sizes found at the different Spanish regions and locations it was not possible to estimate accurately haplotype frequencies via an expectation-maximization (EM) algorithm [342] to evaluate haplotype sharing between local regions/sites.

Overall, in spite of presenting a relatively small sample size, the present Spanish population study has allowed us to see the great potential of NGS-based HLA population studies in order to identify 3-/4-field HLA allele signatures at a regional level as a consequence of both differential regional historic events and the characteristic regional orography that favors more isolation of certain local populations [136][260]. Nonetheless, future studies of larger population sample size at a wider geographic scale will be needed to assess more accurately the HLA diversity in Spanish population in order to confirm these observations and findings of our study as well as to reveal other unknown but significant polymorphism within the HLA genes system.

## **II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION**

In the scope of the present thesis work:

- A first case-control study (**Study 1**) was carried out to examine HLA-disease associations with MS in these Spanish population cohorts: 17<sup>th</sup> IHIW Spanish population healthy control (N=282) versus a cohort of multiple sclerosis (MS) patients in the Spanish population (N=238, recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain). In this sense, the initial main goal was to attempt a fine-mapping of these allele and haplotype associations by full gene resolution level via NGS.
- In addition, a second exercise or test case (**Study 2**) of this case-control study was carried out using the same MS Spanish group but, in this second case, using an alternative healthy control group dataset (N=196) specifically from the Spanish northeastern region of Catalonia, and thus to evaluate possible differences in the findings of HLA-disease association with MS due to plausible regional HLA genetic variation within mainland Spain as a statistical approach to try controlling for any possible existing population stratification (i.e. differences in genetic structure between disease and control groups) as a confounding factor that may affect the results obtained in the first study.

Although data is not shown here, in relation to all Spanish population 3-/4-field and 2-field (either “trimmed” or generated) datasets (from both healthy controls and MS cases) used to conduct this case-control analysis, no overall deviations from expected Hardy-Weinberg Equilibrium Proportions (HWEP) were observed in any of the HLA loci analyzed with the exception of the following:

-A minor but significant departure from HWEP at the HLA-DPA1 locus (see **Table R-2**) previously mentioned and explained in the case of 17<sup>th</sup> IHIW Spanish population healthy control dataset at 3-/4-field resolution.

-As for the MS cases dataset (regardless of the allele resolution level analyzed), and as expected, deviations from HWEP at loci strongly associated with disease (such as HLA-DRB3/4/5, HLA-DRB1, HLA-DQA1 and HLA-DQB1) were observed not because of genotyping or genotype calling error, but due to an existing selective pattern linked with disease in this case. In this sense, it would have been obviously counter-productive to remove these loci from further investigation. Thus, when testing for deviations for HWEP we only took into consideration the respective healthy control datasets used in the present study which all passed this quality data test [949].

## **11. HLA ALLELE LEVEL ANALYSES ON FIRST CASE-CONTROL STUDY**

• Firstly, in the corresponding **Study 1**, associations of HLA alleles with MS risk (see **Table R-11** and **Figure R-4**) and protection (see **Table R-11** and **Figure R-5**) were respectively evaluated at the 3-/4-field of resolution obtained via NGS-based HLA genotyping method. In accordance with previous and extensively documented findings, especially in populations of European descent (reviewed in [650][651]), in the initial nominal analysis (i.e. when the overall cohort was analyzed) (see **Table R-11a** and **Figure R-4a**) the most significant risk association was identified for the allele *HLA-DRB1\*15:01:01:01* (OR=2.46; p=4.10E-07) and additionally for the corresponding *HLA-DRB5\*01:01:01* allele (OR=2.44; p=4.31E-07), since both HLA loci are physically located in very close proximity and thus presenting a very high LD [344]. At the same time, still within HLA class II region other significant risk signals were identified in the present Spanish MS cohort, being also in consonance with previous findings (reviewed in [650][651]): on one hand, the

characteristic allele *HLA-DQB1\*06:02:01* (OR=2.24; p=1.08E-05) and the respective tightly associated allele *HLA-DQA1\*01:02:01:01* (OR=2.21; p=6.11E-06) (forming the respective HLA-DQ heterodimer), which both are known to be also in strong linkage disequilibrium with *HLA-DRB1\*15:01:01:01*; whereas, on the other hand, a tentatively independent and more moderate risk signal linked to *HLA-DPB1\*03:01:01* allele (OR=1.85; p=2.20E-02) was detected as well in this Spanish HLA-MS association study. In relation to HLA class I region, alleles *HLA-A\*03:01:01:01* (OR=1.51; p=4.14E-02), *-C\*07:02:01:03* (OR=1.31; p=1.75E-01) and *-B\*07:02:01* (OR=1.33; p=1.47E-01) were moderately associated with MS; however, it is also known these observed associations are again due to a most likely haplotype effect since they reflect the strong LD between these class I alleles and the highly predisposing *HLA-DRB1\*15:01:01:01* allele as previously described in other studies (e.g. [292][727]). As a next step in the present HLA-MS association analysis, and in order to confirm or discard these suspected haplotype effects on many of these observed risk signals as well as the possible independent association with *HLA-DPB1*, a conditional analysis was carried out with respect to the major risk HLA allele found (*HLA-DRB1\*15:01:01:01*) controlling for potential confounding. This conditional analysis on *HLA-DRB1\*15:01:01:01* (i.e. when considering the stratum lacking *HLA-DRB1\*15:01:01:01*) (see **Table R-11b** and **Figure R-4b**) revealed and confirmed the statistically independent and significant association (where the risk association was even stronger) with *HLA-DPB1\*03:01:01* (OR=2.23; p=8.90E-03), while the rest of abovementioned risk allelic signals (at *HLA-DQA1*, *-DQB1*, *-DRB5* loci) previously detected (and as it is confirmed here to be under a strong haplotype effect reliant on *HLA-DRB1\*15:01:01:01*) were either not statistically measurable (i.e. “binned” category in the  $\chi^2$  statistic for a contingency table analysis of case-control data) or not significant at all in this *DRB1\*15:01:01:01*-negative stratum.



**Table R-11a. Study 1: Significant HLA allele-level (at the 3-/4-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort.**

| Study 1-Nominal Analysis HLA Allele Level |                         | Controls (2n=564) | MS cases (2n=476) |      |             |              |       |                 |                   |
|---|-------------------------|-------------------|-------------------|------|-------------|--------------|-------|-----------------|-------------------|
| HLA Alleles shown in Haplotype Blocks     | HLA Allele              | Allele Freq Ctls  | Allele Freq Cases | OR   | 95%CI lower | 95% CI upper | 1/OR  | p-value         | minus log p-value |
| DPB1                                      | <i>DPB1*03:01:01</i>    | 0.041             | 0.074             | 1.85 | 1.05        | 3.30         | -     | <b>2.20E-02</b> | 1.7               |
| DQB1                                      | <i>DQB1*06:02:01</i>    | 0.091             | 0.183             | 2.24 | 1.53        | 3.30         | -     | <b>1.08E-05</b> | 5.0               |
| DQA1                                      | <i>DQA1*01:02:01:01</i> | 0.103             | 0.202             | 2.21 | 1.54        | 3.18         | -     | <b>6.11E-06</b> | 5.2               |
| DRB1                                      | <i>DRB1*15:01:01:01</i> | 0.094             | 0.204             | 2.46 | 1.70        | 3.58         | -     | <b>4.10E-07</b> | 6.4               |
| DRB345                                    | <i>DRB5*01:01:01</i>    | 0.092             | 0.200             | 2.44 | 1.69        | 3.55         | -     | <b>4.31E-07</b> | 6.4               |
| B   | <i>B*07:02:01</i>       | 0.094             | 0.122             | 1.33 | 0.89        | 2.01         | -     | 1.47E-01        | 0.8               |
| C   | <i>C*07:02:01:03</i>    | 0.096             | 0.122             | 1.31 | 0.87        | 1.97         | -     | 1.75E-01        | 0.8               |
| A   | <i>A*03:01:01:01</i>    | 0.084             | 0.122             | 1.51 | 0.99        | 2.31         | -     | <b>4.14E-02</b> | 1.4               |
| DRB1                                      | <i>DRB1*04:02:01</i>    | 0.021             | 0.006             | 0.3  | 0.05        | 1.13         | 3.33  | <b>5.08E-02</b> | 1.3               |
| DRB1                                      | <i>DRB1*04:01:01:01</i> | 0.024             | 0.006             | 0.26 | 0.05        | 0.93         | 3.85  | <b>2.27E-02</b> | 1.6               |
| DRB4                                      | <i>DRB4*01:03:01:01</i> | 0.152             | 0.097             | 0.59 | 0.40        | 0.88         | 1.69  | <b>6.76E-03</b> | 2.2               |
| C   | <i>C*05:01:01:02</i>    | 0.048             | 0.023             | 0.47 | 0.21        | 0.99         | 2.13  | <b>3.26E-02</b> | 1.5               |
| B   | <i>B*44:03:01:01</i>    | 0.098             | 0.090             | 0.92 | 0.59        | 1.42         | 1.09  | 6.87E-01        | 0.2               |
| B   | <i>B*44:02:01:01</i>    | 0.058             | 0.034             | 0.56 | 0.29        | 1.06         | 1.79  | 6.02E-02        | 1.2               |
| DQB1                                      | <i>DQB1*06:01:01</i>    | 0.017             | 0.004             | 0.24 | 0.03        | 1.15         | 4.17  | <b>4.79E-02</b> | 1.3               |
| DRB5                                      | <i>DRB5*01:02</i>       | 0.015             | 0.002             | 0.13 | 0.00        | 0.98         | 7.69  | <b>2.58E-02</b> | 1.6               |
| B   | <i>B*58:01:01:01</i>    | 0.024             | 0.002             | 0.09 | 0.00        | 0.57         | 11.11 | <b>2.71E-03</b> | 2.6               |
| B   | <i>B*38:01:01</i>       | 0.031             | 0.004             | 0.13 | 0.01        | 0.56         | 7.69  | <b>1.53E-03</b> | 2.8               |
| Bw4 (DLRTLLR)                             | <i>Bw4 (DLRTLLR)</i>    | 0.057             | 0.036             | 0.62 | 0.34        | 1.12         | 1.62  | 1.45E-01        | 0.8               |
| Bw4 (NLRIALR)                             | <i>Bw4 (NLRIALR)</i>    | 0.204             | 0.166             | 0.78 | 0.57        | 1.06         | 1.29  | 1.14E-01        | 0.9               |
| Bw4 (NLRITALR)                            | <i>Bw4 (NLRITALR)</i>   | 0.171             | 0.145             | 0.83 | 0.58        | 1.14         | 1.21  | 2.73E-01        | 0.6               |
| Bw6 (SLRNLRG)                             | <i>Bw6 (SLRNLRG)</i>    | 0.567             | 0.653             | 1.45 | 1.12        | 1.84         | 0.69  | <b>3.70E-03</b> | 2.4               |

HLA alleles associated with risk (RED colored) or protection (GREEN colored) to MS susceptibility are depicted based on the respective most common -bearing haplotypes in which they are embedded. OR, Odds ratio; 95% CI, 95% confidence interval. P values derived from a two-tailed Fischer's exact test. A P value of 0.05 ( $\alpha$ ) or less was considered statistically significant (in bold).

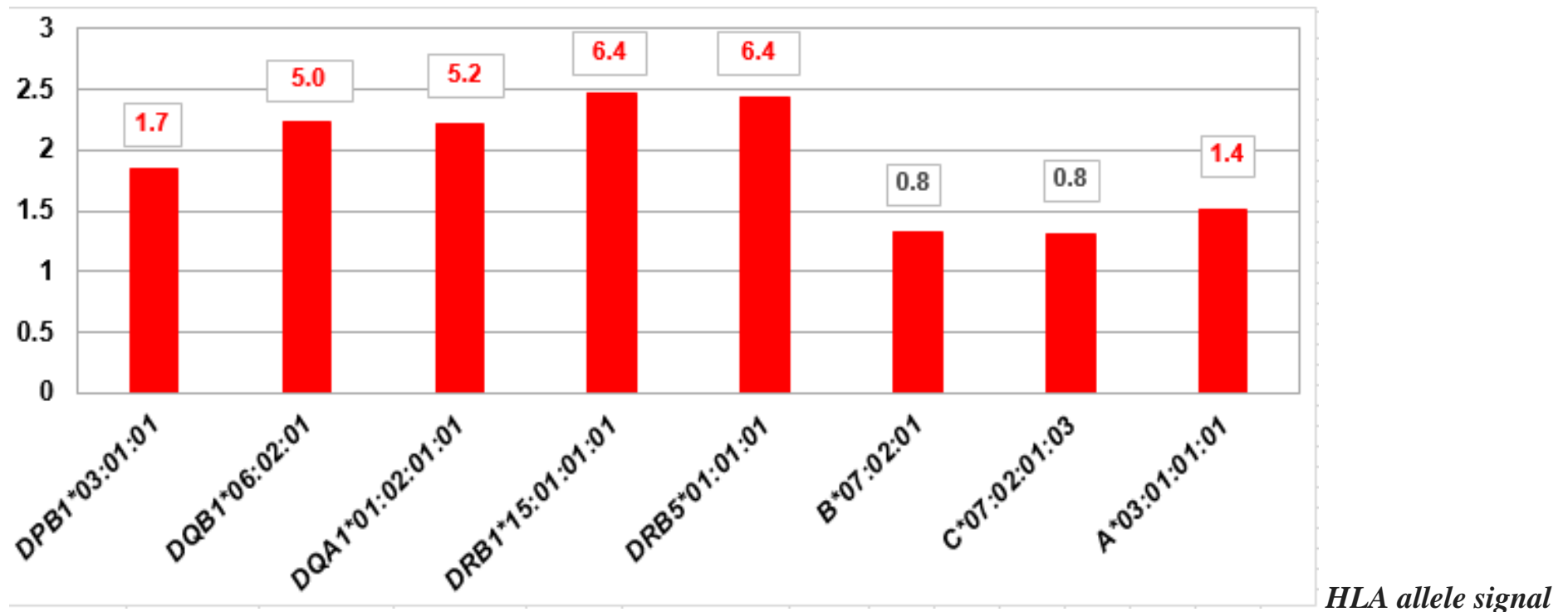
**Table R-11b. Study 1: Significant HLA allele-level (at the 3-/4-field resolution) associations with multiple sclerosis (MS) in the stratified analysis by conditioning on *HLA-DRB1\*15:01:01:01* of the Spanish population cohort.**

| Study 1-Stratified Analysis DRB1*15:01 Negative Stratum HLA Allele Level |                         | Controls (2n=482) | MS cases (2n=298) |        |             |              |       |                 |                   |
|--|-------------------------|-------------------|-------------------|--------|-------------|--------------|-------|-----------------|-------------------|
| HLA Alleles shown in Haplotype Blocks                                    | HLA Allele              | Allele Freq Ctls  | Allele Freq Cases | OR     | 95%CI lower | 95% CI upper | 1/OR  | p-value         | minus log p-value |
| DPB1   | <i>DPB1*03:01:01</i>    | 0.039             | 0.084             | 2.23   | 1.15        | 4.37         | -     | <b>8.90E-03</b> | 2.1               |
| DQB1   | <i>DQB1*06:02:01</i>    | 0.004             | 0.000             | binned | -           | -            | -     | -               | -                 |
| DQA1   | <i>DQA1*01:02:01:01</i> | 0.006             | 0.000             | binned | -           | -            | -     | -               | -                 |
| DRB345   | <i>DRB5*01:01:01</i>    | 0.002             | 0.000             | binned | -           | -            | -     | -               | -                 |
| B  | <i>B*07:02:01</i>       | 0.052             | 0.027             | 0.5    | 0.19        | 1.17         | -     | 9.16E-02        | 1.0               |
| C  | <i>C*07:02:01:03</i>    | 0.054             | 0.030             | 0.55   | 0.22        | 1.22         | -     | 1.20E-01        | 0.9               |
| A  | <i>A*03:01:01:01</i>    | 0.068             | 0.074             | 1.08   | 0.59        | 1.96         | -     | 7.76E-01        | 0.1               |
| DRB1   | <i>DRB1*04:02:01</i>    | 0.023             | 0.007             | binned | -           | -            | -     | -               | -                 |
| DRB1   | <i>DRB1*04:01:01:01</i> | 0.027             | 0.010             | 0.37   | 0.07        | 1.35         | 2.70  | 1.06E-01        | 1.0               |
| DRB4   | <i>DRB4*01:03:01:01</i> | 0.170             | 0.144             | 0.82   | 0.54        | 1.25         | 1.22  | 3.39E-01        | 0.5               |
| C  | <i>C*05:01:01:02</i>    | 0.044             | 0.020             | 0.45   | 0.15        | 1.17         | 2.22  | 8.19E-02        | 1.1               |
| B  | <i>B*44:03:01:01</i>    | 0.104             | 0.087             | 0.83   | 0.48        | 1.39         | 1.20  | 4.51E-01        | 0.3               |
| B  | <i>B*44:02:01:01</i>    | 0.054             | 0.034             | 0.61   | 0.26        | 1.33         | 1.64  | 1.87E-01        | 0.7               |
| DQB1   | <i>DQB1*06:01:01</i>    | 0.017             | 0.003             | binned | -           | -            | -     | -               | -                 |
| DRB5   | <i>DRB5*01:02</i>       | 0.017             | 0.003             | binned | -           | -            | -     | -               | -                 |
| B  | <i>B*58:01:01:01</i>    | 0.027             | 0.003             | 0.12   | 0.00        | 0.82         | 8.33  | <b>1.58E-02</b> | 1.8               |
| B  | <i>B*38:01:01</i>       | 0.035             | 0.003             | 0.09   | 0.00        | 0.59         | 11.11 | <b>3.92E-03</b> | 2.4               |
| Bw4 (DLRTLLR)  | <i>Bw4 (DLRTLLR)</i>    | 0.058             | 0.047             | 0.80   | 0.41        | 1.54         | 1.25  | 5.05E-01        | 0.3               |
| Bw4 (NLRIALR)  | <i>Bw4 (NLRIALR)</i>    | 0.222             | 0.195             | 0.85   | 0.59        | 1.21         | 1.18  | 3.64E-01        | 0.4               |
| Bw4 (NLR TALR)   | <i>Bw4 (NLR TALR)</i>   | 0.174             | 0.134             | 0.73   | 0.49        | 1.10         | 1.36  | 1.38E-01        | 0.9               |
| Bw6 (SLRNLRG)  | <i>Bw6 (SLRNLRG)</i>    | 0.544             | 0.624             | 1.39   | 1.04        | 1.87         | 0.72  | <b>2.72E-02</b> | 1.6               |

HLA alleles associated with risk (RED colored) or protection (GREEN colored) to MS susceptibility are depicted based on the respective most common -bearing haplotypes in which they are embedded. OR, Odds ratio; 95% CI, 95% confidence interval. P values derived from a two-tailed Fischer's exact test. A P value of 0.05 ( $\alpha$ ) or less was considered statistically significant (in bold).

Figure R-4a. Study 1: Bar chart representation of odds ratio (OR) values (Y axis) and main HLA alleles associated (X axis) with risk to multiple sclerosis (MS) susceptibility in the nominal analysis of the Spanish population cohort. - log (p-value) values are displayed in the squares above each respective bar.

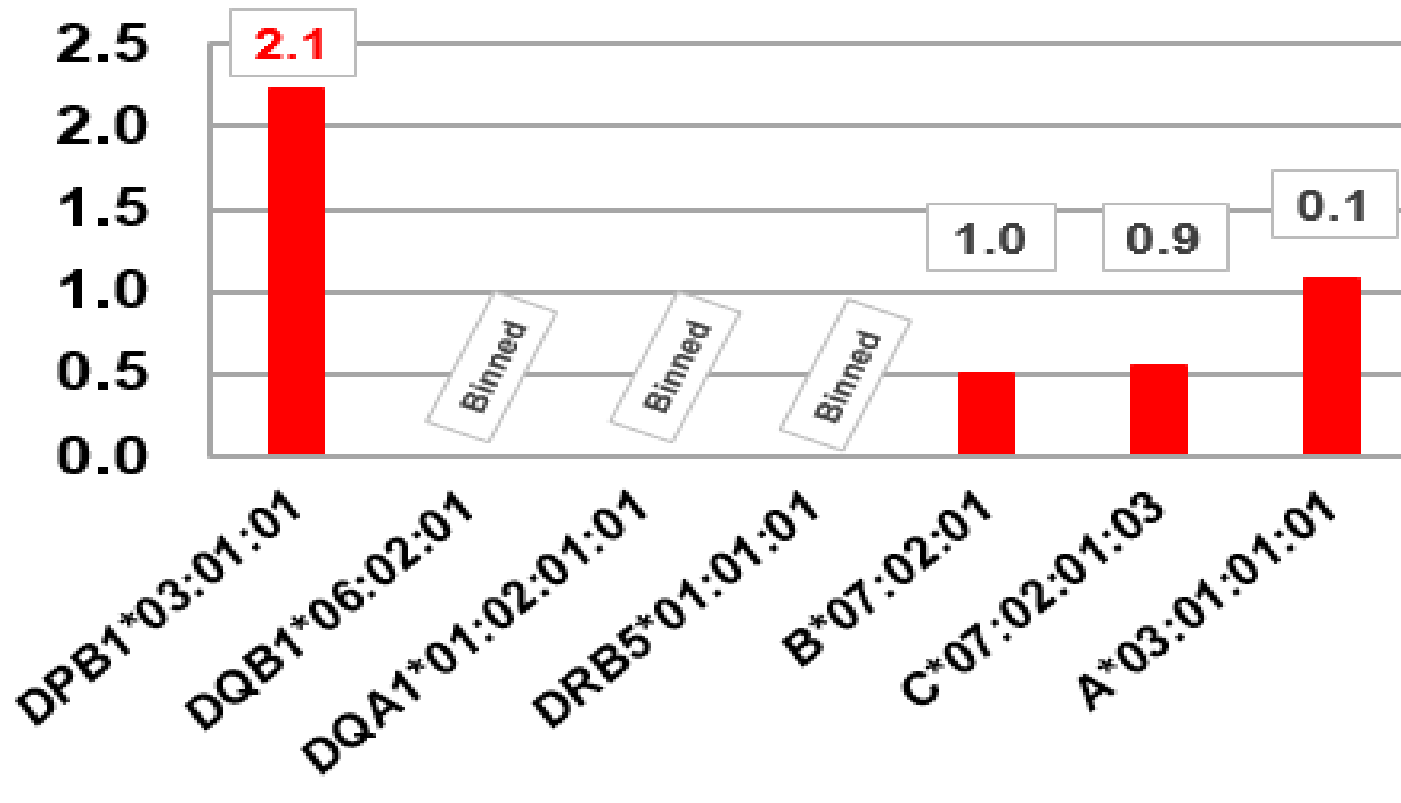
OR



HLA alleles associated with risk (RED colored) to MS susceptibility are depicted based on the respective most common -bearing haplotypes in which they are embedded. OR, Odds ratio; P values derived from a two-tailed Fischer's exact test are shown as the -log (p-value) (values displayed in the squares above each respective bar). A -log (p-value) of 1.3 or higher (equivalent to  $p\text{-value} \leq 0.05$ ) was considered statistically significant (in red bold).

Figure R-4b. Study 1: Bar chart representation of odds ratio (OR) values (Y axis) and main HLA alleles associated (X axis) with risk to multiple sclerosis (MS) susceptibility in the stratified analysis by conditioning on *HLA-DRB1\*15:01:01:01* of the Spanish population cohort.  $-\log(p\text{-value})$  values are displayed in the squares above each respective bar.

OR

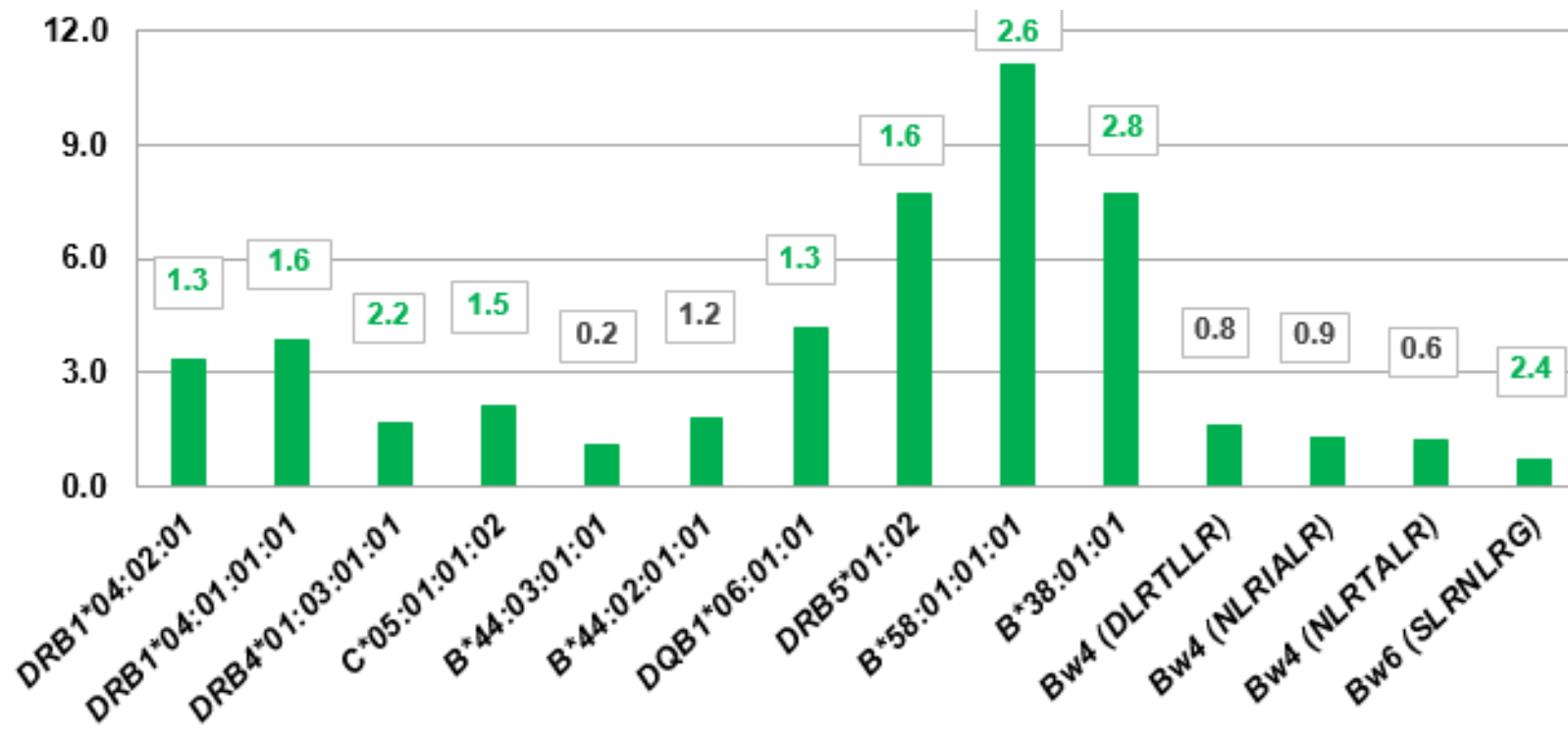


*HLA allele signal*

HLA alleles associated with risk (RED colored) to MS susceptibility are depicted based on the respective most common  $-\log(p\text{-value})$ -bearing haplotypes in which they are embedded. OR, Odds ratio; P values derived from a two-tailed Fischer's exact test are shown as the  $-\log(p\text{-value})$  (values displayed in the squares above each respective bar). A  $-\log(p\text{-value})$  of 1.3 or higher (equivalent to  $p\text{-value} \leq 0.05$ ) was considered statistically significant (in red bold). "Binned" category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to low allele frequency values.

Figure R-5a. Study 1: Bar chart representation of the inverse of odds ratio (OR) values (Y axis) and main HLA alleles associated (X axis) with protection to multiple sclerosis (MS) susceptibility in the nominal analysis of the Spanish population cohort.  $-\log(p\text{-value})$  values are displayed in the squares above each respective bar.

*1/OR*

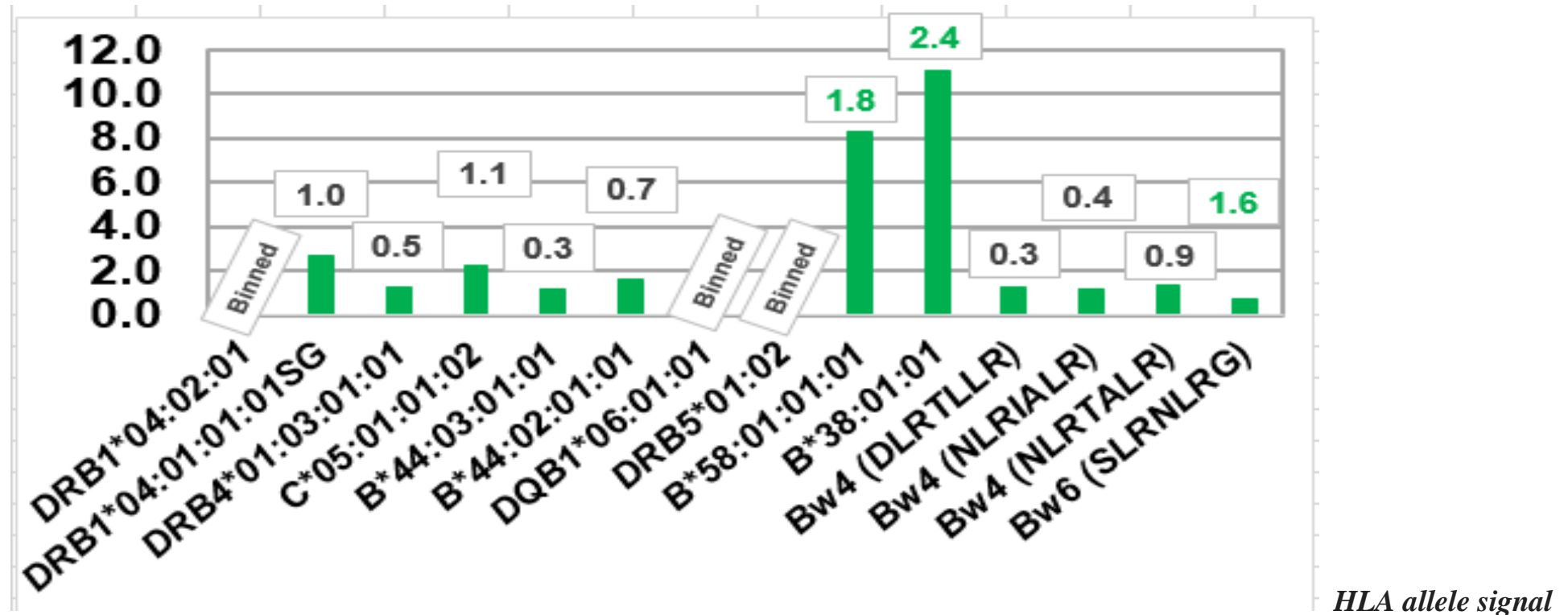


*HLA allele signal*

HLA alleles associated with protection (GREEN colored) to MS susceptibility are depicted based on the respective most common -bearing haplotypes in which they are embedded. OR, Odds ratio; P values derived from a two-tailed Fischer's exact test are shown as the  $-\log(p\text{-value})$  (values displayed in the squares above each respective bar). A  $-\log(p\text{-value})$  of 1.3 or higher (equivalent to  $p\text{-value} \leq 0.05$ ) was considered statistically significant (in green bold).

Figure R-5b. Study 1: Bar chart representation of the inverse of odds ratio (OR) values (Y axis) and main HLA alleles associated (X axis) with protection to multiple sclerosis (MS) susceptibility in the stratified analysis by conditioning on *HLA-DRB1\*15:01:01:01* of the Spanish population cohort.  $-\log(p\text{-value})$  values are displayed in the squares above each respective bar.

*1/OR*



HLA alleles associated with protection (GREEN colored) to MS susceptibility are depicted based on the respective most common  $-\log(p\text{-value})$ -bearing haplotypes in which they are embedded. OR, Odds ratio; P values derived from a two-tailed Fischer's exact test are shown as the  $-\log(p\text{-value})$  (values displayed in the squares above each respective bar). A  $-\log(p\text{-value})$  of 1.3 or higher (equivalent to  $p\text{-value} \leq 0.05$ ) was considered statistically significant (in green bold). "Binned" category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to low allele frequency values.

Referring now to protective effects for MS identified at the 3-/4-field allele level in the present Spanish population study, three main observations can be remarked in relation to the initial nominal analysis (see **Table R-11a** and **Figure R-5a**):

-Firstly, HLA class II alleles *DRB1\*04:02:01* (OR=0.30; p=5.08E-02) and *DRB1\*04:01:01:01* (OR=0.26; p=2.27E-02), as well as the respective associated *DRB4\*01:03:01:01* (OR=0.59; p=6.76E-03) in a very tight LD with these HLA-DRB1 alleles, were certainly protective for MS in the current Spanish dataset. Strikingly, this finding is in line with formerly reported studies in African American [669][683] but not with the large majority of previous studies in populations of European descent (which, conversely, even showed that certain different *HLA-DRB1\*04* alleles exhibit a predisposing effect), and with the only exception of this same result recently described in two different NGS studies of non-Hispanic European American cohorts [292][293] (although with certain nuances that are later commented in the **DISCUSSION** section and in the context of the present thesis work). Furthermore, in the present dataset, another important protective signal for MS detected within the HLA class II region was tentatively coming from the haplotype *HLA-DRB5\*01:02~HLA-DRB1\*15:02:01:02~HLA-DQB1\*06:01:01*. Nonetheless, at the HLA allele level only *HLA-DRB5\*01:02* (OR=0.13; p=2.58E-02) and *-DQB1\*06:01:01* (OR=0.24; p=4.79E-02) alleles were statistically significant in their protective effect to MS susceptibility. Whereas either *HLA-DRB1\*15:02:01:02* allele (AF=0.0137, in healthy controls; AF=0.0000, in cases) or its other intronic variant *DRB1\*15:02:01:01* (AF=0.0017, in healthy controls; AF= 0.0021, in cases) were not statistically measurable since they fall in the “binned” category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to their allele frequencies are very low in the present cohort.

-Secondly, within the HLA class I region, three different protective signals were detected. *HLA-B\*38:01:01* allele (OR=0.13;  $p=1.53E-03$ ) showed a statistically significant and strong protective effect, being this finding consistent with previous reports in population cohorts of European ancestry [658][677]. Moreover, in the present dataset *HLA-B\*58:01:01:01* (OR=0.09;  $p=2.71E-03$ ) was a new class I identified strongly protective allele and statistically significant that, to the best of our knowledge, it has not been previously reported in the literature (reviewed in [650][651]). Since this represented a novel finding, interpretation of this result was taken even with more caution and it was also further evaluated for trying to detect any possible not apparent LD effect or a plausible confounding result due to population substructure. At the same time, a third protective (although more modest comparatively) HLA class I signal was also observed in relation to alleles *HLA-B\*44:02:01:01* (OR=0.56;  $p=6.02E-02$ ) and *C\*05:01:01:02* (OR=0.47;  $p=3.26E-02$ ). In this case, as it has been also well-documented in the literature, *HLA-B\*44:02* allele (presenting the motif Bw4-80T) is in extremely tight LD with *HLA-C\*05:01* allele and thus it is still difficult to discriminate between their tentative individual contributions to protection for MS; yet, and at first, it appears to be relying more on the *HLA-C\*05:01* allele variant [292][650][714][715][727]. It is also noteworthy that the counterpart *HLA-B\*44:03:01:01* allele (OR=0.92;  $p=6.87E-01$ ) is clearly not protective in the present dataset.

-Thirdly, and once again in the context of HLA class I region, it was also of interest to assess the association with MS susceptibility/protection for HLA-B alleles (as a clarification, for this particular analysis other HLA-A and -C alleles were not considered here) that present either serological Bw6 epitope (SLRNLRG motif in amino acid positions 77-83) or Bw4 epitope (here, evaluating only three main Bw4 motif subgroups according to the specific series of amino acids found in positions 77-83: Bw4 (DLRTLLR); Bw4 (NLRIALR); and Bw4 (NLRTALR))



(see **Figure I-27** [728]). Thus, according to the present Spanish population dataset, it can be observed that all three main Bw4 motif subgroups evaluated [Bw4 (DLRTLRL), (OR=0.62; p=1.45E-01); Bw4 (NLRIALR), (OR=0.78; p=1.14E-01); and Bw4 (NLRTALR), (OR=0.83; p=2.73E-01)] were not clearly protective (i.e. at least not statistically significant) in contrast to what has been described in previous studies for both European and African American cohorts [724-727][947]. Conversely, in our dataset Bw6 epitope (SLRNLRG) (that has not been described to interact with any KIR [720]) showed a statistically significant predisposing effect to MS susceptibility (OR=1.45; p=3.70E-03), which has not been previously described either (at least directly referring to the specific group of HLA-B alleles encoding Bw6 epitope) in the literature as far as our knowledge.

In addition to this initial nominal analysis, we then conducted a stratification or conditional analysis in a model adjusting, once again, for the major risk HLA allele found (*HLA-DRB1\*15:01:01:01*) to further evaluate these nominally observed protective effects for MS in the present Spanish dataset. Thus, our aim was to assess whether these protective associations were simply reflecting LD patterns (thus, dependent in this case of the presence of highly predisposing *HLA-DRB1\*15:01:01:01* allele) or were indeed statistically (and thus functionally, at least in theory) independent since they remained nominally significant even in the *DRB1\*15:01:01:01*-negative stratum (see **Table R-11b** and **Figure R-5b**). In detail:

-On one hand, conditional analysis revealed that the nominally observed protective effect of both class II signals (across different alleles in tight LD respectively) abovementioned [*DRB1\*04:02:01* (OR="binned"; p="binned") with *DRB1\*04:01:01:01* (OR=0.37; p=1.06E-01) and with *DRB4\*01:03:01:01* (OR=0.82; p=3.39E-01); and also the *HLA-DRB5\*01:02* (OR="binned"; p="binned") with *-DQB1\*06:01:01* (OR="binned";

p="binned"] can be certainly attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele. Because when the data are stratified on the presence of *DRB1\*15:01:01:01*, the statistical significance and the protective effect of these HLA class II allele signals are clearly diminished (i.e. resulted in loss of the statistical significance associated with protection) in the stratum missing *DRB1\*15:01:01:01*. This interpretation suggests that these observed protective associations in class II alleles with MS in the present dataset may not be necessarily due to their functional protective role (i.e. impairing or avoiding at some level the immunopathogenesis of MS) but may simply reflect related LD patterns and allele/haplotype frequency distributions. Similarly, as it is also shown and explained later in the haplotype level analyses section, alleles *HLA-B\*44:02:01:01* (OR=0.61.; p=1.87E-01) and *C\*05:01:01:02* (OR=0.45.; p=8.19E-02) were also dependent of *DRB1\*15:01:01:01* allele and their protective effect relies as well on negative LD with this highly predisposing allele *DRB1\*15:01:01:01*.

-On the other hand, *HLA-B\*38:01:01* allele (OR=0.09; p=3.92E-03) as well as *HLA-B\*58:01:01:01* (OR=0.12; p=1.58E-02) remained nominally significant even in the *DRB1\*15:01:01:01*-negative stratum, so the protective association cannot be simply attributed to negative LD with the highly predisposing *DRB1\*15:01:01:01*. Interestingly, *HLA-B\*38:01:01* and *B\*58:01:01:01* have in common that they both encode the Bw4 motif subgroup NLRIALR (Bw4-80I). However, this Bw4 motif subgroup NLRIALR (OR=0.85; p=3.64E-01) as well as the other two Bw4 motif subgroups [Bw4 (DLRTLLR), (OR=0.80 p=5.05E-01) and Bw4 (NLRTALR), (OR=0.73; p=1.38E-01)] were not associated in either stratum. Thus, in the present dataset we did not find a clear Bw4 protective association with MS susceptibility that could reflect either LD pattern (i.e. negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele.) or ligand mediated KIR3DL1 signaling (indirectly evaluated here).

-Moreover, as for Bw6 epitope (SLRNLRG) encoded by the respective group of HLA-B alleles analyzed here. In this case, its risk association also stayed relatively strong and statistically significant (OR=1.39; p=2.72E-02) in the stratum lacking *DRB1\*15:01:01:01*. Thus, this given risk association cannot be attributed simply to LD patterns in relation to the highly predisposing allele *DRB1\*15:01:01:01* and, consequently, it appears to be independent and tentatively conferring a synergic effect, although this would need to be further evaluated in future studies.

- As part of this **Study 1** shown here and also in relation to allele level analyses, taking advantage of the very high-resolution HLA genotyping data (up to the 4-field with minimum ambiguities) obtained via NGS it was also of interest to assess the degree of risk association with MS in relation to a particular intronic variant of *HLA-DRB5\*01:01:01* allele recently described [308] and termed as *HLA-DRB5\*01:01:01v1* (with a single substitution of A to G in Intron 2 at Position 7312) (see **Table R-12**). Which was indeed positively associated with MS in a previously reported NGS study of a non-Hispanic European American cohort [293]. So, at the time when our study was conducted, there were available two *HLA-DRB5\*01:01:01* allele genomic reference sequences included in the MIA FORA™ NGS FLEX HLA Genotyping Software version 3.0, via VNC viewer, with reference to IPD-IMGT/HLA database release 3.25.0 (Immucor, Inc. Norcross, GA, USA): *HLA-DRB5\*01:01:01* allele sequence (GenBank accession number AL713966); and, in addition, an intronic variant of this *HLA-DRB5\*01:01:01* allele sequence (although without complete gene coverage, lacking intron-1 sequence) denoted as *HLA-DRB5\*01:01:01v1* (GenBank accession number KU593576) that was described by cloning and sequencing experiments in a previous work [308]. In the present Spanish dataset, and in contrast to this other large European American cohort MS study [293], this very rare intronic variant *HLA-DRB5\*01:01:01v1* occurred at similar frequencies in both MS cases and controls (0.4% cases

versus 0.3% controls, OR=1.23; p=1.00). Whereas, as previously commented, the much more common, comparatively, *HLA-DRB5\*01:01:01* variant occurred at higher frequencies in MS cases than controls (20.0% cases versus 9.2% controls, OR=2.44; p=4.31E-07). Thus, the present Spanish HLA-MS study, with a modest and still limited sample size, was therefore not sufficiently powered to detect this tentative association for this infrequent intronic variant *HLA-DRB5\*01:01:01v1*.

|                                    | MS (2n) | 476   | Controls (2n) | 564   |      |           |          |
|------------------------------------|---------|-------|---------------|-------|------|-----------|----------|
| HLA Allele                         | Count   | Fq    | Count         | Fq    | OR   | 95% CI    | p-value  |
| <b>HLA-DRB1*15:01:01:01</b>        | 97      | 0.204 | 55            | 0.094 | 2.46 | 1.70-3.58 | 4.10E-07 |
| <b>HLA-DRB5*01:01:01 (nt. A)</b>   | 95      | 0.200 | 54            | 0.092 | 2.44 | 1.69-3.55 | 4.31E-07 |
| <b>HLA-DRB5*01:01:01v1 (nt. G)</b> | 2       | 0.004 | 2             | 0.003 | 1.23 | 0.17-8.75 | 1.00     |

**Table R-12. Study 1: Significant associations of *HLA-DRB1\*15:01:01:01*, *HLA-DRB5\*01:01:01* and *HLA-DRB5\*01:01:01v1* (intron variant) alleles in Spanish MS cases and controls.**

## **12. HLA HAPLOTYPE LEVEL ANALYSES ON FIRST CASE-CONTROL STUDY**

- Subsequently, in the corresponding **Study 1**, we also evaluated these abovementioned risk/protection allelic associations to MS susceptibility in the context of HLA class II haplotype blocks (see **Table R-13**) as well as in regards to the fully extended HLA class I and class II haplotypes (see **Table R-14**) at the 3-/4-field of resolution. In detail:

-On one hand (see **Table R-13**), the distribution of class II *HLA-DQB1~DQA1~DRB1~DRB3/4/5* haplotype blocks in MS cases and controls was compared. Based on the risk/protective HLA allele signals previously detected (and even though we still analyzed and reviewed the haplotype distribution data entirely), we particularly assessed (being

of special interest in the scope of the present study) which *HLA-DRB1\*15:01:01:01-*, *HLA-DRB1\*15:02:01:01/:02-*, *HLA-DRB1\*04:01:01:01-*, *HLA-DRB1\*07:01:01:01-* and *HLA-DRB1\*04:04:01*-bearing haplotypes are associated with risk/protection to MS susceptibility. Overall, there is a clear correlation of the MS risk/protection associations found between the individual allele signals previously identified and the corresponding associated HLA class II haplotypes shown here, in which these alleles are embedded displaying specific and distinctive LD patterns.

Firstly, and as it was expected, the *HLA-DQB1\*06:02:01~DQA1\*01:02:01:01~DRB1\*15:01:01:01~DRB5\*01:01:01* haplotype shows a strong association, being statistically very significant (OR=2.39; p=3.45E-06), by conferring risk to MS susceptibility. In addition, it is worth noting that this prototypic or “classic” *HLA-DRB1\*15:01:01:01*-bearing class II haplotype representing the major MS risk signal is the most frequent of its kind in Spanish general population, as it has been also described in other populations of European ancestry [130][292][293][464][727]. Whereas other positively associated *HLA-DRB1\*15:01:01:01*-bearing haplotypes with MS evaluated here exhibit low or very low haplotype frequencies, being more difficult to deduce with certainty any plausible and suitably clear interpretation of these association results obtained at the haplotype level (i.e. since these haplotype frequencies fall into the “binned” category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to their low haplotype frequencies; yet, their statistical parameters were manually calculated and shown here only for purposes of comparison). For instance, even though found at low haplotype frequencies, it is worth pointing out that the *HLA-DRB1\*15:01:01:01*-bearing haplotype lacking *HLA-DQB1\*06:02:01* and containing *HLA-DQB1\*06:03:01* allele instead has still a noticeable higher frequency in MS cases than in healthy controls. Consequently, this could be suggesting that specific presence of *HLA-*

*DQB1\*06:02:01* allele within this prototypic *HLA-DRB1\*15:01:01:01*-bearing class II haplotype may not be as pivotal on the major observed associated risk to MS susceptibility (observed here for Spanish population); and also as previously described in other case-control population studies such as in the case of African Americans [669]. Alternatively, both *HLA-DQB1\*06:02:01* and *HLA-DQB1\*06:03:01* (but not *HLA-DQB1\*06:01:01*, with a protective role identified in the present study, at first, due to related LD patterns and allele/haplotype frequency distributions) could be potentially playing a similar secondary role on MS risk at the HLA haplotype level. Nevertheless, this preliminary observation would need to be further investigated in larger NGS studies for Spanish population.

Moreover, although statistically speaking it was not sufficiently measurable either (once again, due to low haplotype frequency values found in the present Spanish population cohort), in regards to *HLA-DRB1\*15:02:01*-bearing haplotypes it was possible to confirm those protective allele signals previously detected for *HLA-DQB1\*06:01:01* and - *DRB5\*01:02* embedded in this haplotype. Particularly, evaluated here at the 3-/4-field haplotype level, it was striking to find a (tentatively) clear difference in distribution found between haplotypes comprising distinctive 4-field intron variants *HLA-DRB1\*15:02:01:01* and *HLA-DRB1\*15:02:01:02* regarding association with MS susceptibility. These two variants cannot be distinguished at an unsequenced region of Intron 1 2146 bp G-->C SNP, and from a STR (dinucleotide) in Intron 2 (positions 5693 to 5748 bp) with repeats of GT; however, on the other hand, these two 4-field variants differed in Intron 2 at the position 6272 Deletion(.)-->T SNP according to v.3.25.0 IPD-IMGT/HLA database (released July 2016) [87][295][297][362]. Thus, in the present study, the former haplotype was similarly present in both cases and healthy control groups (OR=1.23; p=1.00), while the latter *HLA-DRB1\*15:02:01:02*-bearing haplotype clearly shows a strong association, being statistically very significant (OR=0.08; p=9.91E-03), by conferring

protection to MS susceptibility. Therefore, this nicely exemplifies how NGS HLA genotyping allowed us to detect a relevant 4-field difference at the haplotype level in relation to the association of MS, in this case, for protection.

Also, referring now to other protective effects observed at the HLA class II haplotype level (although again found at relatively low haplotype frequency values, with the only exception of

*DQB1\*02:02:01:01~DQA1\*02:01:01:01~DRB1\*07:01:01:01~DRB4\*01:01:01:01*

haplotype, which is commonly found in the present Spanish general population dataset;

however both MS cases (AF=11.6%) and healthy controls (AF=11.8%) show quite similar frequency distributions (slightly higher in controls but not statistically significant at all)). It is

noteworthy the protective association found in several different *HLA-DRB1\*04:01/04:04-* and *-DRB1\*07:01:01:01*-bearing haplotypes, presenting a diverse set of encoded HLA-DQ

heterodimers within the corresponding haplotype and containing *HLA-DRB4\*01:03:01:01* or *-DRB4\*01:01:01:01/-DRB4\*01:03:01:02N* alleles, respectively, as the secondary DRB loci. As

previously commented at the HLA allele-level analyses, these protective signals may not be necessarily due to their functional protective role (i.e. impairing or avoiding at some level the immunopathogenesis of MS), but may simply reflect related LD patterns and allele/haplotype

frequency distributions relative to the counterpart *HLA-DRB1\*15:01:01:01*-bearing class II haplotypes. Strikingly, the protective role found for *HLA-DRB1\*04:01/04:04*-bearing

haplotypes in the present Spanish population study is clearly in contrast to what it had been previously reported in singular Spanish population groups such as Basques [742] and Canary

Islanders [738], where these allele groups (and corresponding bearing haplotypes) were more prevalent in MS patients. These discrepancies found between studies in Spanish population may

be indicative most likely of existing population stratification which may be causing these notably different observed HLA allele/haplotype distributions; or alternatively, but less likely

at first, the possibility that within the Spanish MS patient general population there could be different MS sub-cohorts that may be defined according to distinctive HLA markers [689][709]. Yet, these admittedly speculative interpretations would need to be further investigated in larger NGS studies with sufficient and adequate clinical data available for Spanish population. Interestingly, this observed heterogeneity has been also found in relation to other *HLA-DRB1\*04* subtypes (and respective carrying haplotypes) and their risk/protective association to MS susceptibility across worldwide populations so far studied [650][651]. Thus, for example, a relevant positive association of the other allele group *HLA-DRB1\*04:05* with MS has been widely described in Japanese and Asian populations [294][678-681] as well as in the unique Sardinian islander population (tentatively influenced by diverse founder populations) [650][662][670][686][687][688]. Also, in a recent NGS HLA study on a large European American cohort, analyses of the *HLA-DRB1\*04* in the absence of *HLA-DRB1\*15:01* haplotypes revealed that the *HLA-DQB1\*03:01:01:01~HLA-DQA1\*03:03:01:01~HLA-DRB1\*04:01:01:01~HLA-DRB4\*01:03:01:01* haplotype was protective, whereas the *HLA-DQB1\*03:02:01~HLA-DQA1\*03:01:01~HLA-DRB1\*04:01:01:01~HLA-DRB4\*01:03:01:01* haplotype was associated with disease susceptibility [293]. However this pattern was not found in the present Spanish population cohort, where both HLA class II haplotypes are clearly protective (see **Table R-13**). Altogether, these findings may suggest complex interactions between HLA loci (in the context of HLA haplotypes as the fundamental unit of genetic control of their immune role), which may also implicate epistasis (in cis- and trans-) among HLA class II loci conferring specific susceptibility and resistance effects respectively [672][674][677][952]. At the same time, these findings raise questions about disease mechanisms that certainly require further examination in functional studies [677]. Moreover, large trans-ethnic NGS HLA studies may contribute to better interrogate and shed light to this



observed HLA-DRB1 heterogeneity and the respective association patterns to MS susceptibility, even within populations of European ancestry [650][651].

Finally, as another important remark to be mentioned in this HLA class II haplotype-level analysis of the present Spanish population NGS HLA-MS study and similarly to the limitations found in other previous studies on populations of European ancestry [292][293][650][651][727]; identification of the true predisposing gene of MS susceptibility within the prototypic susceptibility HLA-DR15 (*HLA-DRB1\*15:01:01:01~DRB5\*01:01:01*) haplotype is handicapped by the intense and exceptionally tight linkage disequilibrium (LD) across these given individual alleles at HLA-DRB1 and HLA-DRB5 loci. Consequently, in the majority of populations of European descent all common *HLA-DRB1\*15:01~DQB1\*06:02* haplotypes carry the *HLA-DRB5\*01:01* allele, while all frequent *HLA-DRB1\*16:01~DQB1\*05:02* haplotypes carry the *HLA-DRB5\*02:02* allele, thus the role of allelic variation at *HLA-DRB5* cannot be suitably assessed [292][293][668][727][743][944]. Nevertheless, on the other hand, the role of allelic variation at *HLA-DRB5* with risk to MS susceptibility may be better elucidated in Asian populations (primarily in Southeast Asia and Oceania) [697]. In which, in contrast to populations of European ancestry [696] and even though incidence of MS appears to be much lower [697], common *HLA-DRB1\*15:02*-bearing haplotypes (such as *HLA-DRB5\*01:02~DRB1\*15:02:01:01~DQA1\*01:03:01:01~DQB1\*06:01:01* (HF=10.7%) and *HLA-DRB5\*01:01:01~DRB1\*15:02:01:01~DQA1\*01:02:01:01~DQB1\*05:02:01* (HF=8.9%) [297]) may allow to discern the individual contributions of these HLA-DRB5 alleles in MS disease susceptibility, thus taking advantage of the particular LD patterns commonly found in these specific ethnic groups.

**Table R-13. Study 1: Significant HLA class II haplotype-level (at the 3-/4-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort.**

| Haplotype HLA-DQB1~DQA1~DRB1~DRB3/4/5                                | MS cases<br>(2n=476) |       | Controls<br>(2n=564) |       | OR    | 95% CI      | <i>p</i> value  |
|--|----------------------|-------|----------------------|-------|-------|-------------|-----------------|
|  | Count                | Fq    | Count                | Fq    |       |             |                 |
| DQB1*06:02:01~DQA1*01:02:01:01~DRB1*15:01:01:01~DRB5*01:01:01        | 87                   | 0.183 | 50                   | 0.086 | 2.39  | 1.65-3.46   | <b>3.45E-06</b> |
| DQB1*06:03:01~DQA1*01:02:01:01~DRB1*15:01:01:01~DRB5*01:01:01        | 6                    | 0.013 | 3                    | 0.005 | 2.50  | 0.615-9.94  | 0.31            |
| DQB1*05:02:01~DQA1*01:02:01:01~DRB1*15:01:01:01~DRB5*01:01:01        | 1                    | 0.002 | 0                    | 0.000 | 2.46* | 0.08-73.39  | 0.45            |
| DQB1*05:02:01~DQA1*01:02:02~DRB1*15:01:01:01~DRB5*01:01:01           | 2                    | 0.004 | 0                    | 0.000 | 4.92* | 0.22-109.45 | 0.20            |
| DQB1*06:01:01~DQA1*01:02:01:01~DRB1*15:01:01:01~DRB5*01:01:01        | 1                    | 0.002 | 1                    | 0.002 | 1.23  | 0.08-19.68  | 1.00            |
| DQB1*06:01:01~DQA1*01:03:01:01~DRB1*15:02:01:01~DRB5*01:02           | 1                    | 0.002 | 1                    | 0.002 | 1.23  | 0.08-19.68  | 1.00            |
| DQB1*06:01:01~DQA1*01:03:01:01~DRB1*15:02:01:02~DRB5*01:02           | 0                    | 0.000 | 8                    | 0.014 | 0.08* | 0.004-1.32  | <b>9.91E-03</b> |
| DQB1*03:02:01~DQA1*03:01:01~DRB1*04:01:01:01~DRB4*01:03:01:01        | 2                    | 0.004 | 8                    | 0.014 | 0.30  | 0.06-1.44   | 0.20            |
| DQB1*03:01:01:01~DQA1*03:03:01:01~DRB1*04:01:01:01~DRB4*01:03:01:01  | 1                    | 0.002 | 5                    | 0.009 | 0.24  | 0.03-2.09   | 0.23            |
| DQB1*03:03:02:01~DQA1*02:01:01:01~DRB1*07:01:01:01~DRB4*01:03:01:02N | 4                    | 0.008 | 9                    | 0.015 | 0.54  | 0.17-1.77   | 0.40            |
| DQB1*02:02:01:01~DQA1*02:01:01:01~DRB1*07:01:01:01~DRB4*01:01:01:01  | 55                   | 0.116 | 69                   | 0.118 | 0.98  | 0.67-1.42   | 0.92            |
| DQB1*03:02:01~DQA1*03:01:01~DRB1*04:04:01~DRB4*01:03:01:01           | 9                    | 0.019 | 12                   | 0.021 | 0.92  | 0.38-2.20   | 1.00            |

Main HLA class II haplotypes associated with risk (OR>1) or protection (OR<1) to MS susceptibility are depicted. OR, Odds ratio; 95% CI, 95% confidence interval. P values derived from a two-tailed Fischer's exact test. A P value of 0.05 ( $\alpha$ ) or less was considered statistically significant (in bold). \*Haldane-Anscombe correction was applied in those situations with zero counts (i.e. just adding 0.5 to each of the cells and then calculate the odds ratio (OR) over these adjusted cell counts).

**Table R-14. Study 1: Significant extended HLA haplotype-level (at the 3-/4-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort (Controls 2n=564; MS cases 2n=476).**

| <i>HLA-A</i>         | <i>HLA-B</i>         | <i>HLA-C</i>         | <i>HLA-DQA1</i>         | <i>HLA-DQB1</i>         | <i>HLA-DRB1</i>         | <i>HLA-DRB3/4/5</i>     | HF in Controls | HF in MS cases | OR    |
|----------------------|----------------------|----------------------|-------------------------|-------------------------|-------------------------|-------------------------|----------------|----------------|-------|
| <i>A*03:01:01:01</i> | <i>B*07:02:01</i>    | <i>C*07:02:01:03</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0205         | 0.0567         | 2.87  |
| <i>A*25:01:01</i>    | <i>B*18:01:01:02</i> | <i>C*12:03:01:01</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0068         | 0.0105         | 1.55  |
| <i>A*02:01:01:01</i> | <i>B*14:01:01</i>    | <i>C*08:02:01:02</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0000         | 0.0042         | 4.92* |
| <i>A*02:01:01:01</i> | <i>B*40:01:02</i>    | <i>C*03:04:01:01</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0017         | 0.0042         | 2.48  |
| <i>A*11:01:01:01</i> | <i>B*51:01:01:01</i> | <i>C*15:02:01:01</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0017         | 0.0042         | 2.48  |
| <i>A*29:02:01:01</i> | <i>B*44:03:01:01</i> | <i>C*16:01:01:01</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:02:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0017         | 0.0042         | 2.48  |
| <i>A*02:01:01:01</i> | <i>B*07:02:01</i>    | <i>C*07:02:01:03</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:03:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0017         | 0.0021         | 1.24  |
| <i>A*02:01:01:01</i> | <i>B*44:02:01:01</i> | <i>C*05:01:01:02</i> | <i>DQA1*01:02:01:01</i> | <i>DQB1*06:03:01</i>    | <i>DRB1*15:01:01:01</i> | <i>DRB5*01:01:01</i>    | 0.0017         | 0.0021         | 1.24  |
| <i>A*11:01:01:01</i> | <i>B*52:01:01:02</i> | <i>C*12:02:02</i>    | <i>DQA1*01:03:01:01</i> | <i>DQB1*06:01:01</i>    | <i>DRB1*15:02:01:02</i> | <i>DRB5*01:02</i>       | 0.0051         | 0.0000         | 0.20* |
| <i>A*29:02:01:01</i> | <i>B*44:02:01:01</i> | <i>C*05:01:01:02</i> | <i>DQA1*03:03:01:01</i> | <i>DQB1*03:01:01:01</i> | <i>DRB1*04:01:01:01</i> | <i>DRB4*01:03:01:01</i> | 0.0051         | 0.0000         | 0.20* |
| <i>A*02:01:01:01</i> | <i>B*44:03:01:01</i> | <i>C*16:02:01</i>    | <i>DQA1*03:01:01</i>    | <i>DQB1*03:02:01</i>    | <i>DRB1*04:02:01</i>    | <i>DRB4*01:03:01:01</i> | 0.0034         | 0.0021         | 0.62  |
| <i>A*02:01:01:01</i> | <i>B*40:01:02</i>    | <i>C*03:04:01:01</i> | <i>DQA1*03:01:01</i>    | <i>DQB1*03:02:01</i>    | <i>DRB1*04:04:01</i>    | <i>DRB4*01:03:01:01</i> | 0.0086         | 0.0000         | 0.11* |
| <i>A*33:03:01</i>    | <i>B*58:01:01:01</i> | <i>C*03:02:02:01</i> | <i>DQA1*05:01:01:03</i> | <i>DQB1*02:01:01</i>    | <i>DRB1*03:01:01:01</i> | <i>DRB3*02:02:01:01</i> | 0.0034         | 0.0000         | 0.31* |
| <i>A*02:01:01:01</i> | <i>B*38:01:01</i>    | <i>C*12:03:01:01</i> | <i>DQA1*01:03:01:02</i> | <i>DQB1*06:03:01</i>    | <i>DRB1*13:01:01:01</i> | <i>DRB3*01:01:02:01</i> | 0.0034         | 0.0021         | 0.62  |

Main extended HLA haplotypes associated with risk (OR>1) or protection (OR<1) to MS susceptibility are depicted. OR, Odds ratio.

\*Haldane-Anscombe correction was applied in those situations with zero counts (i.e. just adding 0.5 to each of the cells and then calculate the odds ratio (OR) over these adjusted cell counts).

-The clonal nature of NGS also allowed us to conduct a very detailed analysis of each of the previously detected HLA allelic risk/protection signals to MS susceptibility in the context of the corresponding completely extended HLA class I and class II haplotype blocks at the 3-/4-field of resolution (even though estimated via EM in the current Spanish unrelated dataset available) (see **Table R-14**). Thus, this analysis helped us to put into perspective the given characteristic LD patterns and also to map precisely the respective individual allele associations. Once again, based on the HLA MS-risk and -protective allelic signals initially identified (and even though we still analyzed and reviewed the extended haplotype distribution data entirely), we particularly assessed (being of special interest in the scope of the present study) those distinctive *HLA-DRB1\*15:01:01:01~DQB1\*06:02:01*-bearing haplotypes; other *HLA-DRB1\*15:01:01:01*-bearing haplotypes observed in the present Spanish population cohort; as well as *HLA-DRB1\*15:02:01:02*, *HLA-DRB1\*04:01:01:01*-, *HLA-DRB1\*04:02:01*- and *HLA-DRB1\*04:04:01*-bearing haplotypes associated with protection to MS susceptibility; in addition to the extended haplotypes comprising the also protective HLA class I *HLA-B\*38:01:01* and *HLA-B\*58:01:01:01* alleles respectively. Overall, all these observed extended HLA haplotype associations are in line with those HLA allele and class II haplotype risk/protection signals previously shown and commented in detail. However, once again, due to the pronounced haplotypic diversity and distinct patterns of LD observed at the 3-/4-field of resolution, most of these extended haplotype associations were not statistically measurable (i.e. larger number of haplotypes relative to alleles at individual loci tends to decrease power due to the additional degrees of freedom required for the analysis [711]) since they fall in the “binned” category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to their haplotype frequencies are very low in the present Spanish population cohort (presenting still a modest sample size). Yet, their statistical parameters (odds ratio, OR) were manually calculated and are

shown here only for purposes of comparison. Thus, the purpose of this analysis was merely to have a relative comparison of estimated risk/protection across these different extended haplotypes.

Interestingly, in contrast to other studies on populations of North-/Central-European ancestry (in which the prototypic susceptibility *HLA-A\*03:01:01:01~B\*07:02:01~C\*07:02:01:03~DQA1\*01:02:01:01~DQB1\*06:02:01~DRB1\*15:01:01:01~DRB5\*01:01:01* extended haplotype is considerably more prevalent in MS patients) [292][293][650][651][727], the present Spanish population cohort shows a (tentatively) greater diversity of extended *HLA-DRB1\*15:01:01:01~DQB1\*06:02:01*-bearing haplotypes positively associated to MS, which encompass different sets of HLA class I alleles commonly found in populations from Mediterranean regions [130][297][464]. Thus, this observation suggests that extended susceptible *HLA-DRB1\*15:01*-bearing haplotypes are heterogeneous and, in turn, these different haplotype-based functional cassettes can be associated with MS susceptibility to some extent and varying across populations, even within populations of European ancestry [650][651][711]. Still, testing whether the point estimates for these ORs are significantly different will require a larger sample set. It is also noteworthy that according to the extended haplotype distributions the risk association of *HLA-DQB1\*06:03:01*-bearing haplotypes appears to be much weaker than in the case of *HLA-DQB1\*06:02:01*-bearing haplotypes on this secondary role of MS risk.

Moreover, the relatively strong protective association of *HLA-A\*02:01* allele in LD with the *HLA-C\*03:04~B\*40:01* haplotype previously reported in other studies in European American cohorts [292][658][712][727] was not clearly detected in the present Spanish population study since at the individual allele level none of those protective signals were statistically significant

(data not shown). In fact, one of the *HLA-DRB1\*15:01:01:01~DQB1\*06:02:01*-bearing haplotypes found here comprises this HLA class I haplotype. On other hand, it can be seen that the protective *HLA-DRB1\*04:04:01*-bearing haplotype also contains this same HLA class I haplotype and, thus, this may explain their associated protective effect, not necessarily due to their functional protective role but simply attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01*. Similarly, *HLA-B\*44:02:01:01* and *-C\*05:01:01:02* allelic protective signals may be driven by allele *HLA-DRB1\*04:01:01:01* at HLA-DRB1 locus since all these are embedded in this same extended haplotype.

Furthermore, we confirmed the presence of statistically independent HLA-B effects (i.e. displaying *HLA-DRB1\*15:01:01:01*-independent associations). Our analysis fine-mapped these to *HLA-B\*58:01:01:01* and *HLA-B\*38:01* again at the extended haplotype-level.

Lastly, since the haplotype diversity dramatically increases when including *HLA-DPA1* and *-DPB1* loci due to existing hotspot of recombination between HLA-DQ and -DP loci [92]. Consequently, extended haplotype associations relative to the risk signal found in *HLA-DPB1\*03:01:01* (going beyond its well-known strong LD with *HLA-DPA1\*01:03:01:03*, for encoding the respective heterodimer) cannot be suitably assessed in the present study. Future studies on the HLA-DP structure-function may provide further evidence to improve our understanding of the exact function of HLA-DP in the pathogenesis of MS.

In summary, knowledge of related extended haplotypic associations up to the 3-/4-field contribute to better elucidate the role of HLA class I and class II genes in MS susceptibility as well as detect and disentangle possible hitchhiking effects due to existing extended LD between neighboring genes. At the same time, it is also possible that different loci and alleles may act in synergy to confer susceptibility or protection to MS risk.

### **13. HLA ALLELE LEVEL ANALYSES ON SECOND CASE-CONTROL STUDY**

• Firstly, in the other so-called **Study 2**, we evaluated the associations of HLA alleles with MS risk and protection (see **Table R-15**) respectively at the 2-field of resolution. The analysis of alleles grouped by this 2-field of resolution led to similar conclusions as in the previous analysis based on 3-/4-field alleles, replicating also these results in associated risk/protection with similar statistical significance.

Moreover, it was of interest to evaluate possible existing differences in the findings of HLA-disease association with MS as a consequence of not being real causative variants but due to plausible regional HLA genetic variation within mainland Spain in the previous **Study 1**. Thus, as a statistical approach to try controlling for any possible existing population stratification (i.e. differences in genetic structure between disease and control groups) as a confounding factor that may affect the results obtained in the first study (**Study 1**), a second control group (not only ethnically- but also regionally-matched as it was from Catalonia) was used here in **Study 2**. In this sense, no major differences are observed between **Study 1** and **Study 2** at the HLA allele-level analyses. Only, as minor exceptions:

-It is noteworthy that in those associations related to protection for MS susceptibility, the statistical significance was slightly diminished in the case of *HLA-DQB1\*06:01* and *DRB5\*01:02* alleles. Yet, this was not indicative of existing population stratification effect.

-Whereas relative to risk signals, in **Study 2** *HLA-DPB1\*03:01* allelic signal was not as statistically significant as in **Study 1**. Nevertheless, this was still based on a very minor difference in regards to allele frequency distribution, thus, not being suggestive either of existing population substructure effect that might be conditioning this association found in both Spanish and Catalan population.

**Table R-15. Study 1&2: Significant HLA allele-level (at the 2-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort and in comparison to the second healthy control group (Cat-Controls from Catalonia).**

| HLA-Allele | Cat-MS cases (2n=476) (2-FL) |       | ESP-Controls (2n=564) (2-FL) |       | Spain MS HLA - Study 1 |           |                 | Cat-Controls (2n=392) (2-FL) |       | Catalonia MS HLA - Study 2 |           |                 |
|------------|------------------------------|-------|------------------------------|-------|------------------------|-----------|-----------------|------------------------------|-------|----------------------------|-----------|-----------------|
|            | Count                        | Fq    | Count                        | Fq    | OR                     | 95% CI    | p value         | Count                        | Fq    | OR                         | 95% CI    | p value         |
| DPB1*03:01 | 35                           | 0.074 | 24                           | 0.041 | 1.85                   | 1.05-3.30 | <b>2.20E-02</b> | 26                           | 0.066 | 1.12                       | 0.64-1.97 | 6.79E-01        |
| DQB1*06:02 | 87                           | 0.183 | 53                           | 0.091 | 2.24                   | 1.53-3.30 | <b>1.08E-05</b> | 30                           | 0.077 | 2.70                       | 1.71-4.34 | <b>5.08E-06</b> |
| DQA1*01:02 | 120                          | 0.252 | 88                           | 0.151 | 1.90                   | 1.38-2.62 | <b>3.54E-05</b> | -                            | -     | -                          | -         | -               |
| DRB1*15:01 | 97                           | 0.204 | 55                           | 0.094 | 2.46                   | 1.70-3.58 | <b>4.10E-07</b> | 32                           | 0.082 | 2.88                       | 1.86-4.55 | <b>4.78E-07</b> |
| DRB5*01:01 | 98                           | 0.206 | 56                           | 0.096 | 2.44                   | 1.69-3.55 | <b>4.31E-07</b> | 34                           | 0.087 | 2.73                       | 1.78-4.27 | <b>1.15E-06</b> |
| B*07:02    | 58                           | 0.122 | 55                           | 0.094 | 1.33                   | 0.89-2.01 | 1.47E-01        | 32                           | 0.082 | 1.56                       | 0.97-2.54 | 5.31E-02        |
| C*07:02    | 66                           | 0.139 | 60                           | 0.103 | 1.41                   | 0.95-2.08 | 7.20E-02        | 36                           | 0.092 | 1.59                       | 1.02-2.52 | <b>3.30E-02</b> |
| A*03:01    | 61                           | 0.128 | 59                           | 0.101 | 1.31                   | 0.88-1.95 | 1.66E-01        | 42                           | 0.107 | 1.22                       | 0.79-1.91 | 3.41E-01        |
| HLA-Allele | Cat-MS cases (2n=476) (2-FL) |       | ESP-Controls (2n=564) (2-FL) |       | Spain MS HLA - Study 1 |           |                 | Cat-Controls (2n=392) (2-FL) |       | Catalonia MS HLA - Study 2 |           |                 |
|            | Count                        | Fq    | Count                        | Fq    | OR                     | 95% CI    | p value         | Count                        | Fq    | OR                         | 95% CI    | p value         |
| DRB1*04:02 | 3                            | 0.006 | 12                           | 0.021 | 0.30                   | 0.05-1.13 | 5.08E-02        | 5                            | 0.013 | 0.49                       | 0.12-2.07 | 4.79E-01        |
| DRB1*04:01 | 3                            | 0.006 | 14                           | 0.024 | 0.26                   | 0.05-0.93 | <b>2.27E-02</b> | 10                           | 0.026 | 0.24                       | 0.04-0.95 | <b>2.04E-02</b> |
| DRB4*01:03 | 58                           | 0.122 | 108                          | 0.185 | 0.61                   | 0.43-0.87 | <b>4.94E-03</b> | 75                           | 0.191 | 0.59                       | 0.40-0.87 | <b>4.68E-03</b> |
| C*05:01    | 34                           | 0.071 | 50                           | 0.086 | 0.82                   | 0.51-1.32 | 3.95E-01        | 40                           | 0.102 | 0.68                       | 0.41-1.12 | 1.08E-01        |
| B*44:03    | 46                           | 0.097 | 57                           | 0.098 | 0.99                   | 0.64-1.52 | 9.58E-01        | 37                           | 0.094 | 1.03                       | 0.64-1.67 | 9.11E-01        |
| B*44:02    | 16                           | 0.034 | 34                           | 0.058 | 0.56                   | 0.29-1.06 | 6.02E-02        | 23                           | 0.059 | 0.56                       | 0.27-1.12 | 7.61E-02        |
| B*38:01    | 2                            | 0.004 | 18                           | 0.031 | 0.13                   | 0.01-0.56 | <b>1.53E-03</b> | 15                           | 0.038 | 0.11                       | 0.01-0.46 | <b>3.13E-04</b> |
| B*58:01    | 1                            | 0.002 | 14                           | 0.024 | 0.09                   | 0.00-0.57 | <b>2.71E-03</b> | 7                            | 0.018 | 0.116                      | 0.01-0.95 | <b>2.61E-02</b> |
| DQB1*06:01 | 2                            | 0.004 | 10                           | 0.017 | 0.24                   | 0.03-1.15 | <b>4.79E-02</b> | 5                            | 0.013 | 0.33                       | 0.06-1.69 | 2.54E-01        |
| DRB5*01:02 | 1                            | 0.002 | 9                            | 0.015 | 0.13                   | 0.00-0.98 | <b>2.58E-02</b> | 5                            | 0.013 | 0.16                       | 0.02-1.40 | 9.64E-02        |

HLA alleles associated with risk (RED colored) or protection (GREEN colored) to MS susceptibility are depicted based on the respective most common -bearing haplotypes in which they are embedded. OR, Odds ratio; 95% CI, 95% confidence interval. P values derived from a two-tailed Fischer's exact test. A P value of 0.05 ( $\alpha$ ) or less was considered statistically significant (in bold).



#### **14. HLA HAPLOTYPE LEVEL ANALYSES ON SECOND CASE-CONTROL STUDY**

- Secondly, in the corresponding **Study 2**, we also evaluated these abovementioned risk/protection allelic associations to MS susceptibility in the context of HLA class II haplotype blocks (see **Table R-16**) as well as in regards to the fully extended HLA class I and class II haplotypes (see **Table R-17**) at the 2-field of resolution. In detail:

-The analysis of HLA class II haplotype blocks grouped by this 2-field of resolution (see **Table R-16**) led to similar conclusions as in the previous analysis based on 3-/4-field data in **Study 1**, replicating also those results in associated risk/protection with (for the most part) similar statistical significance. Nevertheless, the LD displayed between specific non-coding polymorphisms was undetected at this 2-field allele resolution level (either “trimmed” or generated). Consequently, when these haplotypes are reduced to 2-field haplotypes this results in loss of specificity of the haplotype frequency distribution, lowering also the apparent LD between these class II loci. For instance, here it was not possible to identify the specifically protective *HLA-DRB1\*15:02:01:02*-bearing haplotype association above described. This demonstrates the importance of characterizing HLA alleles from full-length HLA gene sequences including UTRs and all intronic regions, allowing assignment of specific haplotypes. Moreover, in some cases (e.g. *HLA-DQB1\*05:02~DRB1\*15:01~DRB5\*01:01*) the grouping or combination of the 3-/4-field allelic variants into a single 2-field variant may lead to a more statistical power or significance. However, this may not be biologically appropriate since these non-coding regions contain relevant sites (establishing also a specific LD pattern) for transcription promoters, inhibitors, alternative splice sites, methylation sites, binding sites for post-translational miRNA degradation and many other functions as yet undetermined [951]. Therefore, statistical significance at the 2-field should not generally take precedence over

**Table R-16. Study 1&2: Significant HLA class II haplotype-level (at the 2-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort and in comparison to the second healthy control group (Controls Cat from Catalonia).**

| Haplotype HLA-DQB1~DRB1~DRB3/4/5 | MS cases<br>(2n=476) |       | Controls ESP<br>(2n=564) |       |       |             |                 | Controls Cat<br>(2n=392) |       |        |             |                 |
|----------------------------------|----------------------|-------|--------------------------|-------|-------|-------------|-----------------|--------------------------|-------|--------|-------------|-----------------|
|                                  | Count                | Fq    | Count                    | Fq    | OR    | 95% CI      | p value         | Count                    | Fq    | OR     | 95% CI      | p value         |
| DQB1*06:02~DRB1*15:01~DRB5*01:01 | 87                   | 0.183 | 43                       | 0.081 | 2.53  | 1.72-3.74   | <b>1.98E-06</b> | 29                       | 0.074 | 2.80   | 1.80-4.36   | <b>1.99E-06</b> |
| DQB1*06:03~DRB1*15:01~DRB5*01:01 | 6                    | 0.013 | 3                        | 0.006 | 2.24  | 0.56-9.02   | 0.32            | 0                        | 0.000 | 10.00* | 0.56-179.52 | <b>0.04</b>     |
| DQB1*05:02~DRB1*15:01~DRB5*01:01 | 4                    | 0.008 | 0                        | 0.000 | 8.98* | 0.47-170.20 | 0.05            | 3                        | 0.008 | 1.10   | 0.24-4.94   | 1.00            |
| DQB1*06:01~DRB1*15:01~DRB5*01:01 | 1                    | 0.002 | 1                        | 0.002 | 1.11  | 0.07-17.85  | 1.00            | 0                        | 0.000 | 1.65*  | 0.06-49.27  | 1.00            |
| DQB1*06:01~DRB1*15:02~DRB5*01:02 | 1                    | 0.002 | 8                        | 0.015 | 0.14  | 0.02-1.10   | <b>0.04</b>     | 5                        | 0.013 | 0.16   | 0.02-1.40   | 0.10            |
| DQB1*03:02~DRB1*04:01~DRB4*01:03 | 2                    | 0.004 | 9                        | 0.017 | 0.24  | 0.05-1.14   | 0.07            | 4                        | 0.010 | 0.41   | 0.08-2.25   | 0.42            |
| DQB1*03:01~DRB1*04:01~DRB4*01:03 | 1                    | 0.002 | 4                        | 0.008 | 0.28  | 0.03-2.49   | 0.38            | 6                        | 0.015 | 0.14   | 0.02-1.13   | 0.05            |
| DQB1*03:03~DRB1*07:01~DRB4*01:03 | 4                    | 0.008 | 8                        | 0.015 | 0.55  | 0.17-1.85   | 0.39            | 12                       | 0.031 | 0.27   | 0.09-0.84   | <b>0.02</b>     |
| DQB1*02:02~DRB1*07:01~DRB4*01:01 | 55                   | 0.116 | 56                       | 0.106 | 1.11  | 0.75-1.64   | 0.69            | 44                       | 0.112 | 1.03   | 0.69-1.57   | 0.91            |
| DQB1*03:02~DRB1*04:04~DRB4*01:03 | 10                   | 0.021 | 12                       | 0.023 | 0.93  | 0.40-2.16   | 1.00            | 14                       | 0.036 | 0.58   | 0.25-1.32   | 0.22            |

Main HLA class II haplotypes associated with risk (OR>1) or protection (OR<1) to MS susceptibility are depicted. OR, Odds ratio; 95% CI, 95% confidence interval. P values derived from a two-tailed Fischer's exact test. A P value of 0.05 ( $\alpha$ ) or less was considered statistically significant (in bold). \*Haldane-Anscombe correction was applied in those situations with zero counts (i.e. just adding 0.5 to each of the cells and then calculate the odds ratio (OR) over these adjusted cell counts).

**Table R-17. Study 1&2: Significant extended HLA haplotype-level (at the 2-field resolution) associations with multiple sclerosis (MS) in the nominal analysis of the Spanish population cohort (HC 2n=564; MS 2n=476) and in comparison to the second healthy control (HC) group (Controls from Catalonia (Cat), 2n=392).**

|         |         |         |            |            |              |                | Spain MS HLA - Study 1 |       | Catalonia MS HLA - Study 2 |       |
|---------|---------|---------|------------|------------|--------------|----------------|------------------------|-------|----------------------------|-------|
| HLA-A   | HLA-B   | HLA-C   | HLA-DQB1   | HLA-DRB1   | HLA-DRB3/4/5 | HF in MS (Cat) | HF in HC (ESP)         | OR    | HF in HC (Cat)             | OR    |
| A*03:01 | B*07:02 | C*07:02 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0567         | 0.0226                 | 2.60  | 0.0204                     | 2.89  |
| A*25:01 | B*18:01 | C*12:03 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0105         | 0.0075                 | 1.40  | 0.0000                     | 8.31* |
| A*02:01 | B*14:01 | C*08:02 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0042         | 0.0000                 | 4.47* | 0.0000                     | 3.30* |
| A*02:01 | B*40:01 | C*03:04 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0042         | 0.0019                 | 2.23  | 0.0000                     | 3.30* |
| A*11:01 | B*51:01 | C*15:02 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0042         | 0.0019                 | 2.23  | 0.0026                     | 1.65  |
| A*29:02 | B*44:03 | C*16:01 | DQB1*06:02 | DRB1*15:01 | DRB5*01:01   | 0.0042         | 0.0019                 | 2.23  | 0.0051                     | 0.82  |
| A*02:01 | B*07:02 | C*07:02 | DQB1*06:03 | DRB1*15:01 | DRB5*01:01   | 0.0021         | 0.0019                 | 1.11  | 0.0000                     | 1.65* |
| A*02:01 | B*44:02 | C*05:01 | DQB1*06:03 | DRB1*15:01 | DRB5*01:01   | 0.0021         | 0.0019                 | 1.11  | 0.0000                     | 1.65* |
| A*11:01 | B*52:01 | C*12:02 | DQB1*06:01 | DRB1*15:02 | DRB5*01:02   | 0.0000         | 0.0057                 | 0.18* | 0.0000                     | 0.82* |
| A*29:02 | B*44:02 | C*05:01 | DQB1*03:01 | DRB1*04:01 | DRB4*01:03   | 0.0000         | 0.0038                 | 0.28* | 0.0000                     | 0.82* |
| A*02:01 | B*44:03 | C*16:02 | DQB1*03:02 | DRB1*04:02 | DRB4*01:03   | 0.0021         | 0.0038                 | 0.56  | 0.0026                     | 0.82  |
| A*02:01 | B*40:01 | C*03:04 | DQB1*03:02 | DRB1*04:04 | DRB4*01:03   | 0.0000         | 0.0086                 | 0.11* | 0.0000                     | 0.82* |
| A*33:03 | B*58:01 | C*03:02 | DQB1*02:01 | DRB1*03:01 | DRB3*02:02   | 0.0000         | 0.0019                 | 0.56* | 0.0000                     | 0.82* |
| A*02:01 | B*38:01 | C*12:03 | DQB1*06:03 | DRB1*13:01 | DRB3*01:01   | 0.0021         | 0.0038                 | 0.56  | 0.0000                     | 1.65* |

Main extended HLA haplotypes associated with risk (OR>1) or protection (OR<1) to MS susceptibility are depicted. OR, Odds ratio.

\*Haldane-Anscombe correction was applied in those situations with zero counts (i.e. just adding 0.5 to each of the cells and then calculate the odds ratio (OR) over these adjusted cell counts)

biological relevance (and respective LD patterns associated to it) found at the HLA genomic level.

On the other hand (see also **Table R-16**), when comparing HLA class II haplotype blocks results obtained between **Study 1** and **Study 2** (even though with the limitation that *HLA-DQA1* locus was not characterized in this second case), no confounding factors (due to population stratification), which may affect the results of the study, appear to be detected. The only minor exceptions are observed in relation to *HLA-DQB1\*06:03~DRB1\*15:01~DRB5\*01:01* (conferring risk) and *HLA-DQB1\*03:03~DRB1\*07:01~DRB4\*01:03* (conferring protection) haplotypes, which seem to be statistically more significant in their corresponding associations in **Study 2**. Nonetheless, since their haplotype frequencies are, overall, very low in both the present Spanish and Catalan population cohorts (presenting still modest sample sizes respectively), interpretation of these results will need to be further investigated and confirmed in future larger studies.

-Finally, relative to the analysis in the context of the corresponding completely extended HLA class I and class II haplotype blocks at the 2-field of resolution (see **Table R-17**), similar conclusions can be taken as from the previous analysis based on 3-/4-field data as part of **Study 1**, replicating also those results in associated risk/protection with very similar OR values shown here. Whereas comparison of **Study 1** and **Study 2** OR results may have revealed certain possible existing population substructure (although not significant) reflected in the different level of association observed to MS susceptibility between these two studies, though being only the case for some of the haplotypes shown here:

-In relation to risk: where those risk extended *HLA-DRB1\*15:01~DQB1\*06:02*-bearing haplotypes containing, respectively, *HLA-B\*18:01* (much stronger risk effect detected in

**Study 2** than in **Study 1**) and *HLA-B\*44:03* (much weaker risk effect found in **Study 2** versus **Study 1**) show opposite trends of association to MS susceptibility between these two studies.

-Regarding protection: where the protective effect (still attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01*) of all those associated extended haplotypes may not be as strong (i.e. ORs higher in **Study 2** versus **Study 1**).

Therefore, when evaluating association to MS susceptibility relative to completely extended HLA class I and class II haplotype blocks (here observed at least at the 2-field of resolution) this level of analyses could be more sensitive to possible existing underlying population substructure (even though no significant) in the given study cohort of interest. Yet, since haplotype frequencies are, overall, very low in both Spanish and Catalan population cohorts (presenting still modest sample sizes respectively) of the present study, interpretation of these results will need to be further corroborated in future larger studies.





## ***DISCUSSION***





## I. NGS-BASED HLA STUDY IN 17TH-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)

In the present study, HLA allelic sequences of 11 major classical HLA genes were characterized (according to the IPD-IMGT/HLA released version 3.25.0 (July 2016), available at the moment of the study) with extensive HLA genomic sequence coverage and phased-alleles with minimum allelic and heterozygous ambiguity per tested locus at the 3-/4-field for a fairly representative Spanish population cohort (N=282) by applying this novel high-throughput and high-resolution NGS-based HLA typing method [187][763-766]. Also, allelic and haplotypic HLA frequency distributions were examined at the 3-/4-field allele resolution level in this Spanish population cohort (denominated as 17th-IHIW Spanish population cohort as a healthy control reference group) [269].

The exceptional advantage offered by NGS technology (over all previous legacy HLA molecular typing methods) for HLA allele characterization is the ability to produce (almost) unambiguous allele level genotypes from full-length and (mostly) phased nucleotide sequences. The technique used here represents the current industry standard for these NGS-based HLA genotyping commercially available methods/kits for both clinical and research applications. Most of these current NGS approaches, as the one used here, enable full-length (mainly in the case of HLA class I loci) or almost full-length (mainly in the case of HLA class II loci) gene typing for HLA loci: *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPBI* using amplification with long-range PCR and shotgun deep-sequencing approach relying on short-read sequencing platforms (and particularly those that present a paired-end sequencing mode) [554]. At the same time, and as previously described in detail in the **INTRODUCTION** section, it is important to underscore that unknown level of underlying HLA diversity at the 4-field remains still unrevealed due to (among other factors): limitations of current

NGS-based (both 2<sup>nd</sup> and 3<sup>rd</sup> generations) HLA genotyping approaches (including the one NGS method used in the present study) [152][158][161][184][202][204][296][368] as well as current IPD-IMGT/HLA database limitations (i.e. where only 10% of known HLA class I and class II alleles have been fully sequenced (full genomic sequences) [146][463]). Yet, the introduction of NGS-based approaches (both 2<sup>nd</sup> and 3<sup>rd</sup> generations of sequencing technologies) for HLA genomic characterization has enabled to start analyzing in-depth (at both very high allele resolution (4-field) level and in terms of high-throughput mode with high number of HLA genes interrogated at a large sample scale) the almost full-length of coding and non-coding sequence regions of this very complex and highly polymorphic human genomic region. Which had not been possible for many decades until very recently (~since 2010 year) due to the many limitations shown by legacy HLA genotyping methods (e.g. SSP, SSO, RT-PCR and SBT) [137].

### **1. HLA ALLELE LEVEL ANALYSES**

At the HLA allele level, NGS-based HLA genotyping data at the 3- and 4-field allele resolution allows a further and in-depth description of the HLA allelic diversity given by silent substitutions and non-coding segments.

Focusing on *HLA* class I loci, it can be observed that *HLA-B* locus presents the highest allele diversity in comparison to *HLA-A* and *-C* loci in relation to the number of distinct alleles (*k*) (i.e. when looking at the collapsed 2-field allele resolution level, which defines a given specific HLA allele with a unique protein sequence) found in this Spanish population cohort. Nevertheless, the 4-field allele resolution level has allowed us (as it can be also generally observed in other NGS HLA population-level studies [267-287]) to reveal a significant diversity at the nucleotide level for *HLA-A* and *-C* loci in contrast to *HLA-B* locus (in which, in general terms, less 4-field allelic variants are found for a given *HLA-B\*XX:XX:XX* generic allele group). This can be illustrated, for

instance, when comparing the respective  $k$  (number of distinct alleles per locus) values in each case of allele resolution level (collapsed 2-field versus 4-field):

- *HLA-A* locus [collapsed 2-field ( $k=28$ ) versus 4-field ( $k=36$ )];
- *HLA-C* locus [collapsed 2-field ( $k=28$ ) versus 4-field ( $k=40$ )];
- *HLA-B* locus [collapsed 2-field ( $k=48$ ) versus 4-field ( $k=53$ )].

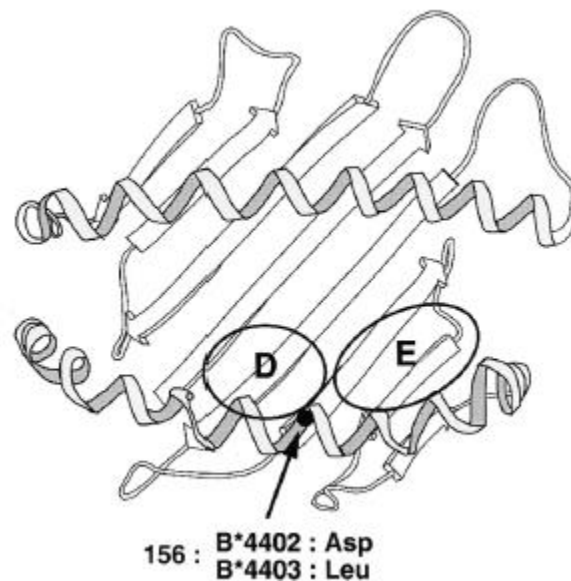
Thus, this may suggest plausible existing distinct functional roles and modes of evolution between these classical HLA class I loci as previously discussed in the literature [104][809].

In relation to HLA class II loci, both *HLA-DPA1* (heterozygosity index at 4-field = 0.84; whereas heterozygosity index at collapsed 2-field = 0.35) and *HLA-DQA1* (heterozygosity index at 4-field = 0.91; whereas heterozygosity index at collapsed 2-field = 0.88) loci exemplify the higher level of heterozygosity found at the 4-field level (molecular variation in non-coding regions in both introns and 5'/3'UTR regions) in comparison to the 2-field level (specific HLA protein-coding alleles). For instance, the only observed variant *HLA-DQA1\*05:01* at the collapsed 2-field level shows, in contrast, a total of three different variants at the 4-field level (*HLA-DQA1\*05:01:01:01*, *HLA-DQA1\*05:01:01:02* and *HLA-DQA1\*05:01:01:03*; representing 39.7%, 53.4% and 6.8% respectively inside this *HLA-DQA1\*05:01* allele group). Conversely, it can be also noted how certain loci (especially focusing our attention now on the HLA class II “B” (encoding beta chain) genes counterpart) such as *HLA-DPB1* (estimated homozygosity  $F=0.184$  at collapsed 2-field in contrast to  $F=0.177$  at 4-field), *HLA-DQB1* ( $F=0.110$  at collapsed 2-field in contrast to  $F=0.087$  at 4-field) and *HLA-DRB1* ( $F=0.074$  at collapsed 2-field in contrast to  $F=0.073$  at 4-field) loci show less differences regarding the allelic diversity found between the collapsed 2-field level variants and the 4-field level variants as described in the present study and likewise in other NGS HLA population-level studies [267-287]. In this respect (and as previously and more extensively commented in the **INTRODUCTION** section, ref. [27][35][56]

[104][178][268][273][288][301][309][428]), based on these observed differences of relative diversity either at the protein- or nucleotide level respectively, one could conjecture that classical HLA class II “A” (*DPAI*, *DQAI*, *DRA*) loci, despite worldwide human population diversity, appear to be under high selective pressure in which synonymous and non-coding allele variants (which, indeed, may be involved in defining regulatory functions related to cell surface expression and/or the stability of the respective peptide binding groove of these HLA class II heterodimers) are predominantly generated over protein-coding allele variants. Thus, from a functional perspective, allele diversification at the protein level (defining contact positions of these alpha subunits for the binding with the respective beta subunits forming heterodimer molecules) of HLA class II “A” loci may be restrained and relatively conserved (similarly to the case of the non-polymorphic  $\beta 2$  microglobulin associated to the given heavy polymorphic  $\alpha$  chain encoded by *HLA-A*, *-B*, *-C* class I genes respectively, establishing classical MHC class I molecules) in order to facilitate and ensure the pairing with the respective wide allele range of classical HLA class II “B” loci (*DPBI*, *DQBI*, *DRB1/3/4/5*), which, in turn, define extensive peptide binding repertoire that is intimately related with their key immunobiological role for antigen presentation on these HLA class II heterodimers [810]. At the same time, certain specific and mutually exclusive pairing patterns have been described in this heterodimerization process, thus defining differentiated allelic pair groups. As an example, in the case of *HLA-DQAI* and *-DQBI* alleles two distinct allele pair groups forming heterodimers have been described, where these different groups also appear to have divergent evolutionary origins and sequences [810].

Also, within HLA class I alleles described in the present study and as a very distinctive and characteristic HLA allele “signature” or “feature” found in Spanish population, it needs to be remarked a very intriguing relative HLA allele frequency distribution and respective ratio found within the *HLA-B\*44* allotype group (also known as *HLA-B\*44* supertype). In particular, regarding

both *HLA-B\*44:02* and *HLA-B\*44:03* subtypes, which only differ by one single residue (defined as a micropolymorphism) located on the  $\alpha 2$  helix of the HLA-B molecule: *HLA-B\*44:02* Asp (codon 156 (GAC), exon 3) and, in contrast, *HLA-B\*44:03* Leu (codon 156 (CTG), exon 3) [811]. It has been widely reported how these two *HLA-B\*44* subtypes induce strong T cell responses [811][812] and, thus, how critical it is their functional impact considering mismatches for clinical transplantation [813][814]. In addition, it has been described that this HLA polymorphism found inside this *HLA-B\*44* allotype group critically influences on TCR recognition, and, furthermore, specifically being involved and defining related peptide specificity based on peptide-binding preferential engagement events, where several molecular characteristics (presented by both HLA molecules and the corresponding antigenic peptide to be loaded) play a key role (e.g. even including the given antigenic peptide flexibility, for its accommodation within a certain HLA molecule peptide-binding groove, can be a major parameter) [815].



**Figure D-1.** Schematic representation of the structure of HLA-B44 binding-peptide groove. The positions of polymorphic residue 156 (small black circle indicated by the black arrow) and pockets D and E are indicated. Figure and respective footnote are obtained and adapted from [816].

Furthermore, more recent studies evaluating the influence of position 156 polymorphisms on both the requirement of tapasin for efficient surface expression of different HLA-B44 subtypes and their peptide nature and binding features have shown that B\*44/Pos.156 variants are highly tapasin-dependent (where tapasin chaperoning seems to be especially needed to acquire peptides of unusual length) [817]. Indeed, also recent studies have postulated that tapasin may modulate MHC I plasticity by dynamically coupling the peptide binding region and  $\alpha 3$  domain of MHC I allosterically, resulting in enhanced peptide selector function [818].

In relation to observed HLA allele frequency distributions at the *HLA-B* locus, these two *HLA-B\*44:02* and *HLA-B\*44:03* subtypes respectively present relatively high frequencies (thus, being generally defined as common allele groups) in most human worldwide populations (and across a wide range of different ethnic backgrounds) so far studied, and are followed (among other *HLA-B\*44* subtypes) by the less frequent *HLA-B\*44:05*, *HLA-B\*44:04* and *HLA-B\*44:06* subtypes (e.g. [130][221][223-227][259][260][267-287][297][299][328][339][464][466][467][474][476][481][496][547][558-561][563][564][571][600][602][603][608][611][614][621][624-630][772][808][819-851][943]). Interestingly, and at the same time, it has been also widely reported a high diversity found in the *HLA-B\*44:02* / *HLA-B\*44:03* ratio of allele frequency distributions among many human populations even within those considered broadly common ethnic groups (e.g. broad population group that comprises those populations of European/Caucasoid ancestry). Thus, it has been observed that *HLA-B\*44:02* subtype appears more prevalent than *HLA-B\*44:03* subtype in the majority of North, Central and East European as well as European American (from North America) populations (e.g. [223][225-227][259][268][474][476][943] (including Polish, Russian, Romanian, Kazakhstani, Bosnian-Herzegovinian, Greek and Turkish ethnic groups in German HSCT donor registry) [481][819-830]), and in addition to some Near-Eastern (including both Jewish and Arabs population groups

of the Levant [466][614][621][832]) and few reported Middle-Eastern [831] populations. Whereas, conversely, Mediterranean and South European populations (e.g. [221][260][269][545][546][558-561][563][564][571][600][602][603][608][624-630][833-835]), in addition to both North and sub-Saharan African and African descent (in which, as an example in some cases, even *HLA-B\*44:10* is one of the most common alleles found [273]) (e.g. [272][275][276][821][496][836-838]), Asian (e.g. [270][278][839-845]), Hispanic (people originally or descent from Central and South America) (e.g. [224][267][277][547][772][821]) and some populations from the Arabian Peninsula [271][285][821], exhibit an inverted ratio where *HLA-B\*44:03* subtype is more notably frequent than its counterpart the *HLA-B\*44:02* subtype. Therefore, in comparison to many other populations of European ancestry, the most striking fact is this stark difference uniquely found in Spanish population, as originally reported in [834][835], where the *HLA-B\*44:03* / *HLA-B\*44:02* is almost 2:1 ratio. In fact, for rest of neighboring/nearby European countries in relation to Spain:

- i) Mainland Portugal [602], also located in the Iberian Peninsula, shows this same clear *HLA-B\*44:03* / *HLA-B\*44:02* predominant ratio as Spain being almost the only two exceptions within the European continent.
- ii) Whereas in the French Bone Marrow Donor Registry (FBMDR) population, frequencies of these two *HLA-B\*44* subtypes are almost evenly distributed with a ~1:1 ratio (with *HLA-B\*44:03* subtype frequency being slightly higher) [846];
- iii) Interestingly, in the case of Italy there are conflicting results. On one hand, some HLA studies in several Italian population cohorts [297][833] show that *HLA-B\*44:03* is comparatively more prevalent. Nevertheless, on the other hand, the most recent and largest study from the Italian Bone Marrow Donor Registry (IBMDR) population shows a clear prevalence of *HLA-B\*44:02* instead [847]. These contradictory observations may be explained



by the fact of the existing, and possibly significant, HLA regional variation (observed North-South gradient) found within Italy, as previously described [848], and/or in addition to a plausible predominance of Italian donors, in this given national registry, who may be more closely related (people originally or descent from) with North/Central European populations (that also present this prevalence of *HLA-B\*44:02*), as also observed, for instance, in Italian American population attending to this *HLA-B\*44:03 / HLA-B\*44:02* ratio [772].

In this regard, it is also noticeable the observed (and still remaining, despite population admixture events over time [852][853]) Spanish genetic influence in Hispanic populations not only based on the similarities of the most common HLA haplotypes found between both these populations (as it is later commented with more detail in the respective **DISCUSSION** sub-section about haplotype analyses, including respective *HLA-B\*44:02* and *-B\*44:03* carrying haplotypes and their frequency distributions at 2- and 4-field in the present Spanish population and related foreign populations), but also according to this characteristic and singular *HLA-B\*44:03/HLA-B\*44:02* allele frequency ratio [221][224][260][267][269][277][545-547][558-561][563][564][571][600][602][603][608][624-630][772][821][834][835].

Also here, it was further investigated, as part of the present thesis work and by doing a thorough revision of the currently available scientific literature (<https://www.ncbi.nlm.nih.gov/pubmed/> and AFND [130][464]), an initial postulate suggested, for instance, by Santos et al. [834]. Where it was evaluated whether the *HLA-B\*44:03 / HLA-B\*44:02* ratio presented by those neighboring North African Arab populations (Moroccan [611-613], Algerian [854], Tunisian [850][851][855][856] and Libyan [849]), that are found geographically close to Spain, did follow or not this same ratio pattern found in Spanish population (and, indeed, in the entire Iberian Peninsula [602]). After reviewing HLA North African population studies in the literature (even though high-resolution HLA typing data reported for these populations was limited; for instance,

there was not 2-field *HLA-B* locus resolved typing data available for any Algerian population cohort [130][464][854]), a clear predominance of *HLA-B\*44:03* over *HLA-B\*44:02* has been reported in Cyrenaica population from Libya [849] and Moroccan populations (showing a significant genetic substrate of Berber ethnic group) [611-613]. Whereas studies on Tunisian general population group have shown that it does not seem to follow this same trend (i.e. predominance of *HLA-B\*44:03*). Nevertheless, it has not been fully investigated yet at the required high-resolution level in Tunisian population groups [851]. In addition, Tunisian population groups may present a highly complex genetic diversity (in the context of a complex demographic history of migrations from some regions of Africa, Europe, and the Middle East (particularly from the Arabian Peninsula)) [857-859] with a remarkable regional genetic variation (establishing also here a North-South gradient, where ancient Berber component is relatively more substantial in the North and Center regions than in the South) as previously reported examining other genetic markers (*Alu*/STR) [860]. Based on reported HLA allele/haplotype frequency distributions, some studies focused on Southern Tunisian population groups [850][856] described a *HLA-B\*44:02* predominance over its counterpart *HLA-B\*44:03*. In contrast, an earlier study carried out in Tunis population (located in Northern region of Tunisia) did show that *HLA-B\*44:03* was the most common subtype [855]. At the same time, and in consonance with a key demographic historical event such as the one defined by the Islamic conquest period along the North African region between 7th–9th centuries, Arabian Peninsula populations also show this predominance of *HLA-B\*44:03* over *HLA-B\*44:02* subtype distribution [271][285][821]. In summary, from the point of view of this *HLA-B\*44:03* / *HLA-B\*44:02* allele frequency ratio, these results additionally confirm the observed strong genetic influence (and thus a relative level of shared genetic substrate and relatedness) of North African Berbers (autochthonous) and Muslim Arabs (settlers originally from the Arabian Peninsula) population ancestries on modern-day Spanish (and, indeed Iberian) gene

pool population as it has been also widely described by different genetic markers, including HLA genes, [563-566][602][611-613][620][808][851][861] and also extensively supported by well-documented historical and demographic facts [555][556][558][578][613][808][851]. In fact, two major historical demographic events may have critically contributed to this observed considerable genetic (including HLA) relatedness between Iberians and North Africans: i) a main pre-Neolithic contribution (i.e. involving gene exchange) from northward Saharan migration, which occurred in 10,000–4,000 BC, when the Berbers relocated to the Northern Mediterranean coast during hyper-arid conditions [563]; ii) both Iberian and North African territories were similarly invaded by Phoenicians, Romans, Germans (Visigoths in Iberia, Vandals in North Africa) and Muslim Eastern Arabs (where these latter may have contributed with a lower but conserved gene flow (due to low admixture rate but establishing a very long period of settlement (8-10 centuries), in addition to a deep social and cultural imprint in both Iberian and North African population groups) [808][851]. Nonetheless, future larger-scale studies across North African and Muslim Eastern Arab (Arabian Peninsula) populations have the potential to increase our understanding of the historical demographic factors influencing the region [858].

At the 4-field allele level, in relation to non-coding (i.e. intron/untranslated regions) variation found within these two *HLA-B\*44* subtypes, it is noteworthy how NGS-based HLA genotyping approach has enabled to reveal this significant underlying level of allelic diversity. In the present Spanish population study, *HLA-B\*44:02:01:01* (AF=5.9%), *HLA-B\*44:03:01:01* (AF=8.5%); in addition to *HLA-B\*44:05:01* (AF=0.2%) and *HLA-B\*44:04* (AF=0.2%) allele variants were found. Moreover, in comparison to other NGS 4-field HLA population studies of very large sample size, and just as some examples:

- In the Argentinian registry population study (N=1,472) [224], detected *HLA-B\*44:02* variants were: *HLA-B\*44:02:01:01* (AF=4.35%), *HLA-B\*44:02:01:03* (AF=0.24%), *HLA-B\*44:02:06*

(AF=0.07%) and *HLA-B\*44:02:46* (AF=0.03%)); whereas *HLA-B\*44:03* variants included: *HLA-B\*44:03:01:01* (AF=5.81%), *HLA-B\*44:03:01:02* (AF=0.03%), *HLA-B\*44:03:01:09* (AF=0.03%) and *HLA-B\*44:03:02* (AF=0.17%). In this respect, several similarities can be observed with the Spanish population cohort in addition to other 4-field *HLA-B\*44:03* and *HLA-B\*44:02* variants that may be presumably more frequent in Amerindian related ethnic groups.

- In a recently described large European American U.S. cohort (N=2,248) [268], detected *HLA-B\*44:02* variants were: *HLA-B\*44:02:01:01* (AF=7.03%) and *HLA-B\*44:02:01:03* (AF=0.86%); whereas *HLA-B\*44:03* variants included: *HLA-B\*44:03:01:01* (AF=4.84%) and *HLA-B\*44:03:02* (AF=0.06%). Here, this European American population group may represent the prototypic North-Central European substrate in regards to this *HLA-B\*44:02* / *HLA-B\*44:03* allele frequency ratio.

Interestingly, this much higher level of 4-field allele diversity found in this pair of *HLA-B\*44* allele groups for both Argentinian and European American population cohorts may reflect how these broad population groups from Latin America and the U.S. have been under constant and highly complex population admixture processes where different migration waves (with ebbs and flows of immigrants from all over the world especially during the last two-five centuries) have been shaping ancestral population proportions, which have been also fluctuating through time [852][853][862][863].

Overall, in accordance with other HLA Spanish population studies [221][260][545][546][558-561][563][564][571][600][608][624-630][834][835], the present NGS HLA Spanish population study [269] shows that *HLA-B\*44:03* subtype is predominant (almost twice as frequent) over *HLA-B\*44:02* subtype in Spanish population. In addition, very infrequent (thus, so far, considered rare alleles) *HLA-B\*44:05* and *HLA-B\*44:04* subtypes are also found, whereas other subtypes such as

*HLA-B\*44:06*, *-B\*44:07* or *-B\*44:10* are almost or apparently absent in Spanish population. At the allelic level, the mutual prevalence of this pair of subtypes *HLA-B\*44:02* and *HLA-B\*44:03* as well as their respective differences in their allele frequency distribution among diverse worldwide human populations make these two allele groups very singular. Based on findings from previous studies [811][815], a combination of natural selection mechanisms (e.g. differential selection of peptide ligands, differential peptide flexibility determined by the respective peptide-MHC molecule complex (pMHC) structure and, consequently, differential determination of T cell repertoire at different possible stages (as thymic positive selection and/or peripheral clonal diversification during immune response)) may explain the functional basis of the maintenance of this pair of allele groups, which appear to be independent in relation to the selective advantage in immunity, among worldwide populations through time. At the same time, distinguishable *HLA-B\*44:02* / *HLA-B\*44:03* ratios (with either predominance of *HLA-B\*44:03*, or predominance of *HLA-B\*44:02* or both subtypes being evenly distributed in a certain group of populations) are found in different groups of populations (even between populations within the same considered broad ethnic group). This may be also indicating the effect of natural selection advantages and/or differential demographic events in populations throughout history (e.g. a population bottleneck, population expansions, migration waves (ebbs and flows), population admixture or founder effect). In the case of Iberian (Spanish and Portuguese) populations, and as far as our knowledge (and based on currently available scientific literature), it still needs to be further investigated the specific series of selective pressure and/or demographic events (e.g. both North African Berber and Muslim Eastern Arab genetic influence) which may have determined this singular predominance of *HLA-B\*44:03* over *HLA-B\*44:02* in comparison to the vast majority of European populations. An additional and very interesting consideration that has been also commented in previous studies [819][823][864], it is how linkage disequilibrium, involving both MHC-linked HLA and non-HLA

polymorphisms that segregate with different *HLA-B\*44* haplotypes, may have also contributed on sustaining the selective advantage that seems to be associated/represented within this pair of *HLA-B\*44* subtypes. However, this supposition of plausible influence by other elements of conserved *HLA-B\*44* haplotypes seems less likely since these HLA and non-HLA polymorphisms are apparently not well-preserved (well-defined) within the given haplotypes described so far in different human populations.

As an additional striking observation also regarding to *HLA-B* locus, the *HLA-B\*73:01* allele (AF=0.2%) was found in the present Spanish population cohort. As a matter of fact, this exceptionally diverged (i.e. structurally divergent from other *HLA-B* alleles) *HLA-B\*73:01* allele [865], despite it has been more recently debated [866], has been presumably identified (based on simulations of whole genome comparisons which showed that introgression from archaic hominins provides a better fit to the data than a model in which the allele arose in Africa before the out-of-Africa event) as a representative example of how certain HLA alleles in present-day human populations would have been originally acquired by admixture with archaic humans (i.e. the so-called adaptive introgression of archaic HLA alleles originally from primitive ancient hominids) [867]. In this specific case it has been postulated that *HLA-B\*73:01* allele would have been introgressed from Denisovans into early modern humans (thus adaptive introgression of archaic alleles may have been significantly shaped modern human immune systems) [104][867]. In modern populations, *HLA-B\*73:01* allele has been found in relatively high frequency (AF=1.7%) in Near Eastern regions such as in Lebanese population [104][132][621], in which *HLA-B\*73:01* allele associates with *HLA-C\*15:05:01* and *-C\*12:02:02* alleles. In the present Spanish population cohort, it was observed that *HLA-B\*73:01* allele displays a specific association with *HLA-C\*15:05:01*.

Moreover, from an epidemiological standpoint and as previously commented, large NGS-based HLA studies can be very informative [121] on populations presenting a high correlated prevalence of certain HLA allele/haplotype subtypes (which have been also found to be ethnic-specific in many instances) associated with a given infectious/autoimmune disease [109][513] or drug-induced hypersensitivity reaction [126][127]. In Spanish population, common allele groups such as *HLA-B27* (*HLA-B\*27:05:02* (AF=3.1%); *HLA-B\*27:02:01* (AF=0.4%); *HLA-B\*27:03* (AF=0.2%) in the present study [269]) and *HLA-B51* (*HLA-B\*51:01:01:01* (AF=6.2%); *HLA-B\*51:08:01* (AF=0.2%) in the present study [269]) have been previously investigated (mostly based on single-locus screening studies) due to their very well-documented associations with the clinical entities of spondyloarthritis (SpA) [123] and Behçet's disease (BD) [633] respectively. However, despite previous studies have described possible related specific polymorphisms located within the MHC region between HLA and non-HLA genes (intrinsically related with inflammation and the immune response) in the disease phenotype and severity, it has not been well-defined yet both the impact of diversity found in these HLA-associated allele subtypes and the plausible implication of the existing diversity found at the bearing extended haplotypes. Furthermore, many of these differential polymorphisms may play a pivotal role mechanistically speaking. In this sense, several hypotheses of the involvement of these respective HLA allele subtypes have been proposed, for example: different or common specificities of peptide binding; correspondence between HLA protein folding and assembly features and association with disease phenotype; or epistatic interactions with neighboring genes. Thus, allele and haplotype frequency distributions data shown in the present thesis work may be useful as a reference source for future epidemiology studies and, thus, to screen more accurately the prevalence of these HLA allele variants and their main LD patterns within the HLA system in both healthy controls and ethnically matched patients in Spain.

Out of the HLA class II alleles, it is very outstanding the high *HLA-DPBI\*04:01:01:01* allele frequency observed in the present Spanish population (AF=36.1%) similarly to what have been described in most of worldwide human populations (e.g. [297]). In fact, a recent study in Japanese population has suggested that *HLA-DPBI\*04:01* allele (presenting high allele frequency and high haplotype homozygosity (HH) parameter values) has recently undergone very strong positive selection [868]. In this sense, this observation may exemplify how certain level of HLA diversity shown by some modern populations may have been founded recently, in part, due to major demographic events (e.g. a population bottleneck, population expansions, migration waves, population admixture or founder effect).

In comparison to legacy molecular HLA genotyping approaches (e.g. SSO, SSP, RT-PCR or even SBT), extensive genomic sequence coverage provided by NGS technology clearly facilitates (in a very timely and cost-effective manner) the almost unambiguous identification of novel alleles [172][179][211], possible null or expression variant alleles [210][458][523][869-871] (<http://hla.alleles.org/alleles/nulls.html>) and also the detection of rare alleles [208][209].

In regards to the two identified and confirmed new alleles (being both exon variants: *HLA-B\*38:20:02* and *HLA-DRB3\*02:71*) in this Spanish population cohort. These findings exemplify (even though our current Spanish population cohort did not have a large sample size (N=282 subjects)) how NGS-based HLA genotyping approach allows the almost thorough interrogation of the entire coding region of classical HLA class I and class II genes (extending beyond those exons that encode the antigen recognition domain (ARD): exons 2 and 3 for the HLA class I genes; and exon 2 for HLA class II) and, consequently, it enables the assessment of both non-synonymous (change of amino acid coding) and synonymous (maintaining the same amino acid residue) nucleotide exchange. Thus, being this information available, it is possible even to interpret and to speculate with more certainty possible molecular implications and effects of those given amino



acid changes on the HLA protein structure itself, antigen-presentation functionality, related peptide specificity, protein stability/integration into the cell membrane or interaction with respective T cell co-receptors [265][266][553].

On the other hand, it is important to remark that while current NGS methods together with their related HLA genotyping software programs are able to describe and report (e.g. [353][354]) the detection of coding (exon) variants with a relatively high level of accuracy [172][179][211], detection and characterization of non-coding variants as well as new alleles that are hybrids of two or more known alleles are still major challenges for the currently available bioinformatics tools and related algorithmic strategies [296]. This is because in many instances it is quite challenging and complex to do this type of assessment due to low or inadequate genomic sequence coverage, lack of phasing and, thus, related ambiguities found, especially, in some of these intronic and untranslated regions. At the same time, and even more importantly, analysis of the data generated by current NGS approaches (either using short-read or long-read platforms) still critically relies on accurate reference database in order to assess the quality of the raw NGS sequences (reads) generated and to perform their subsequent alignment and construction of phased consensus sequences per HLA allele tested [76][87][184][295][296][362]. Thus, the hyper-polymorphic nature of HLA system can represent a very considerable burden to accurately phase and implement sequencing analysis when a reference sequence is either unavailable (although it is expected that massive input of NGS data will be completing all unsequenced segments of HLA reference alleles of the IPD-IMGT/HLA database) or highly variable [87][146][295][362][463]. Where the main concern is that without precise phasing the inferred full-length sequences given by assembling fragments may be still not 100% accurate and reliable. Since assemblies (which, in turn, critically depend upon not just coverage but depth of coverage) produced, if they are based on HLA genomic regions that miss information or present low coverage in the reference database, may be of low

accuracy or incorrectly phased. Consequently, certain given SNPs that could be key in order to assemble full-length sequences may not be available from the database. In fact, it is known that the IPD-IMGT/HLA database [87][295][362] contains genomic sequences for most of the main serological groups for class I (*HLA-A*, *-B* and *-C*), however the coverage for HLA class II genes is much lower, which may affect the accuracy of any assembled HLA class II allele sequence [87][146][295][296][362][463].

Furthermore, despite current NGS-related HLA genotyping software programs may be able to report coding (exon) variants, the own user needs to consider the IPD-IMGT/HLA database released version being used at the time of analysis and how updated it is [87], in order to see if a further evaluation of the validity of this tentative novel allele detected is necessary and in addition to perform a second parallel sequencing test as well as meeting rest of requirements established for submission and official reporting of novel HLA alleles (<https://www.ebi.ac.uk/ipd/imgt/hla/subs/submit.html>). On this task of validating novel HLA coding variants (as it was also done for the present study), BLAST (Basic Local Alignment Search Tool) libraries (provided by European Bioinformatic Institute (EBI) and integrated in the IPD-IMGT/HLA Database) are commonly used [872]. This BLAST tool (<https://www.ebi.ac.uk/ipd/imgt/hla/blast.html>) searches against the nucleotide and protein sequences of HLA alleles and related sequences included in the database, although currently the nucleotide library only includes the coding sequences but not the non-coding reference sequences [87][295][362]. Lastly, and as also previously commented in [356], it should be noted that there are instances when current NGS-related HLA genotyping software programs (due to inherent limitations of the given NGS-based HLA targeting and related analysis software strategies [202][204][296] as well as due to a particular poor performance of either the sequencing or during DNA library preparation steps or even due to initial status of a given DNA sample) automatically

(and in some instances systematically and by default) report incorrect HLA alleles for a given locus or set of loci tested. In this context, unless the user does an exhaustive manual review of the automatic HLA genotype assignments generated by the respective NGS HLA software analysis program and user also interrogates the associated sequencing data, those reported HLA genotypes and alleles (especially at the 4-field) may be inconsistent and inaccurate. At the same time, the user also needs to be aware of that while many current HLA genotyping software programs may allow the user to review/edit automatically initial generated HLA genotype assignments, the respective consensus sequences may not be updated according to the final assigned/corrected HLA allele call made by the user (representing another important limitation when investigating associated sequencing data).

The identification of null alleles (in both donor and recipient) is clinically relevant, especially in the HSCT setting, since nonidentification (being misdiagnosed as normally expressed variants) or misidentification can lead to serious clinical complications for the recipient such as poor engraftment or GvHD [210][869]. Because the initially estimated prevalence of HLA null alleles may be around 0.3% or, most likely, even higher across worldwide human populations [210][523][871], the historical need for a consistent screening strategy for HLA null alleles has been just now covered by the recent implementation of NGS-based HLA genotyping methods in the clinical histocompatibility and immunogenetics (H&I) laboratories. By using this novel sequencing technology, the vast majority of HLA genomic regions are sequenced to ensure detection of all possible existing null alleles. In the present study, two distinct null (or non-expressed) alleles were found with relatively (within the given HLA locus and also in relation to other populations previously studied) intermediate-high frequency in this Spanish population cohort:

- *HLA-C\*04:09N* allele (AF=0.4%), initially described and characterized in [640][873] (Deletion; Exon 7, 1095delA, in codon 341, causes frameshift and loss of stop codon in exon 8, resulting in the peptide containing an additional 32 amino acids).
- And *HLA-DRB4\*01:03:01:02N* allele (AF=1.8%; considering also the particularity that the absence (represented as “*HLA-DRB4\*00:00*”) of *HLA-DRB4* locus accounts here for AF=70.2%; see **RESULTS** section), initially described and characterized in [874-876] (Point Mutation; this *HLA-DRB4\*01:03:01:02N* allele contains an aberrant splice site, located between the end of intron 1 and beginning of exon 2 segments, that causes incorrect splicing and it results in lack of protein sequence (e.g. no protein product has been detected serologically)).

Moreover, in comparison to other reported HLA genotyping population (mostly made up of unrelated subjects) datasets (which many of these correspond to BM/UCB donor registry populations) of both Spanish origin and of European ancestry, it can be observed for these particular two distinct null alleles the following:

- *HLA-C\*04:09N*: relatively similar but still smaller frequencies have been previously reported for this null allele in other comparatively larger cohorts within Spanish population (e.g. Catalonia (N=2,895; AF=0.14%) [221]), European populations (such as England (N=519; AF=0.2%); Germany (N~1,000,000; AF=0.05-0.14%); and the Netherlands (N=1,305; AF=0.1%)) [130][464][480][525][526][554][943], Argentinian registry population (N=1,472; AF=0.14%) [224] and an European American population cohort in the U.S (N=2,248; AF=0.14%) [268] and being AF=0.09% according to the NMDP registry database (presenting a 59.9% of European White component) [523], which is also in the estimated range (0.1–0.25%) of *HLA-C\*04:09N* allele frequency distribution originally calculated by Pinto et al. [877] for individuals of European American (Caucasoid) origin. Whereas, in other group of studies

mostly from Northern (Wales (N=8,412; AF=0.0654%)) [878] and Eastern (Hungary (N=7,345; AF=0.0136%)) [879]; Poland (N=23,595; AF=0.0021%)) [474]) regions of Europe this null *HLA-C* allele appears to be even much less frequent. In light of all of the above reported results, and as previously underscored in [879], it can be noted a plausible but considerable descending South-North and West-East gradient in the European continent in regards to the *HLA-C\*04:09N* allele frequency distribution. Once more, this exemplifies the high level of regional HLA diversity found within the European continent [136] and among populations of European ancestry (e.g. [297]). Lastly and most strikingly, this relatively high value of *HLA-C\*04:09N* allele frequency described in our current Spanish population cohort (despite being limited by its relatively small sample size (N=282)) may be explained by the solid and well-documented fact that generally Spanish population distinctly presents a high allele frequency distribution for the *HLA-B\*44:03* allele group which, in turn, displays a very tight association with this given null allele *HLA-C\*04:09N* (e.g. typical extended haplotype *HLA-A\*23:01~C\*04:09N~B\*44:03~DRB1\*07:01~DQB1\*02:02*) [130][297][464][877][880]. In the present Spanish population study, 2-locus *HLA-B~C* presents high D' value (D'=1) and intermediate high 2-locus (*HLA-B\*44:03:01:01~C\*04:09N*) frequency (HF=0.4%).

- *HLA-DRB4\*01:03:01:02N*: this HLA null allele appears to be relatively less frequent (especially in comparison to other populations of European ancestry; e.g. see [297]) in the present Spanish population cohort. However, it is important to take into consideration that its AF value (1.8%) shown here is in the context of considering not only allele frequency distributions of given *HLA-DRB4* alleles but also the frequency distribution of the absence of *HLA-DRB4* alleles according to the association/exclusion patterns described for respective *HLA-DRB1~HLA-DRB3/4/5* allele families [344]. Taking a closer look to the populations reported in [297], it can be observed how populations of Northern-Central European ancestry

show the highest allele frequency value (AF~3.4%-4.3%) for this *HLA-DRB4\*01:03:01:02N* null allele. Whereas populations of Spanish, Hispanic or Asian-Pacific Islanders origins show values around AF~1.2%-1.9% and, in turn, these are followed by Middle-Eastern (AF=1.0%) and African-descent (AF=0.4%) populations. At the same time, it should be noted that, in general, null HLA class II alleles have not been extensively described yet at the population-level (neither at the donor registry-level) due to the still existing limited genomic (even at the exon level and especially beyond describing ARD-coding exons) coverage offered by commercial/in-house HLA molecular typing methods (including many of the available NGS approaches) routinely used by clinical H&I laboratories. Nevertheless, some first significant efforts in some populations have been dedicated to describe the frequency distribution of this null allele as well as to define in which most frequent haplotypes this null allele is embedded. In this sense, as an example, a HLA Croatian population study (as a fairly good representation of Eastern European population) has revealed a significant relative allele frequency value for this null allele (6.35% in the context of among *HLA-DRB1\*04* positive samples; and 98.21% in the context of among *HLA-DRB1\*07:01~DQB1\*03:03* positive samples) [881]. Moreover, in the German donor registry Zentrales Knochenmarkspender-Register Deutschland (encompassing a broad HLA diversity mostly representative of Northern-Central European populations) [480][525][526][554], the allele frequency value found is also relatively noticeable (AF=3.6%). At the HLA haplotype level, this type of studies [297][881][882] have reported how *HLA-DRB4\*01:03:01:02N* null allele is most commonly (and almost exclusively) associated with the *HLA-DRB1\*07:01:01~DQA1\*02:01:01~DQB1\*03:03:02* class II haplotype group. Although it has been also described to be carried, but relatively less frequently, within HLA class II haplotypes containing mainly *HLA-DRB1\*04:02* (*HLA-*

*DRB4\*01:03:01:02N~DRB1\*04:02~DQB1\*03:02*) and also, but in a lesser extent, -  
*DRB4\*04:03* (*HLA-DRB4\*01:03:01:02N~DRB1\*04:03~DQB1\*03:02*) alleles [881].

In summary, since null alleles generally show relatively low frequency distributions, it is essential to carry out much larger NGS HLA population studies (at the highest allele resolution level, with extensive genomic sequence coverage and considering always the most updated IPD-IMGT/HLA database released version available at the time of the study) and at a wider geographic scale in order to assess more accurately the prevalence of HLA null alleles at the population-level and, thus, acquiring precise local population-specific HLA genotype data, that will definitely contribute to better determine their clinical relevance especially in the transplantation setting as well as improving significantly the donor search process and registry planning strategies (e.g. improved matching strategies and identifying suitably matched donors in a timely and a cost-effective manner) [869][881]. Moreover, as previously and extensively commented in the **INTRODUCTION** section, it will be also important to dedicate efforts on improving the assessment of expression variant alleles [e.g. including encoded soluble proteins (S); low expression levels (L); and questionable expression status (Q)] [458], where NGS-based HLA genotyping approaches may critically contribute to the in-depth description of regulatory non-coding variants defining HLA expression patterns. At the same time, this approach may need to be combined with other techniques (e.g. RNA-seq (applying also here novel NGS technology) or flow-cytometry) that may allow the evaluation of surface HLA expression differences which can also arise due to, among other factors, variations in HLA-bound peptide repertoires, modes of peptide binding, or certain level of specificity related to the respective loading-peptide intracellular machinery in a cell-type-dependent manner [450][883][884].

Previously considered rare alleles, *HLA-C\*12:166* (AF=0.2%), initially described and characterized in [885], and *HLA-B\*15:220* (AF=0.4%), initially described and characterized in

[886][887], were detected in the present Spanish population cohort in several instances. In the case of *HLA-B\*15:220*, this allele has been commonly found in African-descent/related populations [130][464]. Whereas, *HLA-C\*12:166* appears to be, as far as it has been reported up to date [130][464], a rare allele quite characteristic of Spanish population. As previously described in detail in the **INTRODUCTION** section, in-depth characterization of HLA nucleotide diversity by NGS technologies at the population-level and at a very broad geographic scale may allow an accurate designation (also identifying thus under- and overrepresentation situations) of rare and common alleles specific for each given population/region/ethnic group, being this information very invaluable for many clinical applications.

In relation to the latter and moving now to definition of common and well-documented (CWD) HLA alleles [147][148][149][150][479][480], as also previously described in recent NGS-related HLA population studies [137][146][224][300], the 4-field allele resolution level reveals how in certain allele groups, an allele considered rare initially it actually presents a common occurrence while the lowest numbered allele is not the most frequent. For instance, in *HLA-B* locus (e.g. *HLA-B\*35:01:01:01* allele represents only 3.6% of this allele group whereas *HLA-B\*35:01:01:02* allele represents 96.4% of this allele group found in this Spanish population cohort) and *HLA-DRB1* locus (e.g. *HLA-DRB1\*12:01:01:03* allele represents 100% of this allele group whereas *HLA-DRB1\*12:01:01:01* allele is absent in this Spanish population cohort). Thus, NGS technology enables the identification of possible several common alleles that are associated with the same HLA protein. In addition, application of NGS technology for HLA genotyping can significantly contribute for a better characterization of the entire genomic sequence in these critical CWD alleles for many relevant clinical and research purposes [146]. In light of these facts, NGS technology may contribute to create an updated and more comprehensive catalogue of CWD HLA alleles. In



addition, future larger NGS HLA Spanish population studies will allow a more accurate and updated definition of CWD and rare alleles specific for this population.

In comparison to all previously reported studies (that were available in the scientific literature (<https://www.ncbi.nlm.nih.gov/pubmed/>) at the moment of the present study) that have attempted the description of HLA allele and haplotype diversity in Spanish population (including both Spanish general population cohorts [221][260][546][624] and Spanish regional cohorts [545][558-561][563][564][571][600][608][625-630]). The present HLA Spanish population study (comprising a fairly representative cohort of 282 unrelated healthy subjects) signifies the very first evaluation ever done in Spanish population describing HLA allele and (extended) haplotype diversity with very minimum ambiguities (at a very high-resolution, resolving at the 3-/4-field allele resolution level) and with a very extensive level of HLA genomic characterization (i.e. inclusion of additional classical HLA class II genes not well-documented previously such as *HLA-DRB3/4/5*, *HLA-DQA1*, *HLA-DPA1* and *HLA-DPBI*; as well as typifying the full-length and/or spanning to the most possible extended sequence of all *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPBI*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci tested), which are based on full-length and (mostly) phased nucleotide sequences generated via NGS (using a short-read sequencing platform with a paired-end sequencing mode). Despite the current Spanish population cohort here shows a relatively discrete sample size (covering 10 different Spanish Autonomous Communities/Regions (out of a total of 17 in Spain) across the country (and more or less evenly distributed geographically) which are described by 25-26 individuals per region), application of NGS technology for HLA genotyping has enabled to still reveal a great level of HLA allelic nucleotide diversity within the Spanish population as well as to identify singular and specific 4-field extended haplotype associations (later commented with more detail in the respective **DISCUSSION** sub-section about haplotype analyses). In the next pages, there is a first discussion where observed HLA allele

frequency distributions for tested loci in the present Spanish population cohort are compared with respective HLA allele frequency distribution datasets previously reported by other main and considered here the most representative HLA studies in Spanish population previously reported [221][260][564][624]:

i) Firstly, referring to a more in-depth and detailed discussion on those observed HLA alleles, at this 3-/4-field allele resolution level described here, which have been found within each tested HLA locus for the present Spanish population cohort, as well as considering the aforementioned (see **INTRODUCTION** section) Balas et al. study [624] as a very comparable Spanish population (despite being made up of HSC transplant patients) reference HLA allele and haplotype (defined by family segregation analysis) frequency distributions dataset (being one (if not the most) of the most representative for Spanish population) in relation to our current study. Even though corresponding allele resolution level in Balas et al. HLA Spanish population (cohort made up of N=253 Spanish hematopoietic patients) study is referenced by older released versions (between 1998-2010 years) of IPD-IMGT/HLA database and it presents limited HLA genomic sequence coverage (especially in the case of HLA class II genes) for tested HLA loci *HLA-A*, *-B*, *-C*, *-DQB1*, *-DRB1* and *-DRB3/4/5* by using a SBT method [624]. In detail, main findings of this comparative analysis with Balas et al. study [624] are described as follows:

- For *HLA-A* locus, both the present 3-/4-field dataset of Spanish population and Balas et al. study [624] show the same set of 7 most frequent *HLA-A* alleles (*HLA-A\*02:01:01:01*, *-A\*01:01:01:01*, *-A\*03:01:01:01*, *-A\*11:01:01:01*, *-A\*29:02:01:01*, *-A\*24:02:01:01* and *-A\*32:01:01* with allele resolution level as described in the present study), and with similar allele frequency values for each respective allele (most of them higher or close to 5%) in Spanish population. Also in both studies this group of 7 *HLA-A* alleles represents 70-75% interval of all *HLA-A* alleles that have been described in respective Spanish population

cohorts. On the other hand, those considered rare *HLA-A* alleles (in this particular case, when considering those alleles that show values of allele frequency that are less than 1%) more differences than similarities can be observed in relation to which specific rare *HLA-A* alleles are found in each of these two cohorts, despite both studies show a same number of rare alleles at this locus (18 alleles). One of the most plausible explanations of these differences could be due to the present Spanish population study included samples from some Spanish regions (such as Salamanca, Santander, Gran Canaria and some regions of Andalusia) that are not as well represented as in Balas et al. study [624] (where the vast majority of studied subjects were Spanish patients with indication for allogeneic HSCT who, most likely, were originally from Madrid and/or the central region of Spain where the respective HLA clinical H&I laboratory and transplantation institutions are physically located). Referring to the high allelic diversity (especially at the 4-field allele resolution level) revealed in the present Spanish population cohort by application of NGS technology for HLA genotyping. An illustrating example is the *HLA-A\*03:01* allele group that would have been described as a single variant by legacy HLA molecular typing methods. Whereas, in the present NGS HLA study, there are up to three variants (*HLA-A\*03:01:01:01*, *HLA-A\*03:01:01:03* and *HLA-A\*03:01:01:05*; representing 84.2%, 7.0% and 8.8% respectively inside this *HLA-A\*03:01* allele group).

- For *HLA-B* locus, similarly to *HLA-A* locus, the present Spanish population study and Balas et al. study [624] show the same set of 7 most frequent *HLA-B* alleles (*HLA-B\*07:02:01*, *-B\*44:03:01:01*, *-B\*08:01:01:01*, *-B\*51:01:01:01*, *-B\*44:02:01:01*, *-B\*35:01:01:02* and *-B\*18:01:01:01* with allele resolution level as described in the present study), which present an allele frequency value higher or close to 4%. Also in both studies this group of 7 *HLA-B* alleles represents 45-60% range of all defined *HLA-B* alleles in Spanish population. In contrast, within the respective groups that comprise those considered *HLA-B* rare alleles

(presenting less than 1% of allele frequency value) more variability is found when comparing these two study cohorts due to the same aforementioned explanation for *HLA-A* locus (i.e. as there are some differences in which Spanish regions were sampled in each study). In the case of *HLA-B* locus, an example of remarkable allelic diversity found at the 4-field in the present study is the *HLA-B\*18:01* allele group. Which presents just one variant (*HLA-B\*18:01*) at the 2-field whereas there are up to three different variants (*HLA-B\*18:01:01:01*, *HLA-B\*18:01:01:02* and *HLA-B\*18:01:01:03*; representing 56.4%, 41.0% and 2.6% respectively inside this *HLA-B\*18:01* allele group) found in Spanish population.

- In relation to *HLA-C* locus, similar observations can be made as those above mentioned for both *HLA-A* and *-B* loci, when comparing the 3-/4-field HLA dataset of present Spanish population study (that includes a Spanish population sample cohort made up of several 25-26 sample sets from main regions across the country) with Balas et al. study (which mostly covers a Spanish population cohort representative of the central region of mainland Spain) [624]. In this case, a similar set of 8 most frequent *HLA-C* variants (*HLA-C\*07:01:01:01*, *-C\*04:01:01:01*, *-C\*07:02:01:03*, *-C\*12:03:01:01*, *-C\*16:01:01:01*, *-C\*02:02:02:01*, *-C\*05:01:01:02* and *-C\*06:02:01:01* with allele resolution level as described in the present study) is found, which all show an allele frequency value higher or close to 5%. Nevertheless, it is quite striking that a much higher allele diversity is found comparatively at the 4-field in the present Spanish population study. For instance, *HLA-C\*05:01:01* (*HLA-C\*05:01:01:01* (AF=4.0%) and *HLA-C\*05:01:01:02* (AF=4.9%)) and *HLA-C\*08:02:01* allele groups (*HLA-C\*08:02:01:01* (AF=4.0%) and *HLA-C\*08:02:01:02* (AF=1.5%)) present a more diverse distribution at the non-coding region level in the present study in comparison to Balas et al. study (in which *HLA-C\*05:01:01* (8.9%) and *HLA-C\*08:02:01* (5.1%) are only described) [624]. This noted difference of allelic diversity found between these two studies can be

illustrated as, on one hand, the group of high frequency *HLA-C* alleles (showing an allele frequency value higher than 5%) accounts for 78% of all defined *HLA-C* variants in the Spanish population cohort from Balas et al. study [624]. Whereas the same group of high frequency alleles accounts only for 60% of all defined *HLA-C* variants in the case of the present HLA Spanish population study. Moreover, as another example of high allele diversity detected at the 4-field in the present study, it is also noteworthy three different variants that are identified within the *HLA-C\*04:01:01* allele group (*HLA-C\*04:01:01:01*, *HLA-C\*04:01:01:05* and *HLA-C\*04:01:01:06*; representing 70.2%, 2.4% and 27.4% respectively within this allele group). At the same time, similar rare *HLA-C* allele variants (presenting less than 1% of allele frequency value) can be seen when comparing these two studies. Some examples of these infrequent variants are *HLA-C\*07:04:01:01* or *HLA-C\*17:01:01:05* found in the present study. Lastly, in relation to null alleles detected at the *HLA-C* locus. Strikingly, and conversely to the present study, in Balas et al. study only *HLA-C\*07:32N* (Insertion; Exon 3, 560-561 InsCGCAGAT, in codon 163, causes frameshift and premature stop at codon 198) was detected but not *HLA-C\*04:09N* allele. Understanding that *HLA-C\*04:09N* allele is relatively quite frequent in Spanish population (and observing that both studies describe high allele frequency values for *HLA-B\*44:03* allele group), a conceivable explanation could be that, presumably, in this specific Balas et al. study (since it is not specified in the written article) it would have not been possible to determine presence or absence of *HLA-C\*04:09N* allele assuming that this null allele would have been embedded in certain given ambiguities (e.g. as it is explained in [221]).

- The case of *HLA-DPA1* locus clearly exemplifies the higher level of heterozygosity found at the 4-field in comparison to the 2-field level, thanks to the high-resolution capacity and very low ambiguity that can be obtained by the application of NGS-based HLA genotyping

methods as it was performed in the present study. As an example, in the present Spanish population study only two respective variants *HLA-DPA1\*01:03* (78.7%) and *HLA-DPA1\*02:01* (16.7%) are described when just looking at the 2-field. Whereas at the 4-field, up to 6 variants of the *HLA-DPA1\*01:03* allele group (*HLA-DPA1\*01:03:01:01*, *HLA-DPA1\*01:03:01:02*, *HLA-DPA1\*01:03:01:03*, *HLA-DPA1\*01:03:01:04*, *HLA-DPA1\*01:03:01:05* and *HLA-DPA1\*01:03:05*; representing 15.7%, 39.0%, 8.1%, 22.6%, 14.3% and 0.2% respectively inside this allele group) are detected; and up to 4 variants of the *HLA-DPA1\*02:01* allele group (*HLA-DPA1\*02:01:01:01*, *HLA-DPA1\*02:01:01:02*, *HLA-DPA1\*02:01:02* and *HLA-DPA1\*02:01:08*; representing 47.8%, 26.1%, 21.7% and 4.3% respectively inside this allele group) were found. The 6 most prevalent *HLA-DPA1* alleles (presenting allele frequency values higher to 5%) found in the present Spanish population cohort, which represent the 87% of all detected *HLA-DPA1* variants here, are *HLA-DPA1\*01:03:01:02*, *-DPA1\*01:03:01:04*, *-DPA1\*01:03:01:01*, *-DPA1\*01:03:01:05*, *-DPA1\*02:01:01:01* and *-DPA1\*01:03:01:03*. Whereas *HLA-DPA1\*03:01*, *HLA-DPA1\*01:03:05* and *HLA-DPA1\*04:01* seem to be very infrequent (considered as rare alleles presenting less than 1% of allele frequency value) in Spanish population. Moreover, in comparison to other previous studies at both 2-field and 4-field allele resolution levels, the distribution of *HLA-DPA1* alleles described in the present Spanish population study is in consonance with what has been reported for other populations of both European ancestry [130][297][464][806][888][889] and Spanish ancestry [627][630].

- As it has been previously mentioned (see **INTRODUCTION** section), phasing ambiguities are very common in genotype results for *HLA-DPBI* locus making its molecular characterization at high-resolution quite challenging and with important limitations. Unlike previous HLA studies based on legacy methodologies (e.g. SBT, SSO, RT-PCR or SSP), in

the present study it was possible to characterize the *HLA-DPB1* allele diversity in Spanish population minimizing (although not completely eliminating) considerably this type of ambiguity by the application of the aforementioned NGS-based HLA typing methodology. A representative example of allelic diversity found at the 4-field in this locus is the *HLA-DPB1\*04:02* allele group. Which only presents one single variant (*HLA-DPB1\*04:02*) at 2-field level, whereas there are 2 variants (*HLA-DPB1\*04:02:01:01* and *HLA-DPB1\*04:02:01:02*; representing 37.3% and 62.7% respectively inside this allele group) found in Spanish population. Moreover, *HLA-DPB1\*04:01:01:01*, *-DPB1\*02:01:02*, *-DPB1\*04:02:01:02* and *-DPB1\*01:01:01* were found as the most common variants (all of them showing an allele frequency value greater than 5%) representing 66% of the defined *HLA-DPB1* alleles in Spanish population. In contrast, *HLA-DPB1\*105:01*, *HLA-DPB1\*19:01* and *HLA-DPB1\*59:01* are some examples of rare *HLA-DPB1* allele variants (presenting less than 1% of allele frequency value) in Spanish population. Furthermore, the distribution of *HLA-DPB1* alleles described in the present Spanish population study is also similar with what has been observed for other populations of European ancestry at both the 2-field and 4-field allele resolution levels [130][297][464][627][630][806][888][889]. Nevertheless, a striking exception is the case of the allele group *HLA-DPB1\*03:01*. As this allele group appears to be more frequent in European American (Caucasoid) ethnic group in the U.S. population (about 10%) [806][889] and Northern European populations (about 13%) [888] than in the present Spanish population cohort (about 4%). Which, in turn, seems to be more similar to the respective Hispanic ethnic group in the U.S population (about 6%) in relation to the frequency of this allele group *HLA-DPB1\*03:01* [889]. It is also noteworthy the very high prevalence of *HLA-DPB1\*04:01:01:01* allele found in the present Spanish population cohort, as well as at the worldwide population-level, and as it has been previously

discussed in the present thesis work. Thus, this example may illustrate how certain level of HLA diversity shown by some modern populations may have been founded recently, in part, due to major demographic events (e.g. a population bottleneck, population expansions, migration waves, population admixture or founder effect).

- For *HLA-DQA1* locus, similarly to what has been observed for *HLA-DPA1* locus, a striking allelic diversity is found at the 4-field level in the present study. As an example, the only observed variant *HLA-DQA1\*05:01* at 2-field level shows a total of 3 different variants at 4-field level (*HLA-DQA1\*05:01:01:01*, *HLA-DQA1\*05:01:01:02* and *HLA-DQA1\*05:01:01:03*; representing 39.7%, 53.4% and 6.8% respectively within this allele group). Due to this high allelic diversity observed at 4-field level, up to 9 most frequent different *HLA-DQA1* allele variants (*HLA-DQA1\*02:01:01:01*, *-DQA1\*05:05:01:01*, *-DQA1\*01:02:01:01*, *-DQA1\*01:01:01:02*, *-DQA1\*01:03:01:02*, *-DQA1\*03:01:01*, *-DQA1\*05:01:01:02*, *-DQA1\*03:03:01:01* and *-DQA1\*05:01:01:01* in this present study) were found to present an allele frequency value greater than 5% in Spanish population. In fact, these above mentioned common variants account for 81% of all defined *HLA-DQA1* allele variants in the present study. On the other hand, infrequent variants as *HLA-DQA1\*01:05:02*, *HLA-DQA1\*01:04:02* and *HLA-DQA1\*05:03* can be considered as rare *HLA-DQA1* alleles (presenting less than 1% of allele frequency value) in this Spanish population study. In addition, the distribution of *HLA-DQA1* alleles described here it is in consonance with what has been reported for other populations of European ancestry at 2-field and 4-field allele resolution levels [130][297][464][890][891].

- In relation to *HLA-DQB1* locus, and in comparison to Balas et al. study (in which only exon 2 was interrogated) [624], it is noticeable the higher level of resolution, much less level of ambiguity and, thus, more detailed characterization of the allelic diversity that was obtained



in the present study (using NGS-based HLA typing method [187] for sequencing almost the complete sequence of this given locus (targeting by long-range PCR and sequencing exons 1–5 and introns 1–4) in comparison. Nevertheless, both studies still present a comparable set of 7 most frequent *HLA-DQB1* alleles (*HLA-DQB1\*02:02:01:01*, *-DQB1\*02:01:01*, *-DQB1\*03:01:01:03*, *-DQB1\*03:02:01*, *-DQB1\*06:02:01*, *-DQB1\*05:01:01:03* and *-DQB1\*06:03:01* with allele resolution level as described in the present study) that show allele frequency values higher than 5% in Spanish population. In the present study, these common *HLA-DQB1* allele variants represent about 73% of all defined alleles at this locus in Spanish population. Considered rare variants (presenting less than 1% of allele frequency value) in Spanish population such as *HLA-DQB1\*02:02:01:02*, *HLA-DQB1\*03:04:01* and *HLA-DQB1\*03:02:03* are observed in the present study. Furthermore, an example of certain level of diversity found at the 4-field in this locus, it is the *HLA-DQB1\*03:01* allele group. Which only presents one variant (*HLA-DQB1\*03:01*) at the 2-field level, whereas there are up to 3 variants (*HLA-DQB1\*03:01:01:01*, *HLA-DQB1\*03:01:01:02* and *HLA-DQB1\*03:01:01:03*; representing 18.8%, 11.8% and 69.4% respectively within this allele group) found in Spanish population. Also, in this case, the distribution of *HLA-DQB1* alleles described in the present Spanish population study is most generally in line with what has been observed for other populations of European ancestry at both 2-field and 4-field allele resolution levels [130][297][464].

- For *HLA-DRB1* locus, similar observations can be made as it has been above mentioned for *HLA-DQB1* locus. Once again, both studies show a comparable group of most common *HLA-DRB1* variants (*HLA-DRB1\*07:01:01:01*, *-DRB1\*03:01:01:01*, *-DRB1\*15:01:01:01*, *-DRB1\*13:01:01:01*, *-DRB1\*01:01:01* and *-DRB1\*11:04:01* with allele resolution level as described in the present study) that present an allele frequency value higher than 5%. This

group of most frequent *HLA-DRB1* variants in Spanish population accounts for 58% in the present study. In contrast, *HLA-DRB1\*04:07:01*, *HLA-DRB1\*11:01:02*, *HLA-DRB1\*14:04:01* and *HLA-DRB1\*15:06:01* are some examples of rare *HLA-DRB1* allele variants (presenting less than 1% of allele frequency value) in Spanish population. In comparison to other HLA loci described in the present study, it is noteworthy that the diversity found at the 4-field allele resolution level at this *HLA-DRB1* locus appears to be less. As an exception, two variants *HLA-DRB1\*10:01:01:01* (representing 90% inside this *HLA-DRB1\*10:01* allele group) and *HLA-DRB1\*10:01:01:02* (representing 10% inside this *HLA-DRB1\*10:01* allele group) were detected in the present Spanish population study. Likewise, *HLA-DRB3*, *-DRB4*, *-DRB5* loci also seem to present a more relatively restricted non-coding diversity. Nevertheless, it is important to underscore (as it has been previously noted and explained at both the **INTRODUCTION** section and at the beginning of this **DISCUSSION** section [146][152][158][161][184][202][204][296][368][463]) how all these *HLA-DRB* loci present:

- Long intron sequences and abundant presence of DNA regions (mostly non-coding regions) with repetitive and extensive low-complexity and imbalanced sequence composition, such as: homopolymer repeats poly(dA), poly(dT), poly(dG) and poly(dC) (composed of eight or more nucleotides); regions of short-tandem repeats (STRs; comprised of 1–6 bp per repeating unit); or high AT- or GC-rich regions (that often contain mononucleotide repeats of 10 or more bases).

- Thus, many unsequenced regions in the IPD-IMGT/HLA reference allele database for these *HLA-DRB* loci.

Therefore, due to the impossibility to still unambiguously characterize the full-length of these *HLA-DRB* loci [368], there is definitely an underlying diversity at the 4-field level which remains unrevealed.

• Lastly, regarding *HLA-DRB3*, *-DRB4*, *-DRB5* loci, similar results can be observed between both studies. Although, there is a much higher allele resolution level at the 2nd, 3rd and 4th field, especially for *HLA-DRB4* locus, described in the present study and in comparison to Balas et al. study [624]. These *HLA-DRB3/4/5* genes characteristically behave as alleles of a single locus (the so-called “*HLA-DRB3/4/5* superlocus”) as the presence of one of these genes at the haplotype level generally excludes the presence of the other two genes. This is based on the linkage constraints that exist between the *HLA-DRB3/4/5* loci and the given present *HLA-DRB1* locus, where several *HLA-DRB1* allele families can be differentiated [344]. Thus, these genes seem to present relatively low diversity. On the other hand, we observed certain cases of striking diversity at the 4-field level in the present Spanish population study. For instance, *HLA-DRB3\*02:02* allele group presents two distinct allele variants (*HLA-DRB3\*02:02:01:01* and *HLA-DRB3\*02:02:01:02*; representing 41.9% and 58.1% respectively inside this allele group). Furthermore, in the case of *HLA-DRB4* locus we observed up to four different variants for the *HLA-DRB4\*01:03* allele group (*HLA-DRB4\*01:03:01:01*, the null allele *HLA-DRB4\*01:03:01:02N*, *HLA-DRB4\*01:03:02* and *HLA-DRB4\*01:03:03*; representing 84.6%, 9.6%, 4.8% and 1.0% respectively inside this *HLA-DRB4\*01:03* allele group). In contrast, no such diversity was observed for *HLA-DRB5* locus in the present study. Moreover, the respective allele frequency distributions of *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* alleles described in this Spanish population study are in consonance with what has been reported for other populations of

European/Mediterranean ancestry at both 2-field and 4-field allele resolution levels [130][297][464].

ii) Secondly, also at the HLA allele level, referring now to other representative and descriptive HLA Spanish population studies previously reported apart from Balas et al. study [624]. Here, most of these selected studies from the scientific literature have in common that they describe relatively high-resolution (2-field or higher allele resolution level) HLA diversity of, generally, main classical HLA class I (*HLA-A*, *-B*, *-C*) and class II (*HLA-DQB1* and *-DRB1* at least) genes; present a significant sample size of a given Spanish general population cohort or otherwise of a certain Spanish major region; and that show both allele and, also in most of the cases, 2-locus and even extended haplotype estimated frequency distributions. Thus, it is noteworthy, relative to the present study, the following findings and noticeable relative similarities/differences observed in each respective comparison with these other reference HLA studies carried out in Spanish population (as well as those closely related foreign populations showing Spanish ancestry):

- Romòn et al. study [260] describes (comprising the largest reported sample size up to date) the HLA diversity found within the Iberian Peninsula although only based on: 1-field very low-resolution (generic level of resolution) HLA genotyping data (of a reduced set of loci: *HLA-A*, *-B*, *-C*, *-DQB1*, *-DRB1*) obtained from a very large cohort (N=63,484) of the Spanish BM donor registry (Registro Español de Donantes de Médula Ósea, REDMO) in addition to a 1-field HLA genotyping dataset (including only *HLA-A*, *-B* and *-DRB1* loci) from a very large Portuguese regional panel registry (N=59,443). Thus, Romòn et al. [260] mapped HLA variation in the majority of Iberian Peninsula by combining classical population genetic analyses with geographic information approaches (i.e. evaluating the correlation between variation of HLA allele/haplotype frequency distributions and latitude/longitude as

geographical parameters). Conversely, the current NGS HLA study comprises a very high-resolution (3-/4-field) HLA genotyping dataset of all major 11 classical HLA class I and class II genes though, in this case, based on a much more discrete Spanish general population sample size (N=282, covering only 10 out of the 17 Spanish Autonomous Communities/Regions) and even to a more discrete extent at the regional level (n=25-26 per Spanish Autonomous Community/Region). Nonetheless, there are still some commonalities in the main findings of both studies:

- Romòn et al. study [260] has revealed HLA genetic similarities and specific genetic signatures within certain given geographical regions (e.g. on one hand, the Atlantic/Celtic-Galician-North Portuguese Domain; whereas Central Castilian Plateau, Mediterranean Basin and Andalusian regions are genetically close; and another differentiated group is the Higher Ebro Valley area which encompasses Basque Country, Navarre and La Rioja regions) of Spain. These HLA Spanish regional-specific signatures, as previously mentioned in the **INTRODUCTION** section, may have been shaped (among other factors (e.g. natural selection) and at least to a certain extent) due to the singular orography found across the Iberian Peninsula along with well-documented historical facts that determined major demographic events of significant impact and imprint [555][556][558][578], and thus it has made of Spain a country that currently shows an extensive cultural and social diversity within its entire population, which is also reflected at the genetic level (in this given case as for HLA genes). In the current NGS HLA study in Spanish population, although at a much smaller scale (that will certainly need to be further extended and evaluated to be adequately representative), similar general findings to those from Romòn et al. study [260] can be observed especially based on the results shown in **Table R-10**.

-Importantly, as an aspect that had been hardly evaluated before and in particular based on HLA genetic data, Romòn et al. study [260] has also initially described (as also, for instance, it has been previously detected at the European continental level [136][137]) a broad and complex regional variation of the HLA diversity across the Spanish territory where frequency distributions of many HLA alleles (and similarly observed in respective 2-locus HLA haplotypes) vary continuously across the Iberian Peninsula, either increasing or decreasing from the Mediterranean coast (South-East) to the Atlantic domain (North-West) or from the Strait of Gibraltar (South) to the Pyrenees and Bay of Biscay (North). Thus, in the present study (once again, with all due caution because of the very limited Spanish population sample size that was evaluated here) some similar (although some others were opposite to those observations made in Romòn et al. study) observations can be noted, in some cases being more evident than others (when looking at results shown in **Table R-10**; and similarly in respective 2-locus HLA haplotypes (data not shown per Spanish region evaluated in the present thesis work)), such as:

1) Allele frequency distribution of *HLA-DRB1\*07* (but not in the case of *HLA-B\*44*) slightly decreases from the Mediterranean coast (in our current study based on the broader Eastern-Spain sub-group) to the Atlantic domain (in our current study based on the broader Northern-Central Spain sub-group). Whereas, in this same geographical longitude-latitude direction, *HLA-B\*51* and *-DRB1\*13* allele frequency distributions appear to be increased. The latter finding in regards to the *HLA-B\*51* allele group it may be also of significance with very important clinical implications in Spanish population (as well as in other populations of European ancestry), as this *HLA-B* allele has been considered (as it is also later commented with more detail in the respective **DISCUSSION** sub-section about haplotype analyses in the present thesis work) as a relevant negative

predictive factor in order to find a suitable unrelated donor (i.e. full-matched donor (typically 10/10), especially in the HSCT setting) in both Spanish [624] and European (e.g. Swiss) [892] populations. Since it has been widely reported how this very common *HLA-B* allele, especially in populations of European ancestry, is associated with a high rate of *HLA-C* allele disparities (i.e. *HLA-B\*51* displays a very broad distribution with regard to its *HLA-C* association; and thus it does not follow the typical pattern shown by most of the *HLA-B* alleles where they tend to show a preferential association with one or two *HLA-C* allele variants).

2) Allele frequency distribution of *HLA-DRB1\*07* slightly increases from the Strait of Gibraltar (in our current study based on the broader Southern-Spain subgroup) to the Bay of Biscay and the Pyrenees (in our current study based on the broader Northern-Central Spain sub-group). Whereas, in this same geographical longitude-latitude direction, *HLA-DRB1\*11* allele frequency distribution does not appear to be decreased neither significantly increased in our current study unlike Romòn et al. study (where it appears to be decreased) [260].

-HLA allele frequency distributions estimated on the Spanish REDMO donors cohort described in Romòn et al. study [260] were found to be very similar to those obtained on different Spanish population cohorts analyzed in other main studies (for example, such as those coming from Castilla-Leon (N=1,940) [628]; Girona (Catalonia) (N=88) and Guipuzcoa (Basque Country) (N=100) [627]; or Murcia (N=173) [560]). Likewise, results of most common HLA allele frequency distributions shown in the present study are in consonance with reported results by all these main studies previously carried out in Spanish population [221][260][545][546][558-561][563][564][571][600][608][624-630][757].

-In Spanish population [based on findings not exclusively described in this Romòn et al. study [260] but additionally reported in other low-resolution (1-/2-field allele resolution level) HLA studies that also present important and representative sample sizes (e.g. N=5,458 cord blood units at the Umbilical Cord Blood Bank in Barcelona, Catalonia (located in North-East Spain) [221]), it is also noteworthy the singularity detected in the case of *HLA-DQB1* locus, at least when analyzing at lower-resolution levels. Thus, it has been observed how at the 1-/2-fields of resolution the *HLA-DQB1* locus shows a relatively low allelic diversity in most of those Spanish population study cohorts so far evaluated. In fact, this “excess of allele homozygosity” (or “unexpected reduction of observed heterozygosity”) detected at this given locus (e.g. overcount of the genotype *HLA-DQB1\*03:02* (allele belonging to one chromosome), *-DQB1\*03:03* (allele belonging to the other second chromosome) observed in Spanish population) could be explained (in what is the so-called “Wahlund effect” [340][341][777][893], which is a possible causative factor in certain instances of detected significant deviation from HWEP) by an existing, and apparently substantial, population genetic stratification or substructure previously described in other Spanish population cohorts [221][260], as well as similarly reported in other populations (e.g. [268]). Therefore, based on all these initial evidences as well as in line with well-documented geographical and historical facts [556][578], a large portion of modern-day Spanish population may presumably consist of several subpopulations that show important different HLA allele frequency distributions, and being thus especially in relation to allele frequency distributions displayed at the *HLA-DQB1* locus. Interestingly, in our current NGS HLA study if collapsed 2-field HLA genotype data is only considered, *HLA-DQB1* locus shows, indeed, this aforementioned low allele diversity (illustrated by its  $k$  value ( $k=17$ ) and, comparatively, being lower than in other HLA loci tested at the same



collapsed 2-field resolution, for example: *HLA-A* (k=28), *-B* (k=48), *-C* (k=28), *-DPBI* (k=26) or *-DRBI* (k=35)). Nevertheless, when looking at 3-/4-field allele resolution level (as it is enabled by application of a NGS-based HLA genotyping method in the present study) the initially unexpected, but indeed noticeable, non-coding variation detected within this *HLA-DQB1* locus clearly increases its k value (up to k=24) and, consequently, its allele diversity described. Which, in turn, it becomes more equivalent to that one observed in other HLA loci tested at this higher 3-/4-field resolution, for example: *HLA-A* (k=36), *-B* (k=53), *-C* (k=40), *-DPBI* (k=29) or *-DRBI* (k=37)). In summary, this discussion may exemplify how assessment of HLA diversity considering, on one hand, the nucleotide level (up to the 4-field) or, on the other hand, at the protein level (represented by 2-field) may lead into concurred interpretations in some instances, whereas in some other cases not equivalent elucidations may be concluded instead [104][137]. Thus, at this very high allele resolution level (3-/4-field), this plausible population genetic substructure detected in Spain could be more reflected and evident in the specific 4-field differences found inside the haplotype associations (i.e. LD patterns) as well as at the 4-field allele frequency distribution level for certain HLA locus/loci. Unfortunately, the very restricted Spanish population sample size that was evaluated at the regional level (n=25-26 per region) in the present study did not allow us to investigate these considerations accurately. In this sense, it is expected that future studies of both considerably larger sample size, (being thus of higher associated statistical power) and at a wider geographic scale of the Spanish territory will considerably contribute to better understand this population genetic substructure initially detected in Spain.

-In the SOT setting, it is well-known that Spain holds a privileged and leading position worldwide in providing transplant services to its patient population (e.g. with 40 donors and more than 100 transplant procedures per million population in 2015) [642]. Nevertheless,

in the HSCT setting, it is important to remark (as it is mentioned in the Romòn et al. study [260]) that the REDMO Spanish BM and UCB donor national registry currently shows an overrepresentation of donors from major but very few Spanish regions (such as Madrid, Andalusia or just Barcelona, out of the total region of Catalonia (comprising up to the 60% of total donors)) as well as an important paucity of donors from other certain local sparsely populated regions in Spain (e.g. Zamora, Soria, Huesca or Teruel). In addition, there is an important level of variability in the REDMO database in regards to the number of HLA loci and allele resolution level consistently tested and reported per donor, critically limiting the current characterization of Spanish donor registry population. Thus, results of most common HLA allele and extended haplotype frequency distributions of all 11 major HLA classical genes obtained from future larger NGS studies and at a wider geographic scale of the Spanish territory will provide an invaluable refined information for the potential improvement of the current registry in terms of population coverage as well as strategies of the organization/prioritization of donors recruitment. Moreover, development of local donor registries, in addition to improvements of the most national main registry, may contribute to better cover the diverse HLA genetic background found in Spain, which actually presents remarkable geographical (regional) HLA signatures.

- Another important group of previously reported HLA studies (i.e. in terms of sample size and higher allele resolution level) in Spanish population, which should be noted, includes:

-A recent study has characterized most common HLA allele and 5-locus haplotype frequency distributions at *HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1* loci for N=5,458 cord blood units at the Barcelona Umbilical Cord Blood (UCB) Bank (which covers 6 Spanish regions and Andorra) in Barcelona, Catalonia (located in North-East Spain) [130][221][464], and with similar results as shown in the Roura et al. NGS-HLA study on heart failure patients

[545]. As expected and overall, results from that study are in consonance with those results observed within the Eastern-Spain group of the current study (N=78, comprising all subjects from Barcelona, Valencia and Murcia regions).

-It is also noteworthy other recent study that has described HLA allele and haplotype frequencies for *HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1* loci in N=215 unrelated individuals from Gran Canaria Island (belonging in this particular case to the kidney transplant patient waiting list) [564]. Here, although with all due caution (because of the very restricted sample size evaluated in the respective Spanish population subsets of the present study, as for the one subset originally from Gran Canaria Island (n=25)), similar findings in relation to the most common HLA allele frequencies are observed between both studies (results of the current study are shown in **Table R-10**).

- In addition, referring to other group of HLA studies that have been focus on foreign population cohorts but that either culturally, geographically and/or historically are relatively close or related to original Spanish population [555][556][578][894][895]. It is important to remark how many (if not most) of the high frequency allelic and haplotypic distributions reported for each of these foreign populations are similar or at least very in line with the group of most common HLA allele and haplotype frequency distributions reported in the present Spanish population study (which are shown here at much higher allele resolution level in comparison to these other previous reported studies) having all this very important implications from both H&I clinical and research viewpoints. For instance, in the transplantation setting these similarities on HLA allele and haplotype frequency distributions definitely maximize the likelihood of finding compatible (even full-matched) donors at the worldwide scale (e.g. donor search via Bone Marrow Donor Worldwide (BMDW) database) [215][220][896]. At the same time, as some other examples, current and future data from

HLA-disease association studies, pharmacogenetics studies and/or HLA-restricted epitope-based vaccine related immunotherapies as well as respective epidemiology studies on all these (relatively) genetically (from the HLA genes standpoint) close populations can evidently contribute and facilitate development of common diagnostic and therapeutic strategies. Thus, as some main examples of HLA population studies that have shown important level of relatedness with the general population from mainland Spain:

-In relation to the Iberian Portuguese population, Spínola et al. [602] examined HLA polymorphism of *HLA-A*, *-B* and *-DRB1* loci in N=145 individuals across Portugal and described the respective allele and haplotype frequency distributions which are found to be very similar to those described in Spanish population, as observed in the present study (e.g. including the characteristic *HLA-B\*44:03/-B\*44:02* ratio previously discussed and the relatively high allele frequency prevalence of *HLA-B\*51:01:01* allele which, in turn, strikingly displays a very broad pattern of haplotype association distributions with *HLA-C* alleles) and as expected based on obvious geographical proximity (without important restricting orographic barriers between both territories) and well-documented historical facts for many centuries (where even Portugal was under the rule of the Spanish Kingdom during an important period of time, between 1580 and 1640 in the so-called dynastic “Iberian Union”) [555][556][578]. Also it can be observed [602], based on similarities found in HLA allele/haplotype distributions between different populations, how the original ancestry that has led to modern-day Portuguese population (very similar to what occurs in the case of Spanish population) has been genetically influenced by important North-Central Europeans’ (e.g. Germanic tribes/Christian Visigoths) [558] as well as North Africans’ (e.g. North African Berbers newcomers integrated with the Muslim Eastern Arab settlers) [808][851] gene flows throughout the history (especially during these last centuries).

Spínola et al. study detected slight (due to the respective limited sample sized studied) but relevant differences in North–South allele frequency distribution within mainland Portugal [602] as also recently reported in Romòn et al. study [260] and tentatively here as well in our present study.

-Another reference study that should be noted it is the Pingel et al. study [481], which has described the most common allele and haplotype frequency distributions found in the donors pool of the DKMS German Bone Marrow Donor Center but specifically with foreign parentage from different countries, including Spain. Here, once again, characteristic common HLA alleles (e.g. prevalence of *HLA-B\*44:03* allele) and extended haplotypes (e.g. *HLA-A\*01:01~B\*08:01~C\*07:01~DRB1\*03:01*; see **Table R-7** in the present study) of these reported “German DKMS” donors with Spanish (foreign) parentage are very similar to those observed in the Spanish general population (as shown in the present study). Therefore, Pingel et al. study [481] underscores the importance of two main aspects. First, development of local donor registries in each country significantly contributes to increase the likelihood of finding a compatible donor since the probability of identifying a matched donor is higher when both patient and donor are of the same ethnic background [897]. And second, large and diverse pool of donors in each given registry population/country may increase the likelihood of finding a matched donor for a given patient who resides in another country and, if especially, belongs to an ethnic group that is importantly underrepresented within the total given BM registry (e.g. African American or Hispanic individuals in the U.S., that show lower registration rates for becoming potential bone marrow donors) [898][899]. As illustrated in Pingel et al. study [481], the presence of Spanish ancestry in modern-day German donor registry population may be essentially explained by those most recent emigration waves that occurred from Spain to rest of Europe during the XX century

(and that still continues nowadays but in a lesser extent). Spanish emigration to the rest of Europe mostly started after the end of World War II, where Western European countries were in high need of migrant workers to maintain their high levels of economic growth as well as due to the important economic and political constraints in the historical context that Spain was experiencing during those first decades after its Civil War, which had been recently taken place (1936-1939) [895].

-Another remarkable Spanish emigration and demographic event from the past, which is also relatively reflected in the observed HLA allele and haplotype frequency distributions of modern-day populations in spite of population admixture events and ebbs and flows that may have occurred between neighboring and distant worldwide populations (e.g. [852]). It is the one that extensively occurred between Spain and Latin American countries (located in both Central and South America regions) and to a lesser extent between Spain and the U.S. [862][863], during two main periods of time of the considered recent human history [894][895]:

- 1) The first and largest period of time is mostly related to the vast and sustained extension of the Spanish Kingdom into the New World (i.e. the Americas) between XV and XIX centuries.

- 2) Whereas a second period, much more recently, can be also identified and assigned during the XX century due to similar socioeconomic and political factors as above mentioned in the case of Spanish emigration to economically stronger European countries.

Thus, several HLA population studies {such as those reported by Mack et al. (Spanish American group in the U.S. (N=279 individuals of Spanish ancestry)) [772]; also those denominated Hispanic ethnic groups (mostly of Cuban (CARHIS-Caribbean Hispanic, N=115,374 unrelated donors) and Mexican (MSWHIS-Mexican or Chicano, N=276,235

unrelated donors) ancestries) in the U.S., which are reported in the NMDP BM U.S. registry [259][299]; the Hurley et al. study on Argentinian donor registry population [224]); or studies on the Mexican Admixed (Mestizo) population [267][547]} relatively show many similarities with original Spanish general (mainland) population (as described in the present study) in regards to those most common HLA allele and haplotype frequency distributions described. Thus, despite more recent and incessant events of population admixture and ongoing migration waves, the Spanish genetic (including HLA genes and as similarly shown for other genetic markers such as Y-chromosome or mtDNA) [224][852][853] imprint and influence still remains in these foreign countries and regions as it is evident based on these relatively high frequency HLA (both allelic and haplotypic) distributions. This fact may also have important clinical implications such as the donor search process when relying on foreign registries in the transplantation setting. Where the likelihood to find a given donor will be higher in those foreign countries and ethnic groups which at some point in the past history (but that can be considered relatively recent) and for considerably long periods of time (i.e. several centuries) were closely related to original Spanish population (i.e. many former colonies and extended territories around the world which were under the rule of the Spanish Kingdom, between XV and XIX centuries in the CE approximately) as they experienced very important immigration waves from mainland Spain [894].

- Lastly, it should be noted that due to the paucity of precedent HLA Spanish population (and other closely related foreign populations) studies (being, at the same time, representative enough of the respective general population group) that may have evaluated 3-/4-field of allele resolution level and all 11 major classical HLA class I and class II loci and related allele and haplotype frequency distributions. This current comparative review of HLA Spanish population reports from the scientific literature still remains limited (at both the HLA allele

and haplotype frequency distribution levels). Moreover, the comparability between HLA data of the present study and HLA data from other studies is also restricted due to major differences found on the IPD-IMGT/HLA database used in each given past study and also on the HLA genomic coverage and level of ambiguities resolved or not resolved that may ultimately create, in some instances, overestimation/underestimation of certain HLA allele/haplotype frequency distributions [104][137]. Therefore, future NGS high-resolution HLA population studies presenting much larger sample sizes and carried out at a wider geographical scale (in this case for Spanish population) may overcome these current above mentioned limitations/misinterpretations [301].

## **2. HLA HAPLOTYPE LEVEL ANALYSES**

The HLA system is known (as it has been more extensively explained in the **INTRODUCTION** section of the present thesis work) for presenting a singularly strong and extensive linkage disequilibrium (LD) along its genomic region, although with some important exceptions due to existing recombination hotspots (e.g. there is no strong LD between *HLA-DP* and the rest of the class II haplotype because of existing hotspot of recombination between *HLA-DQ* and *-DP* loci). Thus, it has been widely described the existence of blocks (for example, *HLA-B~C* block and *HLA-DR~DQ* block, where these particular neighboring loci especially show a very close physical proximity) of conserved DNA sequence defined by a given associated pattern of specific single nucleotide polymorphisms (SNPs) [97][100][900]. Moreover, these nonrandom associations (as long as in the absence of recombination events) of linked HLA alleles at several loci, for a given chromosome, establish the inheritable unit defined as haplotype. At the population-level it has been also widely reported how HLA haplotype frequency distributions and specific HLA loci combinations within a certain haplotype significantly vary between different worldwide



populations and ethnic groups. In fact, these haplotype frequency distributions can be highly informative and, indeed, they have important clinical implications (e.g. determining negative/positive predictive factors for respective matching assessment and donor search process in the transplantation setting) [215] as well as an important meaning and use in research fields related to anthropology, evolutionary biology of HLA diversity or HLA-disease/drug hypersensitivity associations (e.g. [268][287][293][809]), just as some broad examples.

Application of most widely used NGS technologies (especially referring here to 2<sup>nd</sup> generation sequencing approaches using short-read sequencing platforms; and since 3<sup>rd</sup> generation systems based on long-read sequencing platforms still do not present enough accuracy rate particularly in the base-calling process [158]) on HLA genotyping methods has allowed to overcome important limitations presented by the legacy SBT methodology (considered the gold-standard technique for providing high-resolution up to the 3-/4- field of HLA genotypes) due to the impossibility to resolve heterozygous (i.e. cis/trans or phasing) ambiguities between the two given different alleles of a corresponding HLA locus. Concerning the scope of the present study, most of the current NGS-based HLA genotyping methods rely on initial long-range PCR per HLA gene tested (providing an extensive genomic sequence coverage) and paired-end short-read deep-sequencing platforms (on a high-throughput fashion) as it is the case of the respective protocol and NGS technology employed in the present study [187][763-766]. Thus, this current type of NGS-based HLA typing methods enables assignment of phased allele calls per HLA locus. However, phasing determination is not methodologically possible at the extended HLA haplotype level. Consequently, the assignment of alleles along each same chromosome defining each given haplotype is not resolved by current NGS-based HLA genotyping methodologies in diploid (2n) individual's DNA (i.e. HLA haplotypes are phased-unknown). Therefore, HLA 2-locus and extended haplotypes of Spanish population cohorts (made up of unrelated Spanish healthy

individuals and unrelated MS patients respectively) shown in the present thesis work are phased-unknown. Thus, it was only possible to infer extended HLA haplotypes (and respective frequency distributions) by application of EM algorithm with its inherent limitations previously mentioned [101][332-336][342]. Where, basically, EM estimated haplotypes are known to have a certain level of associated inaccuracy which is especially significant for those rare haplotypes that present very low frequencies ( $n=1$  or 2 counts). In addition, it has been detected an overestimation of LD values of EM estimated haplotypes [331] in comparison to LD estimates from phased HLA haplotype data that can be built (and accurately phased) by nuclear family-based allele segregation analysis [101][337-339].

Yet, and especially in comparison to legacy HLA genotyping methods, this current type of NGS-based methods enables a much higher precision and robustness in the characterization of HLA allele nucleotide diversity as it has been previously mentioned (see **INTRODUCTION** section) [161][178]. Where (almost) full-length and phased HLA allele sequence data with minimum ambiguities, including non-coding regions, significantly contribute to reveal invaluable information (previously almost unknown and not ostensible) relative to specific (in spite of being inferred via EM algorithm) haplotype associations (defined by given LD constraints and patterns) at the 3-/4-field allele resolution level. Thus, as an example, it is possible to detect more easily and more accurately specific HLA haplotype ethnic/population/regional background signatures, which had been hardly described in Spanish population until now, with Balas et al. study [624] as one of the very few studies reported up to date (at the moment of preparing the present study) in the scientific literature). Moreover, haplotype estimates and their distributions, from NGS-based HLA genotyping data in this case (as exemplified in the present MS HLA Spanish population case-control study carried out here), can be very informative for studying association of candidate gene studies and fine-mapping of disease genes. In fact, NGS HLA genotype data (in combination with

statistical analyses approaches such as conditional analysis and, if applicable/feasible, evaluation of that same given disease phenotype of interest in a sufficiently diverse range of ethnic groups (which may display different HLA allele/haplotype associations that may be informative)) may allow a better assessment in order to dissect (i.e. uncouple or deconstruct haplotype associations in order to separate allelic from haplotypic associations) given identified associated HLA haplotypes, which may present strong LD, in order to actually discern those real and distinctive causative HLA variation patterns relative to a given disease phenotype [401][412]. Furthermore, variation in those exons not commonly characterized by previous legacy HLA typing methods and variation in the non-coding regions of the HLA genes may have (although it has not been extensively described and well-defined yet) important functional consequences resulting in abrogation of expression, alternative splicing, altered levels of expression, post-translational regulatory molecule binding, aberrant tissue-specific expression and/or stability at the cell surface which may be of importance for deconvoluting the causative aspects of HLA-associated diseases [207][309].

Referring now in detail to the present study, application of this NGS-based methodology has allowed the assessment of distinctive 3-/4-field HLA haplotype associations (that had not been apparent before as they were unattainable by previous legacy HLA typing methods (e.g. SSO, SSP, SBT or RT-PCR)) especially when evaluating non-coding region variation at both 2-locus and extended HLA haplotype (encompassing 6-locus, 7-locus and 9-locus respectively) distributions. 3-/4-field HLA data of most common extended haplotype frequency distributions is generally shown in the present thesis work document (see **RESULTS** section (**Tables R-5, R-7, R-8, R-9**)). Whereas all respective collapsed 2-field HLA data of most common extended haplotype frequency distributions is mostly not shown in the present thesis work document, with some exceptions that are highlighted in some parts of this **DISCUSSION** section. Thus, in the present Spanish

population cohort, the following observations and findings are noteworthy, which are also compared with results reported by previous studies on other singular HLA Spanish population cohorts and other related foreign populations cohorts as well:

- Firstly, several different sets of 2-locus haplotypes were estimated for the present Spanish population cohort (see **Table R-5**). Worthy of special mention and of interest are some of the frequency distribution results shown in regards to *HLA-B~C* and *HLA-DRB1~DQB1* allele pairs, since it has been extensively reported that these respective allele pairs commonly present some of the strongest LD values, as expected owing to their genomic chromosomal physical proximity [29]. Furthermore, refined characterization up to the 4-field resolution of these specific above mentioned 2-locus association distributions (as well as in respective extended haplotype associations) may have important implications in the transplantation setting (e.g. matching assessment and donor search) [215][220], for anthropological studies [131][132][496][497] and also in the context of evaluating HLA-disease associations [401][412] and, ultimately, possible related diagnostic/risk assessment value of this information (e.g. celiac disease or narcolepsy) [121]. In detail:

1) Some of the most common HLA 2-locus haplotypes observed for these particular allele pair combinations in the present Spanish population study were the following:

-Most common *HLA-B~C* allele pairs:

*HLA-B\*07:02:01~C\*07:02:01:03* (HF=9.6%);

*HLA-B\*08:01:01:01~C\*07:01:01:01* (HF=5.9%);

*HLA-B\*44:03:01:01~C\*16:01:01:01* (HF=5.1%);

*HLA-B\*44:02:01:01~C\*05:01:01:02* (HF=4.7%);

-Most common *HLA-DRB1~DQB1* allele pairs:

*HLA-DRB1\*07:01:01:01~DQB1\*02:02:01:01* (HF=14.5%);

*HLA-DRB1\*03:01:01:01~DQB1\*02:01:01* (HF=13.5%);

*HLA-DRB1\*15:01:01:01~DQB1\*06:02:01* (HF=8.1%);

*HLA-DRB1\*13:01:01:01~DQB1\*06:03:01* (HF=7.1%);

and also, although not as frequent, to be noted the

*HLA-DRB1\*14:54:01~DQB1\*05:03:01:01* (HF=2.3%);

These above listed allele pairs have been also generally found in relatively high frequencies in other Spanish, Portuguese, Mediterranean-European, Latin American and Hispanic populations previously reported (e.g. [130][221][224][259][260][267][297][299][464][522][546-548][558-561][563][564][571][600][602][603][608][624-630][772]).

2) Interestingly, application of NGS technology for HLA genotyping in the present Spanish population study has allowed to describe specific 3-/4-field haplotype associations that were not apparent at the collapsed 2-field allele resolution level that would be typically obtained by legacy HLA typing methods (e.g. SSO, SSP, RT-PCR or SBT). Furthermore, even some of these 3-/4-field haplotype associations were not apparent or not totally explained either in some cases in Balas et al. study [264], due to limitations of that study in regards to the genomic sequence coverage of HLA genes tested and since that study was also referenced by older released versions (between 1998-2010 years) of IPD-IMGT/HLA database. As some main examples of striking 3-/4-field haplotype associations found in the present study for HLA 2-locus haplotypes:

-Within the *HLA-B~C* allele pairs, for example:

*HLA-B\*18:01:01:01* non-coding variant appears to present specific associations with *-C\*03:03:01:01* (HF=0.2%) and, predominantly, *-C\*05:01:01:01* (HF=3.8%); whereas

*HLA-B\*18:01:01:02* non-coding variant appears to display specific associations with

totally different *HLA-C* alleles comparatively, as *-C\*07:01:01:01* (HF=1.1%) and *-C\*12:03:01:01* (HF=1.8%); finally, in the case of *HLA-B\*18:01:01:03* non-coding variant, only one specific association was found with *-C\*02:10:01:02* (HF=0.2%).

-Within the *HLA-DRB1~DQB1* allele pairs, for instance, one of the most striking cases is the very distinctive 2-locus haplotype groups defined respectively by the different *HLA-DQB1\*03:01:01:01/:02/:03* non-coding allele variants (in fact, its presence in a given recipient/patient to be transplanted could be considered as a possible (although perhaps minor) negative predictive value in order to find a respective highly matched/histocompatible donor):

Where *HLA-DQB1\*03:01:01:01* variant is linked with some specific *HLA-DRB1\*04* allele groups (*HLA-DRB1\*04:01:01:01* (HF=0.8%); *-DRB1\*04:07:01* (HF=0.8%); and *-DRB1\*04:08:01* (HF=0.4%)), in addition to *HLA-DRB1\*11:01:01:01* (HF=0.2%) and *-DRB1\*12:01:01:03* (HF=0.9%) alleles; in contrast, the *HLA-DQB1\*03:01:01:02* variant shows a highly conserved association with *HLA-DRB1\*11:04:01* (HF=1.9%); lastly, *HLA-DQB1\*03:01:01:03* variant displays a very broad distribution in relation to its association with *HLA-DRB1* alleles (7 different associated *HLA-DRB1* alleles (inside *HLA-DRB1\*04*, *-DRB1\*11* (predominantly) and *-DRB1\*13* allele groups) were observed in the present study) (see **Table R-5** for more details and information relative to respective HF values found in this case).

-In addition, it is also noteworthy within the *HLA-DPA1~DPB1* allele pairs (that had not been well-described in most of the previous studies carried out in Spanish population [221][260][545][546][558-561][563][564][571][600][608][624-626][628][629], with the exception of very few preliminary studies that were quite limited in allele resolution level [627][630]), which encode the respective HLA-DP heterodimers, for instance:

The different 4-field or non-coding allele variants found inside the *HLA-DPA1\*01:03:01* allele group display singular and specific associations with respective differentiated *HLA-DPBI* alleles (see **Table R-5** for more details and information relative to respective HF values found in this case). These specific 2-locus association distributions found, in turn, may be indicative of certain plausible functional implications for the given HLA-DP heterodimer formed especially in regards, as some tentative examples, to cell-surface expression of HLA molecule, as well as to antigen presentation and definition of its related specificities [806][807]. Thus, in the present study, it can be observed (as similarly described in other initial NGS HLA worldwide population studies [297]) that *HLA-DPA1\*01:03:01:01* variant is preferentially (although not exclusively) linked with *HLA-DPBI\*02:01:02* (HF=11.2%). Whereas the *HLA-DPA1\*01:03:01:02* variant is mostly (although, once again, not exclusively) associated with *HLA-DPBI\*04:01:01:01* (HF=20.3%). In the cases of *HLA-DPA1\*01:03:01:03* and *HLA-DPA1\*01:03:01:04* variants, they appear to show strong associations with *HLA-DPBI\*03:01:01* (HF=3.8%) and *HLA-DPBI\*04:01:01:01* (HF=15.1%) alleles respectively. As for *HLA-DPA1\*01:03:01:05* variant, it presents two main associations more or less equally distributed with *HLA-DPBI\*04:02:01:01* allele (HF=4.0%) and *HLA-DPBI\*04:02:01:02* allele (HF=6.6%) respectively. Another very striking case that should be noted is relative to the denominated HLA-DP3, a common antigen relevant for functional matching algorithms of unrelated HSCT that can be encoded by two differentiated transmembrane (TM) region variants, *HLA-DPBI\*03:01:01* and *-DPBI\*104:01* (defined in the present study according to v.3.25.0 of respective IPD-IMGT/HLA database) [901]. Strikingly, at the coding region level, these two TM *HLA-DPBI* variants only differ (even though, with the most currently exceptions of *HLA-*

*DPB1\*104:01:02/:03/:04* alleles according to latest v.3.39.0 of IPD-IMGT/HLA database [295]) by one single residue located on the TM domain of this encoded HLA-DPB1 beta chain molecule: *HLA-DPB1\*03:01:01* Val (codon 205 (GTG), exon 4) and, in contrast, *HLA-DPB1\*104:01:01* Met (codon 205 (ATG), exon 4). Precisely, this differential residue position is contained within the denominated GxxxG dimerization (i.e. pairing) motif in the TM domain of this given HLA-DPB1 beta chain, which, indeed, has been described to (at least partially) determine its molecular association to the respective TM domain (presenting a dual GxxxGxxGxxxG motif in this case) of the given HLA-DPA1 alpha chain molecule (and as it occurs with other HLA class II heterodimers that are formed) [901-905]. Thus, despite limited functional role of this observed TM region polymorphism has been described until now, it is expected that refined description [e.g. via NGS technology, since *HLA-DPB1* alleles, such as *HLA-DPB1\*104:01*, have not been routinely characterized by legacy HLA typing methods (i.e. only covering exon 2); and consequently they have been considered and included under a respective broad, non-specific and ambiguous allele group (e.g. *HLA-DPB1\*03:01:01* allele group in this same example)] of this type of coding and non-coding HLA polymorphism (especially at the population-level) may shed more light to better understand: its plausible effect on cell-surface stability/regulation of expression of HLA molecule or HLA class II heterodimer structural assembly process and related specificity/molecular restriction, among other possible relevant features. In the present Spanish population study, and in relation to these estimated 2-locus association distributions, it can be observed that *HLA-DPB1\*03:01:01* allele is mostly linked with *HLA-DPA1\*01:03:01:03* non-coding variant (HF=3.8%). Whereas *HLA-DPB1\*104:01* allele is predominantly associated with *HLA-DPA1\*01:03:01:02* non-coding variant (HF=3.7%).



3) Out of these *HLA-B~C* allele pairs described in the present Spanish population study. It was also of special interest to see in detail the 2-locus association distributions, and possible substantial similarities/differences, shown by respective *HLA-B\*44:02* and *-B\*44:03* common subtypes in the present Spanish population. Which both (as previously discussed in the above sub-section) establish this singular dimorphism within the *HLA-B\*44* broad allele family [811], and where there is this distinctive observed prevalence allele frequency ratio *HLA-B\*44:03* / *HLA-B\*44:02* that seems specific of Iberian (Spanish and Portuguese) populations out of the wide group of European populations so far described [130][464][819][834][835]. Thus, as for the present study where it is reported up to the 4-field of allele resolution level (see **Table R-5**):

-*HLA-B\*44:03:01:01* segregates with the following *HLA-C* alleles:

Most commonly with *HLA-C\*16:01:01:01* variant (HF=5.1%) and secondarily with *-C\*04:01:01:01* (HF=2.4%). Moreover, much less frequent associations are also found with (presenting in this particular case a highly conserved linkage) null allele *HLA-C\*04:09N* (HF=0.4%) and *-C\*16:02:01* (HF=0.4%). In addition, being even much less frequent, it is also associated with *HLA-C\*05:01:01:02* (HF=0.2%) and *-C\*15:02:01:01* (HF=0.2%) alleles.

-In contrast, *HLA-B\*44:02:01:01*, comparatively, shows quite different 2-locus association distributions, since it segregates with the following *HLA-C* alleles:

Most commonly and primarily with *HLA-C\*05:01:01:02* variant (HF=4.7%). At the same time, but much less frequently, other associations are found as well with *HLA-C\*02:02:02:01* (HF=0.4%), *-C\*07:04:01:01* (HF=0.4%), *-C\*12:03:01:01* (HF=0.4%) and *-C\*05:09:01* (HF=0.2%) alleles.

4) Moreover, when evaluating *HLA-B~C* haplotype frequency distributions in the present Spanish population study, there is also a very distinctive case to be noted discretely (and as initially described by Balas et al. [624]). Which has been also briefly discussed before (i.e. commented findings relative to the geographical variation detected for *HLA-B\*51* allele frequency distribution in the present study as well as in Romòn et al. study [260] within the Iberian Peninsula). Where it has been detected that *HLA-B\*51:01:01:01* allele displays a very broad distribution in relation to its association with *HLA-C* alleles (e.g. 7 different associated *HLA-C* alleles were observed in the present study; see **Table R-5**). In the present study, and in detail, *HLA-B\*51:01:01:01* allele (which presents a high allele frequency value in this Spanish population cohort described (AF=6.2%)) displays 2-locus *HLA-B~C* association distributions as follows:

-In high frequencies with *HLA-C\*14:02:01:01* allele (HF=2.0%).

-In more intermediate frequencies with *HLA-C\*02:02:02:01* (HF=1.1%), *HLA-C\*15:02:01:01* (HF=1.1%) and *HLA-C\*01:02:01* alleles (HF=0.9%).

And in lower frequencies with *HLA-C\*04:01:01:01* (HF=0.5%), *HLA-C\*03:03:01:01* (HF=0.4%) and *HLA-C\*04:01:01:06* alleles (HF=0.2%).

Interestingly, in Balas et al. study [624] this number of observed linked *HLA-C* alleles is even larger, where *HLA-B\*51:01:01:01* allele additionally segregates with alleles:

-*HLA-C<sub>w</sub>\*16:02* (HF=10%); whereas in the present study these are the only observed associations: *HLA-B\*44:03:01:01~HLA-C\*16:02:01* (HF=0.4%) and *HLA-B\*51:08:01~HLA-C\*16:02:01* (HF=0.2%).

-*HLA-C<sub>w</sub>\*16:01:01* (HF=5%); whereas in the present study these are the only observed associations: *HLA-B\*44:03:01:01~HLA-C\*16:01:01:01* (HF=5.1%), *HLA-B\*44:04~HLA-C\*16:01:01:01* (HF=0.2%) and *HLA-B\*15:03:01:02~HLA-C\*16:01:01:01* (HF=0.2%).

-*HLA-Cw\*05:01:01* (HF=2.5%); whereas in the present study these are the only observed associations: *HLA-B\*44:02:01:01~HLA-C\*05:01:01:02* (HF=4.7%), *HLA-B\*18:01:01:01~HLA-C\*05:01:01:01* (HF=3.8%), *HLA-B\*44:03:01:01~HLA-C\*05:01:01:02* (HF=0.2%) and *HLA-B\*58:01:01:01~HLA-C\*05:01:01:01* (HF=0.2%).

-And *HLA-Cw\*07:01:01* (HF=2.5%); whereas in the present study these are the only observed associations: *HLA-B\*08:01:01:01~HLA-C\*07:01:01:01* (HF=6.0%), *HLA-B\*49:01:01~HLA-C\*07:01:01:01* (HF=3.8%), *HLA-B\*18:01:01:02~HLA-C\*07:01:01:01* (HF=1.1%), *HLA-B\*57:01:01~HLA-C\*07:01:01:01* (HF=0.4%), *HLA-B\*40:02:01~HLA-C\*07:01:01:01* (HF=0.2%) and *HLA-B\*41:01:01~HLA-C\*07:01:01:01* (HF=0.2%).

At the same time, although both studies have many of these *HLA-C* alleles in common (within these *HLA-B\*51:01:01:01~C* allele pairs described), some of these *HLA-B\*51:01:01:01~C* allele pairs show noticeable different haplotype frequency distribution values. For instance, in Balas et al. study [624] the most frequent allele pair found is *HLA-B\*51:01:01:01~Cw\*15:02:01* (HF=24.5%), whereas in the present study this same allele pair only shows an HF=1.1%, and where the given association with *HLA-C\*14:02:01:01* allele is the most frequent found here (HF=2.0%). In this sense, future and larger NGS HLA Spanish population studies may allow an in-depth assessment on this matter to clarify these currently observed differences between these studies. Yet, this comparative analysis between these two studies provides an insight into the very extensive and uniquely broad distribution of *HLA-B~C* segregation patterns presented by this *HLA-B\*51:01:01:01* allele, which may be explained by the strong selection given for *HLA-B~C* haplotype diversity, maximizing thus the available peptide-binding repertoire and, in turn, allowing a fundamental survival advantage against pathogens' diversity [132]. Conversely, as for the transplantation setting and based on all these findings from the present and previous studies, *HLA-B\*51:01:01:01*

allele is considered a primary negative predictive value in order to find a full-match unrelated donor (URD) for a given Spanish hematopoietic patient [624], and as it has been similarly observed in other populations of European ancestry [892].

5) As another example that can illustrate how useful and informative may be the depiction of 3-/4-field HLA haplotypic associations. Previously, during the past decades, it has been extensively reported the role and relevance of single HLA allele/locus in association with diseases (and even in some cases signifying one of the strongest predisposing genetic factor), such as *HLA-B27* in spondyloarthritis, *HLA-B51* in Behçet's disease, *HLA-DQ2/DQ8* heterodimers in celiac disease or *HLA-DRB1* locus in rheumatoid arthritis [906]. Nonetheless, in addition to obtain HLA allelic characterization up to the 4-field of these corresponding associated alleles by NGS-based typing methods. Knowledge of related haplotypic associations up to the 4-field may also contribute to better elucidate the role of HLA genes in each given pathogenesis as well as detect and dissect possible hitchhiking effects due to existing extended LD between neighboring genes. Thus, in the scope of Spanish population and as an example, description of segregation patterns of the *HLA-B\*27* allele group with corresponding *HLA-C* alleles (see **Table R-5** for more details and information relative to respective HF values found in this case) may be informative as a reference source for future epidemiology studies and, thus, to screen more accurately the prevalence of *HLA-B27* allele variants and their main LD patterns within the HLA system in both healthy controls and patients with spondyloarthritis. Also, in addition to better define these haplotype associations, characterization of HLA class II genes “A” and “B” (encoding respective alpha and beta chains of heterodimeric HLA class II molecules) by application of NGS technology will definitely provide insight into the epitope formed (by the assembly of these encoded alpha and beta chains) and its relevance and effect in the given studied HLA-disease association

(e.g. characterization of both *HLA-DQA1* and *HLA-DQB1* genes to delineate the known HLA-DQ allele competition in narcolepsy) [907].

6) Finally, 3-/4-field HLA genotyping data can be also very informative even at the 2-locus haplotype level (especially in those allele pairs (e.g. *HLA-B~C* and *HLA-DRB1~DQB1*) that show strong LD and that, consequently, have remained tightly conserved throughout generations and long periods of time in human history) to define (at least until certain extent [542][543]) genetic relatedness between populations with plausible common shared historical demographic events and/or ancestries, which may be also supported by other genetic markers [542][543] as well as by anthropological, archaeological and linguistical evidences. As an important and striking example observed in the present Spanish population study:

Relatively high frequency (especially in comparison to other populations of European ancestry [130][297][464]) of the *HLA-DQB1\*03:19:01* allele (AF=0.7%) [908] and its more common LD patterns displayed with *HLA-DRB1\*13:04* (HF=0.2%) and *DRB1\*11:02:01* (HF=0.6%) are found in Iberian populations of Spanish (as in the present study and as also observed in [221]) and Portuguese backgrounds [823]. Moreover, and strikingly, these same HLA remarks are also quite commonly found in populations originally from North Africa and also of sub-Saharan African ancestry [130][297][464][823][909]. Furthermore, and also intriguingly, same *HLA-DQB1\*03:19:01* allele related remarks have been described and observed frequently in reported populations from across the Arabian Peninsula as well [130][271][285][464][910]. Altogether, these HLA findings (also supported at the extended haplotype level, which is discussed later in the present thesis work) exemplify the relative significant genetic influence and relatedness between Iberian populations (being thus distinctive from the rest of populations of European ancestry so far described (e.g. [474])) and North African

populations as well as other populations of sub-Saharan African descent. In addition, these findings (relative to the common presence and prevalence of this *HLA-DQB1\*03:19:01* allele in all these Iberian, African and Arab population groups) may also suggest the relevant past Muslim Eastern Arab genetic contribution (at least until certain extent) in the present-day HLA gene pools of both Iberian and North African populations. Based on major historical demographic events (which, in turn, caused a relevant HLA gene flow episode), this interpretation would be strongly supported by the well-documented Muslim Arab (i.e. people originally from Eastern and Southern Syria as well as from the North and Central regions of the Arabian Peninsula, under the suzerainty of the Arab Umayyad Caliphate of Damascus [569]) invasion that took place in most regions of North Africa and, subsequently, of the Iberian Peninsula during the 7<sup>th</sup>-15<sup>th</sup> centuries period in the CE [555][556][558][578][613][808][851].

- Secondly, when considering associations between class I and class II regions in extended haplotypes (i.e. extended HLA haplotypes encompassing 6-locus, 7-locus and 9-locus respectively) and their corresponding frequency distributions that were inferred via EM algorithm in the present NGS HLA Spanish population study, the main remarks found are discussed in the following points:

- 1) In comparison to traditional typing methods (e.g. SSO, SSP, SBT and RT-PCR), NGS-based HLA genotyping methods have enabled high-throughput testing at a very high-resolution (up to the 3-/4-field with minimum ambiguities owing to both high genomic coverage and phasing per locus) of all 11 major HLA class I and class II genes. Moreover, in the case of unrelated subject-based NGS HLA population studies (as in the present study), extended 3-/4-field haplotypes can be inferred (being more accurate at high frequencies than very low frequencies) via EM algorithm. Thus, just as an illustrative example, when

comparing HLA haplotype frequency distributions of 6-locus extended HLA haplotypes at different allele resolution levels, it can be noted:

-Collapsed 2-field HLA genotyping data (that would be typically obtained by legacy HLA typing methods (e.g. SSO, SSP, RT-PCR or SBT)) (data not shown in the present thesis work document) displays a relative limited haplotypic diversity where many of the respective haplotypes are embedded inside broader and less specific haplotypic distributions. Consequently, certain strata of haplotype diversity are not apparent and are unattainable to be described. For example, relatively frequent extended haplotypes found in the present study at this collapsed 2-field such as:

*HLA-A\*30:02~C\*05:01~B\*18:01~DRB3\*02:02~DRB1\*03:01~DQB1\*02:01* (HF=1.9%);  
and *HLA-A\*25:01~C\*12:03~B\*18:01~DRB5\*01:01~DRB1\*15:01~DQB1\*06:02* (HF=0.8%); appear to share both the same identical carried *HLA-B\*18:01* allele variant.

-Whereas, as described in the present Spanish population study, 3-/4-field haplotype distributions, in this given case of 6-locus extended HLA haplotypes, clearly allow to breakdown and, thus, discern a much higher haplotypic variety (e.g. following previous example above mentioned: at the 3-/4-field it can be now distinguished two distinctive non-coding variants of this *HLA-B\*18:01:01* allele that segregate specifically with: *HLA-A\*30:02:01:01~C\*05:01:01:01~B\*18:01:01:01~DRB3\*02:02:01:01~*

*DRB1\*03:01:01:01~DQB1\*02:01:01* (HF=5.0%);

in contrast, *HLA-A\*25:01:01:01~C\*12:03:01:01~B\*18:01:01:02~DRB5\*01:01:01~*

*DRB1\*15:01:01:01~DQB1\*06:02:01* (HF=2.0%)). Thus, inferred combinations of alleles per parental chromosome are very diverse in terms of allele content per locus as well as very diverse and distinctive at the 3-/4-field allele resolution level for certain given loci which differ depending on the LD pattern described in each case.

-In addition (as previously shown in **Figure R-3** of the present study, and as recently reported in other NGS HLA studies [268][286][287][331]), haplotype distributions also become extremely divergent in terms of the multiplicity of *HLA-DP* alleles with which they associate [268][286][287][331]. This seems to be especially owing to the weak LD between *HLA-DP* and the rest of the class II haplotype since existing hotspot of recombination is present between *HLA-DQ* and *-DP* loci [92].

Therefore, these observed prominently increased multiplicity and haplotype diversity when evaluating 3-/4-field allele resolution and, even more, when including *HLA-DP* loci (which, as an example, critically contribute to the increase of mismatches in the donor-recipient transplantation setting) inside haplotype distributions may have direct implications, for example, in relation to the lesser likelihood of finding unrelated donors (URD) in HSCT [233]. At the same time, as already reported in past HSCT donor-recipient pairs retrospective studies that were carried out even prior to the NGS era and at lower allele resolution levels [911]. In spite of common HLA alleles may be found in respective HSCT donor-recipient pairs, these may still show substantial haplotypic diversity and, thus, important level of disparity. Therefore, although, on one hand, it is expected that full-matching up to the 4-field may improve HSCT outcome (which is still under evaluation [215-218]). On the other hand (in this “double-edged sword” situation), it may be more unlikely to locate a corresponding HLA matched donor, especially in the context of a patient presenting a less frequent HLA genotype for a given local/national/foreign donor registry population. Altogether, this evidences and underlines the importance, in order to minimize the detrimental effects of HLA mismatching, to continue developing and optimizing different (as some main examples): HSCT protocols (e.g. considering the use of BM (stem cells directly from the marrow or when they are induced to migrate to the peripheral blood and thus to be collected) and/or the use UCB units);



conditioning regimens; co-infusion approaches; alternatively, performing haploidentical transplantation with a specific conditioning regimen; even possible novel therapeutic strategies; and where it is also fundamental to identify and extend, if it is feasible, permissive HLA mismatches [624][911].

2) As a very noteworthy example of these distinctive 3-/4-field HLA haplotype associations revealed by application of NGS technology. The *HLA-DRB1\*11:04:01* allele (which presents a relatively high frequency in the present Spanish population cohort, AF=4.8%) was found embedded in one extended haplotype (shown here below as extended HLA haplotype encompassing 9-locus) that differs strikingly at the 3-/4-field in several *HLA* loci with its counterpart extended haplotype (more commonly found in South-East European populations), and where both haplotypes show similar haplotype frequencies in the present Spanish population cohort. In detail (see also [297]):

*HLA-A\*24:02:01:01~C\*04:01:01:06~B\*35:02:01~DRB3\*02:02:01:02~DRB1\*11:04:01~DQA1\*05:05:01:01~DQB1\*03:01:01:02~DPA1\*01:03:01:04~DPB1\*04:01:01:01*

(HF=3.0%, in the present study), which is actually highly conserved and relatively common in some modern-day Near-Eastern populations [614][621]; and, on the other hand,

*HLA-*

*A\*01:01:01:01~C\*07:01:01:01~B\*18:01:01:02~DRB3\*02:02:01:02~DRB1\*11:04:01~DQA1\*05:05:01:01~DQB1\*03:01:01:03~DPA1\*01:03:01:05~DPB1\*04:02:01:01*

(HF=2.0%, in the present study), which appears to be more commonly present in modern-day populations originally from the Balkans and Adriatic-Ionian areas situated in South-East Europe (according at least to currently reported HLA genotype datasets at the population-level [130][464][847]). Therefore, non-coding variation (in the latter example according to both *HLA-DPA1* and *-DQB1* loci) shows haplotype-specific patterns where these

arrangements also can define distinctive population-specific HLA signatures/profiles [131][132][137].

3) As previously mentioned, Balas et al. study [624] can be considered as one of the most recently reported and representative works evaluating the distribution of both HLA alleles (*HLA-A*, *-B*, *-C*, *-DQB1*, *-DRB1* and *-DRB3/4/5* loci) and significantly extended haplotypes (defined in Balas et al. study by family segregation analysis) at a relatively high-resolution level in Spanish population [624]. Results of most common HLA allele and top-ranking extended haplotype frequency distributions described in our present study are (until certain extent, and despite several important differences found in terms of study sample size; allele resolution level (i.e. HLA genomic coverage per locus); or released reference version of IPD-IMGT/HLA database used in each study and number of HLA loci tested; which altogether do not allow a completely fair comparison between studies, being thus limited) in accordance with those observed in Balas et al. study [624] as well as with other main studies in Spanish population reported so far such as:

UCB Bank in Barcelona (Catalonia) [221]; 2-locus haplotype frequency distributions of the Iberian Peninsula shown in Romòn et al. study [260]; also, those previous studies from Castilla-Leon [628], Girona (Catalonia) and Guipuzcoa (Basque Country) [627] as well as Murcia [560]; moreover, Arnaiz-Villena, A. et al. study that summarizes main initial HLA findings of Iberian populations as well as other European, Near-/Middle-Eastern and North African populations [558]; other Spanish population cohorts reported in the AFND database [130][464]; and even singular (i.e. for being populations in territories geographically isolated, which cut off from the Iberian Peninsula) HLA datasets concretely reported in Balearic Islands populations [571] and a Gran Canaria Island kidney transplant patient population cohort [564] respectively. Yet, paucity of extensive previous studies covering both all major

HLA class I and class II loci and respective high-resolution 3-/4-field allelic/haplotypic data in other Spanish population cohorts has precluded an in-depth comparison of our results to existing literature.

4) Moreover, it is also of significance the fact that Balas et al. study [624] (where haplotypes and their respective frequency distributions were based on family segregation analysis from a study cohort of patients and respective direct relatives) and our current study (which was made up of unrelated healthy individuals) show analogous results describing the most common HLA alleles and especially regarding the respective most common extended HLA haplotypes found in Spanish population. Thus, this substantial concordance between both studies may allow to confirm more confidently that despite HLA 2-locus and extended haplotypes have been inferred via EM algorithm (i.e. phased-unknown) in the present study, this statistical algorithmic inference seems to be accurate enough as it is quite in line with those phased HLA haplotypes (and corresponding frequency distributions) determined by family segregation analysis (being, indeed, the real HLA haplotype inheritance analysis) in respective Balas et al. study [624].

5) The majority of the previously reported HLA studies in Spanish population have been conducted using lower resolution (e.g. either 1-field or 2-field or by P, G groups) typing methods with an important level of both allelic and phasing ambiguities, covering just a few HLA loci and/or are restricted to small sample size cohorts (and even not quite representative or where even only considering a cohort of patients but not considering a group of healthy individuals) of some population groups and regions [558-561][563][564][571][600][608][625-630]. Just very recently, a few studies, including the present thesis work, [221][260][269][297][545][546][564][624] have overcome many of

these past limitations and have achieved (although still with certain important downsides (mainly, in relation to incomplete coverage of HLA gene sequence), especially in some cases) to gain a better insight of the existing genetic complexity of the Spanish general population and thus of the HLA diversity (both at the allele and haplotypes levels) across the Iberian Peninsula. Within this latter group of studies, it is worthy of note certain important resemblances, but also some other striking differences, regarding main HLA haplotype signatures which have been described in Spanish population.

\*As significant commonalities between these very recent studies, it is noteworthy the following instances:

-Firstly, the top most common extended haplotypes that have been reported comprise *HLA-A\*29:02~C\*16:01~B\*44:03~DRB1\*07:01~DQB1\*02:02* (frequently found in Western European Mediterranean populations, and in particular very characteristic of Iberian populations) as well as *HLA-A\*01:01~C\*07:01~B\*08:01~DRB1\*03:01~DQB1\*02:01* (which is found at high frequencies especially in populations of Northern-Central European ancestry).

-Secondly, haplotype frequency distributions usually over 1% have been also characteristically observed in Spanish population for extended haplotypes such as:

*HLA-A\*30:02~C\*05:01~B\*18:01~DRB1\*03:01~DQB1\*02:01* (frequently found in Western European Mediterranean populations);

or *HLA-A\*03:01~C\*07:02~B\*07:02~DRB1\*15:01~DQB1\*06:02* (found at high frequencies especially in populations of Northern-Central European ancestry).

-Thirdly, when comparing (or searching) this top list of extended haplotypes described in high frequencies for Spanish population with the respective haplotype rankings described on major European and North American population datasets and HSCT donor registry

populations (e.g. [130][223][225-227][259][268][297][299][464][474][481][943]). It can be observed that the majority of most common Spanish (and Iberian [602]) extended haplotypes are also relevant haplotypes in registries and population datasets of predominant European ancestry. Nevertheless, there are also certain common extended haplotypes in Spanish population, which are not as frequent as in other foreign (mostly of predominant European ancestry) registries or reported populations datasets. Thus, these seem to be definitely specific Spanish (Iberian) HLA haplotype signatures (as they show restricted high frequency distributions in comparison to other reported foreign populations), where, in detail:

This occurs with *HLA-A\*30:02:01:01~C\*05:01:01:01~B\*18:01:01:01~DRB1\*03:01:01:01~DQB1\*02:01:01* (HF=5.0%, in the present Spanish population study (ranked #3 in **Table R-7**) (data also shown in [297]);

and even more steeply in the case of *HLA-A\*25:01:01~C\*12:03:01:01~B\*18:01:01:02~DRB1\*15:01:01:01~DQB1\*06:02:01* (HF=2.0%, in the present Spanish population study (ranked #7 in **Table R-7**) (data also shown in [297])) and also for *HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~DRB1\*07:01:01:01~DQB1\*02:02:01:01* haplotype (HF=7.5%, in the present Spanish population study (ranked #2 in **Table R-7**) (data also shown in [297])).

Therefore, this remark illustrates the importance of development of local donor registries (thus, reducing also the dependence on foreign registries, which may be a more complicated option logistically, costly and timely speaking), in addition to improvements (e.g. refined and more permissible matching strategies for unrelated donor search in HSCT) of the most national main donor registry, with the final goal of optimizing donor search by covering better the different HLA genetic substrates found in Spain, which actually presents

remarkable geographical (regional) HLA diversity [260]. Moreover, this increased knowledge and its application in development of local donor registries may be applicable in the SOT setting as well, which presents important challenges due to time constrictions related to donor organ quality (i.e. ischemia time) and suitability for optimal transplantation in addition to the existing most relevant problem of organ shortage. Current strategies to overcome these burdens are limited to desensitization protocols and the recently created kidney exchange programs under the coordination of well-established international networks especially between neighboring countries [912].

-At the same time and as it was initially described in Balas et al. study of a very representative Spanish population patient cohort [624], this other following observation was also found in our current NGS HLA Spanish healthy population cohort study [269]. In contrast to many of those very frequent HLA class I haplotypes that display fairly conserved and strong associations with specific HLA class II haplotypes as previously shown (e.g. most common extended haplotypes carrying *HLA-B\*44:03:01:01~C\*16:01:01:01* or *HLA-B\*18:01:01:01~C\*05:01:01:01*). Interestingly, as some striking examples, two relatively frequent HLA class I haplotypes in Spanish population, such as *HLA-B\*07:02:01~C\*07:02:01:03* and *HLA-B\*44:02:01:01~C\*05:01:01:02*, showed a much higher variability in relation to number of associated HLA class II haplotypes (where these may represent negative predictive factors for finding a suitable donor in the HSCT setting [220]). Thus, in the case of the present study, they were found together with five and four different HLA class II haplotypes respectively (data shown in [297]). Therefore, the present NGS HLA genotype data described here for a representative Spanish population cohort allows an in-depth assessment of the given diversity of HLA bi-locus associations embedded in these described extended haplotypes. Which, as previously mentioned, implies

a greater level of difficulty in donor searches, since the presence of, as an example, uncommon *HLA-C~B* or *-DRB1~DQB1* linkages for a given registry population decreases the likelihood of success in finding a full-matched donor [897]. Moreover, the present NGS HLA work in Spanish population [269] (as well as shown in few other recent HLA studies at very high-resolution (3-, 4-field) [221][624]) has allowed a very comprehensive description of the haplotype organization (not very well defined until now) of the region encompassing *HLA-DRB1~DRB3/4/5* loci. In fact, although the expected linkages in this particular *HLA-DR* class II region may be observed in most cases, as it has been well-documented in past large multi-ethnic studies (by reference studies such as Robbins et al., Holdsworth et al., Bontrop et al. and Andersson [75][86][344][536]), current application of NGS HLA genotyping has allowed a very reliable and accurate determination of presence/absence of these *HLA-DRB3/4/5* loci and, thus, the association displayed by respective *HLA-DRB1* alleles within the genotypes (for instance, as reported in the most recent 17<sup>th</sup>-IHIW Reference B-Lymphoblastoid Cell Lines component [356]). Even though exceptions to those expected LD patterns in *HLA-DRB1~DRB3/4/5* loci associations do not seem to be fairly common in Spanish population, there are preceding HLA studies which have described certain striking outliers of these expected *HLA-DR* class II linkage patterns such as:

The *HLA-DRB1\*15:03:01:01~DRB5\*Absent* haplotype, mainly reported in African American populations [75][668]; while the *DRB1\*15:01:01:01~DRB5\*Absent* haplotype has been mostly detected in some U.S. population groups of European (EUR) and Asian-Pacific Islander (API) descents respectively [259]; finally, another unusual haplotype, the *HLA-DRB1\*10:01:01~DRB5\*01:01:01* has been observed in a recent HLA NGS study carried out in South African population [272].

-As previously remarked at the *HLA-B* allele and 2-locus *HLA-B~C* haplotype levels, the *HLA-B\*51:01:01:01* carrying extended haplotypes represent a very noteworthy and singular case which has been commonly described in main previously (and even more comprehensively in those most recent works [221][269][564][624]) reported HLA Spanish population studies and other Iberian populations [602][603]. In the present NGS HLA study, it can be observed that despite *HLA-B\*51:01:01:01* allele is relatively frequent in our healthy donor Spanish population cohort (AF=6.2% in the present study, ranked #4), there is also a very broad range of evenly distributed haplotypic associations that *HLA-B\*51:01:01:01* allele displays with *HLA-C* (where up to 6 different associations were detected in the present study) and, subsequently, with other HLA loci as well. Consequently, its broad LD pattern shown in displayed extended haplotype associations makes *HLA-B\*51:01:01:01* allele to be classified as a very relevant negative predictive factor in Spanish population (i.e. where a larger, and in a longer period of time, search of donors will be always required to find a 10/10 matched with the same extended haplotype association), as also reported in other previous studies in Swiss and Spanish population cohorts [624][892]. Thus, this pronounced haplotype association diversity that *HLA-B\*51:01:01:01* displays with *HLA-C* alleles is also reflected at the extended haplotype level. In the present study, these are the most likely *HLA-B\*51:01:01:01* carrying extended haplotypes detected (data shown in [297]):

*HLA-A\*02:01:01:01~C\*02:02:02:01~B\*51:01:01:01~DRB1\*11:01:01:01~*

*DQB1\*03:01:01:03* (HF=1.0%);

*HLA-A\*02:01:01:01~C\*14:02:01:01~B\*51:01:01:01~DRB1\*08:01:01~*

*DQB1\*04:02:01* (HF=1.0%);



*HLA-A\*11:01:01:01~C\*04:01:01:01~B\*51:01:01:01~DRB1\*04:07:01~*

*DQB1\*03:01:01:01* (HF=1.0%); and

*HLA-A\*24:02:01:01~C\*02:02:02:01~B\*51:01:01:01~DRB1\*07:01:01:01~*

*DQB1\*02:02:01:01* (HF=1.0%).

In comparison to other HLA studies (also at high levels of allele resolution) in Spanish population [221][564], very similar *HLA-B\*51:01* carrying extended haplotypes are observed. However, in contrast to the present study (presenting HLA data at the 3-, 4-field), it is striking a slightly (since all related *HLA-B\*51:01* haplotype distributions are still evenly spread) but notable predominance of:

*HLA-A\*02:01~C\*14:02~B\*51:01~DRB1\*08:01~DQB1\*04:02* haplotype according to the HLA dataset reported from Barcelona Umbilical Cord Blood (UCB) Bank (which covers 6 Spanish regions and Andorra) in Barcelona, Catalonia (located in North-East Spain) [130][221][464];

whereas, in the singular Gran Canaria Island kidney transplant patient population cohort [564] this other *HLA-A\*02:01:01~C\*12:03:01~B\*51:01:01~DRB1\*11:01:01~DQB1\*03:01:01* haplotype appears to be one of the most relatively frequent.

At the worldwide population level and in relation to global allele frequency distribution found for *HLA-B\*51:01:01:01*, populations with high *HLA-B51* prevalence lie predominantly north of the equator and overlies the ancient trading routes (e.g. the so-called “Silk Road” between the Mediterranean and the Orient regions), spanning Western Europe and even as far East as Japan [913]. In the European continent, Mediterranean populations (Italian and Greek) and in particular Iberian populations [131][221][269][297][464][602] show a relatively significant high frequency for *HLA-B\*51:01* and those same most

common extended haplotypes just described above. Interestingly, similar to the previously mentioned *HLA-DQB1\*03:19:01* case, same *HLA-B\*51:01* allele/haplotype signatures have been characteristically identified in North African Berber and Muslim Eastern Arab descent populations [130][271][285][464][611-613][808][849][855] where their genetic substrates show high level of relatedness and historical demographic influence on Iberian populations' genetic pool [555][556][558][578][602]. Moreover, from an epidemiological standpoint and as previously commented, studies on populations (such as those described in Spanish population [633][914]) presenting a high correlated prevalence of *HLA-B51* subtypes and Behçet's disease (BD) have been of relevance in the immunogenetics field [913]. However, despite some studies have described possible related specific polymorphisms between HLA and non-HLA genes (intrinsically related with inflammation and the immune response) in the disease phenotype and severity [915], it has not been well-defined yet both the impact of diversity found in *HLA-B51* subtypes and the plausible implication of the existing diversity found at the *HLA-B\*51:01:01:01* carrying extended haplotype level. As a matter of fact, there are two main intriguing observations not fully explained yet [913]:

-One is the existing wide spectrum in the relative risk of *HLA-B51* for BD across different ethnic groups, where a contributory, if not primary, role of genetic loci centromeric to *HLA-B51* is likely.

-In addition, there is a striking lack of disease occurrence in Amerindian populations despite some of them show high *HLA-B51* allele frequency distributions [130][464]. The main proposed hypothesis has been that this may indicate that in these New World populations either linkage patterns along or in the proximity of the *HLA* region are unusual (i.e. due to differential recombination events and/or rates in respective haplotype pools

between American continent and Eurasian continent populations) and/or that certain external (e.g. environmental/demographic) risk factors importantly present in Eurasia are absent or significantly different from the Americas.

Therefore, future multi-ethnic NGS HLA studies at large scale in BD may shed light to resolve this matter by contributing to the fine-mapping of BD associated risk factors that need to be disentangled.

-Finally, another very similar case to the previous one just shown on *HLA-B\*51:01:01:01* allele/carrying extended haplotype signature, it is the *HLA-B\*49:01:01* allele/carrying extended haplotype signature almost uniquely found in Mediterranean populations (Italian and Greek) and especially in the Iberian Peninsula [130][221][269][464][602][624], and, in turn, being of special significance in the Canarias Islands [563][564] archipelago, in comparison to other neighboring populations of North-Central European ancestry. Where, once again, a relatively frequent *HLA-B\*49:01:01* allele (AF=4.0%) in Spanish general population displays an evenly broad set of haplotypic frequency distributions (HF=1.0%; data shown in [297]):

*HLA-A\*01:02~C\*07:01:01:01~B\*49:01:01~DRB1\*13:05:01~DQB1\*03:01:01:03;*

*HLA-A\*11:01:01:01~C\*07:01:01:01~B\*49:01:01~DRB1\*13:02:01~DQB1\*06:09:01;*

*HLA-A\*23:01:01~C\*07:01:01:01~B\*49:01:01~DRB1\*03:01:01:01~DQB1\*02:01:01.*

Furthermore, this *HLA-B\*49:01:01* allele/carrying extended haplotype signature serves as an evidence of the also relatively notable genetic relatedness and influence, in this case, of a main sub-Saharan (probably from Central-East African regions such as Ethiopia) gene flow in the Iberian Peninsula as well [130][464]. Most likely, this would have been as a result of an ancient pre-Neolithic contribution from northward Saharan migration during hyper-arid conditions; and also, in a lesser extent, due to the later Islamic conquest period

along the North African region between 7th–9th centuries before arriving to the Iberian Peninsula [555][556][558][578][613][808][851].

\*On the other hand, among the most remarkable dissimilarities found in the reported results between these most recent studies (including the present thesis work) in Spanish population [221][260][269][564][624], it is worth mentioning the following cases:

-As one first striking example, it is in relation to the differently reported frequency distributions of both most common *HLA-B\*35:02:01* and *HLA-B\*38:01:01* carrying extended haplotypes. Which, indeed, may reflect the relevant Sephardic Jewish genetic substrate that is still present in modern-day Spanish general population, mainly as a historical demographic consequence of those notably numerous population groups that remained in the Iberian Peninsula as Crypto-Jews (maintaining a relatively high inbreeding rate (by living in closed communities with minimum mixed marriages) throughout the history) during these last centuries [558][571][572][574][578]. Thus, in the current NGS HLA Spanish population study and at the allele level, *HLA-B\*38:01:01* allele (AF=3.0%) is more frequently found than *HLA-B\*35:02:01* (AF=1.6%). Which, in turn, is in consonance with what has been described in some other most representative HLA Spanish population studies up to date, such as the ones from Barcelona Umbilical Cord Blood (UCB) Bank [221] and Balas et al. [624]. Nevertheless, in contrast to these other studies [221][624] and even in the case of respective Gran Canaria Island [564] and Balearic Islands [571] reported populations, this current Spanish healthy population cohort [269][297] curiously shows a salient predominance of the extended haplotype

*HLA-*

*A\*24:02:01:01~C\*04:01:01:06~B\*35:02:01~DRB3\*02:02:01:02~DRB1\*11:04:01~*

*DQB1\*03:01:01:02* (HF=2.0%) (in comparison to HF=0.34% in [221]; HF=0.23% in [564]; or HF=1.7% in Majorcan Jewish descents (“*Chuetas*”) [571]);

over its counterpart extended haplotype

*HLA-*

*A\*02:01:01:01~C\*12:03:01:01~B\*38:01:01~DRB3\*01:01:02:01~DRB1\*13:01:01:01~*

*DQB1\*06:03:01* (HF=1.0%; and whose association to *HLA-A\*26:01:01:01* was not detected in current study [269][297]) (in comparison to HF=0.10% (and even being HF=0.68%, when associated to *HLA-A\*26:01* instead) in [221]; HF=0.23% (and even being HF=0.70%, when associated to *HLA-A\*26:01* instead) in [564]; or HF=2.6%, when associated to *HLA-A\*24* in Majorcan Jewish descents (“*Chuetas*”) [571], whereas HF=1.7% when associated to *HLA-A\*26* instead).

One plausible explanation behind these observed unusual extended haplotype distributions in the current study may rely on the fact that it can be observed how *HLA-B\*38:01:01* displays a much broader range of *HLA-B~DRB1* haplotypic associations (nine different types detected here) in contrast to the relatively more conserved *HLA-B~DRB1* haplotypic associations shown by *HLA-B\*35:02:01* (only three different types observed in the present study) (see **Table R-5**). Yet, it would be necessary to further investigate these particular haplotype distributions in future larger scale NGS HLA Spanish population studies in order to completely assess their prevalence and regional variation as well as to reveal possible links according to specific historical demographic events of the Jewish (including both Ashkenazi and non-Ashkenazi (specifically Sephardic) backgrounds) genetic imprint in Spanish general population.

-As a second example of main dissimilarities (although here in a lesser extent) found between some of these most recent and representative HLA Spanish population studies, it

is also noteworthy the case of *HLA-B\*13:02:01* carrying extended haplotypes. It has been well-documented that *HLA-B\*13:02* allele can be considered as a fairly almost worldwide (or global) positive predictive factor in transplant donor search due to its strong linkage (almost exclusively) displayed with *HLA-C\*06:02* allele and, principally, as it is embedded in a highly conserved extended *HLA-A\*30:01~C\*06:02~B\*13:02~DRB1\*07:01~DQB1\*02:02* haplotype that shows relatively high frequency distributions across different ethnic groups (especially in Asian, Middle Eastern, European and Hispanic backgrounds, but being much less common in African ancestry) and geographical regions [130][259][299][460][464][492]. Thus, this may be illustrating a plausible extensive convergent HLA evolution case at the population-level and/or, as an alternative, it could represent a very common haplotype bearing a significant ancestral nature (at least according to the Eurasian continental landmass) [104][131][132]. In the present NGS HLA study, as well as in Balas et al. study [624], the only extended haplotype detected has been *HLA-A\*30:01:01~C\*06:02:01:01~B\*13:02:01~DRB4\*01:03:01:01~DRB1\*07:01:01:01~DQB1\*02:02:01:01* (HF=1.5%, in the present study, data shown in [297]). Nonetheless, in Barcelona UCB Bank sample cohort study [221] and even in the case of respective Gran Canaria Island kidney transplant patient cohort [564] additional *HLA-B\*13:02* carrying haplotypic associations, although not as frequent, have been identified as well, thus revealing a supplementary level of diversity. As some main examples [130][464]:

*HLA-A\*02:01~C\*06:02~B\*13:02~DRB1\*07:01~DQB1\*02:02*, presenting HF=0.31% in [221], and HF=0.47% in [564];

*HLA-A\*68:01~C\*07:01~B\*13:02~DRB1\*07:01~DQB1\*02:02*, presenting HF=0.03% in [221], whereas not being detected in [564];

and *HLA-A\*02:02~B\*13:02~C\*08:04~DRB1\*09:01~DQB1\*02:02*, presenting HF=0.03% in [221], whereas not being detected in [564].

Another haplotypic association carrying *HLA-B\*13:02* allele, although so far it has not been detected in Spanish population, it has been described in populations of North-East European descent:

*HLA-A\*02:01:01~B\*13:02:01~C\*02:02:02~DRB1\*07:01:01~DQB1\*02:02:01*,

presenting, as an example, a value of HF=0.01% in the Polish BMDR [474].

-Lastly, there are two other extended haplotypes that should be noted as relevant examples of these dissimilarities found between HLA Spanish population studies most recently described [221][260][269][564][624]:

-One of them is the *HLA-A\*02:01:01:01~C\*07:02:01:03~B\*07:02:01~DRB1\*01:03~DQB1\*05:01:01:03* haplotype. Which in the current study [269][297] shows a value of HF=3.0% (extended haplotype ranked #5 in **Table R-7**, including respective *HLA-DRB3/4/5* allele), whereas in other HLA Spanish population studies (although at a lower allele resolution level) the frequency distribution of this same extended haplotype appears to be much lesser comparatively: for instance, HF=0.38% in Barcelona Umbilical Cord Blood (UCB) Bank [221] and HF=0.47% in Canarias patient population [564]. Moreover, this haplotype (characteristically found here in Spanish population) may be also indicative, once again, of the already discussed historical genetic contribution from North African Berber and Muslim Arab population ancestries in the Iberian Peninsula [558][563-568][578][612][808]. In this sense, it is particularly noteworthy the relatively high haplotype frequency distribution of the bilocus *HLA-DRB1\*01:03~DQB1\*05:01* (HF=4.50%) reported in the Tunisian Berber population of Djerba Island, which is strikingly divergent from other North African populations but closely related with Eastern

Mediterranean population groups [916]. At the same time, it is also striking the relatively common presence of this fairly conserved same extended haplotype in Latin American populations [130][464], illustrating also here those Mediterranean and North African (of both Berber and Muslim Arab backgrounds) genetic components originally coming from the Iberian Peninsula and that were brought mainly by the Spanish colonialism between the 15<sup>th</sup> and 19<sup>th</sup> centuries [267][522][547-552].

-Whereas the other case is related with the extended haplotype group encompassing *HLA-C\*05:01~B\*18:01~DRB1\*03:01~DQB1\*02:01* alleles and that displays different associations with certain *HLA-A* alleles. In the present Spanish population cohort studied [269][297], *HLA-A\*30:02~C\*05:01~B\*18:01~DRB1\*03:01~DQB1\*02:01* was the most frequent haplotype (HF=4.8%; data shown in [297]) detected within this group, which, in turn, is typically found in Western European Mediterranean populations [130][464]. Nonetheless, in some other recent HLA Spanish population studies it has been observed to carry *HLA-A\*02:01* instead (although not as frequently observed), as some main examples, in Barcelona Umbilical Cord Blood (UCB) Bank (HF=0.48%) [221] and in Canarias patient population (HF=0.23%) [564].

6) As previously and widely discussed at the HLA allele level analysis, certain null, rare and novel alleles were identified in the present NGS HLA Spanish population study. Identification of this type of generally infrequent alleles (particularly if they are non-expressed alleles (i.e. resulting in lack of a functional HLA protein) or if they present any non-synonymous mutation across the coding region (and most importantly when affecting the PBR or ARD)) is very relevant in the clinical setting and particularly for assessment of donor-recipient matching in HSCT. Since misidentification of these type of alleles can have a negative impact on graft outcome and lead to a variety of transplant-related outcomes including: overall survival,



disease-free survival, graft rejection, acute and chronic graft-vs-host disease (GvHD), relapse and transplant-related mortality [208-210][458][523][524][869-871]. Furthermore, presence of a given null or rare allele (according to certain population/regional/ethnic group) may be a significant negative predictive value when searching for a full-match URD [208][209]. Therefore, accurate and unambiguous characterization and rapid detection of these null, rare and even novel HLA alleles by NGS high-throughput technologies is significantly contributing to minimize the potential for overlooked (especially when using limited and lower resolution typing methods (SSP, SSO, SBT or RT-PCR)) HLA mismatching on transplant outcome [302][523]. Nevertheless, traditional screening recommendations as well as testing guidelines for typing clinical-based decisions and accrediting regulations in URD registries (and, subsequently, their respective historical datasets) across the world have been mostly focused and delimited to the matching at 8/8 2-field high-resolution of *HLA-A*, *-B*, *-C* and *-DRB1* loci and have been also based on the sequencing of only those DNA encoding segments for their respective ARDs as the minimum standard in allogeneic HCT [214][479][523]. Mainly because, up to date, there have not been yet enough large-scale studies (with also an enough number of strong evidences) able to define how clinically relevant is the mismatching between alleles showing polymorphisms located outside the ARD region in the HSCT setting [217][218][527][529] as well as due to the predominance (until very recently with the implementation of NGS-based typing methods) of more cost-effective legacy molecular HLA typing methods (primarily SSO and/or SBT) [144][152]. Furthermore, both overrepresentation of European descent population groups and underrepresentation of other ethnic groups (African, Asian and Hispanic broad groups as well as considering the large list of respective regional sub-groups) in the majority of, if not all, URD registries is still a major limitation as well [179][221][223-227][474][476][480][523][525][526][554][943].

More representative large NGS HLA typing URD datasets (even though not phased at the haplotype level) may also facilitate assessment of not very well known LD patterns (due to the paucity of worldwide and multi-ethnic population studies until very recently) [101][523] displayed by these rare/null/novel alleles as well as to infer the most common extended bearing 3-/4-field haplotypes respectively. In fact, this current and future available NGS HLA information across worldwide URD registries (e.g. prevalence of null/rare/novel alleles (thus, defining CWD/CWID categories), most common coding positions/regions presenting these mutations and LD patterns data by ethnic group) in relation to a given population/region may be very instrumental for those H&I laboratories still with limited access or logistics for implementation of NGS-based HLA typing methods. As these H&I laboratories may be able to predict (or at least to suspect) presence of these null/rare/novel alleles based on these better documented LD patterns and also to have improved resources for guiding testing decisions and laboratory typing practices (e.g. permanently include exon 4 as part of the screening regimen for HLA class I alleles since this exon region has been defined as a common site outside of the ARD for a mutation resulting in a null allele to occur [523]).

In the present NGS HLA Spanish population study of a healthy cohort, these were the most likely or tentative (i.e. inferred via iterative EM algorithm and/or based on common LD described in large datasets [297]) extended bearing haplotypes respectively found:

-Novel allele-carrying haplotypes identified in the present study:

*HLA-*

*A\*29:02:01:01~C\*12:03:01:01~**B\*38:20:02**~DRB4\*01:03:01:01~DRB1\*07:01:01:01~  
DQB1\*02:02:01:01 (HF=0.19%);*

and

*HLA-*

*A\*11:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB3\*02:71~DRB1\*03:01:01:01~DQB1\*02:01:01* (HF=0.19%).

-Null allele-carrying haplotypes detected in the present study:

*HLA-A\*23:01:01:01~C\*04:09N~B\*44:03:01:01~DRB4\*01:01:01:01~DRB1\*07:01:01:01~DQB1\*02:02:01:01* (HF=0.38%), as similarly described for other European descent population groups in previous studies [523][871][877][880]. As an additional remark, it is worth to mention that, firstly, given the predominance and relatively high allele frequency distribution of *HLA-B\*44:03:01:01* over its counterpart *-B\*44:02:01:01* in Spanish (and Iberian [602]) population (e.g. [221][260][269][564][624]) and, conversely, significantly differing from other European ancestry populations across the world (e.g. [130][223][225-227][259][268][297][299][464][474][481][917]). And, secondly, also considering the very strong association displayed by *HLA-B\*44:03:01:01* allele with the null *HLA-C\*04:09N* allele ( $D' = 1.0$ , see **Table R-5**). Thus, the Iberian population could be potentially considered as a plausible good URD European population pool in order to find CWD/CWID *HLA-C\*04:09N*-bearing haplotypes and in contrast to other European populations presenting much lower occurrences of this null *HLA-C\*04:09N* allele [523][879].

Whereas *HLA-A\*02:01:01:01~C\*06:02:01:01~B\*57:01:01~*

*DRB4\*01:03:01:02N~DRB1\*07:01:01:01~DQB1\*03:03:02:01* (HF=0.38%) and

*HLA-A\*29:02:01:01~C\*04:01:01:01~B\*44:03:01:01~DRB4\*01:03:01:02N~*

*DRB1\*07:01:01:01~DQB1\*03:03:02:01* (HF=0.38%), were the most common haplotypes observed in the present Spanish population cohort encompassing this other null *HLA-DRB4*

allele. In this case, similar trends have been also observed for some other European ancestry populations [876][881][882]. Future and larger population studies focused on *HLA-DRB3/4/5* genes may also shed more light on the clinical relevance of these genes in transplantation setting [118].

-Rare allele-carrying haplotypes considered in the present study:

*HLA-A\*02:01:01:01~**C\*12:166**~B\*52:01:01:02~DRB5\*01:02~DRB1\*15:02:01:02~*

*DQB1\*06:01:01* (HF=0.19%), and as previously found in other Spanish individual [885];

and

*HLA-A\*02:05:01~C\*12:03:01:01~**B\*15:220**~DRB4\*01:03:01:01~DRB1\*07:01:01:01~*

*DQB1\*02:02:01:01* (HF=0.19%), as also similarly described in African descent individuals [886][887].

Certainly, future NGS HLA studies of larger population sample sizes at a wider geographic scale across the Iberian Peninsula will be needed to accurately detect and also to grasp the real prevalence of novel/null/rare alleles as well as to fully describe respective most common bearing haplotypes and segregation patterns in Spanish general population. In-depth knowledge of this population-specific HLA data will improve current local HLA-typing practices and criteria.

7) As mentioned before, assessment of extended HLA haplotypes diversity (and corresponding LD patterns) across worldwide populations and its geographical/regional variation significantly contributes in the analysis of tracking migrations of modern populations as well as gaining a better insight in anthropological studies [131-133]. NGS HLA data from the current Spanish unrelated population study (in combination with estimation of extended haplotypes via iterative EM algorithm) has also allowed us to identify certain

singular genetic traces [131-133] at very high-resolution and coverage across the HLA region that are present in modern-day Spanish population. These observed and distinctive genetic remnants correspond to unique HLA genetic signatures of particular ethnic backgrounds that, in turn, share similar parallel patterns in relation to their respective demographic history and anthropological influence in the Iberian Peninsula throughout its history and still at the present time. As previously described in the **INTRODUCTION** section of the present thesis work, there are two specific demographically non-negligible minorities that arrived to the Iberian Peninsula several centuries ago (starting from the Middle Ages or even prior to that period of time (e.g. written documents mentioning the presence of Jewish communities in Iberia accumulate from the beginning of the Visigoth period onward in the 4<sup>th</sup> century CE)) and are worth to be highlighted in this section of the discussion of the present thesis work:

-Sephardic Jews, where the majority of them as Crypto-Jews had embraced conversion to Catholic faith in order to avoid the 1492 Edict of Expulsion and who were socially well integrated and highly involved mainly in trading and banking economic activities (with a very heterogeneous social status) of the Iberian society [572][574].

-And Spanish Romani Gypsies, who, although with an important nomadic way of life, were mostly settled in Southern regions of the Iberian Peninsula where they were more socially accepted, being highly involved in the development of regional folkloric culture and who generally adopted and practiced Christian traditions [580][581]. Altogether, this facilitated their integration also in the Iberian society differing from other regions of Europe where these Romani itinerant communities were either persecuted and expelled or even enslaved in some instances along history [590][591].

Thus, these two communities (being both part of their respective historical Diasporas: Jewish Diaspora starting from Near-/Middle-East [918]; and Romani Diaspora tentatively starting from North India [594]) retained their continuity within the Iberian general population over past centuries, with a history characterized by geographical isolation and/or religious and sociocultural constraints against intermarriage but still with an important level of integration within the Iberian society and conversely, as a relevant example, to the case of “*moriscos*” (as North African Arab-Berber descendants), who were mostly (if not completely) expelled from the Iberian Peninsula starting in 1609 under the reign of Philip III of Spain [555]. Nonetheless, and at the same time, as it is described with further detail in the following points, due to the long period of time of their settlement in the Iberian Peninsula (about 8-10 centuries) there is a very significant series of genetic imprints and overall genetic substrate left by North African Arab-Berber background that also remains in modern-day Iberian populations [130][271][285][464][558][563-568][578][611-613][808][849][851][855].

In consonance with results and trends described in genetic studies based on genome-wide analyses and using different uniparental and/or biparental genetic markers (e.g. single nucleotide polymorphisms (SNPs), mitochondrial markers (mtDNA), Y-haplogroup or microsatellite markers) [582][585-588][592-594][918][919]. Previously reported low-resolution and limited sample size/gene coverage HLA studies have initially revealed some of these unique signatures (and plausible origins), although not completely still, for each of these two genetic backgrounds respectively: Sephardic/North African/Ashkenazi/Eastern Jewish (and also shared with those culturally non-Jewish but genetically related populations such as Arab populations originally from the Levant) [571][614-621][808][832]; and Spanish/European Romani Gypsies (and tentative founder population from North India) [589][757][758][920-925]. To the best of our knowledge, the present NGS HLA Spanish

population study has enabled the depiction at a very high-resolution 3-/4-field level for the first time of certain characteristic Sephardic Jewish and Spanish Romani Gypsies extended HLA haplotype signatures (and moreover, including the characterization of all 11 major HLA loci). In summary:

-Distinctive Sephardic Jewish, and in general Middle-Eastern (e.g. [621]),

*HLA-B\*35:02:01*

*(HLA-A\*24:02:01:01~C\*04:01:01:06~B\*35:02:01~DRB3\*02:02:01:02~*

*DRB1\*11:04:01~DQB1\*03:01:01:02 (HF=2.0%; in the present study [269][297])); and*

*HLA-B\*38:01:01*

*(HLA-A\*02:01:01:01~C\*12:03:01:01~B\*38:01:01~DRB3\*01:01:02:01~*

*DRB1\*13:01:01:01~DQB1\*06:03:01 (HF=1.0%; in the present study [269][297]))*

bearing extended haplotypes (that are previously and later discussed in more detail in the present thesis work) were found in this studied Spanish population cohort. These haplotypes and their singular LD patterns shown are very common in other populations of historical Jewish origin [614][615] as well as in highly admixed populations such those from the Americas, which present a relevant and complex Mediterranean/Middle-East substrate brought by Spaniards during the colonial period (15<sup>th</sup> -19<sup>th</sup> centuries) and that includes both Arab and Sephardic Jewish components [224][267][522][547]. Thus, in line with previous HLA studies, different main Jewish population groups (Sephardic/North African/Ashkenazi/Eastern Jewish among others) throughout the history and still currently share a common ancestral gene pool (dating back to their common origins more than 5700 years ago) but there are also dissimilarities (observed as genetic heterogeneity and diversity)

found between these Jewish populations due to their isolation for long periods of time at that given geographical location and a gene flow from local neighboring populations enriching that genetic variety [130][464][558][571][572][574-577][614-621][926][927].

-Secondly, the very distinctive Spanish Romani Gypsy (and most likely inferred via EM) full haplotype (9-locus) *HLA-A\*01:01:01:01~C\*15:02:01:01~B\*40:06:01:02~DRB3\*02:02:01:01~DRB1\*14:04:01~DQA1\*01:04:02~DQB1\*05:03:01:01~DPA1\*01:03:01:02~DPB1\*02:01:02* (HF=0.19%; in the present study) [757] was detected in the present Spanish population cohort, and described (as far as our knowledge) for the first time at this very high-resolution allele level. HLA data from previous studies in European Romani Gypsy population groups (e.g. high frequency distributions of HLA lineage alleles (and respective bearing haplotypes) underscored and highlighted in bold above in addition to other commonly found alleles such as *HLA-DRB1\*15:02*, -*DRB1\*16:01* or -*DRB1\*10:01* in Romani Gypsy populations and also different far related East Asian populations) [130][464][589][757][758][920-925] also supports the hypothesis of a common ancestral origin (from Northwestern India) of different nomadic European Romani Gypsy population groups, which maintained a relatively reduced polymorphism most likely due to the given founder effect and as a consequence of a high degree of endogamy and intramarriage throughout their itinerant history. At the same time, in a similar way to what it is observed with the Jewish Diaspora, also these different European Romani Gypsy population groups (e.g. from Andalusia (Southern Spain), Madrid (Central Spain), Czech Republic and Hungary) show certain level of HLA disparity due to the very likely genetic flow contribution from respective local neighboring populations although in a small extent.



These aforementioned examples illustrate, once again, how informative HLA polymorphism at the population-level can be to detect signals of human peopling and demographic history in modern populations [131-133]. Increased and more accurate knowledge of HLA allele/haplotype diversity and prevalence in these two specific demographically non-negligible minorities and genetic backgrounds within present-day Iberian general population may also have important clinical implications such as in the HSCT setting. Furthermore, future larger (both in sample size and at a wider geographical scale) NGS HLA anthropological studies will be also fundamental for subsequent epidemiological work aimed to elucidating the prevalence and association between HLA genes and related diseases as well as drug-induced hypersensitivity reactions more commonly found in these relatively closed and isolated population groups/genetic backgrounds (at least historically for many centuries) [577][758][922][926-929]. Thus, genetic screening in this kind of historically isolated population groups may be very informative, since there may be specific inherited disorders (relatively conserved due to high degree of endogamy and intramarriage practiced) that tend to be observed more frequently within that respective particular group than in the general population. As this clustering of disorders also reflects the common ancestral gene pool of the individuals within these peculiar population groups. At the same time, by studying such diseases within populations in which they are most common, it has been possible to better disentangle and identify the genes responsible for some of these disorders [927].

Finally, in addition to their relevant genetic, cultural and socioeconomic impact in the Iberian Peninsula throughout history and nowadays, it is also noteworthy that there has been recently also a remarkable institutional recognition of all these ethnic communities in Spain [930-932].

**8)** Another set of singular Spanish population groups, which also need to be remarked, are those related with territories geographically isolated, cutting off from the main landmass of

the Iberian Peninsula. These particular Spanish population groups are found respectively in the Balearic Islands (Majorca, Minorca, and Ibiza islands situated in the Mediterranean Sea) and in the Canary Islands (Santa Cruz de Tenerife, Fuerteventura, Gran Canaria, Lanzarote, La Palma, La Gomera, El Hierro and La Graciosa islands located off the North African Atlantic coast). Which, in turn, have experienced certain unique demographic events (and even, hypothetically, some indistinguishable natural selection events) and specific gene flows (also observed at the HLA system level) over the centuries that in some cases clearly differ (or are observed in a greater manner) from those identified across the Iberian Peninsula. Firstly, in the case of Balearic Islands, their strategic geographical location was from ancient times of great importance for the many maritime trading routes that crisscrossed the Mediterranean Sea. Consequently, important Jewish (Majorcan Jewish community known locally as “*Chuetas*”) and other related Near-/Middle-Eastern communities settled for relatively long periods of time and have left their genetic imprint in addition to the gene flow coming from main Iberian populations throughout the history [558][571][574][577][620][622][623]. Similarly, the Canary Islands show a parallel demographic history given their also unique geographical location. Here, despite the relatively predominant genetic resemblance of modern-day Canary Islanders to other Southern European and Mediterranean populations. Historical demographic events (which are reflected in the several genetic population backgrounds and substrates that have been detected) importantly comprise initial settlement and posterior complex admixture of first local indigenous people (who are thought to come originally from North African regions as they share a common ancestral gene pool with those autochthonous Berber North African populations) with sub-Saharan Africans and Europeans which, overall, have shaped a very distinctive genetic makeup [563-568][610-613]. Thus, this characteristic genetic landscape

across the Canary Islands has also relevant medical implications related to disease susceptibility and specifically in the context of allergy, where around 20% of the population presents allergy-related diseases [566][933].

Regarding the depiction of the HLA system and its diversity on these peculiar Balearic and Canary islander population groups respectively, it is noteworthy these previously reported largest HLA datasets in a main Balearic Islands population cohort [571] and, on the other hand, a Gran Canaria Island kidney transplant patient population cohort [564]. In summary:

-As for Balearic Islands population cohort from Crespí et al. study [571] (including Majorcan (N=407), Minorcan (N=94), Ibizan (N=88) populations, and also Majorcan “*Chuetas*” group of Jewish ancestry (N=103)). In general, the most common extended haplotypes (e.g. *HLA-A\*02:01~C\*07:01~B\*08:01~DRB1\*03:01~DQB1\*02:01* or *HLA-A\*29:02~C\*16:01~B\*44:03~DRB1\*07:01~DQB1\*02:02*) show a very similar frequency distribution to those observed in Spanish mainland (within the Iberian Peninsula) population [221][260][269][558][624]. Interestingly, it is also evident the significant Jewish (“*Chuetas*”) substratum present in modern-day Balearic Islands population with the presence of common *HLA-B\*35:02:01* and *HLA-B\*38:01:01* bearing extended haplotypes of Jewish background previously described [130][464][558][571][574][614-621].

-In the case of the Gran Canaria Island kidney transplant patient population cohort (n=215 unrelated subjects) [564], and in addition to the most typical Spanish Iberian HLA haplotypes that are jointly shared and have been already explained [221][260][269][558][624]. Strikingly, the most frequent extended haplotype *HLA-A\*33:01~C\*08:02~B\*14:02~DRB1\*03:01~DQB1\*02:01* (HF=3.50%) described in [564] it is not found as frequent in general Spanish population (e.g. in the present study (HF=0.98%; (data shown in [297])); or in Barcelona UCB Bank study [221] (HF=0.24%)).

In contrast, this haplotype (at least at the *HLA-B~C* association level) appears to be relatively common in Berber North African populations [130][464][611-613] (e.g. *HLA-A\*11:01~Cw\*08:02~B\*14:02~DRB1\*14:01*; (HF=1.6%) in [611]; even though it does not carry the same exact *HLA-A* and *-DRB1* alleles). Moreover, it is also interesting to remark an observed inverted or dissimilar ratio of haplotype frequency distribution in relation to three main *HLA-C\*08:02~B\*14:02* bearing haplotypes found between different Spanish population cohorts [221][269][297][564] due to plausible population substructure or stratification and regional variation previously reported [260]. Where:

*-HLA-A\*33:01~C\*08:02~B\*14:02~DRB1\*03:01~DQB1\*02:01*; with a value of HF=3.50% in Gran Canaria patient population cohort [564]; whereas HF=0.98% in the present study (data shown in [297]); and HF=0.24% in Barcelona UCB Bank study [221].

*-HLA-A\*33:01~C\*08:02~B\*14:02~DRB1\*01:02~DQB1\*05:01*; with a value of HF=1.9% in Gran Canaria patient population cohort [564]; whereas HF=1.96% in the present study (data shown in [297]); and HF=0.92% in Barcelona UCB Bank study [221].

*-And HLA-A\*02:01~C\*08:02~B\*14:02~DRB1\*03:01~DQB1\*02:01*; with a value of HF=0.47% in Gran Canaria patient population cohort [564]; whereas HF=0.98% in the present study (data shown in [297]); and HF=0.14% in Barcelona UCB Bank study [221].

Nevertheless, at the present thesis work these two peculiar insular Spanish population groups were not covered or not sufficiently (where only a very small sample size (n~25) from Gran Canaria island was tested). Therefore, future larger NGS HLA studies of these respective Balearic and Canary Islander populations, where considering each and every of their corresponding islands, will shed more light on the HLA genetic backgrounds and diversity as well as the level of relatedness and demographic history events (e.g. migration patterns and the peopling in each case) that shaped their modern-day genetic population makeup.

9) Referring now, more in detail, to the main similarities and common HLA haplotype signatures found between the current NGS HLA Spanish population study and related (as earlier explained: historically, culturally, geographically, and thus also genetically linked to Spanish population) foreign populations (previously reported), relative to the description of extended HLA haplotype frequency distributions (mainly focusing on extended haplotypes encompassing some of these loci or all *HLA-A~C~B~DRB3/4/5~DRB1~DQB1* (see **Table R-7** for results of the present Spanish population cohort)). The following points are particularly worth to be mentioned:

-In Spanish Americans cohort from Mack et al. study [772], Hispanic ethnic groups (mostly of Cuban and Mexican ancestries) in the U.S., which are reported in the NMDP BM U.S. registry [259][299], and other related studies of these same ethnic groups found in the U.S. also described in the AFND database [130][464]: common Spanish haplotypes, described in the present study, such as *HLA-A\*29:02~C\*16:01~B\*44:03~DRB1\*07:01* (e.g. (HF=2.3%) in [772]; (CARHIS-Caribbean Hispanic group, with HF=2.9%); (MSWHIS-Mexican or Chicano group, with HF=1.4%) [259][299]) or *HLA-A\*01:01~C\*07:01~B\*08:01~DRB1\*03:01* (e.g. (HF=2.1%) in [772]; (CARHIS-Caribbean Hispanic group, with HF=1.7%); (MSWHIS-Mexican or Chicano group, with HF=1.8%) [259][299])

are frequently found. Thus, the most common haplotypes in individuals of Latin American descent currently living in the United States (as similarly observed in other Latin American countries [132]) also present similar high frequencies in populations of European and, especially, of Spanish ancestry.

-With many similarities to the first above mentioned group, and as a representative example of what may be generally observed in the majority of South American populations (e.g.

registry populations) that show important and still current Spanish HLA genetic influence and background [130][464]; Argentinian donor registry population from Hurley et al. NGS HLA study [224] reported the following four most frequent haplotypes at the 3-/4-field allele resolution level:

*HLA-*

*A\*01:01:01:01~C\*07:01:01:01~B\*08:01:01:01/02~DRB1\*03:01:01~DQB1\*02:01:01*

(HF=2.8%), which presents an HF=8.0% (ranked #1) in the present study (including respective *HLA-DRB3/4/5* allele see **Table R-7**);

*HLA-A\*29:02:01:01~C\*16:01:01~B\*44:03:01:01~DRB1\*07:01:01~DQB1\*02:02:01*

(HF=1.9%), which presents an HF=7.5% (ranked #2) in the present study (including respective *HLA-DRB3/4/5* allele see **Table R-7**);

*HLA-*

*A\*03:01:01:01~C\*07:02:01:03~B\*07:02:01:01/03~DRB1\*15:01:01~DQB1\*06:02:01*

(HF=1.3%), which presents an HF=4.5% (ranked #4) in the present study (including respective *HLA-DRB3/4/5* allele see **Table R-7**);

and *HLA-*

*A\*33:01:01:01~C\*08:02:01:01~B\*14:02:01:01~DRB1\*01:02:01~DQB1\*05:01:01*

(HF=1.2%), which presents an HF=2.0% (ranked #7) in the present study (including respective *HLA-DRB3/4/5* allele see **Table R-7**).

When comparing extended HLA haplotype results of these two NGS HLA studies (Argentinian population) [224] (original Spanish population) [269], it is noteworthy and striking not only that they both share the same most common (in high frequency) extended HLA haplotypes but also how these distinctive 3-/4-field HLA haplotype associations (revealed by application of NGS technology) are highly conserved between both

populations. Thus, this exemplifies the strength of LD established and retained at the non-coding variation level as well as at the coding level within the HLA system and at the population-level, even in spite of, for instance, the incessant both recent and past extensive demographic events of population admixture and ongoing migration waves in Latin American countries [852][853].

-In addition to these two previous U.S. Hispanic and South American groups of HLA population studies. It is also of note the relatively high level of relatedness (based on the observed most common extended HLA haplotype distributions) between original Spanish general (mainland) population [269] and reported Mexican Admixed (Mestizo) populations (especially in regards to their genetic component of European (Caucasoid) descent) [130][267][464][522][547-552]. Where it is well-known that the complex and heterogeneous genetic background of modern-day Mexican Admixed (Mestizo) populations is a result of the recent past (i.e. mainly between XV and XX centuries) admixture of predominantly Native Indian (mainly Amerindian) and European (mainly Spanish) as well as, in a lesser extent, Black (mainly sub-Saharan African) and, even, Asian (mainly from South East Asia) genetic substrates where their contribution appear to vary according to the geographic location [130][267][464][522][547-552]. These genetic singularities are also observed in rest of Latin American populations [852][853]. Thus, considering Zúñiga et al. HLA study in N=234 non-related admixed Mexican individuals [547] as a very adequate representative example, some of the most common extended HLA haplotypes of European ancestry (which are found in high frequencies especially in original Spanish general (mainland) population [269] (see **Table R-7**, including respective *HLA-DRB3/4/5* allele)) are as follows:

*HLA-A\*02:01~C\*07:02~B\*07:02~DRB1\*15:01~DQB1\*06:02*, (HF=0.9%);

*HLA-A\*30:02~C\*05:01~B\*18:01~DRB1\*03:01~DQB1\*02:01*, (HF=0.6%);

*HLA-A\*01:01~C\*07:01~B\*08:01~DRB1\*03:01~DQB1\*02:01*, (HF~0.3%);

and, *HLA-A\*29:02~C\*16:01~B\*44:03~DRB1\*07:01~DQB1\*02:02*, (HF=0.4%).

Thus, in the case of Mexican Admixed populations (and as it similarly occurred in other Latin American countries), it is widely accepted that this currently observed (in the context of HLA here) genetic European (Caucasoid) ancestral component originally came principally from Spain. Where the Spaniards (i.e. mainly Spanish conquerors and colonizers from Spanish regions such as Andalucía, Leon, Extremadura, and the two Castillas), in addition to other minor European groups mainly from Portugal and Genoa (Italy), initially arrived to Mexico early in the 16<sup>th</sup> century and they rapidly and extensively settled all over this Central American territory (the so-called, at that time, Viceroyalty of the New Spain). Then, this initial arrival was continued by several massive migration waves of more colonizers (although not only Spanish but also, in a lesser extent, French, German, and English groups as well) during the 17<sup>th</sup> century and it prevailed through the next two centuries[130][132][267][464][522][547-552].

At the same time, still in relation to this genetic component of European (Caucasoid) ancestry, there are also striking Near-/Middle-Eastern HLA genetic signatures or features (e.g. typically found, as some examples, in Lebanese population [621] as well as in both Ashkenazi and non-Ashkenazi (e.g. Sephardic) Jewish communities/regions [130][132][464][614]) also presented by modern-day Mexican Admixed populations (that, indeed, are also shown by modern-day Hispanic ethnic groups in the U.S. [259][299] as well as by original Spanish general population [269]) which are defined by the presence of, as some main examples, these three relatively frequent and distinctive (fairly conserved across worldwide populations) HLA haplotypes:



Firstly,

*HLA-A\*24:02~C\*04:01~B\*35:02~DRB3\*02:02~DRB1\*11:04~DQB1\*03:01*, where:

HF=2.5%, in Lebanese population cohort (N=426) [621];

HF=1.7%, in Hadassah donor registry in Jerusalem (Israel) (N= 55,801) [614];

HF=0.8%, in CARHIS-Caribbean Hispanic group in the U.S. donor registry [259][299];

HF=0.5%, in MSWHIS-Mexican or Chicano group in the U.S. donor registry [259][299];

HF=0.4%, in Mexican Admixed (Mestizo) cohort study [547];

HF=2.0%, in the present Spanish population study, (ranked #7) (see **Table R-7**, including respective *HLA-DRB3/4/5* allele).

This haplotype is also present in other populations of European ancestry but it is found, in general and comparatively, even less frequently [130][297][464][481][621]. Whereas, comparatively, low or very low frequencies of this haplotype are found in populations of African and Asian ancestries [130][297][464][481][621].

Secondly, this other (although not as conserved and frequent) haplotype

*HLA-A\*02:01~C\*12:03~B\*38:01~DRB3\*01:01~DRB1\*13:01~DQB1\*06:03*, where:

HF=0.2%, in Hadassah donor registry in Jerusalem (Israel) (N= 55,801) [614];

HF=0.05%, in Hispanic group, as a whole, (including both CARHIS-Caribbean and MSWHIS-Mexican or Chicano ethnic groups) from the U.S. donor registry [259][299];

HF=1.0%, in the present Spanish population study, (ranked #9) (data not shown, see [297]);

whereas, this specific extended haplotypic association has not been reported in Lebanese population cohort (N=426) [621]; instead, the singular *HLA-B\*38:01* allele is found displaying these other more frequent haplotype associations: *HLA-A\*26:01~C\*12:03~B\*38:01~DRB4\*01:03~DRB1\*04:02~DQB1\*03:02* (HF < 0.01%)

and *HLA-A\*26:01~C\*12:03~B\*38:01~DRB3\*02:02~DRB1\*11:04~DQB1\*03:01* (HF=1.3%); and similarly, the haplotype association trends observed in the case of reported Mexican Admixed (Mestizo) cohort study [547] are the following: *HLA-B\*38:01~C\*12:03~DRB1\*04:02~DQB1\*03:02* (HF=0.4%) and *HLA-B\*38:01~C\*12:03~DRB1\*07:01~DQB1\*02:02* (HF=0.4%). As a matter of fact, these *HLA-B\*38:01* carrying extended haplotypes are also present and are relatively common in Jewish population HLA dataset from the Israeli Hadassah donor registry [614], containing respectively *HLA-: "DRB1\*04:02"* (HF=2.9%); "*DRB1\*11:04*" (HF=0.1%); and "*DRB1\*07:01*" (HF=0.3%).

In addition, the

*HLA-A\*33:01~C\*08:02~B\*14:02~DRB3/4/5\*Absent~DRB1\*01:02~DQB1\*05:01*

haplotype, where:

HF=1.6%, in Lebanese population cohort [621];

HF=1.0%, in Hadassah donor registry in Jerusalem (Israel) [614];

HF=0.6%, in CARHIS-Caribbean Hispanic group in the U.S. donor registry [259][299];

HF=0.8%, in MSWHIS-Mexican or Chicano group in the U.S. donor registry [259][299];

HF=0.4%, in Mexican Admixed (Mestizo) cohort study [547];

HF=2.0%, in the present Spanish population study, (also ranked #7) (see **Table R-7**, including respective *HLA-DRB3/4/5* allele and [297] for further details).

In this case, this haplotype is similarly common in other populations of European ancestry [130][297][464][481][621]. In contrast, comparatively, low or very low frequencies of this haplotype are found in populations of African and, especially, Asian origins [130][297][464][481][621].

Hence, the relatively common presence of these Near-/Middle-Eastern haplotypes in modern-day Mexican Admixed populations could be identifying (at least partially) the original contribution of the Sephardic Jewish (e.g. coming from the Iberian Peninsula, the Near/Middle East, and North Africa) migration waves to the New World (i.e. the Americas) due to relevant historical facts such as the Edict of Expulsion (issued by the joint Catholic Monarchs of Spain) of (Sephardic) Jews from Spain at the end of the 15<sup>th</sup> century [132][572][575][576].

Furthermore, another distinctive and interesting Eurasian-Mediterranean group of haplotypes is the one that carries the allele pair *HLA-DRB1\*14:54~DQB1\*05:03*, where one of the most typical and most common extended haplotype distribution detected is the following:

*HLA-A\*02:01:01:01~C\*12:03:01:01~B\*35:03:01~DRB3\*02:02:01:01~*

*DRB1\*14:54:01~DQB1\*05:03:01:01*, with HF=1.5% and here shown at the allele resolution level (according to released version 3.25.0 of IPD-IMGT/HLA database) as obtained in the present Spanish population study. Where, as some other examples, alternative allele pairs such as *HLA-C\*01:02:01~B\*27:05:02* or *HLA-C\*07:01:01:01~B\*18:01:01:02* may be also found, but in a lesser frequency, within this same extended haplotype. Thus, despite this above mentioned extended haplotype is not very commonly detected, it has been also more frequently observed in other Spanish population cohorts as well as in Mediterranean-European, Near-/Middle-Eastern, Mexican Admixed, Latin American and Hispanic population groups reported so far (e.g. [130][221][224][297][464][522][547-552]). Nevertheless, it should be noted that the relative allele prevalence and frequency (and, thus, associated haplotype frequency distributions at the population-level) of *HLA-DRB1\*14:54:01* may be still underestimated

and not well-defined since it has not been routinely characterized by legacy HLA typing methods (i.e. only covering exon 2) in the majority of previous studies (and with very few exceptions [823]). Consequently, it may have been usually considered and included under the respective broad, non-specific and ambiguous *HLA-DRB1\*14:01:01* allele group. Where, *HLA-DRB1\*14:01:01* (Tyr (codon 112 (TAC), exon 3)) and *HLA-DRB1\*14:54:01* (His (codon 112 (CAC), exon 3)) alleles only differ in one single position in exon 3 as indicated here. Thus, as previously commented, a more widespread application of NGS technology for HLA genotyping may allow a much more accurate assessment of allele prevalence and frequency (and, thus, of associated haplotype frequency distributions and at the population-level as well) of all possible HLA coding (as well as non-coding) variants as the given example here for *HLA-DRB1\*14:54:01* allele (which, in fact, appears to be more prevalent than the respective *HLA-DRB1\*14:01:01* allele in reported worldwide populations (e.g. [297])).

Lastly, as part of this discussion sub-section relative to Mexican Admixed (Mestizo) populations and its HLA genetic relatedness with original Spanish general population (as the examined cohort of the present study). It is also noteworthy that an important portion of the most common haplotypes described in Mexican Admixed (Mestizo) populations [130][132][267][464][522][547-552] (as it is also observed in South American populations [130][224][464] and other Hispanic ethnic groups [130][259][299][464] respectively) contain alleles uniquely found (considered “quasi-specific”) in corresponding regional and characteristic indigenous/native American (mainly Amerindian) ethnic groups. Therefore, overall, modern-day Latin American populations present a complex, stratified and highly variable HLA allele and haplotype frequency distribution. Firstly, owing to modern and incessant population admixture between neighboring regions/countries but also with

important immigrant foreign groups (in large and several migration waves that have been occurring in Latin American countries during these last centuries) from distant regions of the world (thus including European, African and Asian migrant groups) that have been shaping and influencing on Amerindian ancestral population genetic proportions, which have been also fluctuating through time. And secondly, due to the heterogeneous and still not well-defined origin of this just mentioned intricate cluster of ancestral Native American populations/groups (and their respective genetic substrates) that originally comprised and described all the different North and South American Indian groups [852][853]. Where the main and most accepted theories (supported by different genetic markers, including HLA, as well as extensive anthropological evidences) about their origins and the peopling of the Americas are based on: 1) classical three-waves theory from Asia through the Bering land bridge (Amerindians (most North and South American Indians; 12,000 years BCE), Na-Dene (Athabascans, Navajo, Apache; 8,000 years BCE) and Eskimo-Aleuts (6,000 years BCE)); and also 2) Trans-Pacific route (even considered two-way) of American peopling from Asia and Polynesia, where even prehistoric contacts between Amerindians and Pacific Islanders are also (as in the case of the 1) classical three-waves theory) strongly suggested by genetic data (including HLA) as well as anthropological, archaeological, linguistic, and other cultural traits [934-938]. Thus, these singular HLA haplotypes of diverse Native American backgrounds are not only found in respective original isolated (some of them still nowadays) indigenous ethnic groups but these haplotypes are also importantly and commonly (i.e. found in relative high frequencies) present in modern-day admixed Latin American general populations. In this sense, this has important implications on health-related issues (e.g. transplantation, pharmacogenetics or immunotherapies depending on HLA profiling) at the population-level, where it needs to be considered not only the given

major local and original Latin American population groups but also those respective emigrant Latin American population communities around the world [600][852][939]. Nevertheless, the paucity of high resolution HLA typing data is still, and especially, manifest in all those present-day descendant groups of the originally indigenous Native (North and South) American groups (with very few exceptions; e.g. [130][224][267][464][471][504][938]). In this regard, and in the context of Spanish population, it is well-known the recent (i.e. especially for the past ~20-30 years) and significant series of demographic events of Latin American population groups emigrating to Europe, mainly due to socioeconomic factors, and especially to Spain because of evident cultural and linguistic resemblances. Where, out of the 13% of the most current Spanish population census (from last 2019 publication) which comprises the percentage of immigrant population (including main components such as European (4.7%) and African (2.1%) immigrant groups), approximately 5.1% corresponds to the Central, Caribbean and Southern American group, and thus it represents the largest immigrant population group in Spain [260][595][600]. Therefore, this important demographic factor in the current Spanish general population has led, as previously reported [600], to further evaluate and to attempt the development of virtual transplantation waiting lists (by carrying out HLA profiling (where future studies using NGS technology may have a great positive impact) of representative cohorts of these Central, Caribbean and Southern American immigrants subgroups across Spain) which may be useful especially in the HSCT setting for therapeutic uses and as part of plausible worldwide transplantation programs. In addition, this may also serve for specific epidemiology programs on HLA-linked diseases and drug hypersensitivities in Latin American (including both migrant as well as original) population groups. Interestingly, in the present study (although it has not been previously discussed in

the previous **DISCUSSION** sub-section regarding HLA allele distributions analyses of the present Spanish population cohort, since it was better described at the haplotype level) it can be noted that two haplotypes carrying *characteristic* Native American alleles (and also distinctive allele pairs in strong LD) were also identified (yet at very low (and probably not very accurate) estimated frequencies). The most likely extended haplotypes (defined here at the 3-/4-field) are as follows:

***HLA-A\*68:17~C\*03:04:01:02~B\*40:02:01~DRB4\*01:03:01:01~DRB1\*04:04:01~DQA1\*03:01:01~DQB1\*03:02:01***; (HF=0.2%, in the present Spanish population study).

Where *HLA-A\*68:17* allele was firstly identified in Kolla Amerindians of North-West Argentina [940][941]. Moreover, presence of alleles (and respective carrying haplotypes) such as *HLA-B\*40:02:01* and *-DRB1\*04:04:01* (very typically linked to *-DQB1\*03:02:01* allele) have been also commonly found in many different original Amerindian populations from Mexico, Mesoamerica and South America [130][464][504][600][934-938] as well as reported on previous studies of Amerindian immigrants in Spain [600].

And also, ***HLA-A\*03:01:01:01~C\*04:01:01:01~B\*40:02:01~DRB4\*01:03:01:01~DRB1\*04:07:01~DQA1\*05:01:01:02~DQB1\*03:01:01:01***; (HF=0.2%, in the present Spanish population study). Where, once again, presence of alleles (and respective carrying haplotypes) such as *HLA-B\*40:02:01* and *-DRB1\*04:07:01* (also linked to *-DQB1\*03:01:01:01* allele) have been commonly found in many different original Amerindian populations from Mexico, Mesoamerica and South America [130][464][504][934-938] as well as reported on previous studies of Amerindian immigrants in Spain [600].

Consequently, this illustrates (despite of the discrete sample size examined in the current study) the relatively significant (although in low frequencies) HLA genetic contribution and presence of this Native American genetic substrate within current modern-day Spanish general population, which adds to the observed significant HLA diversity in this population and to the complexity of population stratification (being even more clear the evidence of the diverse sub-populations that inhabit Spain today) and regional variation within the Iberian Peninsula previously mentioned [260].

-Referring now to the also relevant and distinctive HLA genetic relatedness (previously commented at the HLA allelic and 2-locus haplotype levels, and now here based on the observed common extended HLA haplotype distributions) shown between original Spanish general population and some specific neighboring foreign populations from countries that are geographically situated very close to the Spanish territory. It is noteworthy the particular case of both modern-day Portuguese and North African populations. Where, in contrast to other neighboring populations to Spain (such as the French population, which is more genetically related to other North-Central European populations), Portuguese and North African Berber (which, in turn, are also partially related with original Muslim Eastern Arab populations) populations share with the Spaniards a unique series of well-documented major demographic history events [555][556][569][578][808][851] which might have contributed and shaped (at least until certain extent) the currently observed similar HLA gene pools of these populations [130][221][269][464][558][823]. Thus, as previously discussed, the striking example of this singular *HLA-DQB1\*03:19:01* allele (and the characteristic LD patterns that displays) [908] may illustrate the distinctive relatedness (at least to some relative extent) of Iberian populations (clearly differing from other populations of European ancestry so far described (e.g. [474])) [221][269][558][602][823] with African descent and



local populations (in this case from both North and sub-Saharan regions) [130][297][464][823][909] and, in turn, also Muslim Eastern Arab (originally from the Arabian Peninsula) [130][271][285][464][910] descent and native population groups. In the present study, these are the tentative (i.e. EM estimated) *HLA-DQB1\*03:19:01* carrying extended haplotypes detected in Spanish population:

Firstly, according to the more frequently observed association with the *HLA-DRB1\*11:02:01* allele:

*HLA-A\*02:05:01~C\*07:01:01:01~B\*49:01:01~DRB3\*02:24~DRB1\*11:02:01~DQA1\*05:01:01:02~DQB1\*03:19:01*; (HF=0.2%, in the present Spanish population study).

*HLA-*

*A\*24:02:01:01~C\*07:01:01:01~B\*49:01:01~DRB3\*02:02:01:01~DRB1\*11:02:01~DQA1\*05:05:01:01~DQB1\*03:19:01*; (HF=0.2%, in the present Spanish population study).

*HLA-*

*A\*26:01:01:01~C\*07:01:01:01~B\*49:01:01~DRB3\*02:02:01:01~DRB1\*11:02:01~DQA1\*05:05:01:01~DQB1\*03:19:01*; (HF=0.2%, in the present Spanish population study).

Secondly, according in this case to the more rare (not as frequent) association with *HLA-DRB1\*13:04* allele:

*HLA-*

*A\*02:01:01:01~C\*12:03:01:01~B\*18:01:01:02~DRB3\*02:02:01:01~DRB1\*13:04~DQA1\*05:05:01:01~DQB1\*03:19:01*; (HF=0.2%, in the present Spanish population study).

Hence, this may exemplify (despite of the discrete sample size examined in the current Spanish population study) the relatively significant and conserved (although in low frequencies) HLA genetic contribution of these North African Berber (indigenous) and Muslim Eastern Arab (settlers, originally coming from the Arabian Peninsula) genetic substrates within present-day Spanish general population and in consonance with very well-documented historical facts (most likely as a result of an ancient pre-Neolithic contribution from northward Saharan migration during hyper-arid conditions; and also due to the later Islamic conquest period along the North African region between 7th–9th centuries before arriving to the Iberian Peninsula) [555][556][558][578][613][808][851]. At the same time, it should be noted that (as previously mentioned for other examples such as *HLA-DPB1\*03:01:01/-DPB1\*104:01* or *HLA-DRB1\*14:01:01/-DRB1\*14:54:01*): *HLA-DQB1\*03:01:01* (Thr (codon 185 (ACC), exon 3)) and *HLA-DQB1\*03:19:01* (Ile (codon 185 (ATC), exon 3)) alleles only differ in one single position in exon 3 (which is not routinely characterized by legacy HLA typing methods). Thus, future population studies with a more widespread application of NGS technology for HLA genotyping may also allow a much more accurate assessment of allele prevalence and frequency of this singular *HLA-DQB1\*03:19:01* allele.

-Lastly, in relation to the present Spanish population cohort shown here [269] and in comparison to foreign populations concretely of European ancestry (where Pingel et al. study [481], certain main European/North American population groups reported in the AFND database [130][464], the U.S. NMDP registry database (EURCAU-European Caucasian unrelated donors group; N=1,242,890) [259][299] and also a recent NGS HLA study in a large European American population cohort of the U.S. [268][297] can be considered as some main representative references that show the most common extended

HLA haplotypes found in this case). These are the three main remarks that are worth to be mentioned:

- Firstly, some of the most common extended HLA haplotype distributions (at the 3-/4-field) described in the present Spanish population cohort study (as well as in other previously reported lower resolution HLA studies carried out in Spanish population (e.g. [221][624])) such as:

*HLA-A\*01:01:01:01~C\*07:01:01:01~B\*08:01:01:01~DRB3\*01:01:02:01~*

*DRB1\*03:01:01:01~DQB1\*02:01:01;* (HF=8.0%, in the present Spanish population study (ranked #1) (see **Table R-7**); and, as a representative example: HF=9.3%, ranked #1 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]);

*HLA-*

*A\*03:01:01:01~C\*07:02:01:03~B\*07:02:01~DRB5\*01:01:01~DRB1\*15:01:01:01~*

*DQB1\*06:02:01;* (HF=4.5%, in the present Spanish population study (ranked #4) (see **Table R-7**); and, as a representative example: HF=5.2%, ranked #2 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]);

and, also

*HLA-*

*A\*02:01:01:01~C\*07:02:01:03~B\*07:02:01~DRB5\*01:01:01~DRB1\*15:01:01:01~*

*DQB1\*06:02:01;* (HF=5.0%, in the present Spanish population study (ranked #3) (see **Table R-7**); and, as a representative example: HF=2.5%, ranked #4 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]);

are also very common (i.e. found in high frequencies) in main European and North American (as part of the very broad and diverse Caucasoid ethnic group of populations) unrelated donor registries and various related large population datasets/studies reported so far. Thus, this could be illustrating (tentatively and at least to some extent) the relevant and specific HLA gene flow episodes that occurred from North-Central Europe to the Iberian Peninsula due to major demographic history events in the past. Initially, with the presence of Germanic tribes in the Iberian Peninsula starting from the 5<sup>th</sup> century in the CE after the collapse of the (Western) Roman Empire. And subsequently, with the presence of respective descendant Christian Visigoth population groups, especially after the completion of the “*Reconquista*” (against the North African Muslim Arab-Berber populations groups for ruling the entire territory of the Iberian Peninsula) at the end of the 15<sup>th</sup> century. Consequently, these descendant Christian Visigoth population groups became the main genetic substrate of the more modern Iberian general population [555][556][558][578][602].

- Secondly, as one of the main dissimilarities observed between reported Iberian populations (including both Spanish (e.g. [130][221][269][464][624]) and Portuguese populations [602]) and other populations of (Northern-Central-Eastern) European descent (e.g. [130][223][225-227][259][268][297][299][464][474][481][917]), it should be noted the very striking findings that are particularly related to the respective *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* carrying extended HLA haplotype frequency distributions detected. Where, following the same previous main comparison example between the present NGS HLA Spanish population cohort study [269][297] and the corresponding NGS HLA European American cohort study in the U.S. [268][297] as well

as when considering other reported main worldwide populations [130][464][614][466][831][832]:

-Relative to *HLA-B\*44:02:01:01* carrying extended HLA haplotype frequency distributions. In Spanish population these specific extended haplotypes are found in much lower relative frequencies and in a more spread distribution than in other reported populations of European ancestry (e.g. [130][223][225-227][259][268][297][299][464][474][481][917]) and also, as far as our knowledge after reviewing reported studies in the literature, than in some Near-Eastern (Jewish and Arab populations of the Levant) [466][614][621][832] and Middle-Eastern [831] population cohorts. In detail, these are the main and the most likely (i.e. EM estimated) *HLA-B\*44:02:01:01* carrying extended haplotypes found in the present Spanish population study:

*HLA-A\*02:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB5\*01:01:01~*

*DRB1\*15:01:01:01~DQB1\*06:02:01* haplotype (HF=1.0%, in the present Spanish population study (ranked #9) (data shown in [297]); and, as a representative example: HF=0.3%, ranked #28 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]).

*HLA-A\*02:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB3\*02:02:01:02~*

*DRB1\*12:01:01:03~DQB1\*03:01:01:01* haplotype (HF=1.0%, in the present Spanish population study (ranked #9) (data shown in [297]); and, as a representative example: HF=0.2%, ranked #30 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]).

And *HLA-A\*02:01:01:01~C\*12:03:01:01~B\*44:02:01:01~DRB3\*01:01:02:01~*

*DRB1\*03:01:01:01~DQB1\*02:01:01* haplotype (HF=1.0%, in the present Spanish

population study (ranked #9) (data shown in [297]); whereas this haplotype, as far as our knowledge and based on our review of certain main HLA references of reported populations of European ancestry [130][223][225-227][259][268][297][299][464][474][481][917], does not seem to be as frequently present in populations of European descent).

Conversely, among the *HLA-B\*44:02:01:01* carrying extended HLA haplotypes that are most commonly found, the extended haplotype *HLA-A\*02:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB4\*01:03:01:01~DRB1\*04:01:01:01~DQB1\*03:01:01:01* shows a predominant frequency distribution (HF=3.5%, ranked #3) in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297] (as a representative example of populations of European ancestry). Whereas, this same *HLA-B\*44:02:01:01* carrying extended haplotype seems to be hardly present (with very low HF values described so far) in Spanish population. For instance, it is not detected in the present Spanish population cohort study (data shown in [297]) while in a different recent Spanish population study (“Barcelona UCB Bank sample cohort”) [130][221][464] shows a very low HF value of 0.45%.

-Referring now to *HLA-B\*44:03:01:01* carrying extended HLA haplotype frequency distributions, an inverse situation is found in this case. In the present Spanish population study, and as similarly reported in other previous studies on populations of the Iberian Peninsula (e.g. [130][221][269][464][602][624]), *HLA-B\*44:03:01:01* carrying extended HLA haplotypes are found in much higher relative frequencies in comparison to those frequency distributions described in other reported populations of European ancestry (e.g. [130][223][225-227][259][268][297][299][464][474][481][917]). In

closer detail, these are the main and most likely (i.e. EM estimated) *HLA-B\*44:03:01:01* bearing extended haplotypes found in the present Spanish population study:

*HLA-A\*29:02:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~*

*DRB1\*07:01:01:01~DQB1\*02:02:01:01* haplotype (HF=7.5%, in the present Spanish population study (ranked #2) (data also shown in [297]); and, as a representative example: HF=1.8%, ranked #5 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]).

*HLA-A\*23:01:01:01~C\*04:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~*

*DRB1\*07:01:01:01~DQB1\*02:02:01:01* haplotype (HF=2.0%, in the present Spanish population study (ranked #7) (data also shown in [297]); and, as a representative example: HF=0.6%, ranked #18 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]).

*HLA-A\*02:01:01:01~C\*16:01:01:01~B\*44:03:01:01~DRB4\*01:01:01:01~*

*DRB1\*07:01:01:01~DQB1\*02:02:01:01* haplotype (HF=1.0%, in the present Spanish population study (ranked #9) (data also shown in [297]); and, as a representative example: HF=0.2%, ranked #30 in the respective NGS HLA study of a large European American population cohort in the U.S. [268][297]).

Interestingly, some additional striking and representative examples are also noteworthy in relation to these distinctive dissimilarities found in *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* carrying extended HLA haplotype frequency distributions that are observed when comparing Iberian populations and the rest of populations of European ancestry:

-As a first example, and as an additional evidence of the significant genetic imprint left by the Spanish settlers in modern-day Latin American populations and Hispanic

ethnic groups. The extended haplotype frequency distributions shown on Hurley et al. NGS HLA study of the Argentinian donor registry population [224] exemplifies this characteristic trend of *HLA-B\*44:03/HLA-B\*44:02* carrying extended haplotype frequency ratio same as the one found in the populations from the Iberian Peninsula. Where, as an illustrative example, it can be observed a considerably higher HF value for *HLA-*

*A\*29:02:01:01~C\*16:01:01~B\*44:03:01:01~DRB1\*07:01:01~DQB1\*02:02:01*

haplotype (HF=1.9%) than its counterparts *HLA-*

*A\*02:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB1\*13:01:01~DQB1\*06:03:01*

haplotype (HF=0.5%) or *HLA-*

*A\*02:01:01:01~C\*05:01:01:02~B\*44:02:01:01~DRB1\*07:01:01~DQB1\*02:02:01*

haplotype (HF=0.4%).

-As a second remarkable example, in consonance with what has been already described at the allele level and 2-locus *HLA-B~C* haplotype frequency distribution, reported extended *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* carrying extended HLA haplotype frequency distributions in worldwide populations (and in addition to the aforementioned *HLA-B\*51:01:01:01*, *HLA-B\*49:01:01*, *HLA-DQB1\*03:19:01*, *HLA-DRB1\*01:03~DQB1\*05:01:01:03* genetic allele/carrying haplotype signatures) also point to the very plausible genetic influence (and, thus, relatively significant genetic relatedness) of both North African Berber (indigenous) [130][464][611-613][849][855] and Muslim Eastern Arab (settlers, originally coming from the Arabian Peninsula) [130][271][285][464][808] genetic substrates in Iberian populations' genetic pool (e.g. [221][269][602][624][834][835]). For instance, characteristic Iberian extended haplotypes *HLA-A\*29:02~C\*16:01~B\*44:03* either



typically associated with *DRB1\*15:01~DQB1\*06:02* or *DRB1\*07:01~DQB1\*02:02* class II haplotypes are found in relatively high frequencies in Tunisian and Moroccan population groups from North Africa as well as studied population cohorts of Saudi Arabian or Kuwaiti descent from the Arabian Peninsula. It is also noteworthy another haplotype *HLA-A\*30:02~C\*05:01~B\*18:01~DRB1\*03:01~DQB1\*02:01* which is also commonly found in North African and in Mediterranean populations (particularly in the Iberian Peninsula). In parallel, and interestingly, it can be also observed a clear distinction and lower level of HLA relatedness (also according to this *HLA-B\*44:02:01:01/HLA-B\*44:03:01:01* ratio of distributions) between North African Berber/Muslim Arabian Peninsula groups and Arabs of the Levant [808][851]; similarly to what it is observed on how Iberian populations show distinctive and unique HLA genetic signatures not commonly found in other populations of European ancestry in Northern and Central regions of Europe (e.g. [130][223][225-227][259][268][297][299][464][474][481]).

Therefore, in accordance with previously reported HLA population studies (in spite of presenting lower HLA resolution and small limited sample sizes describing only few different population cohorts) [819][822][823][834][835], the comparison and analysis made and discussed in this section of the present thesis work have led us to three main conclusions to be noted:

- 1) In line with the previous observations remarked at the *HLA-B\*44:02/B\*44:03* allele frequency ratio level and also regarding respective 2-locus haplotype distributions, there is a notably diverse and differentiated distribution of these corresponding *HLA-B\*44:02/B\*44:03* carrying extended haplotypes found across worldwide human populations. Moreover, there are also significant and striking differences observed within

the Caucasoid (i.e. European ancestry) ethnic group of populations, where two main population groups can be distinguished: Iberian populations (presenting *HLA-B\*44:03* prevalence over *HLA-B\*44:02*) and rest of populations of European descent (showing the opposite trend). Thus, this exemplifies the evident and significant regional variation of the extended HLA haplotype diversity and, in turn, of the respective haplotypic distributions observed within the European (Caucasoid) populations, as described (attending to other HLA characteristics) in other previous studies [136].

2) At the same time, and as previously remarked in Santos et al study [834], despite these existing differentiated *HLA-B\*44:02/B\*44:03* carrying extended haplotype distributions within Caucasoid ethnic group of populations, the respective observed haplotype associations (i.e. respective LD patterns) given for *HLA-B\*44:02* and *-B\*44:03* are relatively well-conserved and similar among these different Caucasoid populations so far studied. Whereas at the worldwide population level, when main general European (Caucasoid), African, Asian, Hispanic and Near-/Middle-Eastern ethnic groups (ancestries) of populations are compared, some additional levels of diversity and unique/specific HLA haplotype associations can be found in each given broad ethnic group and, at the same time, also defining sets of certain more related ethnic groups of populations.

3) Finally, in this particular case evaluated, this detected existence of differential *HLA-B\*44:02/B\*44:03* carrying extended haplotype distributions among worldwide populations (and even observed within sub-regions/sub-groups of the same considered broad ethnic group, as described here in the case of Caucasoid populations) may be originated by both natural selective pressures (taking also into account the particularly distinctive functional micropolymorphism shown by these *HLA-B\*44:02/B\*44:03*

subtypes [811]) and stochastic demographic events (e.g. migrations that may be associated with either HLA gene flows/drifts, founder effects or bottlenecks). Which, in turn, cannot be always and totally distinguished owing to the unique functional nature, complex genetic features and particular inheritance patterns presented by the HLA system [104][132][542][543].

- Thirdly, it is particularly striking that the most common haplotype in North European and European American populations (extended haplotype at the 4-field level, *HLA-A\*01:01:01:01~C\*07:01:01:01~B\*08:01:01:01~DRB3\*01:01:02:01~DRB1\*03:01:01:01~DQB1\*02:01:01*) (e.g. (HF=9.3%), [268]), although commonly found, it is not as highly frequent in Spanish population (e.g. (HF=8.0%) [269]. In fact, Spanish population (and, overall, Iberian populations [260][602][603]) seems to show a HLA haplotypic diversity with a distinctive and more spread haplotype frequency distribution (i.e. presenting other different haplotypes almost equally prevalent as this one) in contrast to these other populations of European ancestry (e.g. [130][223][225-227][259][268][297][299][464][474][481][917]). This may be also reflecting a singular combination of natural selective pressures and historical demographic events that took place in the Iberian Peninsula different from the ones that may have shaped HLA haplotype diversity in other populations of European descent.

In summary, and as previously mentioned, all these described similarities found between original Spanish general population and other foreign related major populations (e.g. originally from Latin American countries, North African/Muslim Eastern Arab descent or of European ancestry) regarding HLA allele and haplotype frequency distributions definitely maximize the likelihood of finding compatible and highly matched unrelated donors (URD) for given Spanish patients (and vice versa) in the HSCT setting [215][220][529][896]. At the

same time, these HLA haplotype similarities may be of importance, and applicable to all those related populations as described here with a common Spanish/Iberian HLA genetic substrate, in current and future pharmacogenetics investigations, HLA-disease association and related epidemiology studies as well as discovery of novel HLA-based immunotherapies and development of haplo-banks of iPSCs [137][207][221][222][488-495][545][546].

### **3. FINAL REMARKS**

In the present study, 3-/4-field (obtaining the highest allele resolution level with minimum allele and phase ambiguities) HLA allele and extended haplotype frequencies for a relatively representative Spanish population healthy cohort (N=282 subjects, denominated 17th-IHIW Spanish population cohort) have been described. To date, and at an unprecedented scale, this can be also considered the largest study ever done in Spanish healthy population involving 3-/4-field HLA genotype data for all the 11 major classical HLA class I and class II loci (*HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5*) as well as respective allele and full extended haplotype frequency distributions. The current NGS HLA genotype dataset (including both allele and extended haplotype frequency distributions) may serve as a significant and useful reference source for multiple clinical and research applications in the histocompatibility and immunogenetics field and, in particular, in the context of Spanish population. In summary:

- In the SOT setting and as previously mentioned in [252][256-258][260][829], this very high-resolution allele and haplotype frequencies dataset may be instrumental and greatly informative as a first reference source for improved virtual panel reactive antibody (vPRA) calculations in Spanish population as well as for performing respective virtual crossmatching (VXM) analyses (to reliably predict recipient and donor compatibility assessment in a timely manner, especially for SOT involving deceased donors [512][516][519]), which could significantly contribute in

refining the organ allocation and assignment for patients in the waiting list (especially for those hypersensitized patients). Furthermore, as another example, description of both “A” (encoding respective alpha chain) and “B” (encoding respective beta chain) genes within classical HLA class II loci may allow a more precise definition (where differences particularly found within the ARD-coding exons of these genes may be relevant) of the epitope formed by the assembly of these encoded alpha and beta chains [262][263][554].

- This current 3-/4-field HLA allele and haplotype frequencies data may also provide invaluable information (once again, only as a first initial reference source) for improving bone marrow (BM) and umbilical cord blood (UCB) strategic donor recruitment and planning in Spanish registries as well as future matching criteria strategies for HSCT [215][554]. Since larger acquisition of high-resolution local population-specific HLA genotype data will definitely improve donor searches for individual patients using also updated HLA matching algorithms (e.g., haplotype frequency-based matching algorithms which can be designed to predict matched donors based on this type of large NGS HLA datasets from registries). In addition, NGS technology allows to identify suitably matched donors in a timely and a cost-effective manner and, importantly, to also achieve the almost unambiguous and highly reliable identification of novel alleles [172][179][211], possible null or expression variant alleles [210][458][523][869-871] and rare alleles [208][209], and thus the possibility of taking them into account for HSCT donor search purposes. In the context of Spanish population, presence of rare HLA class I or II allele/s, presence of the *HLA-B\*51:01:01:01* allele, presence of *HLA-DQB1\*03:01:01:01/:02/:03* allele variants and/or certain observed infrequent *HLA-B~C* or *HLA-DRB1~DQB1* associations in a given hematopoietic (HSCT) patient may represent some of the main negative predictive factors when considering the respective search and selection process of the most possible full-matched donor likely to be found [215][220][624][892][896].

• Other immunogenetics fields that will definitely and significantly benefit from the present NGS HLA Spanish dataset are those which imply anthropological studies (i.e. drawing more precise demographic and/or selection signatures within the genetic pool of Spanish population) [131][132][137]. Where fine ultra-high allele resolution by NGS and determination of 3-/4-field haplotypic associations have allowed us to identify more accurately specific patterns displayed within Spanish population and to better detect genetic imprints and substrates of either more ancient demographic events or some other more recent or stable throughout history. Main novelties brought by the present HLA NGS study have also enabled us to confirm and to further describe some of the previously reported HLA trends/signatures in Spanish population and certain specific ethnic groups/regions across the country [555][556][558][563][564][578]. In summary, 3-/4-field haplotypic associations and specific patterns displayed within the present Spanish population cohort have permitted the detection and description of genetic imprints from both:

-Early ancestral contributions throughout the history, where it is found a complex European-Mediterranean overall genetic substrate made up of North-Central European, North African Berber, Muslim Eastern Arab and Sephardic Jew genetic components, in addition to a remarkable presence of still relatively isolated Romani (“Gypsy”) genetic ancestry in a portion of the Spanish general population (especially across the Southern region of Spain (Andalusia)).

-And also from some other more recent and current demographic events, mainly Latin American (bearing Amerindian genetic background) and Eastern European-Mediterranean (especially from Romania) ethnic groups migrating to Spain.

-Furthermore, the present NGS HLA study (although with a limited sample size) may also contribute (until certain extent) to obtain a better depiction of the underlying population

substructure (i.e. stratification) and regional variation of the HLA genetic system diversity (regarding the detected differential distribution of certain allele groups and haplotypes) in modern-day Spanish population existing sub-groups across the country [260][604-606]. Thus, the current presented data complement and refine the existing estimates of HLA diversity in the Spanish population, increase population and geographic coverage by NGS data, and add granularity to clinically and genetically relevant HLA data.

- Moreover, NGS HLA studies describing HLA diversity in Spanish population may significantly contribute to the fine-mapping of risk/protection factors in HLA-disease association studies by defining associations of HLA alleles/haplotypes with certain diseases, particularly those with an autoimmune, inflammatory and/or neurological component (such as MS illustrated here); as well as in the design of epidemiology programs (i.e. evaluating prevalence of HLA-associated diseases in Spanish and related populations) and pharmacogenetics investigations (i.e. defining associations of HLA alleles/haplotypes with certain drug-induced hypersensitivity reactions) [120][121][126][127][142][170][207][288-294]. Also certain epidemic events and outbreaks worldwide or in given populations or regions may be better clinically understood, evaluated and monitored at least partially based on large-scale and comprehensive NGS HLA population data (from both healthy and patient cohorts) [513][514]. In the present thesis work, and as an exercise (i.e. test case) to exemplify and put in practice these aforementioned potential applications, we made use of this useful healthy control Spanish population HLA dataset to further evaluate the role of HLA in relation to MS genetic determinants and specifically for Spanish population in this given case.

- At the same time, it is also expected that all this knowledge (provided by this current Spanish HLA dataset as well as additional future larger HLA datasets, and not only at the Spanish population-level but also at the worldwide population-level) can contribute to establish

---

strategies for improving the efficacy of both current and novel immunotherapies and selection criteria of personalized therapeutic approaches [488-490]. Where, in detail:

- 1) Updated and extended Spanish panel of the most common HLA alleles (in order to reach the majority of targeted patients from the corresponding population) as potential novel targets for development of more efficient peptide/epitope-based vaccines in therapies for treating cancer and infectious diseases, and in particular for those patients presenting a highly resistant and/or refractory profile.
  - 2) Updated and extended Spanish panel of the most common full (i.e. encompassing at least the 11 major classical HLA class I and class II loci) extended HLA haplotypes, in this case, may serve to define and construct the most suitable (in terms of histocompatibility barriers to be considered) and representative (considering an adequate size and comprised diversity of the given therapeutic cell registry/bank) Spanish HLA haplo-homozygous allo-donors pool (in terms of both general population and regionally) of cell-based therapy products (therapeutic allogeneic genetically engineered T cells).
- In the regenerative medicine field, and similarly to the above mentioned application, moderate sized banks of iPSCs lines, once again, exclusively selected from HLA haplo-homozygous allo-donors can be constructed. Where a diverse set of these HLA haplo-homozygous iPSCs lines respectively carrying the most common HLA haplotypes of Spanish population can be constructed starting from a respective existing UCB registry population which has already been characterized for HLA genotypes [221][222][491-495][545][546] as well as based on the NGS HLA data from the present study.
  - On the other hand, current limited sample size of the present Spanish population study cohort still inevitably restricts the description of the vast genomic diversity found within the non-coding and untranslated regions of HLA genes as well as the regional variation of HLA allele



and haplotype frequency distributions found within Spain. As mentioned in [260], future studies of both considerably larger sample size, (being thus of higher associated statistical power) and at a wider geographical scale of the Spanish territory will continue shed light and will allow a much more accurate and meticulous description of HLA allele and haplotype diversity and their specific frequency distributions in Spanish population. In addition, future larger (and at a wider scale geographically speaking) NGS HLA Spanish population studies will certainly allow a more accurate and updated definition of distinctive CWD/CWID (including serological equivalents) as well as relevant null and rare (and novel) alleles (and their respective LD patterns and characteristic bearing haplotypes) for this population.

## **II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION**

Case-control studies are an essential analysis tool for genetic association studies. As described in the present thesis work, these case-control studies generally involve samples of unrelated individuals with the (disease) phenotype of interest and a corresponding control sample of unrelated, unaffected/healthy (or randomly ascertained) individuals obtained from the same (ethnic and, highly recommended, geographic/regional) population [773]. Moreover, at the moment of designing this type of studies (being especially critical due to the ultra-high diversity found in the HLA system and, in fact, at the population-level), it is very important to aim sufficient statistical power based on a substantial sample size for the given study (and thus to avoid misleading HLA-disease association results (susceptibility/protection)) [773]. Originally and as a general recommendation in traditional immunogenetic (e.g. HLA and KIR genes) studies, it has been established a minimum of  $2N=200$  chromosomes ( $N=100$  study subjects) for each of the case and

control cohorts [773]. However, much larger sample sizes (N~ at the scale of thousands study subjects) for both are not only preferable but highly needed, and now even more due to the very high allele resolution level provided by NGS-based HLA genotyping methods (i.e. 4-field data with minimum genotyping ambiguity and maximum coverage of HLA loci sequences) [137][301]. Among the advantages of case-control studies it is elimination of the necessity to collect data from family members, which is often logistically difficult and costly. Moreover, immunogenetic studies based on unrelated subjects may facilitate reaching to a statistically significant sample size that, at the same time, can also be representative (encompassing most of the main genetic diversity and characteristic features) of the given population/s of study. Nevertheless, case-control studies can be very sensitive to population stratification within the sample cohorts. In fact, this is a particularly important issue in immunogenetic data (e.g. HLA and KIR systems), where allele/haplotype frequency distributions vary considerably between human ethnic groups and even at the regional level within the same (considered “genetically homogenous”) population [136][137][260]. Thus, if samples are not collected with meticulous attention to homogeneity of ethnic/ancestry background, there is a very likely risk of misinterpreting genetic difference between cases and controls (and thus corresponding HLA-disease association results (susceptibility/protection)). Since, heterogeneity between cases and controls due to allele/haplotype frequency differences related to population stratification may be mistaken for disease association with a particular locus.

In the scope of the present thesis work:

- A first case-control study was carried out to examine HLA-disease associations with MS in these Spanish population cohorts: 17<sup>th</sup> IHIW Spanish population healthy control versus cohort of multiple sclerosis (MS) patients in the Spanish population (recruited at the Department of Neurology, Hospital Clínic, Barcelona, Catalonia, Spain). In this sense, the initial main goal was

to attempt a fine-mapping of these 3-/4-field allele and haplotype associations by full gene resolution level via NGS.

- In addition, a second exercise (i.e. test case) of this case-control study was carried out using the same MS Spanish group but, in this second case, using an alternative healthy control group dataset specifically from the Spanish northeastern region of Catalonia, and thus to evaluate (although limited to the 2-field allele resolution level) possible differences in the findings of HLA-disease association with MS due to plausible regional HLA genetic variation within mainland Spain as a statistical approach to try controlling for any possible existing population stratification (i.e. differences in genetic structure between disease and control groups). Therefore, this example has shown the relevance of defining the composition, especially in relation to the level of representability and possible population substructure, of the respective population healthy and diseased cohorts (ideally to be ethnically and regionally matched) selected for case-control studies.

#### **4. HLA ALLELE LEVEL ANALYSES**

At the HLA allele level, NGS-based HLA genotyping data at the 3-/4-field allele resolution enables an in-depth description of the HLA allelic diversity (practically unambiguous) given by silent substitutions and non-coding segments. In the scope of the present NGS HLA-MS study, our results have allowed us to depict a very high-resolution map of HLA risk and protection for multiple sclerosis (MS) reported here for the first time in Spanish population.

On one hand, we found that the architecture of HLA genetic risk for MS in the present Spanish population cohort (as a prototypic European-Mediterranean population clearly distinctive from other North-/Central European populations) was outstandingly dominated by the well-characterized *HLA-DRB1\*15:01:01:01* allele and its most common and classically associated bearing haplotype *HLA-A\*03:01:01:01~B\*07:02:01~C\*07:02:01:03~*

*DQA1\*01:02:01:01~DQB1\*06:02:01~DRB1\*15:01:01:01~DRB5\*01:01:01*. At the same time, and conversely to other Northern/Central European population MS groups [292][650][651][658][677][727], the present Spanish population cohort also shows a striking diversity of extended *HLA-DRB1\*15:01:01:01~DQB1\*06:02:01*-bearing haplotypes positively associated to MS, which encompass different sets of HLA class I alleles more commonly found in populations from Mediterranean regions [130][297][464]. Consequently, this finding underscores that HLA-MS associations (and especially at the haplotype level) may also vary between each given population group even within groups of European ancestry. Moreover, in contrast to other previous studies on European [650][651][658][670][677] and certain Spanish population [738][742] MS cohorts, the present Spanish population dataset examined via NGS was not dominated in its association to MS risk by additional series of class II alleles (consistent, as previously reported [650][651][658][670][677], with most risk signals being supposedly driven by alleles at *HLA-DRB1* locus such as *HLA-DRB1\*04:05*, *-DRB1\*08:01*, *-DRB1\*03:01*, and *-DRB1\*13:03*). Thus, a relatively simple landscape of HLA association with MS risk was revealed in our NGS HLA Spanish population study. In this sense, a plausible explanation to this would be that previous HLA studies (either based on not completely accurate and still limited imputation of HLA alleles [658][677] and/or with, tentatively, possible existing population stratification not well detected or not suitably controlled [650][651][658][670][677][738][742]) might have incorrectly identified this heterogeneous *HLA-DRB1* picture of association to MS risk, not being truly causative of MS pathogenesis in the majority of instances. Furthermore, our findings underscore the importance of evaluating variants at the highest possible resolution to identify with certainty the primary associations. Yet, increasing study sample sizes across different both regional and ethnic groups (where considering also different but more uniform MS clinical sub-groups of patients) as well as subsequent functional studies will be needed to completely unravel the causal

variants and genes and for definitely confirming (and, if so, better understand their role) or ruling out these primary and tentative secondary *HLA-DRB1* risk signals.

As an additional statistical approach of the assessment of HLA genetic risk for MS in the present Spanish population cohort, we applied stratification or conditioning analysis on the highly predisposing *HLA-DRB1\*15:01:01:01* allele to adjust for LD in order to dissect and interpret these nominal risk associations abovementioned. Thus, contrasting and supporting our interpretation of nominal and stratified results based also on the fully characterized class II and extended HLA haplotypes data generated thanks to the clonal nature of NGS, we were able to fine-mapped the following two main factors:

i) *HLA-DRB5\*01:01:01~HLA-DRB1\*15:01:01:01* was significantly associated with predisposition. Nevertheless, as previously mentioned, identification of the true predisposing gene of MS susceptibility within this prototypic susceptibility HLA-DR15 (*HLA-DRB1\*15:01:01:01~DRB5\*01:01:01*) haplotype is handicapped by the intense and exceptionally tight LD across these given individual alleles at *HLA-DRB1* and *HLA-DRB5* loci. Consequently, in the majority of populations of European descent (including the present Spanish population cohort) the role of allelic variation at *HLA-DRB5* in MS risk cannot be suitably assessed [292][293][668][727][743][944], where large trans-ethnic (especially from Asian-Pacific population groups) NGS HLA studies may contribute to better interrogate and shed light to this. Moreover, the present Spanish HLA-MS study, with a modest and still limited sample size, was also not sufficiently powered to detect the tentative association for the infrequent intronic variant *HLA-DRB5\*01:01:01v1*, which has been previously described in a much larger European American cohort [293]. Still, as it has been initially reported mostly from a functional standpoint (e.g. in relation to peptide repertoires or expression levels) [668][743][944], to separately delineate the specific genomic variation and resulting functional

implications and, thus, plausible disease causing mechanisms (e.g. disease-associated variants influencing via epistasis, defining the structural characteristics of the peptide-binding groove or shaping the T cell repertoire in the thymus leading to an inflammatory demyelinating process) of both *HLA-DRB1* and *HLA-DRB5* loci is a fundamental future step in the definite elucidation and decryption of the role of HLA in MS pathogenesis.

ii) At the same time, nominal and stratified analyses in our study identified a significant MS risk signal relative to *HLA-DPBI\*03:01:01* allele being, remarkably, independent from the highly predisposing *HLA-DRB1\*15:01:01:01* factor. However, apart from its well-known strong LD with *HLA-DPA1\*01:03:01:03* allele (thus, encoding the respective HLA-DP heterodimer), it was difficult to assess its respective and specific extended haplotypic association (when using the present NGS HLA dataset) given the fact of the dramatic increase of haplotype diversity displayed when including *HLA-DPA1* and *-DPBI* loci due to existing hotspot of recombination between HLA-DQ and -DP loci [92], causing here a great loss of statistical power for the HLA-MS association analyses at the haplotype level. Consequently, LD is low between *HLA-DP* genes and all *-DRB1*, *-DQA1* and *-DQB1* loci. Indeed, this low LD may also explain that the *HLA-DPBI\*03:01:01* risk association to MS susceptibility is unlikely to be caused by complex interactions between *HLA-DRB1* alleles [945]. Associated MS risk to HLA-DP region, and to *HLA-DPBI\*03:01* allele in particular (especially for populations of European descent), as an independent signal from *HLA-DRB1\*15:01* factor has been also previously reported in several both HLA-imputation [652][658][945] and direct DNA sequencing for HLA genotyping [684][686][948] studies as well as in some functional studies [954]. In this sense, effect of HLA-DP on MS susceptibility has been postulated by various plausible disease causing mechanisms. Mainly:

-In relation to the binding and presentation of peptide antigens to CD4+ T helper cells, for instance, one GWAS study explained this *HLA-DPB1\*03:01* positive independent association based on a particular SNP rs9277489 where the most statistically significant amino acid mapped to Leu65 of HLA-DPβ1 located in the peptide-binding groove [658]. Also, another HLA genotyping and functional study suggested that *HLA-DPB1\*03:01* allele may be implicated in epitope spreading (i.e. neo-autoreactivity) in MS, in which HLA-DP may play an important role in the development, spread, and propagation of self-recognition during the clinical progression of MS (particularly in the early stages of self-recognition when autoreactivity is characterized by extensive plasticity) [954]. Moreover, other HLA genotyping study in Australian Caucasoid MS patients and Cantonese MS patients identified alleles *HLA-DPB1\*03:01* and *HLA-DPB1\*13:01*, respectively, as independent risk factors for MS in non-*HLA-DQB1\*06:02* patients [684]. As explained by the authors, this finding may reflect communality in a particular nucleotide motif corresponding to a particular peptide-binding or T-cell recognition epitope, since both *HLA-DPB1\*03:01* and *DPB1\*13:01* share amino acid residues 1-34 in the HLA-DPβ1 molecule within a polymorphic region responsible in part for T-cell recognition [684].

-On the other hand, the effect of HLA-DP on MS may be due to differences in levels of expression rather than differences in peptide presentation. From several GWAS studies a particular associated SNP rs9277535 (550A/G variant) to MS susceptibility (as well as to other autoimmune diseases such as systemic lupus erythematosus (SLE) in a Chinese population [953]) has been previously identified [652][945]. This SNP rs9277535 lies in the 3' untranslated (3'-UTR) region of this class II gene *HLA-DPB1*, in which transcriptional factors (such as microRNAs) may bind to regulate expression [416]. Furthermore, together with the SNP variant rs3077 corresponding to the paired *HLA-DPA1*, rs9277535 allele G has

been correlated with the down-regulated level of HLA-DP mRNA, while allele A has been associated with the increased level of HLA-DP mRNA [953]. Thus, allele G has been found predisposing to MS (i.e. MS patients carrying the *HLA-DPB1\*03:01* allele also carry the rs9277535G allele) [945], as similarly observed in the predisposition to chronic hepatitis B (HBV) infection [322] and SLE [953]. Nevertheless, these detected SNP associations (SNPs in the 3'UTR of *HLA-DPB1* (rs9277535) and *-DPA1* (rs3077)) appear to be stronger in Asians than in other European and African-American populations, at least according to chronic hepatitis B and outcomes study [322]. So far, these consistent results suggest that though in different diseases, the role of expression variants of HLA-DP might be the same.

Our present data indicates that *HLA-DPB1\*03:01:01* allele is positively associated with MS risk and independent of other MS associations in the HLA complex evaluated here, confirming previous reports [652][658][684][686][945][948]. Still, future studies on the HLA-DP structure-function (i.e. revealing how exactly the gene variants influence the gene expression like HLA-DP mRNA regulation and the gene function like antigen presentation to CD4+ T cells) may provide further evidence to improve our understanding of the exact function of HLA-DP in the pathogenesis of MS. In addition, application of long-read sequencing approaches for the characterization of fully phased HLA haplotypes may also contribute to better dissect any possible *HLA-DPB1\*03:01:01*-linked loci potentially involved in MS susceptibility [159].

Lastly, in addition to standard case-control association analyses (using both nominal and stratified approaches), we also validated our risk association findings using a second healthy control group (both ethnically and regionally matched) in order to control for population stratification (as a confounding factor that may affect the HLA-MS association results) at the HLA allele level. Overall, no population substructure was detected in the present study.



A relatively more complex association landscape can be observed in relation to the protective HLA effects to MS susceptibility detected in the present Spanish population dataset. Resultant association analysis at the 3-/4-field of resolution, following stratification, also proved to be effective at identifying the independent protective effects of specific HLA alleles and haplotypes. Our findings also underscore here the importance of evaluating variants at the highest possible resolution, as well as attending to the LD patterns displayed at the haplotype level, to identify with certainty the primary associations. In summary, two main HLA-MS protective association patterns can be observed in the present study:

i) Those corresponding to HLA class II *DRB1* signals: *HLA-DRB1\*04:01:01:01* (where *HLA-B\*44:02:01:01* and *-C\*05:01:01:02* allelic protective signals seem to be also driven by this same *HLA-DRB1* signal), *-DRB1\*04:02:01* and *-DRB1\*04:04:01* which all are tightly associated with the secondary DRB *HLA-DRB4\*01:03:01:01* and, thus, it was not possible to be disentangled (given also the high haplotype diversity found and the relatively low frequency of each haplotype “variant” in the present study presenting a modest sample size); and, separately, *HLA-DRB5\*01:02~DQB1\*06:01:01* signal (without the *HLA-DRB1\*15:02:01* signal, not statistically associated at the allelic level). In both cases, these protective effects can be certainly attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele.

ii) A series of protective signals driven by HLA class I alleles and, in this case, being all independent from the highly predisposing *HLA-DRB1\*15:01:01:01* factor. On one hand, the *HLA-B\*38:01:01* allele signal being consistent with previous reports in population cohorts of European ancestry [658][677]. Nevertheless, no amino acid position in *HLA-B* has been mapped yet explaining the *HLA-B\*38:01* protective effect at least in these previous SNP-based studies [658]. Moreover, in the present dataset *HLA-B\*58:01:01:01* was found as a novel class I

protective allele signal. To the best of our knowledge, this protective association has not been previously reported in the literature (reviewed in [650][651]). Based on our stratified analysis, the functional protective role of these two independent HLA-B effects detected here, at least in part, may not simply reflect related LD patterns and allele/haplotype frequency distributions but, indeed, some kind of underlying mechanism impairing or avoiding at some level the immunopathogenesis of MS. Interestingly, although limited to the interrogation of *HLA-B* alleles without evaluating other Bw4/Bw6-bearing alleles at the *HLA-A* and *-C* loci, in the present study we found that Bw4 motif subgroups NLRIALR, DLRTLLR and NLR TALR were not associated in either nominal or stratified analysis. Thus, in the present dataset we did not find a clear Bw4 protective association with MS susceptibility that could reflect either LD pattern (i.e. negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele.) or ligand mediated KIR3DL1 signaling (indirectly evaluated here), being this in clear contrast to what has been described in previous studies for both European and African American cohorts [724-727][947]. On the other hand, we described that the Bw6 epitope (SLRNLRG), encoded by the respective group of HLA-B alleles analyzed here, shows a risk association that cannot be attributed simply to LD patterns in relation to the highly predisposing allele *DRB1\*15:01:01:01* and, consequently, it appears to be independent and tentatively conferring a synergic effect, although this would need to be further evaluated in future studies.

Altogether, these findings are not perfectly in line with other previous studies on European [650][651][658][670][677] and certain Spanish population [738][742] MS cohorts (e.g. in the present study, previously reported allele signals such as *HLA-A\*02:01* or *HLA-DRB1\*01:01* did not show a statistically significant protective effect). Nonetheless, as aforementioned, inaccurate imputation of HLA alleles in SNP-based studies, limited HLA allelic resolution in genotyping studies and/or possible existing population stratification may be potential factors that might have

confound these previous results in other studies. Yet, future increasing study sample sizes across different both regional and ethnic groups (where considering also different but more uniform MS clinical sub-groups of patients) as well as subsequent functional studies will be needed to completely unravel the protective variants and genes and their roles.

Once again, in addition to standard case-control association analyses (using both nominal and stratified approaches), we also validated our protective association findings using a second healthy control group (both ethnically and regionally matched) in order to control for population stratification (as a confounding factor that may affect the HLA-MS association results) at the HLA allele level. Overall, no population substructure was detected in the present study.

Finally, it is noteworthy that our present study did also have certain limitations which are important to be considered and that can be summarized as the following:

- i) Many immune-related genes in the MHC were not analyzed in this study; given the complex LD known for the MHC, our analyses do not exclude these genes (e.g. complement *C4A/C4B*, *MICA/MICB* and other cytokine-related genes) as potentially playing roles in MS susceptibility, as previously well noted in the literature [292][652][653][727].
- ii) Some intronic and 5'/3'-untranslated variants (especially in HLA class II loci) were not fully resolved and identified (e.g. *HLA-DRB1\*15:01:01:01/:02/:03*) due to current limitations from short-read sequencing platforms on characterizing repetitive and extensive low-complexity and imbalanced genomic sequence composition, such as: homopolymer repeats poly(dA), poly(dT), poly(dG) and poly(dC); regions of short-tandem repeats (STRs); or high AT- or GC-rich regions [296]. In this sense, application of long-read sequencing approaches for the characterization of fully phased HLA haplotypes may also contribute to better dissect any possible HLA-linked loci potentially involved in MS susceptibility [159].

iii) We described here, replicating previous studies, that the extended haplotype of the HLA class II region (*HLA-DRB5\*01:01~DRB1\*15:01~DQA1\*01:02~DQB1\*06:02*), which has been further refined to *HLA-DRB5\*01:01~DRB1\*15:01*, confers the strongest risk for developing MS. However, we did not interrogate in the present study the previously underestimated genomic variation in certain HLA genes considered to be monomorphic or highly conserved (at least at the protein level) such as *HLA-DRA* locus. In this sense, some recent studies (mostly based on long-read sequencing platforms) have shown how SNPs and polymorphism clusters within the introns and 3'UTR region of *HLA-DRA* locus (in spite of being not as polymorphic as the other class II alpha genes (*HLA-DQA1* and *HLA-DPA1*)) define distinct gene lineages, which, in turn, facilitate the identification and definition of *HLA-DRA~DRB3/4/5~DRB1~DQB1* haplotype patterns [429] as well as to fully understand the DR $\alpha$ ~DR $\beta$  antigen presenting heterodimer genomic structure and variation. Moreover, another recent study has also shown that a splice acceptor variant in *HLA-DRA* (as a result of an alternative splicing event driven by the single nucleotide polymorphism (SNP) rs8084) affects the conformation and cellular localization of the class II DR alpha-chain [946]. In addition, this short HLA-DRA isoform can be loaded into the peptide-binding site of canonical HLA class II heterodimers, functioning itself as a putative full-length antigen. In that study, authors also experimentally showed that short HLA-DRA expression is up-regulated by IFN- $\alpha$  stimulation, supporting a possible functional link between short HLA-DRA presentation and diseases characterized by inflammatory responses [946].

iv) In the present study, we did not attempt association analyses of individual amino acids in the HLA class I and class II genes studied here. Even though this analysis step can potentially reveal functionally important aspects of MS disease association (as previously shown in other studies [292][658][727]), still the picture, as an example, at *HLA-DRB1* however appears to be

more complex as there has not been yet a single model based on amino acids that could explain the entire locus effect. Moreover, some disease associated amino acid residues simply “tag” an allele, recapitulating an already well-established allele association [292][727]. Also, the peptide binding properties of HLA molecules are obviously determined by multiple amino acid residues. Consequently, a given strong association detected in one single amino acid may not be biologically meaningful in terms of HLA functionality and its potential impact on MS pathogenesis. In general, the potential role of individual amino acids in disease associations can be best evaluated by comparing alleles that differ in disease risk, and differ in only one amino acid position (e.g. *HLA-DRB1\*15:01* versus *HLA-DRB1\*15:02*, ideally in trans-ethnic studies). In this sense, amino acids may allow new hypotheses to be formulated and necessarily further evaluated in future corresponding functional studies. Thus, this association analysis of individual amino acids fell out of the scope of the present thesis work.

## **5. HLA HAPLOTYPE LEVEL ANALYSES**

At the HLA haplotype level, NGS-based HLA genotyping data at the 3-/4-field allele resolution allowed a further and in-depth description of specific LD patterns (even in fully extended HLA class I and class II haplotypes) to better dissect those HLA allele signals initially detected with risk/protection to MS susceptibility in the present study. Thus, differences found in non-coding polymorphisms were clearly advantageous for breaking down HLA LD patterns, which, in turn, were very pertinent for mapping with more precision MS causative variants by eliminating at the same time false signals resulting from ‘hitchhiking’ alleles. Moreover, NGS-based HLA genotyping studies (as the one presented here) may overcome limitations encountered in previous studies (either based on lower resolution HLA genotyping techniques or HLA-imputation approaches using SNPs panels), in which certain questionable or inconsistent findings in relation

to some given HLA allele association signals might have been tagging in reality to untested or not fully resolved HLA loci in the past.

From the HLA class II haplotype analyses at the 3-/4-field of the present study, our findings led to the conclusion, as one main example, that the major MS risk is attributed to *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* signal but not to *HLA-DQB1\*06:02:01* allele (as found in other MS population studies [669]). Thus, when trans-ethnic studies may be not feasible to be conducted, NGS-based HLA haplotype level analyses may contribute, at least to some extent, to dissect class II signals even when these are displaying a strong LD. Another very interesting finding, which indeed illustrates how informative it is the description of specific haplotype LD patterns at the 3-/4-field of resolution, is relative to the *HLA-DRB1\*15:02:01:02*-bearing haplotype that clearly shows a strong association by conferring protection to MS susceptibility, being in contrast to the neutral signal from the counterpart *HLA-DRB1\*15:02:01:01*-bearing haplotype. This finding is also in line with the results from the Finn et al. study [698], where they showed that the difference in risk association with MS of *HLA-DRB1\*15:01* versus *-DRB1\*15:02* is not due to a lack of antigen presentation by DRβ1\*15:02, at least in the context of putative myelin peptides, and suggested that other mechanisms involving *HLA-DRB1\*15:01* may account for increased susceptibility to MS and vice versa. In this sense, distinctive haplotypic variants indeed bear a series of antigen specificities corresponding to the encoded HLA molecules, which are comprising by those respective haplotype blocks, establishing also, in turn, disease-associated variants that define the structural characteristics of the peptide-binding groove and/or shaping the T cell repertoire in the thymus leading to an inflammatory demyelinating process [668].

In relation to the fully extended HLA haplotype analyses at the 3-/4-field in the present study, the current extended haplotype distribution data enabled to clearly identify which risk/protective HLA class I signals were indeed independent or otherwise were associated as a consequence of

their respective LD pattern with HLA class II (mainly *HLA-DRB1*) causative variants. At the same time, in addition to the most common associated bearing haplotype *HLA-A\*03:01:01:01~B\*07:02:01~C\*07:02:01:03~DQA1\*01:02:01:01~DQB1\*06:02:01~DRB1\*15:01:01:01~DRB5\*01:01:01* found conferring risk to MS susceptibility [292][650][651][658][677][727], the present Spanish population cohort shows a striking diversity of extended *HLA-DRB1\*15:01:01:01~DQB1\*06:02:01*-bearing haplotypes positively associated to MS, which encompass different sets of HLA class I alleles commonly found in populations from Mediterranean regions [130][297][464]. This finding confirms the existing heterogeneity among risk extended *HLA-DRB1\*15:01*-bearing haplotypes as previously described [711]. Moreover, these distinctive extended *HLA-DRB1\*15:01*-bearing haplotypes (stratified by HLA class I tagging) appear to differ for risk to MS susceptibility. Yet, future increasing study sample sizes across different both regional and ethnic groups (where considering also different but more uniform MS clinical sub-groups of patients) as well as subsequent functional studies will furnish the basis for MHC-associated susceptibility in MS, in which the entire MHC haplotype is certainly the fundamental unit of genetic control of immune response in health and disease [711]. Moreover, based on these haplotypic associations shown in the present study, our findings also arise the importance of evaluating in future functional studies not only the autoantigen presentation to T cells in the context of *HLA-DRA~DRB1/DRB5* or *HLA-DQA1~DQB1* [743][944][946] but also in the context of *HLA-DPA1~DPB1* [948], where there may be a plausible synergism between HLA-DP and -DR gene products playing a role in the genetic susceptibility to MS.

Once again, in addition to standard case-control association analyses (using both nominal and stratified approaches), we also validated our risk and protective haplotype association findings using a second healthy control group (both ethnically and regionally matched) in order to control for population stratification (as a confounding factor that may affect the HLA-MS association

results) at the HLA haplotype level. Overall, no significant population substructure was detected in the present study, although at the extended HLA haplotype level the distribution of haplotypic variants seems to be more sensitive to this possible existing underlying factor.

Finally, as a noteworthy limitation, since the present Spanish NGS HLA-MS study shows a modest and still limited sample size it was therefore not sufficiently powered to entirely detect and depict the 3-/4-field haplotypic MS associations and, critically, their totally real statistical significance. Consequently, many positively and negatively associated class II and fully extended haplotypes with MS evaluated in the present study exhibit low or very low haplotype frequencies, being more difficult to deduce with certainty any plausible and suitably clear interpretation of these association results obtained at the haplotype level (i.e. since these haplotype frequencies fall into the “binned” category in the  $\chi^2$  statistic for a contingency table analysis of case-control data due to their low haplotype frequencies; yet, their statistical parameters were manually calculated and shown here only for purposes of comparison). Thus, the difference between common and rare HLA haplotype frequency distributions was too excessive in this case. An additional challenge corresponded to those HLA regions (particularly for HLA-DP) presenting too low LD due to, in most cases, being separated by a recombination hotspot with a high recombination rate [92]. This particular current limitation may be overcome in future NGS studies using long-read sequencing approaches to characterize fully phased haplotypes [159]. Yet, altogether, this represents a serious difficulty for the majority of NGS HLA-disease association studies where the vast HLA haplotypic diversity found in a still insufficient study sample size generally mean the loss or, better said, the non-achievement of statistical power for those HLA allele/haplotype variants of interest to be examined. In this sense, notably increasing study sample sizes across different both regional and ethnic groups (where considering also different but more uniform MS clinical sub-groups of patients) may be a plausible solution to this statistical limitation inherently related to HLA



diversity. Nevertheless, selecting a suitable control group both ethically and regionally matched as well as conducting a fine control approach for population stratification in the given NGS study are absolutely required. Another statistical strategy that may be attempted is the grouping of HLA diversity from 3-/4-field to 2-field resolution in order to achieve the necessary power. However, this may not be biologically appropriate since these non-coding regions contain relevant sites (establishing also a specific LD pattern) for transcription promoters, inhibitors, alternative splice sites, methylation sites, binding sites for post-translational miRNA degradation and many other functions as yet undetermined [951]. Therefore, statistical significance at the 2-field should not generally take precedence over biological relevance (and respective LD patterns associated to it) found at the very HLA genomic level. Lastly, although the present Spanish NGS study was based on EM-estimated haplotype data obtained from an unrelated population cohort, alternative sufficiently large family-based NGS studies may be able to use this HLA haplotype frequency distribution data by exploring HLA haplotype/allele associations with MS using transmission disequilibrium test (TDT) and multiallelic TDT (mTDT), which simultaneously assesses linkage and association. Thus, taking advantage since haplotypes produce more definitive transmissions than do the alleles encompassing them, and this tends to increase power. Nonetheless, the larger number of haplotypes relative to alleles at individual loci tends to decrease power due to the additional degrees of freedom required for the corresponding analysis [711].

## **6. FINAL REMARKS**

The combination of a relatively large sample size with NGS-based HLA genotyping allowed us to obtain an enhanced dissection of the critical role of the HLA in MS susceptibility. In summary, the refined *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* signal was significantly associated with predisposition as expected. A second independent risk allele *HLA-DPB1\*03:01:01* was also

identified, being in consonance with previous studies in populations of European descent. Protective effects from several distinctive HLA class II signals (including *HLA-DRB1\*04:01:01:01*, *-DRB1\*04:02:01* and *-DRB1\*04:04:01*, which all are tightly associated with the secondary DRB *HLA-DRB4\*01:03:01:01*; and, separately, *HLA-DRB5\*01:02~DRB1\*15:02:01:02~DQB1\*06:01:01* signal) were attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele. While the HLA class I alleles *HLA-B\*38:01:01*, previously identified in other studies, and newly *HLA-B\*58:01:01:01* showed moderately protective effects independently from each other and from the HLA class II associated factors. Altogether, the present study demonstrates the effectiveness of very high resolution extended coverage genotyping for dissecting HLA alleles/haplotypes associated with MS disease susceptibility using both nominal and stratified analyses. To the best of our knowledge, this is the first HLA-MS unrelated study using NGS in Spanish population. In this study, both susceptible and protective candidate HLA alleles/haplotypes were mapped with more precision by eliminating at the same time false signals resulting from ‘hitchhiking’ alleles. The mapping to specific HLA allele and respective haplotype structures may allow to design future research focused in their functional features facilitating the understanding of the mechanisms involved in MS pathogenesis.



## ***CONCLUSIONS***



## **I. NGS-BASED HLA STUDY IN 17TH-IHIW SPANISH POPULATION COHORT (HEALTHY CONTROL GROUP)**

1) To the best of our knowledge, this is the first and largest study performed using NGS for the genomic characterization of HLA diversity found in Spanish population. In the present NGS study, we were able to describe allelic diversity at the 3-/4-field resolution of major HLA genes *HLA-A*, *-B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1* and *-DRB3/4/5* (enabling full sequencing of class I loci and extended coverage of class II loci) with minimum level of ambiguities and also to estimate extended haplotype frequencies.

2) NGS HLA sequencing in the present Spanish population cohort has shown striking and highly informative 3-/4-field genotyping results including the description of previously unknown haplotype associations in non-coding regions up to the 4-field allele resolution level, the detection of rare, null and novel polymorphisms as well as the more accurate evaluation of allele and haplotype distributions and prevalence in Spanish population.

3) Overall, results of the present study may contribute as a useful and first reference source for future population studies, for HLA-disease association and pharmacogenetics studies as a healthy control group dataset, for improved virtual panel reactive antibody (vPRA) calculations in Spanish population and for improving donor recruitment strategies of bone marrow and umbilical cord blood registries. Moreover, fine ultra-high allele resolution by NGS and determination of 3-/4-field haplotypic associations have allowed us to identify more accurately specific patterns displayed within Spanish population (including a significant regional variation and population substructure) and to better detect genetic imprints and substrates of either more ancient demographic events or some other more recent or stable throughout history. Data from the present and from future larger NGS studies may also contribute to establish strategies for improving the

efficacy of both current and novel immunotherapies and selection criteria of personalized therapeutic approaches. Lastly, knowledge of the most common extended HLA haplotypes at the 3-/4-field resolution in Spanish population may also serve to construct the most representative Spanish HLA haplo-homozygous bank for allogeneic transplantation of induced pluripotent stem cells (iPSC) derived cell therapies such as novel cellular adoptive therapies based on genetically engineered T cells.

## **II. NGS-BASED HLA CASE-CONTROL STUDY OF MULTIPLE SCLEROSIS IN SPANISH POPULATION**

At the same time, the present thesis work also allowed to interrogate allele and haplotype associations of 11 major classical HLA genes with multiple sclerosis disease in Spanish population based on a very comprehensive analysis using 3-/4-field HLA genotype data for the first time.

These are the main conclusions of the present NGS HLA-MS study in Spanish population:

- 1) Overall, very high-resolution HLA genotyping data allows fine-mapping of susceptibility and protective factors and exclusion of bystander (“hitchhiking”) alleles from contiguous loci.
- 2) The refined *HLA-DRB5\*01:01:01~DRB1\*15:01:01:01* signal was significantly associated with predisposition.
- 3) Nominal and stratified analyses identified a second significant MS risk signal relative to *HLA-DPB1\*03:01:01* allele, being independent from the highly predisposing *HLA-DRB1\*15:01:01:01* factor.
- 4) Protective effects from several distinctive HLA class II signals (several *HLA-DRB1\*04-* and the *HLA-DRB1\*15:02:01:02*-bearing haplotypes) were attributed to negative LD with the highly predisposing *HLA-DRB1\*15:01:01:01* allele.

5) *HLA-B\*38:01:01* and *-B\*58:01:01:01* alleles confer protection and operate independently of the presence of *HLA-DRB1\*15:01:01:01* risk factor.

6) In the present dataset, we did not find a clear Bw4 (relative to only HLA-B alleles and according to motif subgroups NLRIALR, DLRTLLR and NLRTALR, respectively) protective association by itself with MS susceptibility. On the other hand, we described that the Bw6 epitope (SLRNLRG), encoded by the respective group of HLA-B alleles analyzed here, shows a risk association that cannot be attributed simply to LD patterns in relation to the highly predisposing allele *DRB1\*15:01:01:01*.

Finally, the lesson from the study of HLA polymorphism over the last several decades has been that each incremental technological advance that leads to higher resolution has yielded further insights into the cause or mechanisms of disease. With the advent of highest-resolution NGS technologies, there is an opportunity to more comprehensively define the role of HLA in health and disease.





***BIBLIOGRAPHY***



- [1] MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401, (1999) 921–923.
- [2] Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot Jr, C.C., Wright, M.W. and Wain, H.M. Gene map of the extended human MHC. *Nature Reviews Genetics*, 5(12), (2004) p.889.
- [3] Klein, J. (1986). *Natural History of the Major Histocompatibility Complex*. Wiley-Interscience, New York.
- [4] Doherty, Peter C., and Rolf M. Zinkernagel. "A biological role for the major histocompatibility antigens." *The Lancet* 305, no. 7922 (1975): 1406-1409.
- [5] Harel-Bellan, A., A. Quillet, C. Marchiol, R. DeMars, T. Tursz, and D. Fradelizi. "Natural killer susceptibility of human cells may be regulated by genes in the HLA region on chromosome 6." *Proceedings of the National Academy of Sciences* 83, no. 15 (1986): 5688-5692.
- [6] Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev* (2002); 190:95–122.
- [7] Gorer P. The detection of a hereditary genetic difference in the blood of mice by means of human group A serum. *J Genet* (1936) 32: 17–31.
- [8] Gorer P. The detection of antigenic differences in mouse erythrocytes by the employment of immune sera. *Br J Exp Pathol* (1936) 17: 42–50.
- [9] Gorer P. The genetic and antigenic basis of tumour transplantation. *J Pathol Bacteriol* (1937) 44: 691–7.
- [10] Klein, Jan. "Biology of the mouse histocompatibility-2 complex: principles of immunogenetics applied to a single system." (1975): 620.
- [11] Gorer PA, Lyman S, Snell GD. Studies on the genetic and antigenic basis of tumour transplantation. Linkage between a histocompatibility gene and "fused" in mice. *Proc R Soc Lond B Biol Sci* (1948) 135: 499–505.
- [12] Snell GD, Cherry M, Demant P. Evidence that the H-2 private specificities can be arranged in two mutually exclusive systems possibly homogenous with the two subsystems of HL-A. *Transplant Proc* (1971) 3: 183–6.
- [13] Dausset J. Iso-leuco-anticorps. *ActaHaematol* (1958) 20: 156–66.
- [14] van Rood JJ, Eernisse JG, van Leeuwen A. Leucocyte antibodies in sera from pregnant women. *Nature* (1958) 181:1735–6.
- [15] Payne R, Rolfs MR. Fetomaternal leukocyte incompatibility. *J Clin Invest* (1958) 37: 1756–62.
- [16] Thorsby, Erik. "A short history of HLA." *Tissue Antigens* 74, no. 2 (2009): 101-116.
- [17] Marino, Susana G., Andrés Jaramillo, and Marcelo A. Fernández-Viña. "The Human Major Histocompatibility Complex and DNA-Based Typing of Human Leukocyte Antigens for Transplantation." In *Handbook of Human Immunology*, pp. 546-569. CRC Press (2008).
- [18] Janeway Jr, Charles A., Paul Travers, Mark Walport, and Mark J. Shlomchik. "The major histocompatibility complex and its functions." In *Immunobiology: The Immune System in Health and Disease*. 5th edition. Garland Science, 2001.
- [19] Parham, Peter. "Immunogenetics of killer cell immunoglobulin-like receptors." *Molecular Immunology* 42, no. 4 (2005): 459-462.

- [20] Martin, Annalise M., Jerzy K. Kulski, Campbell Witt, Pierre Pontarotti, and Frank T. Christiansen. "Leukocyte Ig-like receptor complex (LRC) in mice and men." *Trends in Immunology* 23, no. 2 (2002): 81-88.
- [21] Yeager, Meredith, and Austin L. Hughes. "Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution." *Immunological reviews* 167, no. 1 (1999): 45-58.
- [22] Doherty, Peter C., and Rolf M. Zinkernagel. "A biological role for the major histocompatibility antigens." *The Lancet* 305, no. 7922 (1975): 1406-1409.
- [23] Apanius, Victor, Dustin Penn, Patricia R. Slev, L. Ramelle Ruff, and Wayne K. Potts. "The nature of selection on the major histocompatibility complex." *Critical Reviews™ in Immunology* 17, no. 2 (1997).
- [24] Margulies, D. H. "Major histocompatibility complex (MHC) molecules: Structure, function, and genetics." *Fundamental Immunology* (2008): 570-613.
- [25] Erlich, Henry A., and Ulf B. Gyllensten. "Shared epitopes among HLA class II alleles: gene conversion, common ancestry and balancing selection." *Immunology Today* 12, no. 11 (1991): 411-414.
- [26] De Bakker, Paul IW, Gil McVean, Pardis C. Sabeti, Marcos M. Miretti, Todd Green, Jonathan Marchini, Xiayi Ke et al. "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC." *Nature Genetics* 38, no. 10 (2006): 1166.
- [27] Shiina, Takashi, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K. Kulski. "The HLA genomic loci map: expression, interaction, diversity and disease." *Journal of Human Genetics* 54, no. 1 (2009): 15.
- [28] Trowsdale, John, Jiannis Ragoussis, and R. Duncan Campbell. "Map of the human MHC." *Immunology Today* 12, no. 12 (1991): 443-446.
- [29] Stewart, C. Andrew, Roger Horton, Richard JN Allcock, Jennifer L. Ashurst, Alexey M. Atrazhev, Penny Coggill, Ian Dunham et al. "Complete MHC haplotype sequencing for common disease gene mapping." *Genome Research* 14, no. 6 (2004): 1176-1187.
- [30] Mungall, A. J., S. A. Palmer, S. K. Sims, C. A. Edwards, J. L. Ashurst, L. Wilming, M. C. Jones et al. "The DNA sequence and analysis of human chromosome 6." *Nature* 425, no. 6960 (2003): 805.
- [31] Xie, Tao, Lee Rowen, Begoña Aguado, Mary Ellen Ahearn, Anup Madan, Shizhen Qin, R. Duncan Campbell, and Leroy Hood. "Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse." *Genome Research* 13, no. 12 (2003): 2621-2636.
- [32] Horton, Roger, Richard Gibson, Penny Coggill, Marcos Miretti, Richard J. Allcock, Jeff Almeida, Simon Forbes et al. "Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project." *Immunogenetics* 60, no. 1 (2008): 1-18.
- [33] Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet et al. "The Ensembl gene annotation system." *Database* 2016 (2016).
- [34] Shiina, Takashi, Antoine Blancher, Hidetoshi Inoko, and Jerzy K. Kulski. "Comparative genomics of the human, macaque and mouse major histocompatibility complex." *Immunology* 150, no. 2 (2017): 127-138.
- [35] Monos, Dimitrios S., and Robert J. Winchester. "The Major Histocompatibility Complex." In *Clinical Immunology*, pp. 79-92. (2019).
- [36] Trowsdale, John. "The gentle art of gene arrangement: the meaning of gene clusters." *Genome Biology* 3, no. 3 (2002): comment2002-1.
- [37] Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. and Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* 80, 487-498 (2002).

- [38] Zhang, E. Y., Knipp, G. T., Ekins, S. and Swaan, P. W. Structural biology and function of solute transporters: implications for identifying and designing substrates. *Drug Metab. Rev.* 34, 709–750 (2002).
- [39] Radosavljevic, M. and Bahram, S. *In vivo* immunogenetics: from *MIC* to *RAET1* loci. *Immunogenetics* 55, 1–9 (2003).
- [40] Hopper, A. K. and Phizicky, E. M. tRNA transfers to the limelight. *Genes Dev.* 17, 162–180 (2003).
- [41] Jack, L. J. and Mather, I. H. Cloning and analysis of cDNA encoding bovine butyrophilin, an apical glycoprotein expressed in mammary tissue and secreted in association with the milk-fat globule membrane during lactation. *J. Biol. Chem.* 265, 14481–14486 (1990).
- [42] Giorgi, D., Friedman, C., Trask, B. J. and Rouquier, S. Characterization of nonfunctional V1R-like pheromone receptor sequences in human. *Genome Res.* 10, 1979–1985 (2000).
- [43] Ziegler, A., Dohr, G. and Uchanska-Ziegler, B. Possible roles for products of polymorphic MHC and linked olfactory receptor genes during selection processes in reproduction. *Am. J. Reprod. Immunol.* 48, 34–42 (2002).
- [44] Coleman, J. E. Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins. *Annu. Rev. Biochem.* 61, 897–946 (1992).
- [45] Gruss, H. J. and Dower, S. K. The TNF ligand superfamily and its relevance for human diseases. *Cytokines Mol. Ther.* 1, 75–105 (1995).
- [46] Mallya, M., Campbell, R. D. and Aguado, B. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. *Genomics* 80, 113–123 (2002).
- [47] Milner, C. M. and Campbell, R. D. Structure and expression of the three MHC-linked *HSP70* genes. *Immunogenetics* 32, 242–251 (1990).
- [48] Gleimer, M. and Parham, P. Stress management: MHC class I and class I-like molecules as reporters of cellular stress. *Immunity* 19, 469–477 (2003).
- [49] Yu, C. Yung. "Molecular genetics of the human MHC complement gene cluster." *Experimental and Clinical Immunogenetics* 15, no. 4 (1998): 213-230.
- [50] Alfonso, C. and Karlsson, L. Nonclassical MHC class II molecules. *Annu. Rev. Immunol.* 18, 113–142 (2000).
- [51] Carrington M, Norman P. The KIR Gene Cluster [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); (2003) May 28.
- [52] Pablo Gomez-Prieto CP-L, Diego Rey Enrique Moreno, Antonio Arnaiz-Villena. HLA-G, -F and -E. Mehra NK, editor. Polymorphism, Function and Evolution in The HLA Complex in Biology and Medicine A Resource Book. Jaypee Brothers Medical Publishers Ltd; (2010). p. 159–74.
- [53] Stephens, Henry AF. "MICA and MICB genes: can the enigma of their polymorphism be resolved?." *Trends in immunology* 22, no. 7 (2001): 378-385.
- [54] Malissen, Marie, Michèle Damotte, Daniel Birnbaum, Jeannine Trucy, and Bertrand R. Jordan. "HLA cosmid clones show complete, widely spaced human class I genes with occasional clusters." *Gene* 20, no. 3 (1982): 485-489.
- [55] Jordan, B. R., D. Caillol, M. Damotte, T. Delovitch, P. Ferrier, B. Kahn-Perles, F. Kourilsky, C. Layet, P. Bouteiller Le, and F. A. Lemonnier. "HLA class I genes: from structure to expression, serology and function." *Immunological Reviews* 84 (1985): 73-92.

- [56] Little, A. M., and P. Parham. "Polymorphism and evolution of HLA class I and II genes and molecules." *Reviews in Immunogenetics* 1, no. 1 (1999): 105-123.
- [57] Sullivan, L. C., C. S. Clements, J. Rossjohn, and A. G. Brooks. "The major histocompatibility complex class Ib molecule HLA-E at the interface between innate and adaptive immunity." *Tissue Antigens* 72, no. 5 (2008): 415-424.
- [58] Geraghty, Daniel E., Beverly H. Koller, and Harry T. Orr. "A human major histocompatibility complex class I gene that encodes a protein with a shortened cytoplasmic segment." *Proceedings of the National Academy of Sciences* 84, no. 24 (1987): 9145-9149.
- [59] Carosella, Edgardo D., Philippe Moreau, Joël Le Maoult, Magali Le Discorde, Jean Dausset, and Nathalie Rouas-Freiss. "HLA-G molecules: from maternal-fetal tolerance to tissue acceptance." *Advances in Immunology* 81 (2003): 199-252.
- [60] Bjorkman, Pamela J., M. A. Saper, B. Samraoui, William S. Bennett, J. L. Strominger, and D. C. Wiley. "Structure of the human class I histocompatibility antigen, HLA-A2." *Nature* 329, no. 6139 (1987): 506.
- [61] Rammensee H-G, Bachmann J, Stevanovic S. MHC ligands and peptide motifs. New York: Springer, (1997).
- [62] Parham, P., C. E. Lomen, D. A. Lawlor, J. P. Ways, N. Holmes, H. L. Coppin, R. D. Salter, A. M. Wan, and P. D. Ennis. "Nature of polymorphism in HLA-A, -B, and -C molecules." *Proceedings of the National Academy of Sciences* 85, no. 11 (1988): 4005-4009.
- [63] Klein, J. A. N., and Akie Sato. "The HLA system." *New England Journal of Medicine* 343, no. 10 (2000): 702-709.
- [64] Connolly, Janet M., Ted H. Hansen, Amie L. Ingold, and Terry A. Potter. "Recognition by CD8 on cytotoxic T lymphocytes is ablated by several substitutions in the class I alpha 3 domain: CD8 and the T-cell receptor recognize the same class I molecule." *Proceedings of the National Academy of Sciences* 87, no. 6 (1990): 2137-2141.
- [65] Foroni, Iris, Ana Rita Couto, Bruno Filipe Bettencourt, Margarida Santos, Manuela Lima, and Jácome Bruges-Armas. "HLA-E, HLA-F and HLA-G—the non-classical side of the MHC cluster." In: *HLA and Associated Important Diseases. IntechOpen*, (2014).
- [66] Braud, V. Jones, E. Y. and McMichael, A. The human major histocompatibility complex class Ib molecule HLA-E binds signal sequence-derived peptides with primary anchor residues at positions 2 and 9. *Eur J Immunol*, Vol. 27, (5), (1997), pp. 1164-1169
- [67] Goodridge, J. P. Burian, A. Lee, N. and Geraghty, D. E.. HLA-F complex without peptide binds to MHC class I protein in the open conformer form. *J Immunol*, Vol. 184, (11), (2010), pp. 6199-6208
- [68] Carosella, E. D. Favier, B. Rouas-Freiss, N. Moreau, P. and Lemaoult, J. Beyond the increasing complexity of the immunomodulatory HLA-G molecule. *Blood*, Vol. 111, (10), (2008), pp. 4862-4870
- [69] Germain, Ronald N., and David H. Margulies. "The biochemistry and cell biology of antigen processing and presentation." *Annual Review of Immunology* 11, no. 1 (1993): 403-450.
- [70] Viret, C., and Jr CA Janeway. "MHC and T cell development." *Reviews in Immunogenetics* 1, no. 1 (1999): 91-104.
- [71] Ljunggren, Hans-Gustaf, and Klas Kärre. "In search of the 'missing self': MHC molecules and NK cell recognition." *Immunology Today* 11 (1990): 237-244.

- [72] Béziat, Vivien, Hugo G. Hilton, Paul J. Norman, and James A. Traherne. "Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology." *Immunology* 150, no. 3 (2017): 248-264.
- [73] Serenius, Bo, Kenth Gustafsson, Eva Widmark, Eva Emmoth, Goran Andersson, Dan Larhammar, Lars Rask, and Per A. Peterson. "Molecular map of the human HLA-SB (HLA-DP) region and sequence of an SB alpha (DP alpha) pseudogene." *The EMBO Journal* 3, no. 13 (1984): 3209-3214.
- [74] Marsh, Steven GE, E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernández-Viña et al. "Nomenclature for factors of the HLA system, 2010." *Tissue Antigens* 75, no. 4 (2010): 291-455.
- [75] Robbins, F., Hurley, C.K., Tang, T., Yao, H., Lin, Y.S., Wade, J., Goeken, N. and Hartzman, R.J. Diversity associated with the second expressed HLA-DRB locus in the human population. *Immunogenetics*, 46(2), (1997), pp.104-110.
- [76] [Mark Kunkel](#), [Jamie Duke](#), [Deborah Ferriola](#), [Curt Lind](#) and [Dimitri Monos](#). "Molecular Methods for Human Leukocyte Antigen Typing: Current Practices and Future Directions" In *Manual of Molecular and Clinical Laboratory Immunology, Eighth Edition*, pp. 1069-1090. American Society of Microbiology, 2016.
- [77] Senju, Satoru, Akinori Kimura, Michio Yasunami, Nobuhiro Kamikawaji, Hideyuki Yoshizumi, Yasuharu Nishimura, and Takehiko Sasazuki. "Allele-specific expression of the cytoplasmic exon of HLA-DQB1 gene." *Immunogenetics* 36, no. 5 (1992): 319-325.
- [78] Kaufman, James F., Charles Auffray, Alan J. Korman, Deborah A. Shackelford, and Jack Strominger. "The class II molecules of the human and murine major histocompatibility complex." *Cell* 36, no. 1 (1984): 1-13.
- [79] Rammensee HG. Chemistry of peptides associated with MHC class I and class II molecules. *Curr Opin Immunol.* (1995);7:85–96.
- [80] James, Eddie A., Antonis K. Moustakas, John Bui, Randi Nouv, George K. Papadopoulos, and William W. Kwok. "The binding of antigenic peptides to HLA-DR is influenced by interactions between pocket 6 and pocket 9." *The Journal of Immunology* 183, no. 5 (2009): 3249-3258.
- [81] Painter, Corrie A., and Lawrence J. Stern. "Conformational variation in structures of classical and non-classical MHCII proteins and functional implications." *Immunological Reviews* 250, no. 1 (2012): 144-157.
- [82] Cresswell, Peter. "Assembly, transport, and function of MHC class II molecules." *Annual Review of Immunology* 12, no. 1 (1994): 259-291.
- [83] Luckheeram, Rishi Vishal, Rui Zhou, Asha Devi Verma, and Bing Xia. "CD4+ T cells: differentiation and functions." *Clinical and Developmental Immunology* 2012 (2012).
- [84] Klein, Ludger, Bruno Kyewski, Paul M. Allen, and Kristin A. Hogquist. "Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see)." *Nature Reviews Immunology* 14, no. 6 (2014): 377.
- [85] Rajalingam, Raja, Michael Cecka, and Elaine F. Reed. "Molecular HLA typing methods used in clinical laboratories." In *Molecular Diagnostics*, pp. 367-379. Academic Press, 2010.
- [86] Holdsworth, R., Hurley, C.K., Marsh, S.G.E., Lau, M., Noreen, H.J., Kempenich, J.H., Setterholm, M. and Maiers, M.. The HLA dictionary 2008: a summary of HLA-A,-B,-C,-DRB1/3/4/5, and-DQB1 alleles and their association with serologically defined HLA-A,-B,-C,-DR, and-DQ antigens. *Tissue Antigens*, 73(2), (2009), pp.95-170.
- [87] Robinson, James, Anup R. Soormally, James D. Hayhurst, and Steven GE Marsh. "The IPD-IMGT/HLA Database—New developments in reporting HLA variation." *Human Immunology* 77(3), (2016): 233-237.



- [88] Schmitz J.L. HLA Typing Using Molecular Methods. In: Coleman W.B., Tsongalis G.J. (Eds) *Molecular Diagnostics. Humana Press* (2006).
- [89] Prugnolle, Franck, Andrea Manica, Marie Charpentier, Jean François Guégan, Vanina Guernier, and François Balloux. "Pathogen-driven selection and worldwide HLA class I diversity." *Current Biology* 15, no. 11 (2005): 1022-1027.
- [90] Malek Kamoun, Jill A Hollenbach, Steven J Mack, Thomas M Williams. *Molecular Pathology in Clinical Practice. Molecular HLA Typing*. 2016; 867-885.
- [91] Erlich, Henry A., and Ulf B. Gyllensten. "Shared epitopes among HLA class II alleles: gene conversion, common ancestry and balancing selection." *Immunology Today* 12, no. 11 (1991): 411-414.
- [92] Cullen, M., Perfetto, S.P., Klitz, W., Nelson, G. and Carrington, M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *The American Journal of Human Genetics*, 71(4), (2002), pp.759-776.
- [93] Parham, Peter, and Tomoko Ohta. "Population biology of antigen presentation by MHC class I molecules." *Science* 272, no. 5258 (1996): 67-74.
- [94] Marsh, Steven GE. "HLA nomenclature and the IMGT/HLA Sequence Database." In Novartis Foundation symposium, pp. 165-176. Chichester; New York; John Wiley; 1999, 2003.
- [95] Lau, M., Park, M.S., Terasaki, P.I., 1997. International Cell Exchange 1974–1996, a 23-year documentation. In: Terasaki, P.I., Gjertson, D.W. (Eds.), *HLA. UCLA Tissue Typing Laboratory, Los Angeles*, (1997), pp. 85–124.
- [96] Nunes, Eduardo, Helen Heslop, Marcelo Fernandez-Vina, Cynthia Taves, Dawn R. Wagenknecht, A. Bradley Eisenbrey, Gottfried Fischer et al. "Definitions of histocompatibility typing terms: harmonization of histocompatibility typing terms working group." *Human Immunology* 72, no. 12 (2011): 1214-1216.
- [97] Single, R. M., and G. Thomson. Linkage Disequilibrium: Population Genetics of Multiple Loci. In: *Encyclopedia of Evolutionary Biology, Academic Press* (2016), pp. 400-404.
- [98] Hedrick, P.W. Gametic disequilibrium measures: Proceed with caution. *Genetics* (1987) 117,331–341.
- [99] Lewontin, R.C. On measures of gametic disequilibrium. *Genetics* (1988) 120,849–852.
- [100] Yunis, Edmond J., Charles E. Larsen, Marcelo Fernandez-Viña, Zuheir L. Awdeh, T. Romero, John A. Hansen, and Chester A. Alper. "Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks." *Tissue Antigens* 62, no. 1 (2003): 1-20.
- [101] Askar M, Daghestani J, Thomas D, Leahy N, Dunn P, Claas F, Doran S, Saji H, Kanangat S, Karoichane M, Tambur A, Monos D, El-Khalifa M, Turner V, Kamoun M, Mustafa M, Ramon D, Gandhi M, Vernaza A, Gorodezky C, Wagenknecht D, Gautreaux M, Hajeer A, Kashi Z and Fernandez-Vina M. 16th IHIW: Global distribution of extended HLA haplotypes. *International Journal of Immunogenetics* 40, no. 1 (2013): 31-38.
- [102] Takahata, Naoyuki, Yoko Satta, and J. Klein. "Polymorphism and balancing selection at major histocompatibility complex loci." *Genetics* 130, no. 4 (1992): 925-938.
- [103] Sanchez-Mazas, Alicia, Jean-François Lemaître, and Mathias Currat. "Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, no. 1590 (2012): 830-839.
- [104] Meyer, Diogo, Vitor RC Aguiar, Bárbara D. Bitarello, Débora YC Brandt, and Kelly Nunes. "A genomic perspective on HLA evolution." *Immunogenetics* 70, no. 1 (2018): 5-27.

- [105] Meyer, Diogo, and Glenys Thomson. "How selection shapes variation of the human major histocompatibility complex: a review." *Annals of Human Genetics* 65, no. 1 (2001): 1-26.
- [106] Bronson, Paola G., Steven J. Mack, Henry A. Erlich, and Montgomery Slatkin. "A sequence-based approach demonstrates that balancing selection in classical human leukocyte antigen (HLA) loci is asymmetric." *Human Molecular Genetics* 22, no. 2 (2012): 252-261.
- [107] Cagliani R and Sironi M. Pathogen-driven selection in the human genome. *Int J Evol Biol* (2013):1–6
- [108] Penn DJ, Damjanovich K and Potts WK. MHC Heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A* 99(17), (2002):11,260–11,264.
- [109] Matzaraki, Vasiliki, Vinod Kumar, Cisca Wijmenga, and Alexandra Zhernakova. "The MHC locus and genetic susceptibility to autoimmune and infectious diseases." *Genome Biology* 18, no. 1 (2017): 76.
- [110] Illing, Patricia T., Anthony W. Purcell, and James McCluskey. "The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions." *Immunogenetics* 69, no. 8-9 (2017): 617-630.
- [111] L. Cavalli-Sforza, P. Menozzi, and A. Piazza, *The History and Geography of Human Genes*, Princeton University Press, (1994) Princeton, NJ, USA.
- [112] Ingulli, Elizabeth. "Mechanism of cellular rejection in transplantation." *Pediatric Nephrology* 25, no. 1 (2010): 61.
- [113] Masouridi-Levrat, Stavroula, Federico Simonetta, and Yves Chalandon. "Immunological basis of bone marrow failure after allogeneic hematopoietic stem cell transplantation." *Frontiers in Immunology* 7 (2016): 362.
- [114] Negrin, Robert S. "Graft-versus-host disease versus graft-versus-leukemia." *ASH Education Program Book* 2015, no. 1 (2015): 225-230.
- [115] Kissmeyer-Nielsen F, Thorsby E. Human transplantation antigens. *Transplant Rev* (1970): 4: 1–176.
- [116] Vredevoe DL, Terasaki PI, Mickey MR, Glasscock R, Merrill JP, Murray JE. Serotyping of human lymphocyte antigens. III. Long term kidney homograft survivors. In: Amos DB, van Rood JJ, eds. *Histocompatibility Testing* (1965). Copenhagen: Munksgaard, 1965, 25–35.
- [117] Nowak, J. "Role of HLA in hematopoietic SCT." *Bone Marrow Transplantation* 42, no. S2 (2008): S71.
- [118] Fernández-Viña MA, Klein JP, Haagensohn M, Spellman SR, Anasetti C, Noreen H, Baxter-Lowe LA, Cano P, Flomenberg N, Confer DL, Horowitz MM, Oudshoorn M, Petersdorf EW, Setterholm M, Champlin R, Lee SJ and de Lima M. Multiple mismatches at the low expression HLA loci DP, DQ, and DRB3/4/5 associate with adverse outcomes in hematopoietic stem cell transplantation. *Blood* 121, no. 22 (2013): 4603-4610.
- [119] Madden, K., and Chabot-Richards, D. HLA testing in the molecular diagnostic laboratory. *Virchows Archiv* 474, no. 2 (2019): 139-147.
- [120] Trowsdale, J. and Knight, J.C. Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, 14, (2013), pp.301-323.
- [121] Shieh, M., Chitnis, N. and Monos, D. Human Leukocyte Antigen and Disease Associations: A Broader Perspective. *Clinics in Laboratory Medicine*, 38(4), (2018), pp.679-693.
- [122] Schlosstein, Lee, Paul I. Terasaki, Rodney Bluestone, and Carl M. Pearson. "High association of an HL-A antigen, W27, with ankylosing spondylitis." *New England Journal of Medicine* 288, no. 14 (1973): 704-706.

- [123] Díaz-Peña R, López-Vázquez A, López-Larrea C (2012) Old and new HLA associations with ankylosing spondylitis. *Tissue Antigens* 80(3):205–213
- [124] Reveille JD (2014) An update on the contribution of the MHC to as susceptibility. *Clin Rheumatol* 33(6):749–757
- [125] Holoshitz, Joseph. "The quest for better understanding of HLA-disease association: scenes from a road less travelled by." *Discovery Medicine* 16, no. 87 (2013): 93.
- [126] Bharadwaj, M., Illing, P., Theodossis, A., Purcell, A.W., Rossjohn, J. and McCluskey, J. Drug hypersensitivity and human leukocyte antigens of the major histocompatibility complex. *Annual Review of Pharmacology and Toxicology*, 52, (2012), pp.401-431.
- [127] Illing, P.T., Purcell, A.W. and McCluskey, J. The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions. *Immunogenetics*, 69, (2017),(8-9), pp.617-630.
- [128] Panneerchelvam, S. and Norazmi, M.N. HLA Polymorphism in Anthropology. *INTECH Open Access Publisher*, (2012), pp.p1-18.
- [129] Riccio ME, Buhler S, Nunes JM, Vangenot C, Cuénod M, Currat M, Di D, Andreani M, Boldyreva M, Chambers G, Chernova M, Chiaroni J, Darke C, Di Cristofaro J, Dubois V, Dunn P, Edinur HA, Elamin N, Eliaou JF, Grubic Z, Jaatinen T, Kanga U, Kervaire B, Kolesar L, Kunachiwa W, Lokki ML, Mehra N, Nicoloso G, Paakkanen R, Voniatis DP, Papasteriades C, Poli F, Richard L, Romón Alonso I, Slavčev A, Sulcebe G, Suslova T, Testi M, Tiercy JM, Varnavidou A, Vidan-Jeras B, Wennerström A and Sanchez-Mazas A. 16th IHIW: analysis of HLA population data, with updated results for 1996 to 2012 workshop data (AHPD project report). *International Journal of Immunogenetics* 40, no. 1 (2013): 21-30.
- [130] González-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H.T., Silva, A.L.S.D., Silva, A.L.T.E., Ghataoraya, G.S., Alfirovic, A., Jones, A.R. and Middleton, D. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research*, 43(D1), (2014), pp.D784-D788. Last accessed October 21, 2019.
- [131] Sanchez-Mazas, A., Fernandez-Viña, M., Middleton, D., Hollenbach, J.A., Buhler, S., Di, D., Rajalingam, R., Dugoujon, J.M., Mack, S.J. and Thorsby, E. Immunogenetics as a tool in anthropological studies. *Immunology*, 133(2), (2011), pp.143-164.
- [132] Fernandez Vina MA, Hollenbach JA, Lyke KE, Szein MB, Maiers M, Klitz W, Cano P, Mack S, Single R, Brautbar C, Israel S, Raimondi E, Khoriaty E, Inati A, Andreani M, Testi M, Moraes ME, Thomson G, Stastny P and Cao K. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, no. 1590 (2012): 820-829.
- [133] Buhler, S. and Sanchez-Mazas, A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLOS One*, 6(2), (2011), p.e14643.
- [134] Solberg, O.D., Mack, S.J., Lancaster, A.K., Single, R.M., Tsai, Y., Sanchez-Mazas, A. and Thomson, G. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Human Immunology*, 69(7), (2008), pp.443-464.
- [135] Von Salomé, J., Gyllensten, U. and Bergström, T.F. Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics*, 59(4), (2007), pp.261-271.
- [136] Sanchez-Mazas, A., Buhler, S. and Nunes, J.M. A new HLA map of Europe: regional genetic variation and its implication for peopling history, disease-association studies and tissue transplantation. *Human Heredity*, 76(3-4), (2013), pp.162-177.

- [137] Sanchez-Mazas, A. and Meyer, D. The relevance of HLA sequencing in population genetics studies. *Journal of Immunology Research*, (2014).
- [138] Monos, D.S., Tekolf, W.A., Shaw, S. and Cooper, H.L. Comparison of structural and functional variation in class I HLA molecules: the role of charged amino acid substitutions. *The Journal of Immunology*, 132(3), (1984), pp.1379-1385.
- [139] Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B. and Erlich, H.A. Analysis of enzymatically amplified  $\beta$ -globin and HLA-DQ $\alpha$  DNA with allele-specific oligonucleotide probes. *Nature*, 324(6093), (1986), p.163.
- [140] Olerup, O. and Zetterquist, H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. *Tissue Antigens*, 39(5), (1992), pp.225-235.
- [141] Cereb, N., Maye, P., Lee, S., Kong, Y. and Yang, S.Y. Locus-specific amplification of HLA class I genes from genomic DNA: locus-specific sequences in the first and third introns of HLA-A,-B, and-C alleles. *Tissue Antigens*, 45(1), (1995), pp.1-11.
- [142] Hosomichi, K., Shiina, T., Tajima, A. and Inoue, I. The impact of next-generation sequencing technologies on HLA research. *Journal of Human Genetics*, 60(11), (2015), p.665.
- [143] Dunn, P. P. J. "Human leucocyte antigen typing: techniques and technology, a critical appraisal." *International Journal of Immunogenetics* 38, no. 6 (2011): 463-473.
- [144] Erlich, H. "HLA DNA typing: past, present, and future." *Tissue Antigens* 80, no. 1 (2012): 1-11.
- [145] Parham, P., Benjamin, R.J., Chen, B.P., Clayberger, C., Ennis, P.D., Krensky, A.M., Lawlor, D.A., Littman, D.R., Norment, A.M., Orr, H.T. and Salter, R.D. Diversity of class I HLA molecules: functional and evolutionary interactions with T cells. In: *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 54, pp. 529-543), (1989). *Cold Spring Harbor Laboratory Press*.
- [146] Lind, C., D. Ferriola, K. Mackiewicz, A. Papazoglou, A. Sasson, and D. Monos. "Filling the gaps—the generation of full genomic sequences for 15 common and well-documented HLA class I alleles using next-generation sequencing technology." *Human Immunology* 74, no. 3 (2013): 325-329.
- [147] Cano, P., Klitz, W., Mack, S.J., Maiers, M., Marsh, S.G., Noreen, H., Reed, E.F., Senitzer, D., Setterholm, M., Smith, A. and Fernández-Viña, M. Common and well-documented HLA alleles: report of the Ad-Hoc committee of the American Society for Histocompatibility and Immunogenetics. *Human Immunology*, 68(5), (2007), pp.392-417.
- [148] Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, Reed EF, Setterholm M, Smith AG, Tilanus MG, Torres M, Varney MD, Voorter CE, Fischer GF, Fleischhauer K, Goodridge D, Klitz W, Little AM, Maiers M, Marsh SG, Müller CR, Noreen H, Rozemuller EH, Sanchez-Mazas A, Senitzer D, Trachtenberg E, Fernandez-Vina M. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens*, 81(4), (2013), pp.194-203.
- [149] Sanchez-Mazas A, Nunes JM, Middleton D, Sauter J, Buhler S, McCabe A, Hofmann J, Baier DM, Schmidt AH, Nicoloso G, Andreani M, Grubic Z, Tiercy JM and Fleischhauer K. Common and well-documented HLA alleles over all of Europe and within European sub-regions: a catalogue from the European Federation for Immunogenetics. *HLA* 89, no. 2 (2017): 104-113.
- [150] He Y, Li J, Mao W, Zhang D, Liu M, Shan X, Zhang B, Zhu C, Shen J, Deng Z, Wang Z, Yu W, Chen Q, Guo W, Su P, Lv R, Li G, Li G, Pei B, Jiao L, Shen G, Liu Y, Feng Z, Su Y, Xie Y, Di W, Liu X, Yang X, Wang J, Qi J, Liu Q, Han Y, He J, Cai J, Zhang Z, Zhu F and Du D. HLA common and well-documented alleles in China. *HLA* 92, no. 4 (2018): 199-205.

- [151] Metzker, M.L. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1), (2010), p.31.
- [152] Erlich, H.A. HLA typing using next generation sequencing: An overview. *Human Immunology*, 76(12), (2015), pp.887-890.
- [153] Edited by Monos, D. and Maiers, M. Issue: Single Molecule DNA Sequencing. *Human Immunology*, 76(12), (2015), pp.883-982.
- [154] Zhou M, Gao D, Chai X, Liu J, Lan Z, Liu Q, Yang F, Guo Y, Fang J, Yang L, Du D, Chen L, Yang X, Zhang M, Zeng H, Lu J, Chen H, Zhang X, Wu S, Han Y, Tan J, Cheng Z, Huang C and Wang W. Application of high-throughput, high-resolution and cost-effective next generation sequencing-based large-scale HLA typing in donor registry. *Tissue Antigens*, 85(1), (2015), pp.20-28.
- [155] Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, Heron S, Huynh A, McLaughlin L, Rogers M, Slavich L, Walker R and Monos DS. Determining performance characteristics of an NGS-based HLA typing method for clinical applications. *HLA*, 87(3), (2016), pp.141-152.
- [156] Weimer, E.T., Montgomery, M., Petraroia, R., Crawford, J. and Schmitz, J.L. Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. *The Journal of Molecular Diagnostics*, 18(5), (2016), pp.668-675.
- [157] Gandhi, M.J., Ferriola, D., Huang, Y., Duke, J.L. and Monos, D. Targeted next-generation sequencing for human leukocyte antigen typing in a clinical laboratory: metrics of relevance and considerations for its successful implementation. *Archives of Pathology and Laboratory Medicine*, 141(6), (2017), pp.806-812.
- [158] Schadt, E.E., Turner, S. and Kasarskis, A. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2), (2010), pp.R227-R240.
- [159] Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin CS, Bowman B, Marks P, Braund H, Madrigal JA, Latham K and Marsh SG. HLA typing for the next generation. *PLoS One*, 10(5), (2015), p.e0127153.
- [160] Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R. and Schatz, M. Third-generation sequencing and the future of genomics. *BioRxiv*, (2016), p.048603.
- [161] Bravo-Egana, V. and Monos, D. The impact of next-generation sequencing in immunogenetics: current status and future directions. *Current Opinion in Organ Transplantation*, 22(4), (2017), pp.400-406.
- [162] Chen, C.Y. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Frontiers in Microbiology*, 5, (2014), p.305.
- [163] Besser, J., Carleton, H.A., Gerner-Smidt, P., Lindsey, R.L. and Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, 24(4), (2018), pp.335-341.
- [164] Ameer, A., Kloosterman, W.P. and Hestand, M.S. Single-molecule sequencing: towards clinical applications. *Trends in Biotechnology*, 37(1), (2019), pp.72-85.
- [165] Turner TR, Guijarro C, Robinson J, et al. Single molecule real-time (SMRT) sequencing of full-length HLA-DRB1. *HLA*. (2016);87:276.
- [166] Mayor NP, Brosnan N, Midwinter W, et al. A multiplexed typing strategy for the HLA class II genes HLA-DRB1, -DQB1 and -DPB1 using DNA barcodes and SMRT® DNA sequencing. *HLA*. (2016);87:276.
- [167] Robinson J, Guijarro C, Leen G, et al. Addressing the bioinformatics challenges of high throughput HLA typing using SMRT® DNA sequencing. *HLA*. (2016);87:271.

- [168] Turner, T.R., Hayhurst, J.D., Hayward, D.R., Bultitude, W.P., Barker, D.J., Robinson, J., Madrigal, J.A., Mayor, N.P. and Marsh, S.G.E. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. *HLA* 91(2), (2018), pp.88-101.
- [169] Lang, K., Surendranath, V., Quenzel, P., Schöfl, G., Schmidt, A.H. and Lange, V., 2018. Full-length HLA class I genotyping with the MinION nanopore sequencer. In: *HLA Typing, Methods Mol Bio*, (2018), (pp. 155-162). *Humana Press*, New York, NY.
- [170] Ton, K.N., Cree, S.L., Gronert-Sum, S.J., Merriman, T.R., Stamp, L.K. and Kennedy, M.A. Multiplexed nanopore sequencing of HLA-B locus in Māori and Pacific Island samples. *Frontiers in Genetics*, 9, (2018), p.152.
- [171] Liu, C., Xiao, F., Hoisington-Lopez, J., Lang, K., Quenzel, P., Duffy, B. and Mitra, R.D. Accurate typing of human leukocyte antigen class I genes by Oxford Nanopore sequencing. *The Journal of Molecular Diagnostics*, 20(4), (2018), pp.428-435.
- [172] Albrecht, V., Zweiniger, C., Surendranath, V., Lang, K., Schöfl, G., Dahl, A., Winkler, S., Lange, V., Böhme, I. and Schmidt, A.H. Dual redundant sequencing strategy: Full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *HLA*, 90(2), (2017), pp.79-87.
- [173] Suzuki S, Ranade S, Osaki K, Ito S, Shigenari A, Ohnuki Y, Oka A, Masuya A, Harting J, Baybayan P, Kitazume M, Sunaga J, Morishima S, Morishima Y, Inoko H, Kulski JK and Shiina T. Reference grade characterization of polymorphisms in full-length HLA class I and II genes with short-read sequencing on the Ion PGM system and long-reads generated by Single Molecule, Real-time Sequencing on the PacBio platform. *Frontiers in Immunology* 9 (2018): 2294.
- [174] Major, E., Rigo, K., Hague, T., Berces, A. and Juhos, S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLOS One*, 8(11), (2013), p.e78410.
- [175] Kidd JM, Sharpton TJ, Bobo D, Norman PJ, Martin AR, Carpenter ML, Sikora M, Gignoux CR, Nemat-Gorgani N, Adams A, Guadalupe M, Guo X, Feng Q, Li Y, Liu X, Parham P, Hoal EG, Feldman MW, Pollard KS, Wall JD, Bustamante CD and Henn BM. Exome capture from saliva produces high quality genomic and metagenomic data. *BMC Genomics*, 15(1), (2014), p.262.
- [176] Meyer, D. and Nunes, K. Editorial: HLA imputation, what is it good for? *Human Immunology*, 78(3), (2017), pp.239-241.
- [177] Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, Ng SC, Rosati E, Hübenenthal M, Ellinghaus D, Jung ES, Lieb W, Abedian S, Malekzadeh R, Cheon JH, Ellul P, Sood A, Midha V, Thelma BK, Wong SH, Schreiber S, Yamazaki K, Kubo M, Boucher G, Rioux JD, Lenz TL, Brant SR and Franke A. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Human Molecular Genetics* 28, no. 12 (2018): 2078-2092.
- [178] Carapito, R., Radosavljevic, M. and Bahram, S Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. *Human Immunology*, 77(11), (2016), pp.1016-1023.
- [179] Yin Y, Lan JH, Nguyen D, Valenzuela N, Takemura P, Bolon YT, Springer B, Saito K, Zheng Y, Hague T, Pasztor A, Horvath G, Rigo K, Reed EF and Zhang Q. Application of high-throughput next-generation sequencing for HLA typing on buccal extracted DNA: results from over 10,000 donor recruitment samples. *PLOS One*, 11(10), (2016), p.e0165810.
- [180] Bentley, G., Higuchi, R., Høglund, B., Goodridge, D., Sayer, D., Trachtenberg, E.A. and Erlich, H.A. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*, 74(5), (2009), pp.393-403.

- [181] Holcomb CL, Höglund B, Anderson MW, Blake LA, Böhme I, Egholm M, Ferriola D, Gabriel C, Gelber SE, Goodridge D, Hawbecker S, Klein R, Ladner M, Lind C, Monos D, Pando MJ, Pröll J, Sayer DC, Schmitz-Agheguian G, Simen BB, Thiele B, Trachtenberg EA, Tyan DB, Wassmuth R, White S and Erlich HA. A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue Antigens*, 77(3), (2011), pp.206-217.
- [182] Gabriel, C., Danzer, M., Hackl, C., Kopal, G., Hufnagl, P., Hofer, K., Polin, H., Stabentheiner, S. and Pröll, J. Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Human Immunology*, 70(11), (2009), pp.960-964.
- [183] Moonsamy, P.V., Williams, T., Bonella, P., Holcomb, C.L., Höglund, B.N., Hillman, G., Goodridge, D., Turenchalk, G.S., Blake, L.A., Daigle, D.A. and Simen, B.B. High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ system for simplified amplicon library preparation. *Tissue Antigens*, 81(3), (2013), pp.141-149.
- [184] Grumbt, B., Eck, S.H., Hinrichsen, T. and Hirv, K. Diagnostic applications of next generation sequencing in immunogenetics and molecular oncology. *Transfusion Medicine and Hemotherapy*, 40(3), (2013), pp.196-206.
- [185] Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D and Monos D. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Human Immunology*, 71(10), (2010), pp.1033-1042.
- [186] Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, Hayashi Y, Paumen M, Katsuyama Y, Mitsunaga S, Ota M, Kulski JK and Inoko H. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens*, 80(4), (2012), pp.305-316.
- [187] Wang, C., Krishnakumar, S., Wilhelmy, J., Babrzadeh, F., Stepanyan, L., Su, L.F., Levinson, D., Fernandez-Viña, M.A., Davis, R.W., Davis, M.M. and Mindrinis, M. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proceedings of the National Academy of Sciences*, 109(22), (2012), pp.8676-8681.
- [188] Gabriel, C., Fürst, D., Faé, I., Wenda, S., Zollikofer, C., Mytilineos, J. and Fischer, G.F. HLA typing by next-generation sequencing—getting closer to reality. *Tissue Antigens*, 83(2), (2014), pp.65-75.
- [189] Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G and Schmidt AH. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*, 15(1), (2014), p.63.
- [190] Hosomichi, K., Jinam, T.A., Mitsunaga, S., Nakaoka, H. and Inoue, I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC Genomics*, 14(1), (2013), p.355.
- [191] Goldenstein, M.F., Vinhal, F., Moral, F., Fidelis, R. and Rodrigues, A.L. P001 Validation of ALLType™ next-generation sequencing in Brazilian samples. *Human Immunology*, 79, (2018), pp.58-59.
- [192] Profaizer, T. and Kumánovics, A. Human Leukocyte Antigen Typing by Next-Generation Sequencing. *Clinics in Laboratory Medicine*, 38(4), (2018), pp.565-578.
- [193] Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, Hov JR, Sauer S, Schimmeler M, Ziemann M, Görg S, Jacob F, Karlsen TH and Franke A. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Research*, 43(11), (2015), pp.e70-e70.
- [194] Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, Jayaraman J, Wroblewski EE, Trowsdale J, Rajalingam R, Oksenberg JR, Chiaroni J, Guethlein LA, Traherne JA, Ronaghi

- M and Parham P. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *The American Journal of Human Genetics*, 99(2), (2016), pp.375-391.
- [195] Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, Dunn T, Mann T, Alicata C, Hollenbach JA, Chang W, Shults Won M, Gunderson KL, Abi-Rached L, Ronaghi M and Parham P. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Research*, 27(5), (2017), pp.813-823.
- [196] Jiao, Y., Li, R., Wu, C., Ding, Y., Liu, Y., Jia, D., Wang, L., Xu, X., Zhu, J., Zheng, M. and Jia, J. High-sensitivity HLA typing by Saturated Tiling Capture Sequencing (STC-Seq). *BMC Genomics*, 19(1), (2018), p.50.
- [197] Krishnakumar, S., Li, M., Wang, C., Osada, M., Kuehn, R., Fukushima, M. and Thorstenson, Y. Validation of the MIA FORA NGS FLEX Assay Using Buccal Swabs as the Sample Source. Immucor, Inc. (2017).
- [198] Ambardar, S., Gupta, R., Trakroo, D., Lal, R. and Vakhlu, J. High throughput sequencing: an overview of sequencing chemistry. *Indian Journal of Microbiology*, 56(4), (2016), pp.394-404.
- [199] Buermans, H.P.J. and Den Dunnen, J.T. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10), (2014), pp.1932-1941.
- [200] Ehrenberg, P.K., Geretz, A., Baldwin, K.M., Apps, R., Polonis, V.R., Robb, M.L., Kim, J.H., Michael, N.L. and Thomas, R. High-throughput multiplex HLA genotyping by next-generation sequencing using multi-locus individual tagging. *BMC Genomics*, 15(1), (2014), p.864.
- [201] Osoegawa, K., Mack, S.J., Udell, J., Noonan, D.A., Ozanne, S., Trachtenberg, E. and Prestegaard, M. HLA Haplotype Validator for quality assessments of HLA typing. *Human Immunology*, 77(3), (2016), pp.273-282.
- [202] Juhos, S., Rigó, K. and Horváth, G. On Genotyping Polymorphic HLA Genes-Ambiguities and Quality Measures Using NGS. In: Next Generation Sequencing-Advances, Applications and Challenges. *IntechOpen*. (2016).
- [203] Fazekas, A.J., Steeves, R. and Newmaster, S.G. Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques*, 48(4), (2010), pp.277-285.
- [204] Duke, J.L., Mosbrugger, T.L., Ferriola, D., Chitnis, N., Hu, T., Tairis, N., Margolis, D.J. and Monos, D.S. Resolving MiSeq-generated ambiguities in HLA-DPB1 typing by using the Oxford Nanopore technology. *The Journal of Molecular Diagnostics*, 21(5), (2019), pp. 852-861.
- [205] Bachtel ND, Umviligihozo G, Pickering S, Mota TM, Liang H, Del Prete GQ, Chatterjee P, Lee GQ, Thomas R, Brockman MA, Neil S, Carrington M, Bwana B, Bangsberg DR, Martin JN, Kallas EG, Donini CS, Cerqueira NB, O'Doherty UT, Hahn BH, Jones RB, Brumme ZL, Nixon DF and Apps R. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nature Genetics*, 41(12), (2009), p.1290.
- [206] Kennedy, A.E., Ozbek, U. and Dorak, M.T. What has GWAS done for HLA and disease associations?. *International Journal of Immunogenetics*, 44(5), (2017), pp.195-211.
- [207] Clark, P.M., Kunkel, M. and Monos, D.S. The dichotomy between disease phenotype databases and the implications for understanding complex diseases involving the major histocompatibility complex. *International Journal of Immunogenetics*, 42(6), (2015), pp.413-422.
- [208] Kempenich, J.H., Setterholm, M. and Maiers, M. Haplotype associations of 90 rare alleles from the National Marrow Donor Program. *Tissue Antigens*, 67(4), (2006), pp.284-289.



- [209] Gonzalez-Galarza, F. F., S. J. Mack, J. Hollenbach, M. Fernandez-Vina, M. Setterholm, J. Kempenich, S. G. E. Marsh, A. R. Jones, D. Middleton, and HLA Rare Allele Consortium. 16th IHIW: Extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles. *International Journal of Immunogenetics* 40, no. 1 (2013): 60-65.
- [210] Elsner, H.A. and Blasczyk, R. Immunogenetics of HLA null alleles: implications for blood stem cell transplantation. *Tissue Antigens*, 64(6), (2004), pp.687-695.
- [211] Schiller, J., Barakat, M., Fisher, T., Gaba, S., Haase, T., Waalkes, C. and Nytes, J. Novel HLA Alleles Identified by High Resolution NGS Typing Impact HCT Donor Selection. *Biology of Blood and Marrow Transplantation*, 24(3), (2018), p.S421.
- [212] Allen, E.S., Yang, B., Garrett, J., Ball, E.D., Maiers, M. and Morris, G.P. Improved accuracy of clinical HLA genotyping by next-generation DNA sequencing affects unrelated donor search results for hematopoietic stem cell transplantation. *Human Immunology*, 79(12), (2018), pp.848-854.
- [213] Baier, D.M., Hofmann, J.A., Fischer, H., Rall, G., Stolze, J., Ruhner, K., Lange, V., Sauter, J. and Schmidt, A.H. Very low error rates of NGS-based HLA typing at stem cell donor recruitment question the need for a standard confirmatory typing step before donor work-up. *Bone Marrow Transplantation*, (2019), 54(6), p.928.
- [214] Dehn J, Spellman S, Hurley CK, Shaw BE, Barker JN, Burns LJ, Confer DL, Eapen M, Fernandez-Vina MA, Hartzman R, Maiers M, Marino SR, Mueller C, Perales MA, Rajalingam R. and Pidala J. Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from NMDP/CIBMTR. *Blood*, (2019) pp.blood-2019001212.
- [215] Tiercy, J.M. How to select the best available related or unrelated donor of hematopoietic stem cells?. *Haematologica*, 101(6), (2016), pp.680-687.
- [216] Vazirabad, I., Chhabra, S., Nytes, J., Mehra, V., Narra, R.K., Szabo, A., Jerkins, J.H., Dhakal, B., Hari, P. and Anderson, M.W. Direct HLA Genetic Comparisons Identify Highly Matched Unrelated Donor-Recipient Pairs with Improved Transplantation Outcome. *Biology of Blood and Marrow Transplantation*, 25(5), (2019), pp.921-931.
- [217] Mayor NP, Hayhurst JD, Turner TR, Szydlo RM, Shaw BE, Bultitude WP, Sayno JR, Tavarozzi F, Latham K, Anthias C, Robinson J, Braund H, Danby R, Perry J, Wilson MC, Bloor AJ, McQuaker IG, MacKinnon S, Marks DI, Pagliuca A, Potter MN, Potter VT, Russell NH, Thomson KJ, Madrigal JA and Marsh SGE. Recipients receiving better HLA-matched hematopoietic cell transplantation grafts, uncovered by a novel HLA typing method, have superior survival: a retrospective study. *Biology of Blood and Marrow Transplantation*, 25(3), (2019), pp.443-450.
- [218] Hurley, C.K., Spellman, S., Dehn, J., Barker, J.N., Devine, S., Fernandez-Vina, M., Gautreaux, M., Logan, B., Maiers, M., Mueller, C., Perales, M.A., Yu N and Pidala J. Regarding “Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study”. *Biology of Blood and Marrow Transplantation*, (2019), 25(8):e268-e269.
- [219] Hou, L., Vierra-Green, C., Lazaro, A., Brady, C., Haagenson, M., Spellman, S. and Hurley, C.K. Limited HLA sequence variation outside of antigen recognition domain exons of 360 10 of 10 matched unrelated hematopoietic stem cell transplant donor-recipient pairs. *HLA*, 89(1), (2017), pp.39-46.
- [220] Tiercy, J.M. and Claas, F. Impact of HLA diversity on donor selection in organ and stem cell transplantation. *Human Heredity*, 76(3-4), (2013), pp.178-186.

- [221] Enrich, E., Campos, E., Martorell, L., Herrero, M.J., Vidal, F., Querol, S. and Rudilla, F.. HLA-A,-B,-C,-DRB1 and-DQB1 allele and haplotype frequencies: An analysis of umbilical cord blood units at the Barcelona Cord Blood Bank. *HLA*, (2019).
- [222] Damien, P. and Allan, D.S. Regenerative therapy and immune modulation using umbilical cord blood-derived cells. *Biology of Blood and Marrow Transplantation*, 21(9), (2015), pp.1545-1554.
- [223] Guerra, S.G., Hamilton-Jones, S., Brown, C.J., Navarrete, C.V. and Chong, W. Next generation sequencing of 11 HLA loci characterises a diverse UK cord blood bank. *Human Immunology*, (2020).
- [224] Hurley, C.K., Hou, L., Lazaro, A., Gerfen, J., Enriquez, E., Galarza, P., Rodriguez Cardozo, M.B., Halagan, M., Maiers, M., Behm, D. and Ng, J. Next generation sequencing characterizes the extent of HLA diversity in an Argentinian registry population. *HLA*, 91(3), (2018), pp.175-186.
- [225] Hou, L., Enriquez, E., Persaud, M., Steiner, N., Oudshoorn, M. and Hurley, C.K. Next generation sequencing characterizes HLA diversity in a registry population from the Netherlands. *HLA*, 93(6), (2019), pp.474-483.
- [226] Lande, A., Andersen, I., Egeland, T., Lie, B.A. and Viken, M.K. HLA-A,-C,-B,-DRB1,-DQB1 and-DPB1 allele and haplotype frequencies in 4514 healthy Norwegians. *Human Immunology*, 79(7), (2018), pp.527-529.
- [227] Schöfl, G., Lang, K., Quenzel, P., Böhme, I., Sauter, J., Hofmann, J.A., Pingel, J., Schmidt, A.H. and Lange, V. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics*, 18(1), (2017), p.161.
- [228] Dehn, J., Setterholm, M., Buck, K., Kempenich, J., Beduhn, B., Gragert, L., Madbouly, A., Fingerson, S. and Maiers, M. HapLogic: a predictive human leukocyte antigen-matching algorithm to enhance rapid identification of the optimal unrelated hematopoietic stem cell sources for transplantation. *Biology of Blood and Marrow Transplantation*, 22(11), (2016), pp.2038-2046.
- [229] Petersdorf EW, Gooley TA, Malkki M, Bacigalupo AP, Cesbron A, Du Toit E, Ehninger G, Egeland T, Fischer GF, Gervais T, Haagenson MD, Horowitz MM, Hsu K, Jindra P, Madrigal A, Oudshoorn M, Ringdén O, Schroeder ML, Spellman SR, Tiercy JM, Velardi A, Witt CS, O'Huigin C, Apps R and Carrington M. HLA-C expression levels define permissible mismatches in hematopoietic cell transplantation. *Blood*, 124(26), (2014), pp.3996-4003.
- [230] Petersdorf, E.W., Malkki, M., O'huigin, C., Carrington, M., Gooley, T., Haagenson, M.D., Horowitz, M.M., Spellman, S.R., Wang, T. and Stevenson, P. High HLA-DP expression and graft-versus-host disease. *New England Journal of Medicine*, 373(7), (2015), pp.599-609.
- [231] S. Morishima, T. Shiina, S. Suzuki, S. Ogawa, A. Sato-Otsubo, K. Kashiwase, F. Azuma, T. Yabe, M. Satake, S. Kato, Y. Kodera, T. Sasazuki, Y. Morishima and Japan Marrow Donor Program. Evolutionary basis of HLA-DPB1 alleles affects acute GVHD in unrelated donor stem cell transplantation. *Blood*, 131, (2018), pp. 808-817
- [232] Schöne, B., Bergmann, S., Lang, K., Wagner, I., Schmidt, A.H., Petersdorf, E.W. and Lange, V. Predicting an HLA-DPB1 expression marker based on standard DPB1 genotyping: linkage analysis of over 32,000 samples. *Human Immunology*, 79(1), (2018), pp.20-27.
- [233] Linjama, T., Räther, C., Ritari, J., Peräsaari, J., Eberhard, H.P., Korhonen, M. and Koskela, S. Extended HLA Haplotypes and Their Impact on DPB1 Matching of Unrelated Hematologic Stem Cell Transplant Donors. *Biology of Blood and Marrow Transplantation*, 25(10), (2019), pp. 1956-1964.

- [234] Crivello, P., Zito, L., Sizzano, F., Zino, E., Maiers, M., Mulder, A., Toffalori, C., Naldini, L., Ciceri, F., Vago, L. and Fleischhauer, K. The impact of amino acid variability on alloreactivity defines a functional distance predictive of permissive HLA-DPB1 mismatches in hematopoietic stem cell transplantation. *Biology of Blood and Marrow Transplantation*, 21(2), (2015), pp.233-241.
- [235] Arrieta-Bolaños, E., Crivello, P., Shaw, B.E., Ahn, K.W., Wang, H.L., Verneris, M.R., Hsu, K.C., Pidala, J., Lee, S.J., Fleischhauer, K. and Spellman, S.R. In silico prediction of nonpermissive HLA-DPB1 mismatches in unrelated HCT by functional distance. *Blood Advances*, 2(14), (2018), pp.1773-1783.
- [236] Meurer, T., Arrieta-Bolaños, E., Metzger, M., Langer, M.M., Van Balen, P., Falkenburg, J.F., Beelen, D.W., Horn, P.A., Fleischhauer, K. and Crivello, P. Dissecting genetic control of HLA-DPB1 expression and its relation to structural mismatch models in hematopoietic stem cell transplantation. *Frontiers in Immunology*, 9, (2018).
- [237] Mariano, L., Zhang, B.M., Osoegawa, K., Lowsky, R. and Fernandez-Vina, M. Assessment by Extended-Coverage Next-Generation Sequencing Typing of DPA1 and DPB1 Mismatches in Siblings Matching at HLA-A,-B,-C,-DRB1, and-DQ Loci. *Biology of Blood and Marrow Transplantation*, (2019).
- [238] Howard, C.A., Fernandez-Vina, M.A., Appelbaum, F.R., Confer, D.L., Devine, S.M., Horowitz, M.M., Mendizabal, A., Laport, G.G., Pasquini, M.C. and Spellman, S.R. Recommendations for donor human leukocyte antigen assessment and matching for allogeneic stem cell transplantation: consensus opinion of the Blood and Marrow Transplant Clinical Trials Network (BMT CTN). *Biology of Blood and Marrow Transplantation*, 21(1), (2015), pp.4-7.
- [239] Hricik D.E. Primer on Transplantation. 3rd Ed. Hoboken, NJ: *Wiley-Blackwell*; (2011).
- [240] Shen, S.W., Chang, C.K., Gao, Y.S., Hsu, P.J., Cheng, S.C., Liu, F.Y. and Lo, S.C. Establishment of calculated panel reactive antibody and its potential benefits in improving the kidney allocation strategy in Taiwan. *Journal of the Formosan Medical Association*, 116(12), (2017), pp.956-963.
- [241] Hart, A., Smith, J.M., Skeans, M.A., Gustafson, S.K., Wilk, A.R., Robinson, A., Wainright, J.L., Haynes, C.R., Snyder, J.J., Kasiske, B.L. and Israni, A.K. OPTN/SRTR 2016 annual data report: kidney. *American Journal of Transplantation*, 18, (2018), pp.18-113.
- [242] Claas, F.H., Dankers, M.K., Oudshoorn, M., van Rood, J.J., Mulder, A., Roelen, D.L., Duquesnoy, R.J. and Doxiadis, I.I. Differential immunogenicity of HLA mismatches in clinical transplantation. *Transplant Immunology*, 14(3-4), (2005), pp.187-191.
- [243] Argani, H. Anti-HLA Antibody: The Role of Epitopes in Organ Transplantation. *Middle East Society for Organ Transplantation*, 1, (2018), pp.38-42.
- [244] Duquesnoy, R.J. A structurally based approach to determine HLA compatibility at the humoral immune level. *Human Immunology*, 67(11), (2006), pp.847-862.
- [245] Duquesnoy, R.J., Kamoun, M., Baxter-Lowe, L.A., Woodle, E.S., Bray, R.A., Claas, F.H.J., Eckels, D.D., Friedewald, J.J., Fuggle, S.V., Gebel, H.M. and Gerlach, J.A. Should HLA mismatch acceptability for sensitized transplant candidates be determined at the high-resolution rather than the antigen level?. *American Journal of Transplantation*, 15(4), (2015), pp.923-930.
- [246] Kallon, D., Navarrete, C.V., Sage, D.A., Stanworth, S., Mufti, G.J., Marsh, J.C.W. and Brown, C.J. Impact of Human Leucocyte Antigen epitope matched platelet transfusions in alloimmunised aplastic anaemia patients. *Transfusion Medicine*, (2019).
- [247] Duquesnoy, R.J. and Marrari, M. HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. II. Verification of the algorithm and determination of the relative immunogenicity of amino acid triplet-defined epitopes. *Human Immunology*, 63(5), (2002), pp.353-363.

- [248] Geneugelijk, K., Thus, K.A. and Spierings, E. Predicting alloreactivity in transplantation. *Journal of Immunology Research*, 2014.
- [249] Tambur, A.R. HLA-Epitope Matching or Eplet Risk Stratification: The Devil Is in the Details. *Frontiers in Immunology*, (9), (2018), p.2010.
- [250] Tambur, A.R. and Claas, F.H.J. HLA epitopes as viewed by antibodies: what is it all about?. *American Journal of Transplantation*, 15(5), (2015), pp.1148-1154.
- [251] Cecka, J.M., Reed, E.F. and Zachary, A.A.. HLA High-Resolution Typing for Sensitized Patients: A Solution in Search of a Problem?. *American Journal of Transplantation*, 15(4), (2015), pp.855-856.
- [252] Alzahrani, M., Qahtani, Z., Harbi, H., Kebasi, S., Essa, O. and Al Attas, R. March. Virtual Crossmatch: Reality of Perception. In: *Transplantation Proceedings, Elsevier*. 51(2), (2019), pp. 488-491.
- [253] Lachmann, N., Todorova, K., Schulze, H., & Schönemann, C. (2013). Luminex® and its applications for solid organ transplantation, hematopoietic stem cell transplantation, and transfusion. *Transfusion Medicine and Hemotherapy*, 40(3), 182-189.
- [254] Gebel, H. M., & Bray, R. A. HLA antibody detection with solid phase assays: great expectations or expectations too great?. *American Journal of Transplantation*, 14(9), (2014), 1964-1975.
- [255] Vogiatzi, P. Some considerations on the current debate about typing resolution in solid organ transplantation. *Transplantation Research*, 5(1), (2016), p.3.
- [256] Kotowski, M., Bogacz, A., Bartkowiak-Wieczorek, J., Bukowska, A., Surowiec, N., Dziewanowski, K., Czerny, B., Grześkowiak, E., Ostrowski, M., Machaliński, B. and Sieńko, J. The Importance of New Generation Sequencing (NGS) HLA Typing in Renal Transplantation—Preliminary Report. In: *Transplantation proceedings Elsevier*. 50(6), (2018), pp. 1605-1615.
- [257] Smith, A.G., Pereira, S., Jaramillo, A., Stoll, S.T., Khan, F.M., Berka, N., Mostafa, A.A., Pando, M.J., Usenko, C.Y., Bettinotti, M., Pyo, C.W, Nelson, W.C., Willis, A., Askar, M. and Geraghty, D.E. Comparison of sequence-specific oligonucleotide probe vs next generation sequencing for HLA-A, B, C, DRB1, DRB3/B4/B5, DQA1, DQB1, DPA1, and DPB1 typing: Toward single-pass high-resolution HLA typing in support of solid organ and hematopoietic cell transplant programs. *HLA*, (2019), 94(3):296-306.
- [258] Huang, Y., Dinh, A., Heron, S., Gasiewski, A., Kneib, C., Mehler, H., Mignogno, M.T., Morlen, R., Slavich, L., Kentzel, E., Frackelton, E.C., Duke JL, Ferriola D, Mosbrugger T, Timofeeva OA, Geier SS and Monos D. Assessing the utilization of high-resolution 2-field HLA typing in solid organ transplantation. *American Journal of Transplantation*. 19(7), (2019), pp. 1955-1963
- [259] Gragert, L., Madbouly, A., Freeman, J. and Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology*, 74(10), (2013), pp.1313-1320.
- [260] Romòn, I., Montes, C., Ligeiro, D., Trindade, H., Sanchez-Mazas, A., Nunes, J.M. and Buhler, S. Mapping the HLA diversity of the Iberian Peninsula. *Human Immunology*, 77(10), (2016), pp.832-840.
- [261] Tambur AR, Campbell P, Claas FH, Feng S, Gebel HM, Jackson AM, Mannon RB, Reed EF, Tinckam K, Askar M, Chandraker A, Chang PP, Colvin M, Demetris AJ, Diamond JM, Dipchand AI, Fairchild RL, Ford ML, Friedewald J, Gill RG1, Glotz D1, Goldberg H, Hachem R, Knechtle S, Kobashigawa J, Levine DJ, Levitsky J, Mengel M, Milford E, Newell KA, O'Leary JG, Palmer S, Randhawa P, Smith J, Snyder L, Starling RC, Sweet S, Taner T, Taylor CJ, Woodle S, Zeevi A and Nickerson P. Sensitization in transplantation: assessment of risk (STAR) 2017 working group meeting report. *American Journal of Transplantation*, 18(7), (2018), pp.1604-1614.

- [262] Yamamoto, F., Wang, L., Chang, C.J. and Tyan, D.B. Mapping and definition of HLA class I and II serologic epitopes using an unbiased reverse engineering strategy. *Human Immunology* 80(9), (2019), pp. 668-702.
- [263] Chang, C.J., Yamamoto, F. and Tyan, D.B. A Reverse-Engineering Strategy Utilizing the Integration of Single Antigen Beads and NGS HLA Genotypes to Detect Potential Antibody Inducing Epitopes. *Human Immunology* 80(9), (2019), pp. 661-667.
- [264] June Jones AMJ, Lucas DP, Bettinotti M. Donor specific antibody assessments using HLA P groups: the devil is in the detail. *Human Immunology*, 79(Supplement), (2018), pp.12
- [265] Edited by Monos, D. and Drake, J. Special Issue: HLA Functional Elements Outside the Antigen Recognition Domains. *Human Immunology*, 80(1), (2019), pp.1-90.
- [266] Harton J, Jin L, Hahn A and Drake J. Immunological Functions of the Membrane Proximal Region of MHC Class II Molecules. *F1000Research*, 5(F1000 Faculty Rev), (2016), pp. 368
- [267] González-Quezada, B.A., Creary, L.E., Munguia-Saldaña, A.J., Flores-Aguilar, H., Fernández-Viña, M.A. and Gorodezky, C. Exploring the ancestry and admixture of mexican oaxaca mestizos from southeast mexico using next-generation sequencing of 11 HLA loci. *Human Immunology*. 80(3), (2019), pp.157-162.
- [268] Creary, L.E., Gangavarapu, S., Mallempati, K.C., Montero-Martín, G., Caillier, S.J., Santaniello, A., Hollenbach, J.A., Oksenberg, J.R. and Fernández-Viña, M.A. Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population. *Human Immunology* 80(10), (2019), pp. 807-822.
- [269] Montero-Martín G, Mallempati KC, Gangavarapu S, Sánchez-Gordo F, Herrero-Mata MJ, Balas A, Vicario JL, Sánchez-García F, González-Escribano MF, Muro M, Moya-Quiles MR, González-Fernández R, Oejo-Vinyals JG, Marín L, Creary LE, Osoegawa K, Vayntrub T, Caro-Oleas JL, Vilches C, Planelles D and Fernández-Viña MA. High-resolution characterization of allelic and haplotypic HLA frequency distribution in a Spanish population using high-throughput next-generation sequencing. *Human Immunology*, 80(7), (2019), pp. 429-436.
- [270] Geretz, A., Ehrenberg, P.K., Bouckenoghe, A., Viña, M.A.F., Michael, N.L., Chansinghakule, D., Limkittikul, K. and Thomas, R. Full-length next-generation sequencing of HLA class I and II genes in a cohort from Thailand. *Human Immunology*, 79(11), (2018), pp. 773-780.
- [271] Hajeer, A.H., Al Balwi, M.A., Aytül Uyar, F., Alhaidan, Y., Alabdulrahman, A., Al Abdulkareem, I. and Al Jumah, M. HLA-A,-B,-C,-DRB1 and-DQB1 allele and haplotype frequencies in Saudis using next generation sequencing technique. *Tissue Antigens*, 82(4), (2013), pp.252-258.
- [272] Thorstenson YR, Creary LE, Huang H, Rozot V, Nguyen TT, Babrzadeh F, Kancharla S, Fukushima M, Kuehn R, Wang C, Li M, Krishnakumar S, Mindrinos M, Fernandez Viña MA, Scriba TJ, Davis MM. Allelic resolution NGS HLA typing of Class I and Class II loci and haplotypes in Cape Town, South Africa. *Human Immunology*, 79(12), (2018), pp.839-847.
- [273] Goeury, T., Creary, L.E., Brunet, L., Galan, M., Pasquier, M., Kervaire, B., Langaney, A., Tiercy, J.M., Fernández-Viña, M.A., Nunes, J.M. and Sanchez-Mazas, A. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*, 91(1), (2018), pp.36-51.
- [274] Seshasubramanian, V., Sathishkannan, A.D., Naganathan, C., Periathiruvadi, S., Nandakumar, Y. and Narayan, S. P101 Distribution of HLA-A, -B, -C, -DRB1 and -DQB1 alleles and haplotypes among different South Indian population. *HLA*, 93(5), (2019), pp.337-338.
- [275] Arlehamn, C.S.L., Copin, R., Leary, S., Mack, S.J., Phillips, E., Mallal, S., Sette, A., Blatner, G., Siefers, H., Ernst, J.D. and TBRU-ASTRA Study Team. Sequence-based HLA-A, B, C, DP, DQ, and DR

- typing of 100 Luo infants from the Boro area of Nyanza Province, Kenya. *Human Immunology*, 78(4), (2017), pp.325-326.
- [276] Grifoni, A., Sidney, J., Carpenter, C., Phillips, E., Mallal, S., Scriba, T.J., Sette, A. and Arlehamn, C.S.L. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 159 individuals from the Worcester region of the Western Cape province of South Africa. *Human Immunology*, 79(3), (2018), pp.143-144.
- [277] Weiskopf, D., Grifoni, A., Arlehamn, C.S.L., Angelo, M., Leary, S., Sidney, J., Frazier, A., Mack, S.J., Phillips, E., Mallal, S. and Cerpas, C. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 339 adults from Managua, Nicaragua. *Human Immunology*, 79(1), (2018), pp.1-2.
- [278] Grifoni A, Weiskopf D, Lindestam Arlehamn CS, Angelo M, Leary S, Sidney J, Frazier A, Phillips E, Mallal S, Mack SJ, Tippalagama R, Goonewardana S, Premawansa S, Premawansa G, Wijewickrama A, De Silva AD, Sette A. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 714 adults from Colombo, Sri Lanka. *Human Immunology*, 79(2), (2018), pp.87-88.
- [279] Sonon P, Sadissou I, Tokplonou L, M'po KKG, Glitho SSC, Agniwo P, Ibikounlé M, Massaro JD, Massougbdji A, Moreau P, Sabbagh A, Mendes-Junior CT, Moutairou KA, Castelli EC, Courtin D, Donadi EA. HLA-G,-E and-F regulatory and coding region variability and haplotypes in the Beninese Toffin population sample. *Molecular Immunology*, 104, (2018), pp.108-127.
- [280] Lima, T.H.A., Buttura, R.V., Donadi, E.A., Veiga-Castelli, L.C., Mendes-Junior, C.T. and Castelli, E.C. HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. *Human Immunology*, 77(10), (2016), pp.841-853.
- [281] Castelli, E.C., Mendes-Junior, C.T., Sabbagh, A., Porto, I.O., Garcia, A., Ramalho, J., Lima, T.H., Massaro, J.D., Dias, F.C., Collares, C.V. and Jamonneau, V. HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Human Immunology*, 76(12), (2015), pp.945-953.
- [282] Castelli EC, Gerasimou P, Paz MA, Ramalho J, Porto IOP, Lima THA, Souza AS, Veiga-Castelli LC, Collares CVA, Donadi EA, Mendes-Junior CT and Costeas P. HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographically distinct population samples of Brazil and Cyprus. *Molecular Immunology*, 83, (2017), pp.115-126.
- [283] Veiga-Castelli, L., Oliveira, M.L.D., Pereira, A., Debortoli, G., Marcorin, L., Fracasso, N., Silva, G., Souza, A., Massaro, J., Simões, A.L. and Sabbagh, A. HLA-G Polymorphisms Are Associated with Non-segmental Vitiligo among Brazilians. *Biomolecules*, 9(9), (2019), p.463.
- [284] Kountouris, E., Li, M. and Mindrinos, M. OR25 Construction of high resolution haplotype database for human leukocyte antigen loci. *Human Immunology*, 78, (2017), p.24.
- [285] Ameen, R., Al Shemmari, S. and Askar, M. Next-generation sequencing characterization of HLA in multi-generation families of Kuwaiti descent. *Human Immunology*, 79(3), (2018), pp.137-142.
- [286] Osoegawa K, Mallempati KC, Gangavarapu S, Oki A, Gendzekhadze K, Marino SR, Brown NK, Bettinotti MP, Weimer ET, Montero-Martín G, Creary LE, Vayntrub TA, Chang CJ, Askar M, Mack SJ and Fernández-Viña MA. HLA alleles and haplotypes observed in 263 US families. *Human Immunology* 80(9), (2019), pp. 644-660.
- [287] Askar M, Madbouly A, Zhrebker L, Willis A, Kennedy S, Padros K, Rodriguez MB, Bach C, Spriewald B, Ameen R, Shemmari SA, Tarassi K, Tsirogianni A, Hamdy N, Mossallam G, Hönger G, Spinnler R, Fischer G, Fae I, Charlton R, Dunk A, Vayntrub TA, Halagan M, Osoegawa K and Fernández-Viña M. HLA Haplotypes In 250 Families: The Baylor Laboratory Results And A Perspective On A Core NGS Testing Model For The 17th International HLA And Immunogenetics Workshop. *Human Immunology* 80(11), (2019), pp. 897-905.

- [288] Kishore, A. and Petrek, M. Next-generation sequencing based HLA typing: deciphering immunogenetic aspects of sarcoidosis. *Frontiers in Genetics*, 9, (2018), p.503.
- [289] Judson, M.A. A sarcoidosis clinician's perspective of MHC functional elements outside the antigen binding site. *Human Immunology*, 80(1), (2019), pp.85-89.
- [290] Misra, M.K., Damotte, V. and Hollenbach, J.A. The immunogenetics of neurological disease. *Immunology*, 153(4), (2018), pp.399-414.
- [291] Hollenbach, J.A., Norman, P.J., Creary, L.E., Damotte, V., Montero-Martin, G., Caillier, S., Anderson, K.M., Misra, M.K., Nemat-Gorgani, N., Osoegawa, K. and Santaniello, A. A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proceedings of the National Academy of Sciences*, 116(15), (2019), pp.7419-7424.
- [292] Mack, S.J., Udell, J., Cohen, F., Osoegawa, K., Hawbecker, S.K., Noonan, D.A., Ladner, M.B., Goodridge, D., Trachtenberg, E.A., Oksenberg, J.R. and Erlich, H.A. High resolution HLA analysis reveals independent class I haplotypes and amino-acid motifs protective for multiple sclerosis. *Genes & Immunity*, 20(4), (2019), p.308. (Repeated in [727]).
- [293] Creary, L.E., Mallempati, K.C., Gangavarapu, S., Caillier, S.J., Oksenberg, J.R. and Fernández-Viña, M.A. Deconstruction of HLA-DRB1\*04:01:01 and HLA-DRB1\*15:01:01 class II haplotypes using next-generation sequencing in European-Americans with multiple sclerosis. *Multiple Sclerosis Journal*, 25(6), (2019), pp.772-782.
- [294] Ogawa, K., Okuno, T., Hosomichi, K., Hosokawa, A., Hirata, J., Suzuki, K., Sakaue, S., Kinoshita, M., Asano, Y., Miyamoto, K. and Inoue, I. Next-generation sequencing identifies contribution of both class I and II HLA genes on susceptibility of multiple sclerosis in Japanese. *Journal of Neuroinflammation*, 16(1), (2019), pp.1-9.
- [295] The IPD-IMGT/HLA database, <https://www.ebi.ac.uk/ipd/imgt/hla/>, last accessed July 15, 2020.
- [296] Klasberg, S., Surendranath, V., Lange, V. and Schöfl, G., Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping. *Transfusion Medicine and Hemotherapy*, (2019) pp.1-13.
- [297] The 17th IHIW NGS HLA data, <http://17ihiw.org/17th-ihw-ngs-hla-data/>, last accessed July 15, 2020.
- [298] Klitz, W., Maiers, M., Spellman, S., Baxter-Lowe, L.A., Schmeckpeper, B., Williams, T.M. and Fernandez-Viña, M. New HLA haplotype frequency reference standards: high-resolution and large sample typing of HLA DR-DQ haplotypes in a sample of European Americans. *Tissue Antigens*, 62(4), (2003), pp.296-307.
- [299] NMDP Haplotype Frequencies Webpage, <https://bioinformatics.bethematchclinical.org/hla-resources/haplotype-frequencies/be-the-match-registry-haplotype-frequencies/>, last accessed July 15, 2020.
- [300] Fernandez-Vina, M.A., Wang, C., Krishnakumar, S., Levinson, D.F., Davis, R.W. and Mindrinos, M. LBP19: extended coverage by next generation sequencing methods refines the characterization of the common and well documented HLA alleles. *Human Immunology*, 76(4), (2015), p.226.
- [301] Sanchez-Mazas, A. and Nunes, J.M. Does NGS typing highlight our understanding of HLA population diversity? Some good reasons to say yes and a few to say be careful. *Human Immunology*, 80(1), (2019), pp.62-66.
- [302] Brown, N.K., Kheradmand, T., Wang, J. and Marino, S.R. Identification and characterization of novel HLA alleles: Utility of next-generation sequencing methods. *Human Immunology*, 77(4), (2016), pp.313-316.

- [303] Adamek, M., Klages, C., Bauer, M., Kudlek, E., Drechsler, A., Leuser, B., Scherer, S., Opelz, G. and Tran, T.H. Seven novel HLA alleles reflect different mechanisms involved in the evolution of HLA diversity: description of the new alleles and review of the literature. *Human Immunology*, 76(1), (2015), pp.30-35.
- [304] Qi, J., Wang, T.J., Chen, L.P., Wang, X.F., Wang, M.N. and Wu, J.H., 2018. Utility of next-generation sequencing methods to identify the novel HLA alleles in potential stem cell donors from Chinese Marrow Donor Program. *International journal of Immunogenetics*, 45(4), (2018), pp.225-229.
- [305] Gourraud, P.A., Khankhanian, P., Cereb, N., Yang, S.Y., Feolo, M., Maiers, M., Rioux, J.D., Hauser, S. and Oksenberg, J. HLA diversity in the 1000 genomes dataset. *PLOS One*, 9(7), (2014), p.e97282.
- [306] Robinson, J., Guethlein, L.A., Cereb, N., Yang, S.Y., Norman, P.J., Marsh, S.G. and Parham, P. Distinguishing functional polymorphism from random variation in the sequences of > 10,000 HLA-A,-B and-C alleles. *PLOS Genetics*, 13(6), (2017).
- [307] Lima, T.H., Souza, A.S., Porto, I.O., Paz, M.A., Veiga-Castelli, L.C., Oliveira, M.L.G., Donadi, E.A., Meyer, D., Sabbagh, A., Mendes-Junior, C.T. and Castelli, E.C. HLA-A promoter, coding, and 3' UTR sequences in a Brazilian cohort, and their evolutionary aspects. *HLA*, 93(2-3), (2019), pp.65-79.
- [308] Barsakis, K., Babrzadeh, F., Chi, A., Mallempati, K., Pickle, W., Mindrinos, M. and Fernández-Viña, M.A. Complete nucleotide sequence characterization of DRB5 alleles reveals a homogeneous allele group that is distinct from other DRB genes. *Human Immunology*, 80(7), (2019), pp.437-448.
- [309] Petersdorf, E.W. and O'Huigin, C. The MHC in the era of next-generation sequencing: Implications for bridging structure with function. *Human Immunology*, 80(1), (2018), pp. 67-78
- [310] Hoarau JJ, Cesari M, Caillens H, Cadet F, Pabion M. HLA-DQA1 genes generate multiple transcripts by alternative splicing and polyadenylation of the 3' untranslated region. *Tissue Antigens*, 63, (2004), pp.58–71.
- [311] Davis DM, Mandelboim O, Luque I, Baba E, Boyson J, Strominger JL. The transmembrane sequence of human histocompatibility leukocyte antigen (HLA)-C as a determinant in inhibition of a subset of natural killer cells. *Journal of Experimental Medicine*. 189(8), (1999), pp. 1265–74.
- [312] Drake LA, Drake JR. A triad of molecular regions contribute to the formation of two distinct MHC class II conformers. *Molecular Immunology*. 74(59), (2016), pp. 70.
- [313] Van Hateren A, James E, Bailey A, Phillips A, Dalchau N, Elliott T. The cell biology of major histocompatibility complex class I assembly: towards a molecular understanding. *Tissue Antigens*. 76(4), (2010), pp. 259–75.
- [314] Kampstra, A.S., van Heemst, J., Janssen, G.M., de Ru, A.H., van Lummel, M., van Veelen, P.A. and Toes, R.E. Ligandomes obtained from different HLA-class II-molecules are homologous for N-and C-terminal residues outside the peptide-binding cleft. *Immunogenetics*, (2019), pp.1-12.
- [315] Heldt, C., Listing, J., Sözeri, O., Bläsing, F., Frischbutter, S. and Müller, B. Differential expression of HLA class II genes associated with disease susceptibility and progression in rheumatoid arthritis. *Arthritis & Rheumatism*, 48(10), (2003), pp. 2779-2787.
- [316] Liu, B., Fu, Y., Wang, Z., Zhou, S., Sun, Y., Wu, Y. and Xu, A. HLA-DRB1 May Be Antagonistically Regulated by the Coordinately Evolved Promoter and 3'-UTR under Stabilizing Selection. *PLOS One*, 6(10), (2011), p.e25794.
- [317] Perfetto, C., Zacheis, M., McDaid, D., Meador III, J.W. and Schwartz, B.D. Polymorphism in the promoter region of HLA-DRB genes. *Human Immunology*, 36(1), (1993), pp.27-33.
- [318] Thomas R, Apps R, Qi Y, Gao X, Male V, O'Huigin C, O'Connor G, Ge D, Fellay J, Martin JN, Margolick J, Goedert JJ, Buchbinder S, Kirk GD, Martin MP, Telenti A, Deeks SG, Walker BD, Goldstein D,



- McVicar DW, Moffett A and Carrington M.. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nature Genetics*, 41(12), (2009), p.1290.
- [319] Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, Pereyra F, Goldstein D, Wolinsky S, Walker B, Young HA, Carrington M.. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*, 472(7344), (2011), p.495.
- [320] Bachtel, N.D., Umvilighozo, G., Pickering, S., Mota, T., Liang, H., Del Prete, G.Q., Chatterjee, P., Lee, G.Q., Thomas, R., Brockman, M.A. and Neil, S., 2018. HLA-C downregulation by HIV-1 adapts to host HLA genotype. *PLoS Pathogens*, 14(9), (2018).
- [321] Apps R1, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, Yuki Y, Del Prete GQ, Goulder P, Brumme ZL, Brumme CJ, John M, Mallal S, Nelson G, Bosch R, Heckerman D, Stein JL, Soderberg KA, Moody MA, Denny TN, Zeng X, Fang J, Moffett A, Lifson JD, Goedert JJ, Buchbinder S, Kirk GD, Fellay J, McLaren P, Deeks SG, Pereyra F, Walker B, Michael NL, Weintrob A, Wolinsky S, Liao W, Carrington M.. Influence of HLA-C expression level on HIV control. *Science*, 340(6128), (2013), pp.87-91.
- [322] Thomas, R., Thio, C.L., Apps, R., Qi, Y., Gao, X., Marti, D., Stein, J.L., Soderberg, K.A., Moody, M.A., Goedert, J.J. and Kirk, G.D. A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. *Journal of Virology*, 86(12), (2012), pp.6979-6985.
- [323] Ingram, K.J., Merkens, H., O'Shields, E.F., Kiger, D. and Gautreaux, M.D. New HLA alleles discovered by next generation sequencing in routine histocompatibility lab work in a medium-volume laboratory. *Human Immunology*. 80(7), (2019), pp. 465-467
- [324] Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L, Spellman S, Guethlein LA, Trachtenberg EA, Cooley S, Bochtler W, Mueller CR, Robinson J, Marsh SG and Maiers M. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens*, 82(2), (2013), pp.106-112.
- [325] Milius RP, Heuer M, Valiga D, Doroschak KJ, Kennedy CJ, Bolon YT, Schneider J, Pollack J, Kim HR3, Cereb N, Hollenbach JA, Mack SJ, Maiers M. Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping. *Human Immunology*, 76(12), (2015), pp.963-974.
- [326] Mack SJ, Milius RP, Gifford BD, Sauter J, Hofmann J, Osoegawa K, Robinson J, Groeneweg M, Turenchalk GS, Adai A, Holcomb C, Rozemuller EH, Penning MT, Heuer ML, Wang C, Salit ML, Schmidt AH, Parham PR, Müller C, Hague T, Fischer G, Fernandez-Viña M, Hollenbach JA, Norman PJ and Maiers M. Minimum information for reporting next generation sequence genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Human Immunology*, 76(12), (2015), pp.954-962.
- [327] Mack, S.J. A gene feature enumeration approach for describing HLA allele polymorphism. *Human Immunology*, 76(12), (2015), pp.975-981.
- [328] Creary, Lisa E., Chia-Jung Chang, Gonzalo Montero Martin, Kalyan C. Mallempati, Sridevi Gangavarapu, Kazutoyo Osoegawa, Tamara A. Vayntrub, and Marcelo A. Fernandez-Vina. OR24 HLA allele and haplotype frequencies characterized using next-generation sequencing methods in unrelated world-wide populations: Summary from the 17th international HLA and immunogenetics workshop. *Human Immunology* 79(Supplement), (2018), pp. 30.
- [329] Simanovsky, A.L., Madbouly, A., Halagan, M., Maiers, M. and Louzoun, Y. Single haplotype admixture models using large scale HLA genotype frequencies to reproduce human admixture. *Immunogenetics*, (2019).

- [330] Meyer, D., Single, R.M., Mack, S.J., Erlich, H.A. and Thomson, G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*, 173(4), (2006), pp.2121-2142.
- [331] Osoegawa, K., Mack, S.J., Prestegard, M. and Fernández-Viña, M.A. Tools for building, analyzing and evaluating HLA haplotypes from families. *Human Immunology* 80(9), (2019), pp. 633-643.
- [332] Niu, T. Algorithms for inferring haplotypes. *Genetic Epidemiology*, 27, (2004), pp. 334
- [333] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977): 1-22.
- [334] Excoffier, L. and Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5), (1995), pp.921-927.
- [335] Long, J.C., Williams, R.C. and Urbanek, M. An EM algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56(3), (1995), p.799.
- [336] Polańska, J. The EM algorithm and its implementation for the estimation of frequencies of SNP-haplotypes. *International Journal of Applied Mathematics and Computer Science*, 3(13), (2003), pp.419-429.
- [337] Perlin, M. W., Burks, M. B., Hoop, R. C. and Hoffman, E. P. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. *American Journal of Human Genetics*, 55(4), (1994), 777–787.
- [338] Becker, T. and Knapp, M. Efficiency of haplotype frequency estimation when nuclear family information is included. *Human Heredity*, 54(1), (2002), pp.45-53.
- [339] Ikeda, N., Kojima, H., Nishikawa, M., Hayashi, K., Futagami, T., Tsujino, T., Kusunoki, Y., Fujii, N., Suegami, S., Miyazaki, Y., Middleton, D., Tanaka, H. and Saji H. Determination of HLA-A,-C,-B,-DRB1 allele and haplotype frequency in Japanese population based on family study. *Tissue Antigens*, 85(4), (2015), pp.252-259.
- [340] Guo, S. W., and Thompson, E. A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, (1992), 361-372.
- [341] Mack, S.J., Gourraud, P.A., Single, R.M., Thomson, G. and Hollenbach, J.A. Analytical methods for immunogenetic population data. In: *Immunogenetics*, (2012), (pp. 215-244). *Humana Press*.
- [342] Schäfer, C., Schmidt, A.H. and Sauter, J. Hapl-o-Mat: open-source software for HLA haplotype frequency estimation from ambiguous and heterogeneous data. *BMC Bioinformatics*, 18(1), (2017), p.284.
- [343] Holcomb, C.L., Rastrou, M., Williams, T.C., Goodridge, D., Lazaro, A.M., Tilanus, M. and Erlich, H.A. Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens*, 83(1), (2014), pp.32-40.
- [344] Andersson, G. Evolution of the human HLA-DR region, *Frontiers in Bioscience*. 3 (1998) d739-745.
- [345] Madbouly, A., Gragert, L., Freeman, J., Leahy, N., Gourraud, P.A., Hollenbach, J.A., Kamoun, M., Fernandez-Vina, M. and Maiers, M. Validation of statistical imputation of allele-level multilocus phased genotypes from ambiguous HLA assignments. *Tissue Antigens*, 84(3), (2014), pp.285-292.
- [346] Alper, C.A. and Larsen, C.E. Pedigree-Defined Haplotypes and Their Applications to Genetic Studies. In: *Haplotyping*, (2017), (pp. 113-127). *Humana Press*, New York, NY.

- [347] Edited by Mack, S. and Fernandez-Vina, M. Articles from Special Issue: 17th International HLA and Immunogenetics Workshop and Conference. *Human Immunology*, (2017-2019).
- [348] The 17th IHIW website, <http://17ihiw.org>, last accessed July 15, 2020.
- [349] Osoegawa K, Vayntrub TA, Wenda S, De Santis D, Barsakis K, Ivanova M, Hsu S, Barone J, Holdsworth R, Diviney M, Askar M, Willis A, Railton D, Laflin S, Gendzekhadze K, Oki A, Sacchi N, Mazzocco M, Andreani M, Ameen R, Stavropoulos-Giokas C, Dinou A, Torres M, Dos Santos Francisco R, Serra-Pages C, Goodridge D, Balladares S, Bettinotti MP, Iglehart B, Kashi Z, Martin R, Saw CL, Ragoussis J, Downing J, Navarrete C, Chong W, Saito K, Petrek M, Tokic S, Padros K, Beatriz Rodriguez M, Zakharova V, Shragina O, Marino SR, Brown NK, Shiina T, Suzuki S, Spierings E, Zhang Q, Yin Y, Morris GP, Hernandez A, Ruiz P, Khor SS, Tokunaga K, Geretz A, Thomas R, Yamamoto F, Mallempati KC, Gangavarapu S, Kanga U, Tyagi S, Marsh SGE, Bultitude WP, Liu X, Cao D, Penning M, Hurley CK, Cesbron A, Mueller C, Mytilineos J, Weimer ET, Bengtsson M, Fischer G, Hansen JA, Chang CJ, Mack SJ, Creary LE, Fernandez-Viña MA. Quality control project of NGS HLA genotyping for the 17th International HLA and Immunogenetics Workshop. *Human Immunology* 80(4), (2019), pp. 228-236.
- [350] The 17th IHIW NGS HLA tools, <http://17ihiw.org/17th-ihw-tools/>, last accessed July 15, 2020.
- [351] Chang, C.J., Osoegawa, K., Milius, R.P., Maiers, M., Xiao, W., Fernandez-Viña, M. and Mack, S.J. Collection and storage of HLA NGS genotyping data for the 17th International HLA and Immunogenetics Workshop. *Human immunology*, 79(2), (2018), pp.77-86.
- [352] Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P. and Thomson, G. PyPop update--a software pipeline for large-scale multilocus population genomics. *Tissue Antigens*, 69, (2007), pp. 192–197.
- [353] Matern, B.M., Groeneweg, M., Voorter, C.E.M. and Tilanus, M.G.J. Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database. *HLA*, 91(1), (2018), pp.29-35.
- [354] Surendranath, V., Albrecht, V., Hayhurst, J.D., Schöne, B., Robinson, J., Marsh, S.G.E., Schmidt, A.H. and Lange, V. TypeLoader: A fast and efficient automated workflow for the annotation and submission of novel full-length HLA alleles. *HLA*, 90(1), (2017), pp.25-31.
- [355] Misra MK, Augusto DG, Martin GM, Nemat-Gorgani N, Sauter J, Hofmann JA, Traherne JA, González-Quezada B, Gorodezky C, Bultitude WP, Marin W, Vierra-Green C, Anderson KM, Balas A, Caro-Oleas JL, Cisneros E, Colucci F, Dandekar R, Elfishawi SM, Fernández-Viña MA, Fouda M, González-Fernández R, Große A, Herrero-Mata MJ, Hollenbach SQ, Marsh SGE, Mentzer A, Middleton D, Moffett A, Moreno-Hidalgo MA, Mossallam GI, Nakimuli A, Oksenberg JR, Oppenheimer SJ, Parham P, Petzl-Erler ML, Planelles D, Sánchez-García F, Sánchez-Gordo F, Schmidt AH, Trowsdale J, Vargas LB, Vicario JL, Vilches C, Norman PJ and Hollenbach JA. Report from the Killer-cell Immunoglobulin-like Receptors (KIR) component of the 17th International HLA and Immunogenetics Workshop. *Human Immunology*, 79(12), (2018), pp. 825-833.
- [356] Creary LE, Guerra SG, Chong W, Brown CJ, Turner TR, Robinson J, Bultitude WP, Mayor NP, Marsh SGE, Saito K, Lam K, Duke JL, Mosbrugger TL, Ferriola D, Monos D, Willis A, Askar M, Fischer G, Saw CL, Ragoussis J, Petrek M, Serra-Pagés C, Juan M, Stavropoulos-Giokas C, Dinou A, Ameen R, Al Shemmari S, Spierings E, Gendzekhadze K, Morris GP, Zhang Q, Kashi Z, Hsu S, Gangavarapu S, Mallempati KC, Yamamoto F, Osoegawa K, Vayntrub T, Chang CJ, Hansen JA and Fernández-Viña MA. Next-generation HLA typing of 382 International Histocompatibility Working Group reference B-Lymphoblastoid cell lines: report from the 17th International HLA and Immunogenetics Workshop. *Human Immunology*. 80(7), (2019), pp. 449-460.
- [357] Ivanova M, Creary LE, Al Hadra B, Lukanov T, Mazzocco M, Sacchi N, Ameen R, Al-Shemmari S, Moise A, Ursu LD, Constantinescu I, Vayntrub T, Fernández-Viña MA, Shivarov V, Naumova E. 17th IHIW component “Immunogenetics of Ageing”–New NGS data. *Human Immunology*. 80(9), (2019), pp.703-713.

- [358] Liu, C. A long road/read to rapid high-resolution HLA typing: the nanopore perspective. *Human Immunology*, (2020).
- [359] Klasberg, S., Lang, K., Günther, M., Schober, G., Massalski, C., Schmidt, A.H., Lange, V. and Schöfl, G. Patterns of non-ARD variation in more than 300 full-length HLA-DPB1 alleles. *Human Immunology*, 80(1), (2019), pp.44-52.
- [360] Voorter, C.E., Groeneweg, M., Groeneveld, L. and Tilanus, M.G. Uncommon HLA alleles identified by hemizygous ultra-high Sanger sequencing: haplotype associations and reconsideration of their assignment in the Common and Well-Documented catalogue. *Human Immunology*, 77(2), (2016), pp.184-190.
- [361] Voorter CEM, Matern B, Tran TH, Fink A, Vidan-Jeras B, Montanic S, Fischer G, Fae I, de Santis D, Whidborne R, Andreani M, Testi M, Groeneweg M, Tilanus MGJ.. Full-length extension of HLA allele sequences by HLA allele-specific hemizygous Sanger sequencing (SSBT). *Human Immunology*, 79(11), (2018), pp.763-772.
- [362] Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., Marsh, S.G.E. IPD-IMGT/HLA Database. *Nucleic Acids Research* (2019), pii: gkz950. doi: 10.1093/nar/gkz950.
- [363] Maccari, G., Robinson, J., Hammond, J.A. and Marsh, S.G.E. The IPD Project: a centralised resource for the study of polymorphism in genes of the immune system. *Immunogenetics*, (2019), pp.1-7.
- [364] Abraham, J.P., Barker, D.J., Robinson, J., Maccari, G. and Marsh, S.G. The IPD Databases: Cataloguing and Understanding Allele Variants. In: HLA Typing. Methods in Molecular Biology, *Humana Press*, (2018), (pp. 31-48).
- [365] Halagan, M., Wang, W., Bashyal, P., Brelford, J., Kennedy, C., Heuer, M., Milius, B., Bolon, Y.T., Mack, S.J. and Maiers, M. P076 A community resource using gene feature enumeration to generate accurate allele calls and sequence annotations for HLA and KIR. *Human Immunology*, 79(Supplement), (2018), p.117.
- [366] The 18th IHIW website, <https://www.ihiw18.org/>, last accessed Sept 15, 2019.
- [367] Goodwin, S., McPherson, J.D. and McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), (2016), p.333.
- [368] Hu, T., Tairis, N.G., Mosbrugger, T., Jayaraman, P. and Monos, D.S. P053 Optimizing the targeting and sequencing of MHC class II region. *Human Immunology*, 80 (Supplement), (2019), p.94.
- [369] Petersdorf, E.W., Malkki, M., Horowitz, M.M., Spellman, S.R., Haagenson, M.D. and Wang, T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. *Blood*, 121(10), (2013), pp.1896-1905.
- [370] Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, Xu R, Chen G, Zhang Y, Zheng X, Jin X, Gao J, Mei J, Sheng Y, Li Q, Liang B, Shen J, Shen C, Jiang H, Zhu C, Fan X, Xu F, Yue M, Yin X, Ye C, Zhang C, Liu X, Yu L, Wu J, Chen M, Zhuang X, Tang L, Shao H, Wu L, Li J, Xu Y, Zhang Y, Zhao S, Wang Y, Li G, Xu H, Zeng L, Wang J, Bai M, Chen Y, Chen W, Kang T, Wu Y, Xu X, Zhu Z, Cui Y, Wang Z, Yang C, Wang P, Xiang L, Chen X, Zhang A, Gao X, Zhang F, Xu J, Zheng M, Zheng J, Zhang J, Yu X, Li Y, Yang S, Yang H, Wang J, Liu J, Hammarström L, Sun L, Wang J and Zhang X1. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nature Genetics*, 48(7), (2016), p.740.
- [371] Olson, J.A., Gibbens, Y., Tram, K., Kempenich, J., Novakovich, J., Buck, K. and Dehn, J. Identification of a 10/10 matched donor for patients with an uncommon haplotype is unlikely. *HLA*, 89(2), (2017), pp.77-81.
- [372] Aslam, H.M., Iqbal, S.M., Shaikh, H., Faizee, F.A., Merchant, A.A., Shaheen, M. and Hashmi, S.K. Haploidentical Stem Cell Transplantation: A Gateway to Infrequent Availability of HLA-Matched Related Donors. *Case Reports in Medicine*, (2018).

- [373] Andreani, M., Gaspari, S. and Locatelli, F. Human leucocyte antigen diversity: A biological gift to escape infections, no longer a barrier for haploidentical Hemopoietic Stem Cell Transplantation. *International Journal of Immunogenetics*, (2019).
- [374] Cano, P., Wendel, J., Iozzino, C., Boasi, R., Ramirez, Y. and Blaney, K. P92 Documentation of rare HLA alleles. *HLA* 91(5), (2018), 395.
- [375] Cano, P. O15 Evolutionary fate of new alleles. *HLA* 91(5), (2018), 329.
- [376] Klitz, W., Hedrick, P. and Louis, E.J. New reservoirs of HLA alleles: pools of rare variants enhance immune defense. *Trends in Genetics*, 28(10), (2012), pp.480-486.
- [377] Phillips, K.P., Cable, J., Mohammed, R.S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K.J., Van Oosterhout, C. and Radwan, J. Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences*, 115(7), (2018), pp.1552-1557.
- [378] Lenz, T.L. Adaptive value of novel MHC immune gene variants. *Proceedings of the National Academy of Sciences*, 115(7), (2018), pp.1414-1416.
- [379] Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D. and 1000 Genomes Project. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29), (2011), pp.11983-11988.
- [380] Miyake, J., Kaneshita, Y., Asatani, S., Tagawa, S., Niioka, H. and Hirano, T. Graphical classification of DNA sequences of HLA alleles by deep learning. *Human Cell*, 31(2), (2018), pp.102-105.
- [381] Liu, Z., Cui, Y., Xiong, Z., Nasiri, A., Zhang, A. and Hu, J. DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. *Scientific Reports*, 9(1), (2019), p.794.
- [382] Chen, B., Khodadoust, M.S., Olsson, N., Wagar, L.E., Fast, E., Liu, C.L., Muftuoglu, Y., Sworder, B.J., Diehn, M., Levy, R., Davis, M.M., Elias, J.E., Altman, R.B. and Alizadeh A.A. Predicting HLA class II antigen presentation through integrated deep learning. *Nature Biotechnology*, (2019), pp.1-12.
- [383] Gao, J., Zhu, C., Zhu, Z., Tang, L., Liu, L., Wen, L. and Sun, L. The human leukocyte antigen and genetic susceptibility in human diseases. *Journal of Bio-X Research*, 2(3), (2019), pp.112-120.
- [384] Blackwell, J.M., Jamieson, S.E. and Burgner, D. (2009). HLA and infectious diseases. *Clinical Microbiology Reviews*, 22(2), pp.370-385.
- [385] Cruz-Tapias, P., Castiblanco, J. and Anaya, J.M. HLA association with autoimmune diseases. In *Autoimmunity: From Bench to Bedside* [Internet]. *El Rosario University Press*, (2013).
- [386] Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., Sougnez, C., Cibulskis, K., Kiezun, A., Hachohen, N., Brusica, V., Wu, C.J. and Getz, G. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature Biotechnology*, 33(11), (2015), p.1152.
- [387] McGranahan, N., Rosenthal, R., Hiley, C.T., Rowan, A.J., Watkins, T.B., Wilson, G.A., Birnbak, N.J., Veeriah, S., Van Loo, P., Herrero, J. and Swanton, C. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell*, 171(6), (2017), pp.1259-1271.
- [388] Chowell, D., Morris, L.G., Grigg, C.M., Weber, J.K., Samstein, R.M., Makarov, V., Kuo, F., Kendall, S.M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D.M., Zehir, A., Solit, D., Berger, M., Zhou, R., Rizvi, N.A. and Chan, T.A. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*, 359(6375), (2018), pp.582-587.

- [389] Chowell D, Krishna C, Pierini F, Makarov V, Rizvi NA, Kuo F, Morris LGT, Riaz N, Lenz TL and Chan TA. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nature Medicine* (2019). doi: 10.1038/s41591-019-0639-4
- [390] Boegel, S., Löwer, M., Bukur, T., Sahin, U. and Castle, J.C. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology*, 3(8), (2014), p.e954893.
- [391] Redwood, A.J., Pavlos, R.K., White, K.D. and Phillips, E.J. HLAs: Key regulators of T-cell-mediated drug hypersensitivity. *HLA*, 91(1), (2018), pp.3-16.
- [392] Ostrov DA, Grant BJ, Pompeu YA, Sidney J, Harndahl M, Southwood S, Oseroff C, Lu S, Jakoncic J, de Oliveira CA, Yang L, Mei H, Shi L, Shabanowitz J, English AM, Wriston A, Lucas A, Phillips E, Mallal S, Grey HM, Sette A, Hunt DF, Buus S and Peters B. Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire. *Proceedings of the National Academy of Sciences*. 109(25), (2012), pp. 9959–9964.
- [393] Illing PT, Vivian JP, Dudek NL, Kostenko L, Chen Z, Bharadwaj M, Miles JJ, Kjer-Nielsen L, Gras S, Williamson NA, Burrows SR, Purcell AW, Rossjohn J and McCluskey J. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature*. 486(7404), (2012), pp. 554–558
- [394] Phillips, E.J. and Mallal, S.A. Active suppression rather than ignorance: tolerance to abacavir-induced HLA-B\*57:01 peptide repertoire alteration. *The Journal of Clinical Investigation*, 128(7), (2018), pp.2746-2749.
- [395] Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*. 30, (1987), pp. 1205–1213.
- [396] van Dronghen, V. and Holoshitz, J. Human leukocyte antigen–disease associations in rheumatoid arthritis. *Rheumatic Disease Clinics*, 43(3), (2017), pp.363-376.
- [397] Raychaudhuri, S., Sandor, C., Stahl, E.A., Freudenberg, J., Lee, H.S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J. and Siminovitch, K.A. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nature Genetics*, 44(3), (2012), p.291.
- [398] Vandiedonck, C. and Knight, J.C. The human Major Histocompatibility Complex as a paradigm in genomics research. *Briefings in Functional Genomics and Proteomics*, 8(5), (2009), pp.379-394.
- [399] Dendrou, C.A., Petersen, J., Rossjohn, J. and Fugger, L. HLA variation and disease. *Nature Reviews Immunology*, 18(5), (2018), p.325.
- [400] Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F. and Freund, C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Frontiers in Immunology*, 8, (2017), p.292.
- [401] Lokki, M.L. and Paakkanen, R. The complexity and diversity of major histocompatibility complex challenge disease association studies. *HLA*, 93(1), (2019), pp.3-15.
- [402] Fernando, M.M., Stevens, C.R., Walsh, E.C., De Jager, P.L., Goyette, P., Plenge, R.M., Vyse, T.J. and Rioux, J.D. Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLOS Genetics*, 4(4), (2008), p.e1000024.
- [403] Kulski, J.K., Shiina, T. and Dijkstra, J.M. Genomic Diversity of the Major Histocompatibility Complex in Health and Disease. *Cells*, 8(10), (2019), 1270.
- [404] Arango, M.T., Perricone, C., Kivity, S., Cipriano, E., Ceccarelli, F., Valesini, G. and Shoenfeld, Y. HLA-DRB1 the notorious gene in the mosaic of autoimmunity. *Immunologic Research*, 65(1), (2017), pp.82-98.

- [405] Naisbitt DJ, Olsson-Brown A, Gibson A, Meng X, Ogese MO, Taylor A and Thomson P. Immune dysregulation increases the incidence of delayed-type drug hypersensitivity reactions. *Allergy*, (2019).
- [406] Tamouza R, Krishnamoorthy R, Giegling I, Leboyer M and Rujescu D. The HLA 8.1 Ancestral Haplotype in schizophrenia: dual implication in neuro-synaptic pruning and autoimmunity? *Acta Psychiatrica Scandinavica*, (2019).
- [407] Gambino, C.M., Aiello, A., Accardi, G., Caruso, C. and Candore, G. Autoimmune diseases and 8.1 ancestral haplotype: An update. *HLA*, 92(3), (2018), pp.137-143.
- [408] Fiorillo, M.T., Paladini, F., Tedeschi, V. and Sorrentino, R. HLA class I or class II and disease association: catch the difference if you can. *Frontiers in Immunology*, 8, (2017), p.1475.
- [409] Misra, M.K., Damotte, V. and Hollenbach, J.A. Structure-based selection of human metabolite binding P4 pocket of DRB1\*15:01 and DRB1\*15:03, with implications for multiple sclerosis. *Genes & Immunity*, 20(1), (2019), p.46.
- [410] Askar M, Sayer D, Wang T, Haagenson M, Spellman SR, Lee SJ, Madbouly A, Fleischhauer K, Hsu KC, Verneris MR, Thomas D, Zhang A, Sobecks RM, Majhail NS and Center for International Blood and Marrow Transplant Research Immunology Working Committee. Analysis of Single Nucleotide Polymorphisms in the Gamma Block of the Major Histocompatibility Complex in Association with Clinical Outcomes of Hematopoietic Cell Transplantation: A Center for International Blood and Marrow Transplant Research Study. *Biology of Blood and Marrow Transplantation*, 25(4), (2019), pp.664-672.
- [411] Maskalan, M., Grubic, Z., Seiwerth, R.S., Vrhovac, R., Mikulic, M., Kamenaric, M.B., Jankovic, K.S., Durakovic, N. and Zunec, R. The MHC gamma block matching: Impact on unrelated hematopoietic stem cell transplantation outcome. *Human Immunology*, (2019).
- [412] Svejgaard, A. and Ryder, L.P. HLA and disease associations: detecting the strongest association. *Tissue Antigens*, 43(1), (1994), pp.18-27.
- [413] de Bakker, P.I. and Raychaudhuri, S. Interrogating the major histocompatibility complex with high-throughput genomics. *Human Molecular Genetics*, 21(R1), (2012), pp.R29-R36.
- [414] Kiezun, A., Garimella, K., Do, R., Stitzel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., Hultman, C.M, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM and Sunyaev SR.. Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, 44(6), (2012), p.623.
- [415] Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shores, N., Whitton, H., Ryan, R.J., Shishkin, A.A., Hatan, M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA and Bernstein BE. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539), (2015), p.337.
- [416] Clark, P.M., Chitnis, N., Shieh, M., Kamoun, M., Johnson, F.B. and Monos, D. Novel and haplotype specific microRNAs encoded by the major histocompatibility complex. *Scientific Reports*, 8(1), (2018), p.3832.
- [417] Karakatsanis, L.P., Pavlos, G.P., Iliopoulos, A.C., Pavlos, E.G., Clark, P.M., Duke, J.L. and Monos, D.S. Assessing information content and interactive relationships of subgenomic DNA sequences of the MHC using complexity theory approaches based on the non-extensive statistical mechanics. *Physica A: Statistical Mechanics and its Applications*, 505, (2018), pp.77-93.
- [418] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B,

- Stamatoyannopoulos JA, Weng Z, White KP and Hardison RC. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17), (2014), pp.6131-6138.
- [419] D'Antonio, M., Reyna, J., Jakubosky, D., Donovan, M.K., Bonder, M.J., Matsui, H., Stegle, O., Nariai, N., D'Antonio-Chronowska, A. and Frazer, K.A. Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *BioRxiv*, (2019) p.564161
- [420] Aguiar, V.R., César, J., Delaneau, O., Dermitzakis, E.T. and Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLOS Genetics*, 15(4), (2019).
- [421] Vandiedonck, C., Taylor, M.S., Lockstone, H.E., Plant, K., Taylor, J.M., Durrant, C., Broxholme, J., Fairfax, B.P. and Knight, J.C. Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Research*, 21(7), (2011), pp.1042-1054.
- [422] Nakatani, K., Ueta, M., Khor, S.S., Hitomi, Y., Okudaira, Y., Masuya, A., Wada, Y., Sotozono, C., Kinoshita, S., Inoko, H. and Tokunaga, K. Identification of HLA-A\*02:06:01 as the primary disease susceptibility HLA allele in cold medicine-related Stevens-Johnson syndrome with severe ocular complications by high-resolution NGS-based HLA typing. *Scientific Reports*, 9(1), (2019), pp.1-8.
- [423] Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, Hartigan CR, Zhang W, Braun DA, Ligon KL, Bachireddy P, Zervantonakis IK, Rosenbluth JM, Ouspenskaia T, Law T, Justesen S, Stevens J, Lane WJ, Eisenhaure T, Lan Zhang G, Clauser KR, Hacoheh N, Carr SA, Wu CJ and Keskin DB. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology*, (2019).
- [424] Correale, P., Saladino, R.E., Nardone, V., Giannicola, R., Agostino, R., Pirtoli, L., Caraglia, M., Botta, C. and Tagliaferri, P. Could PD-1/PDL1 immune checkpoints be linked to HLA signature? *Immunotherapy*. 11(18), (2019), 1523-1526.
- [425] Duke, J.L., Ferriola, D., Mosbrugger, T.L. and Monos, D.S. OR41 Novel HLA alleles identified by next generation sequencing using blood-derived DNA of patients with hematologic disease may influence the search process for the appropriate donor in hematopoietic stem cell transplantation. *Human Immunology*, 80(Supplement), (2019), p.46.
- [426] Pierini, F. and Lenz, T.L. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Molecular Biology and Evolution*, 35(9), (2018), pp.2145-2158.
- [427] Buhler, S., Nunes, J.M. and Sanchez-Mazas, A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*, 68(6-7), (2016), pp.401-416.
- [428] Goeury, T., Creary, L.E., Fernandez-Vina, M.A., Vu-Trieu, A., Nunes, J.M. and Sanchez-Mazas, A. P103 Fine-scale DNA sequence analysis of nine HLA genes suggests a dual origin of the Vietnamese Cham. *HLA*, 93(5), (2019), pp.337-338.
- [429] Matern, B.M., Olieslagers, T.I., Voorter, C.E.M., Groeneweg, M., Tilanus, M.G.J. Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA Haplotypes. *HLA*, (2019).
- [430] van Deutekom, H.W., Mulder, W. and Rozemuller, E.H. Accuracy of NGS HLA typing data influenced by STR. *Human Immunology*. 80(7), (2019), pp.461-464.
- [431] Schwaiger, F.W. and Epplen, J.T. Exonic MHC-DRB Polymorphisms and Intronic Simple Repeat Sequences: Janus' Faces of DNA Sequence Evolution. *Immunological Reviews*, 143(1), (1995), pp.199-224.
- [432] Sospedra, M., Muraro, P.A., Stefanová, I., Zhao, Y., Chung, K., Li, Y., Giulianotti, M., Simon, R., Mariuzza, R., Pinilla, C. and Martin, R. Redundancy in antigen-presenting function of the HLA-DR and-DQ



- molecules in the multiple sclerosis-associated HLA-DR2 haplotype. *The Journal of Immunology*, 176(3), (2006) pp.1951-1961.
- [433] Barsakis, K., Babrzadeh, F., Chi, A., Mindrinou, M.N. and Vina, M.A.F. OR34 Examination of HLA DP intron/exon variation identifies two DPB1 evolutionary groups including intron 2 STR variants that may regulate DPB1 expression levels. *Human Immunology*, 77(Supplement), (2016), p.31.
- [434] Truong, L., Matern, B., Groeneweg, M., Tilanus, M.G., D'Orsogna, L., Martinez, P. and De Santis, D. P046 Characterization of novel single-nucleotide polymorphisms in the promoter of HLA-DP region. *Human Immunology*, 80(Supplement), (2019), p.88.
- [435] Arnaiz-Villena, A., Palacio-Gruber, J., de Salamanca, M.E., Juárez, I., Campos, C., Nieto, J., Muñiz, E. and Martin-Villa, J.M. HLA-G,-A haplotypes in Amerindians (Ecuador): HLA-G\*01:05N World distribution. *Human Immunology*, 79(2), (2018), pp.89-90.
- [436] Alizadeh, M., Bannasar, F., Fraboulet, S., Verite, F., Semana, G. P102 HLA class I-b gene polymorphism and linkage disequilibrium with classical HLA-class I genes. *HLA*, 93(5), (2019), pp.337-338.
- [437] Albrecht, V., Paech, C., Putke, K., Klasberg, S., Massalski, C., Schöne, B., Fuhrmann, M., Schöfl, G., Schmidt, A., Lange, V. and Klaußmeier, A. P061 Full-length characterisation of 150 novel MICA and MICB alleles. *Human Immunology*, 80(Supplement), (2019), p.98.
- [438] Arnaiz-Villena, A., Palacio-Gruber, J., Muñiz, E., Rey, D., Recio, M.J., Campos, C., Martinez-Quiles, N., Martin-Villa, J.M. and Martinez-Laso, J. HLA-DMB in Amerindians: Specific linkage of DMB\*01:03:01/DRB1 alleles. *Human Immunology*, 77(5), (2016), pp.389-394.
- [439] Alvarado-Guerri, R., Cabrera, C.M., Garrido, F. and López-Nevot, M.Á. TAP1 and TAP2 polymorphisms and their linkage disequilibrium with HLA-DR,-DP, and-DQ in an eastern Andalusian population. *Human Immunology*, 66(8), (2005), pp.921-930.
- [440] Djilali-Saiah, I., Benini, V., Daniel, S., Assan, R., Bach, J.F. and Caillat-Zucman, S. Linkage disequilibrium between HLA class II (DR, DQ, DP) and antigen processing (LMP, TAP, DM) genes of the major histocompatibility complex. *Tissue Antigens*, 48(2), (1996), pp.87-92.
- [441] Lang, K., Wagner, I., Schöne, B., Schöfl, G., Birkner, K., Hofmann, J.A., Sauter, J., Pingel, J., Böhme, I., Schmidt, A.H. and Lange, V. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics*, 17(1), (2016), p.374.
- [442] Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, Lang K, Hofmann J, Gnahn C, Heyn B, Marin WM, Dandekar R, Hollenbach JA, Schetelig J, Pingel J, Norman PJ, Sauter J, Schmidt AH and Lange V. Allele-level KIR genotyping of more than a million samples: workflow, algorithm and observations. *Frontiers in immunology*, 9, (2018), p.2843.
- [443] Maniangou, B., Retière, C. and Gagne, K. Next-generation sequencing technology a new tool for killer cell immunoglobulin-like receptor allele typing in hematopoietic stem cell transplantation. *Transfusion Clinique et Biologique*, 25(1), (2018), pp.87-89.
- [444] van de Pasch, L.A., van Ham, K., Vendelbosch, S., Penning, M.T. and Rozemuller, E.H. OR36 NGS Allele-level typing strategy for nine KIR genes. *Human Immunology*, 79(Supplement), (2018), p.42.
- [445] Closa, L., Vidal, F., Herrero, M.J. and Caro, J.L. Design and validation of a multiplex KIR and HLA class I genotyping method using next generation sequencing. *Frontiers in Immunology*, 9, (2018).
- [446] Edgerly, C.H. and Weimer, E.T. The Past, Present, and Future of HLA Typing in Transplantation. In: HLA Typing. Methods in Molecular Biology, vol 1802. *Humana Press*, (2018), (pp. 1-10).
- [447] Ramsuran, V., Hernández-Sánchez, P.G., O'hUigin, C., Sharma, G., Spence, N., Augusto, D.G., Gao, X., García-Sepúlveda, C.A., Kaur, G., Mehra, N.K. and Carrington, M.. Sequence and phylogenetic analysis

- of the untranslated promoter regions for HLA class I genes. *The Journal of Immunology*, 198(6), (2017), pp.2320-2329.
- [448] Apps, R., Meng, Z., Del Prete, G.Q., Lifson, J.D., Zhou, M. and Carrington, M.. Relative expression levels of the HLA class-I proteins in normal and HIV-infected cells. *The Journal of Immunology*, 194(8), (2015), pp.3594-3600.
- [449] Ramsuran, V., Kulkarni, S., O'huigin, C., Yuki, Y., Augusto, D.G., Gao, X. and Carrington, M.. Epigenetic regulation of differential HLA-A allelic expression levels. *Human Molecular Genetics*, 24(15), (2015), pp.4268-4275.
- [450] Raghavan, M., Yarzabek, B., Zaitouna, A.J., Krishnakumar, S. and Ramon, D.S. Strategies for the measurements of expression levels and half-lives of HLA class I allotypes. *Human Immunology*, 80(4), (2019), pp. 221-227.
- [451] Johansson, T., Yohannes, D.A., Koskela, S., Partanen, J. and Saavalainen, P. HLA RNAseq reveals high allele-specific variability in mRNA expression. *BioRxiv*, (2018), p.413534.
- [452] Kaur G, Gras S, Mobbs JI, Vivian JP, Cortes A, Barber T, Kuttikkatte SB, Jensen LT, Attfield KE, Dendrou CA, Carrington M, McVean G, Purcell AW, Rossjohn J and Fugger L. Structural and regulatory diversity shape HLA-C protein expression levels. *Nature Communications*, 8, (2017), p.15924.
- [453] V. Dubois, J.M. Tiercy, M.P. Labonne, A. Dormoy, L. Gebuhrer, A new HLA-B44 allele (B\*44020102S) with a splicing mutation leading to a complete deletion of exon 5. *Tissue Antigens* 63 (2004) pp. 173-180.
- [454] Ramagopalan, S.V., Maugeri, N.J., Handunnetthi, L., Lincoln, M.R., Orton, S.M., Dymont, D.A., DeLuca, G.C., Herrera, B.M., Chao, M.J., Sadovnick, A.D. and Ebers, G.C. Expression of the multiple sclerosis-associated MHC class II Allele HLA-DRB1\*1501 is regulated by vitamin D. *PLoS Genetics*, 5(2), (2009), p.e1000369.
- [455] Badders, J.L., Jones, J.A., Jeresano, M.E., Schillinger, K.P. and Jackson, A.M. Variable HLA expression on deceased donor lymphocytes: not all crossmatches are created equal. *Human Immunology*, 76(11), (2015), pp.795-800.
- [456] Montgomery, M.C., Liu, C., Petrarroia, R. and Weimer, E.T. Using Nanopore Whole-Transcriptome Sequencing for Human Leukocyte Antigen Genotyping and Correlating Donor Human Leukocyte Antigen Expression with Flow Cytometric Crossmatch Results. *The Journal of Molecular Diagnostics*, (2019)..
- [457] Wiebe, C. and Nickerson, P. Strategic use of epitope matching to improve outcomes. *Transplantation*, 100(10), (2016), 2048-2052.
- [458] Eiz-Vesper, B. and Blasczyk, Relevance of HLA expression variants in stem cell transplantation. *New Advances in Stem Cell Transplantation*, (2012), pp.39-58.
- [459] Gaudieri, S., Dawkins, R.L., Habara, K., Kulski, J.K. and Gojobori, T. Nucleotide diversity within the human major histocompatibility complex: function of hitchhiking effect, duplications, indels and recombination. In: *Major Histocompatibility Complex*, Springer, (2000), pp. 186-200.
- [460] Klein, J. "Evolution of MHC". In: *Encyclopedia of Immunology* (2<sup>nd</sup> Ed.). *Academic Press*; (1998), pp. 1700-1702.
- [461] Yin, Y., Reed, E.F. and Zhang, Q. Integrate CRISPR/Cas9 for protein expression of HLA-B\*38:68Q via precise gene editing. *Scientific Reports*, 9(1), (2019), p.8067.
- [462] Jang, Y., Choi, J., Park, N., Kang, J., Kim, M., Kim, Y. and Ju, J.H. Development of immunocompatible pluripotent stem cells via CRISPR-based human leukocyte antigen engineering. *Experimental & Molecular Medicine*, 51(1), (2019), p.3.

- [463] Sverchkova, A., Anzar, I., Stratford, R. and Clancy, T. Improved HLA typing of Class I and Class II alleles from next-generation sequencing data. *HLA*. 94(6), (2019), pp. 504-513.
- [464] Gonzalez-Galarza, FF, McCabe A, Santos, EJMD, Jones, J., Takeshita, L., Ortega-Rivera, ND., Cid-Pavon, GMD., Ramsbottom, K, Ghattaoraya, G., Alfirevic, A., Middleton, D., Jones, AR. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, (2019), pii: gkz1029. doi: 10.1093/nar/gkz1029. Last accessed October 21, 2019.
- [465] Mbunwe, E., Duke, J.L., Ferriola, D., Mosbrugger, T., Damianos, G., Dinou, A., Kotsopoulou, I., Ranciaro, A., Thompson, S., Beggs, W., Mpoloka, S.W, Mokone, G.G., Nyambo, T., Meskel, D.W., Belay, G., Fokunang, C., Njamnshi, A.F., Carrington, M., Maiers, M., Tishkoff, S. and Monos, D. P072 HLA types in ethnically diverse sub-saharan african populations. *Human Immunology*, 80(Supplement), (2019), p.105.
- [466] Orford, G., Jones, J., Corbett, S., Kishawi, A., Hammad, A. and Middleton, D. Next generation HLA typing and haplotypes by descent in Gaza individuals. *Human Immunology*, (2019).
- [467] Al-Yafei, Z., Goeury, T., Alvares, M., Al Seiyari, M., Sanchez-Mazas, A. and Elghazali, G., HLA-B locus: High levels of heterozygosity and a significant departure from neutrality towards excess homozygotes. *HLA*, (2019).
- [468] Pradana, K.A., Widjaya, M.A. and Wahjudi, M. Indonesians Human Leukocyte Antigen (HLA) Distributions and Correlations with Global Diseases. *Immunological investigations*, (2019), pp.1-31.
- [469] Hirata, J., Hosomichi, K., Sakaue, S., Kanai, M., Nakaoka, H., Ishigaki, K., Suzuki, K., Akiyama, M., Kishikawa, T., Ogawa, K. and Masuda, T. Genetic and phenotypic landscape of the major histocompatibility complex region in the Japanese population. *Nature Genetics*, 51(3), (2019), p.470.
- [470] Kwok, J., Tang, W.H., Chu, W.K., Liu, Z., Yang, W., Lee, C.K. and Middleton, D. HLA-DQB1,-DQA1,-DPB1, and-DPA1 genotyping and haplotype frequencies for a Hong Kong Chinese population of 1064 individuals. *Human Immunology*, (2019).
- [471] Williams, R.C., Knowler, W.C., Shuldiner, A.R., Gosalia, N., Van Hout, C., Regeneron Genetics Center, Hanson, R.L., Bogardus, C. and Baier, L.J. Next generation sequencing and the classical HLA loci in full heritage Pima Indians of Arizona: Defining the core HLA variation for North American Paleo-Indians. *Human Immunology*, (2019).
- [472] Haimila, K., Penttilä, A., Arvola, A., Auvinen, M.K. and Korhonen, M. Analysis of the adequate size of a cord blood bank and comparison of HLA haplotype distributions between four populations. *Human Immunology*, 74(2), (2013), pp.189-195.
- [473] Seshasubramanian, V., Venugopal, M., DS Kannan, A., Naganathan, C., Manisekar, N.K., Kumar, Y.N., Narayan, S. and Periathiruvadi, S. Application of high-throughput next-generation sequencing for HLA typing of DNA extracted from postprocessing cord blood units. *HLA*. 94(2), (2019), pp. 141-146.
- [474] Nestorowicz, K., Bogacz, A., Bukowska, A., Chraplak, M., Czerwiński, J., Góralski, M., Gronkowski, M., Jopek, K., Kniżewski, Ł., Kolasiński, M., Kowalski, M.L, Nowak, J., Sowinski, M., Wróblewska-Kabba, S., Tymoniuk, B. and Dudkiewicz, M. High-resolution allele frequencies for NGS based HLA-A, B, C, DQB1 and DRB1 typing of 23 595 bone marrow donors recruited for the polish central potential unrelated bone marrow donor registry. *Human Immunology*, (2020).
- [475] Hassall KB, Latham K, Robinson J, Gymer A, Goodall R, Merlo D, Marsh SGE, Mayor NP. Extending the sequences of HLA class I alleles without full-length genomic coverage using Single Molecule Real-Time (SMRT) DNA Sequencing. *HLA*, (2020).

- [476] Van Sandt, V., Senev, A., Vanden-Driessche, M., Torfs, A., Kerkhofs, J., Hidajat, M., Daniëls, L., and Emonds, M.P.. P195 HLA typing with 11-loci-NGS: Benefits in donor selection. *HLA*, 93(5), (2019), pp.381.
- [477] Rao, P., DeOliveira, A., Young, C.M., Leagans, K., Hanshew, W.E. and Chen, D.F. P064 Evaluation of MIA FORA NGS FLEX HLA typing kit. *Human Immunology*, 79(Supplement), (2018), p.108.
- [478] Periathiruvadi, S., Devi, S.A., Narayan, S., Bhardwaj, S. and Krishnakumar, S. P252 Conceptualization to implementation in 90 days of full gene HLA typing using mia fora HLA flex kits at Jeevan stem cell foundation, India. *Human Immunology*, 78(Supplement), (2017), p.239.
- [479] Hurley CK, Kempenich J, Wadsworth K, Sauter J, Hofmann JA, Schefzyk D, Schmidt AH, Galarza P, Cardozo MBR, Dudkiewicz M, Houdova L, Jindra P, Sorensen BS, Jagannathan L, Mathur A, Linjama T, Torosian T, Freudenberger R, Manolis A, Mavrommatis J, Cereb N, Manor S, Shriki N, Sacchi N, Ameen R, Fisher R, Dunckley H, Andersen I, Alaskar A, Alzahrani M, Hajeer A, Jawdat D, Nicoloso G, Kupatawintu P, Cho L, Kaur A, Bengtsson M and Dehn J. Common, Intermediate and Well-Documented HLA Alleles in World Populations: CIWD Version 3.0.0. *HLA*, (2020).
- [480] Eberhard, H.P., Schmidt, A.H., Mytilineos, J., Fleischhauer, K. and Müller, C.R. Common and well-documented HLA alleles of German stem cell donors by haplotype frequency estimation. *HLA*, 92(4), (2018), pp.206-214. (*Repeated in [526]*).
- [481] Pingel, J., Solloch, U.V., Hofmann, J.A., Lange, V., Ehninger, G. and Schmidt, A.H.. High-resolution HLA haplotype frequencies of stem cell donors in Germany with foreign parentage: how can they be used to improve unrelated donor searches?. *Human Immunology*, 74(3), (2013), pp.330-340.
- [482] Smith, N.T., Ngo, V., Carmazzi, Y., Krishnakumar, S., Li, M., Wang, C., Guerrero, E., Mindrinos, M. and Cao, K.. P073 Next generation sequencing of all classical HLA-class I and II genes. *Human Immunology*, 77(Supplement), (2016), p.90.
- [483] OmniType kit (single tube multiplex for 11 HLA loci), (Omixon, Budapest, Hungary). <https://www.omixon.com/omnitype-eap/>, last accessed October 21, 2019.
- [484] MIA FORA™ NGS MFLEX HLA Typing Kit (novel single tube multiplex for 11 HLA loci), (Immucor, Inc. Norcross, GA, USA). <https://sistemaparaevento.com.br/evento/abto2019/programacao/palestrante/115629/>, last accessed October 21, 2019.
- [485] Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. The structure of haplotype blocks in the human genome. *Science*, 296(5576), (2002), pp. 2225-2229.
- [486] Degli-Esposti, M.A., Leaver, A.L., Christiansen, F.T., Witt, C.S., Abraham, L.J. and Dawkins, R.L. Ancestral haplotypes: conserved population MHC haplotypes. *Human Immunology*, 34(4), (1992), pp.242-252.
- [487] Vayntrub, T, Mack, SJ and Fernández-Viña, MA. (Editorial) Preface: 17th International HLA and Immunogenetics Workshop. *Human Immunology*, (2020).
- [488] Rafei H, Mehta RS and Rezvani K (2019) Cellular Therapies in Cancer. *Front. Immunol.* 10:2788. doi: 10.3389/fimmu.2019.02788, (2019).
- [489] Ping, Y., Liu, C. and Zhang, Y.. T-cell receptor-engineered T cells for cancer treatment: current status and future directions. *Protein & Cell*, 9(3), (2018), pp.254-266.

- [490] Harris, K.M., Davila, B.J., Bollard, C.M. and Keller, M.D. Virus-specific T cells: current and future use in primary immunodeficiency disorders. *The Journal of Allergy and Clinical Immunology: In Practice*, 7(3), (2019), pp.809-818.
- [491] Taylor, C.J., Bolton, E.M., Pocock, S., Sharples, L.D., Pedersen, R.A. and Bradley, J.A. Banking on human embryonic stem cells: estimating the number of donor cell lines needed for HLA matching. *The Lancet*, 366(9502), (2005), pp.2019-2025.
- [492] Gourraud, P.A., Gilson, L., Girard, M. and Peschanski, M. The role of human leukocyte antigen matching in the development of multiethnic “haplobank” of induced pluripotent stem cell lines. *Stem Cells*, 30(2), (2012), pp.180-186.
- [493] Solomon, S., Pitossi, F. and Rao, M.S. Banking on iPSC-is it doable and is it worthwhile. *Stem Cell Reviews and Reports*, 11(1), (2015), pp.1-10.
- [494] Singh, V.K., Kalsan, M., Kumar, N., Saini, A. and Chandra, R. Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery. *Frontiers in Cell and Developmental Biology*, 3, (2015), p.2.
- [495] Lee, S., Huh, J.Y., Turner, D.M., Lee, S., Robinson, J., Stein, J.E., Shim, S.H., Hong, C.P., Kang, M.S., Nakagawa, M. and Kaneko, S. Repurposing the cord blood Bank for haplobanking of HLA-homozygous iPSCs and their usefulness to multiple populations. *Stem Cells*, 36(10), (2018), pp.1552-1566.
- [496] Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, Koech D, Lancaster A, Nelson M, Meyer D, Single R, Hartzman RJ, Plowe CV, Kazura J, Mann DL, Sztein MB, Thomson G, Fernández-Viña MA. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*, 63(4), (2004), pp.293-325. (Repeated in Reference [836]).
- [497] Mack, S.J., Bugawan, T.L., Moonsamy, P.V., Erlich, J.A., Trachtenberg, E.A., Paik, Y.K., Begovich, A.B., Saha, N., Beck, H.P., Stoneking, M. and Erlich, H.A. Evolution of Pacific/Asian populations inferred from HLA class II allele frequency distributions. *Tissue Antigens*, 55(5), (2000), pp.383-400.
- [498] Paganini, J., Abi-Rached, L., Gouret, P., Pontarotti, P., Chiaroni, J. and Di Cristofaro, J. HLA-Ib worldwide genetic diversity: New HLA-H alleles and haplotype structure description. *Molecular Immunology*, 112, (2019), pp.40-50.
- [499] Abi-Rached, L., Gouret, P., Yeh, J.H., Di Cristofaro, J., Pontarotti, P., Picard, C. and Paganini, J. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS One*, 13(10), (2018).
- [500] Jordier, F., Gras, D., De Grandis, M., D'Journo, X.B., Thomas, P.A., Chanez, P., Picard, C., Chiaroni, J., Paganini, J. and Di Cristofaro, J. HLA-H: transcriptional activity and HLA-E mobilization. *Frontiers in Immunology*, 10, (2019).
- [501] Radwan, J., Babik, W., Kaufman, J., Lenz, T.L. and Winternitz, J. Advances in the Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics*, (2020).
- [502] Manczinger, M., Boross, G., Kemény, L., Müller, V., Lenz, T.L., Papp, B. and Pál, C. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLOS Biology*, 17(1), (2019) p.e3000131.
- [503] Kaufman, J. Generalists and specialists: a new view of how MHC class I molecules fight infectious pathogens. *Trends in Immunology*, 39(5), (2018), pp.367-379.
- [504] Barquera R, Zuniga J, Flores-Rivera J, Corona T, Penman BS, Hernández-Zaragoza DI, Soler M, Jonapá-Gómez L, Mallempati KC, Yescas P, Ochoa-Morales A, Barsakis K, Aguilar-Vázquez JA, García-Lechuga M, Mindrinos M, Yunis M, Jiménez-Alvarez L, Mena-Hernández L, Ortega E, Cruz-Lagunas A,

- Tovar-Méndez VH, Granados J, Fernández-Viña M and Yunis E. Diversity of HLA Class I and Class II blocks and conserved extended haplotypes in Lacandon Mayans. *Scientific Reports*, 10(1), (2020), pp.1-22.
- [505] Krause-Kyora B, Nutsua M, Boehme L, Pierini F, Pedersen DD, Kornell SC, Drichel D, Bonazzi M, Möbus L, Tarp P, Susat J, Bosse E, Willburger B, Schmidt AH, Sauter J, Franke A, Wittig M, Caliebe A, Nothnagel M, Schreiber S, Boldsen JL, Lenz TL and Nebel A. Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nature Communications*, 9(1), (2018), pp.1-11.
- [506] Alizadeh, M., Picard, C., Frassati, C., Walencik, A., Cesbron Gauthier, A., Bannasar, F., Verite, F., Semana, G. A new set of reagents and related software used for NGS based classical and non-classical HLA typing showing evidence for a greater HLA haplotype diversity. *Human Immunology*, (2020).
- [507] Orenbuch R., Filip I., Rabadan R. HLA Typing from RNA Sequencing and Applications to Cancer. In: Boegel S. (eds) *Bioinformatics for Cancer Immunotherapy. Methods in Molecular Biology*, vol 2120. *Humana*, New York, NY, (2020).
- [508] Aguiar V.R.C., Masotti C., Camargo A.A., Meyer D. HLApers: HLA Typing and Quantification of Expression with Personalized Index. In: Boegel S. (eds) *Bioinformatics for Cancer Immunotherapy. Methods in Molecular Biology*, vol 2120. *Humana*, New York, NY, (2020).
- [509] Liu C., Berry R. Rapid High-Resolution Typing of Class I HLA Genes by Nanopore Sequencing. In: Boegel S. (eds) *Bioinformatics for Cancer Immunotherapy. Methods in Molecular Biology*, vol 2120. *Humana*, New York, NY, (2020).
- [510] Zeiser, R. and Vago, L. Mechanisms of immune escape after allogeneic hematopoietic cell transplantation. *Blood*, 133(12), (2019), pp.1290-1297.
- [511] Klussmeier, A., Massalski, C., Putke, K., Schäfer, G., Sauter, J., Schefzyk, D., Pruschke, J., Hoffmann, J., Fürst, D., Carapito, R., Bahram, S., Schmidt AH, and Lange V. High-Throughput MICA/B Genotyping of Over Two Million Samples: Workflow and Allele Frequencies. *Frontiers in Immunology*, 11, (2020), p.314.
- [512] De Santis, D., Truong, L., Martinez, P. and D'Orsogna, L. Rapid High Resolution HLA genotyping by MinION Oxford Nanopore Sequencing for Deceased Donor Organ Allocation. *HLA*, (2020).
- [513] Sanchez-Mazas, A. A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Medical Weekly*, 150(1516), (2020).
- [514] Sanchez-Mazas, A. HLA studies in the context of coronavirus outbreaks. *Swiss Medical Weekly*, 150(1516), (2020).
- [515] Hurley, C. K. "Naming HLA diversity: A review of HLA nomenclature." *Human Immunology* (2020).
- [516] Stockton, J.D., Nieto, T., Wroe, E., Poles, A., Inston, N., Briggs, D. and Beggs, A.D. Rapid, highly accurate and cost-effective open-source simultaneous complete HLA typing & phasing of Class I & II alleles using Nanopore sequencing. *HLA*, (2020).
- [517] Profaizer T, Pole A, Monds C, Delgado JC and Lázár-Molnár E. Clinical utility of next generation sequencing based HLA typing for disease association and pharmacogenetic testing. *Human Immunology*, (2020).
- [518] Duygu B, Matern BM, Wieten L, Voorter CEM, Tilanus MGJ. Specific amino acid patterns define split specificities of HLA-B15 antigens enabling conversion from DNA-based typing to serological equivalents. *Immunogenetics*, (2020).
- [519] Mosbrugger, T.L., Dinou, A., Duke, J.L., Ferriola, D., Mehler, H., Pagkrati, I., Georgios Damianos, Eric Mbunwe, Mahdi Sarmady, Ioannis Lyratzakis, Sarah A. Tishkoff, Anh Dinh, Dimitri S. Monos. Utilizing nanopore sequencing technology for the rapid and comprehensive characterization of eleven HLA loci; addressing the need for deceased donor expedited HLA typing, *Human Immunology*, (2020).

- [520] Barquera, R., Krause, J., An ancient view on host pathogen interaction across time and space. *Current Opinion in Immunology*, 65, (2020), pp. 65-69.
- [521] Chitnis NS, Shieh M, Monos D. Regulatory noncoding RNAs and the major histocompatibility complex. *Hum Immunol*. 2020
- [522] Barquera,R, Hernández-Zaragoza, D.I., Bravo-Acevedo, A., Arrieta-Bolaños, E., Clayton, S., Acuña-Alonzo, V., Martínez-Álvarez, J.C., López-Gil, C., Adalid-Sáinz, C., Vega-Martínez, M.R, Escobedo-Ruíz, A., Juárez-Cortés, E.D., Immel, A., Pacheco-Ubaldo, H., González-Medina, L., Lona-Sánchez, A., Lara-Riegos, J., Sánchez-Fernández, M.G.J, Díaz-López, R., Guizar-López, G.U., Medina-Escobedo, C.E., Arrazola-García, M.A., Montiel-Hernández, G.D., Hernández-Hernández, O., Ramos-de la Cruz, F.R, Juárez-Nicolás, F., Pantoja-Torres, J.A., Rodríguez-Munguía, T.J., Juárez-Barreto, V., Delgado-Aguirre, H., Escutia-González, A.B., Goné-Vázquez, I., Benítez-Arvizu, G., Arellano-Prado, F.P., García-Arias, V.E., Rodríguez-López, M.E., Méndez-Mani, P., García-Álvarez, R., González-Martínez, M.R., Aquino-Rubio, G., Escareño-Montiel, N., Vázquez-Castillo, T.V., Uribe-Duarte, M.G., Ruíz-Corral, M.J., Ortega-Yáñez, A., Bernal-Felipe, N., Gómez-Navarro, B., Arriaga-Perea, A.J., Martínez-Bezies, V., Macías-Medrano, R.M., Aguilar-Campos, J.A., Solís-Martínez, R., Serrano-Osuna, R., Sandoval-Sandoval, M.J., Jaramillo-Rodríguez, Y., Salgado-Adame, A., Juárez-de la Cruz, F., Novelo-Garza, B., Pavón-Vargas, M.A., Salgado-Galicia, N., Bortolini, M.C., Gallo, C., Bedoya, G., Rothhammer, F., González-José, R., Ruiz-Linares, A., Canizales-Quinteros, S., Romero-Hidalgo, S., Krause, J., Zúñiga, J., Yunis, E.J., Bekker-Méndez, C. and Granados, J. The immunogenetic diversity of the HLA system in Mexico correlates with underlying population genetic structure. *Human Immunology* (2020).
- [523] Bauer, M, Kempenich, J, Wadsworth, K., Malmberg, C., Beduhn, B. and Dehn, J. Frequencies and haplotype associations of non-expressed HLA alleles in ethnically diverse populations on the National Marrow Donor Program's Be The Match Registry. *Human Immunology*, (2020).
- [524] Lee SJ, Klein J, Haagenon M, Baxter-Lowe LA, Confer DL, Eapen M, Fernandez-Vina M, Flomenberg N, Horowitz M, Hurley CK, Noreen H, Oudshoorn M, Petersdorf E, Setterholm M, Spellman S, Weisdorf D, Williams TM and Anasetti C. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood* 110(13), (2007), pp. 4576-4583.
- [525] Eberhard, H.P, Fleischhauer, K., Mytilineos, J., Schmidt, A., Müller, C. Limited efficacy of using linkage to identify null alleles in Germany, *Biol. Blood Marrow Transplant* 22, (2016), S421.
- [526] Eberhard, H.P., Schmidt, A.H., Mytilineos, J., Fleischhauer, K., Muller, C.R. Common and well-documented HLA alleles of German stem cell donors by haplotype frequency estimation, *HLA* 92(4), (2018), pp.206–214. (Repeated in [480]).
- [527] Roelen, D., de Vaal, Y., Vierra-Green, C., Waldvogel, S., Spellman, S., Claas, F. and Oudshoorn, M. HLA mismatches that are identical for the antigen recognition domain are less immunogenic. *Bone Marrow Transplantation*, 53(6), (2018), pp.729-740.
- [528] Vince, N., Douillard, V., Geffard, E., Meyer, D., Castelli, E.C., Mack, S.J., Limou, S. and Gourraud, P.A.. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. *Genetic Epidemiology* (2020).
- [529] Madrigal, J.A. and Barber, L.D. Matching inside and outside the HLA molecule in allogeneic hematopoietic stem cell transplantation. *Haematologica*, 101(10), (2016), p.1131.
- [530] Dorak, M.T., Shao, W., Machulla, H.K.G., Lobashevsky, E.S., Tang, J., Park, M.H. and Kaslow, R.A. Conserved extended haplotypes of the major histocompatibility complex: further characterization. *Genes and Immunity*, 7(6), (2006), p.450.

- [531] Alcoceba-Sánchez, M. Estudio de polimorfismos genéticos en la evolución clínica de pacientes sometidos a transplante alogénico de progenitores hematopoyéticos. *Doctoral Dissertation, Tesis Doctoral, Universidad de Salamanca*, (2010).
- [532] Kindt, T.J., Goldsby, R.A., Osborne, B.A., and Kuby, J. “The Major Histocompatibility Complex and Antigen Presentation. Chapter 8”. In: *Kuby Immunology (6<sup>th</sup> Edition)*. Macmillan, (2007).
- [533] Mak, T.W., Saunders, M.E. and Jett, B.D. “MHC: The Major Histocompatibility Complex. Chapter 10”. In: *Primer to the Immune Response*. Newnes, (2013).
- [534] Bellanti, JA (Ed). *Immunology IV: Clinical Applications in Health and Disease*. I Care Press, (2012).
- [535] Kobayashi, K.S., and Van Den Elsen, P.J.. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nature Reviews Immunology* 12(12), (2012), pp.813-820.
- [536] Bontrop, R.E., Otting, N., de Groot, N.G. and Doxiadis, G.G. Major histocompatibility complex class II polymorphisms in primates. *Immunol. Rev.*, 167, (1999), pp.339–350.
- [537] Ho, C.S., Ramon, D. and Jaramillo, A. OR27: Bw4/Bw6 ON HLA-A AND HLA-C: THE FORGOTTEN SEROLOGICAL PROPERTIES OF HLA CLASS I ANTIGENS. *Human Immunology*, 75, (2014), p.22.
- [538] Müller, C.A., Engler-Blum, G., Gekeler, V., Steiert, I., Weiss, E. and Schmidt, H. Genetic and serological heterogeneity of the supertypic HLA-B locus specificities Bw4 and Bw6. *Immunogenetics*, 30(3), (1989), pp.200-207.
- [539] Ghodke, Y., Joshi, K., Chopra, A. and Patwardhan, B. HLA and disease. *European Journal of Epidemiology*, 20(6), (2005), pp.475-488.
- [540] Simmonds, M.J. and Gough, S.C.L. The HLA region and autoimmune disease: associations and mechanisms of action. *Current Genomics*, 8(7), (2007), pp.453-465.
- [541] Vodo, D., Sarig, O. and Sprecher, E. The genetics of pemphigus vulgaris. *Frontiers in Medicine*, 5, (2018), p.226.
- [542] Guha, P., Srivastava, S.K., Bhattacharjee, S. and Chaudhuri, T.K.. Human migration, diversity and disease association: a convergent role of established and emerging DNA markers. *Frontiers in Genetics*, 4, (2013), p.155.
- [543] Bos, D.H., Gopurenko, D., Williams, R.N. and DeWoody, J.A. Inferring population history and demography using microsatellites, mitochondrial DNA, and major histocompatibility complex (MHC) genes. *Evolution: International Journal of Organic Evolution*, 62(6), (2008), pp.1458-1468.
- [544] Magi, A., Semeraro, R., Mingrino, A., Giusti, B. and D’Aurizio, R. Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in Bioinformatics*, 19(6), (2018), pp.1256-1272.
- [545] Roura, S., Rudilla, F., Gastelurrutia, P., Enrich, E., Campos, E., Lupón, J., Santiago-Vacas, E., Querol, S. and Bayés-Genís, A. Determination of HLA-A,-B,-C,-DRB1 and-DQB1 allele and haplotype frequencies in heart failure patients. *ESC Heart Failure*, 6(2), (2019), pp.388-395.
- [546] Alvarez-Palomo, B., Garcia-Martinez, I., Gayoso, J., Raya, A., Veiga, A., Abad, M.L., Eiras, A., Guzman-Fulgencio, M., Luis-Hidalgo, M., Eguizabal, C. Santos, S., Balas, A., Alenda, R., Sanchez-Gordo, F., Ponce-Verdugo, L., Villa, J., Carreras, E., Vidal, F., Madrigal, A., Herrero-Mata, M.J., Rudilla, F., and Querol, S. Evaluation of the Spanish Population Coverage of a Prospective HLA Haplobank of Induced Pluripotent Stem Cells. *BMC Stem Cell Research & Therapy*, (2020), Online pre-print.
- [547] Zúñiga, J., Yu, N., Barquera, R., Alosco, S., Ohashi, M., Lebedeva, T., Acuña-Alonzo, V., Yunis, M., Granados-Montiel, J., Cruz-Lagunas, A., Vargas-Alarcón, G., Rodriguez-Reyna, T.S., Fernandez-Vina, M.,



- Granados, J. and Yunis, E.J. HLA class I and class II conserved extended haplotypes and their fragments or blocks in Mexicans: implications for the study of genetic diversity in admixed populations. *PLOS One*, 8(9), (2013), p.e74442.
- [548] Furuzawa-Carballeda J, Zuñiga J, Hernández-Zaragoza DI, Barquera R, Marques-García E, Jiménez-Alvarez L, Cruz-Lagunas A, Ramírez G, Regino NE, Espinosa-Soto R, Yunis EJ, Romero-Hernández F, Azamar-Llamas D, Coss-Adame E, Valdovinos MA, Torres-Landa S, Palacios-Ramírez A, Breña B, Alejandro-Medrano E, Hernández-Ávila A, Granados J and Torres-Villalobos G. An original Eurasian haplotype, *HLA-DRB1\*14:54-DQB1\*05:03*, influences the susceptibility to idiopathic achalasia. *PLOS One*, 13(8), (2018).
- [549] Juárez-Nicolás F, Barquera R, Martínez-Álvarez JC, Hernández-Zaragoza DI, Ortega-Yáñez A, Arrieta-Bolaños E, Clayton S, Bravo-Acevedo A, Arrazola-García MA, Immel A, Juárez-Barreto V, Benítez-Arvizu G, Vega-Martínez MDR, García-Álvarez R, Martínez-Bezies V, Escutia-González AB, Díaz-López R, Guizar-López GU, Salgado-Galicia N, Zúñiga J, Yunis EJ, Bekker-Méndez C and Granados J. Genetic diversity of HLA system in a population from Guerrero, Mexico. *Human Immunology*, (2019).
- [550] Carlos Serrano Sánchez, Mestizaje e historia de la población en México (con un esbozo antropológico de los lacandones de Chiapas). In: Ángel Martín Municio, Pedro García Barreno (Eds.), Polimorfismo génico (HLA) en poblaciones Hispanoamericanas, *Real Academia de Ciencias Exactas, Físicas y Naturales*, Madrid, (1996), pp. 173–193.
- [551] Gorodezky C, Alaez C, Vazquez-Garcia MN, de la RG, Infante E, Balladares S, Toribio R, Perez-Luque E, Muñoz L. The genetic structure of Mexican Mestizos of different locations: tracking back their origins through MHC genes, blood group systems, and microsatellites. *Human Immunology*, (62), (2001), pp. 979–991.
- [552] Hollenbach, J.A., Thomson, G., Cao, K., Fernandez-Vina, M., Erlich, H.A., Bugawan, T.L., Winkler, C., Winter, M. and Klitz, W. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Human Immunology*, 62(4), (2001), pp.378-390.
- [553] Reche, P.A. and Reinherz, E.L. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *Journal of Molecular Biology*, 331(3), (2003), pp.623-641.
- [554] Fürst, D., Neuchel, C., Tsamadou, C., Schrezenmeier, H. and Mytilineos, J. HLA matching in unrelated stem cell transplantation up to date. *Transfusion Medicine and Hemotherapy*, 46(5), (2019), pp.326-336.
- [555] Chapman C. A History of Spain. *Jovian Press*; (2017).
- [556] García de Cortázar F: Atlas de Historia de España. Madrid, Spain: *Editorial Planeta*; (2005).
- [557] Bertranpetit, J. and Cavalli-Sforza, L.L. A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics*, 55(1), (1991), pp.51-67.
- [558] Arnaiz-Villena, A., Martínez-Laso, J. and Alonso-García, J. Iberia: population genetics, anthropology, and linguistics. *Human Biology*, 71(5), (1999), p.725.
- [559] Sanchez-Velasco P, Gomez-Casado E, Martinez-Laso J, Moscoso J, Zamora J, Lowy E, Silvera C, Cemborain A, Leyva-Cobián F, Arnaiz-Villena A. HLA alleles in isolated populations from North Spain: origin of the Basques and the ancient Iberians. *Tissue Antigens*, 61(5), (2003), pp.384-392.
- [560] Muro, M., Marín, L., Torío, A., Moya-Quiles, M.R., Minguela, A., Rosique-Roman, J., Sanchis, M.J., Garcia-Calatayud, M.C., García-Alonso, A.M. and Álvarez-López, M.R.. HLA polymorphism in the Murcia population (Spain): in the cradle of the archaeological Iberians. *Human Immunology*, 62(9), (2001), pp. 910–921.

- [561] Longás, J., Martínez-Laso, J., Rey, D., Areces, C., Casado, E.G., Parga-Lozano, C., Luna, F., de Salamanca, M.E., Moral, P. and Arnaiz-Villena, A. Las Alpujarras region (South East Spain) HLA genes study: evidence of a probable success of 17th century repopulation from North Spain. *Molecular Biology Reports*, 39(2) (2012) 1387-1394.
- [562] Mann VB, Glick TF, Dodds JD, editors. Convivencia: Jews, Muslims, and Christians in Medieval Spain. *G. Braziller* (1992).
- [563] Arnaiz-Villena, A., Muñiz, E., Campos, C., Gomez-Casado, E., Tomasi, S., Martínez-Quiles, N., Martín-Villa, M. and Palacio-Gruber, J. Origin of Ancient Canary Islanders Guanches: presence of Atlantic/Iberian HLA and Y chromosome genes and Ancient Iberian language. *International Journal of Modern Anthropology*, 1(8), (2015), pp. 67-93.
- [564] Martínez-Laso, J., Ramírez-Puga, A., Rivas-García, E., Fernández-Tagarro, E., Auyanet-Saavedra, I., Guerra-Rodríguez, R., Díaz-Novo, N. and García-Cantón, C. North African-Mediterranean HLA genetic contribution in a population of the kidney transplant waiting list patients of Canary origin (Gran Canaria). *HLA*, 92(1), (2018), pp. 12-23.
- [565] Fregel R, Ordóñez AC, Santana-Cabrera J, Cabrera VM, Velasco-Vázquez J, Alberto V, Moreno-Benítez MA, Delgado-Darias T, Rodríguez-Rodríguez A, Hernández JC, Pais J, González-Montelongo R, Lorenzo-Salazar JM, Flores C, Cruz-de-Mercadal MC, Álvarez-Rodríguez N, Shapiro B, Arnay M and Bustamante CD. Mitogenomes illuminate the origin and migration patterns of the indigenous people of the Canary Islands. *PLoS One*, 14(3), (2019), p.e0209125.
- [566] Guillen-Guio, B., Lorenzo-Salazar, JM, González-Montelongo, R., Díaz-de Usera, A., Marcelino-Rodríguez, I., Corrales, A., Cabrera de León, A., Alonso, S., and Flores, C. Genomic Analyses of Human European Diversity at the Southwestern Edge: Isolation, African Influence and Disease Associations in the Canary Islands. *Molecular Biology and Evolution*, 35(12), (2018), pp. 3010–3026.
- [567] Maca-Meyer, N., Arnay, M., Rando, J.C., Flores, C., González, A.M., Cabrera, V.M. and Larruga, J.M. Ancient mtDNA analysis and the origin of the Guanches. *European Journal of Human Genetics*, 12(2), (2004), p.155.
- [568] Fregel, R., Gomes, V., Gusmão, L., González, A.M., Cabrera, V.M., Amorim, A. and Larruga, J.M. Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. *BMC Evolutionary Biology*, 9(1), (2009), p.181.
- [569] Safran, J.M.. The second Umayyad Caliphate: the articulation of caliphal legitimacy in al-Andalus (Vol. 33). *Harvard CME* (2000).
- [570] Cortés G. Historia de los judíos mallorquines y de sus descendientes cristianos. Palma de Mallorca. *Font*, (1985), pp. 70–169.
- [571] Crespí, C., Milà, J., Martínez-Pomar, N., Etxagibel, A., Muñoz-Saa, I., Priego, D., Luque, A., Pons, J., Picornell, A., Ramon, M., Castro, J.A. and Matamoros, N.. HLA polymorphism in a Majorcan population of Jewish descent: comparison with Majorca, Minorca, Ibiza (Balearic Islands) and other Jewish communities. *Tissue Antigens*, 60(4), (2002), pp.282-291.
- [572] Gerber, J.S. Jews of Spain: A History of the Sephardic Experience. *Simon and Schuster* (1994).
- [573] Dadson TJ. The Assimilation of Spain's Moriscos: Fiction or Reality?. *Journal of Levantine Studies*.1(2), (2011), pp.11-30.
- [574] Nogueiro, I., Teixeira, J.C., Amorim, A., Gusmão, L. and Alvarez, L. Portuguese crypto-Jews: the genetic heritage of a complex history. *Frontiers in Genetics*, 6, (2015), p.12.
- [575] Johnson, P. A history of the Jews. *Hachette UK*, (2013).

- [576] Roniger, L. The Western Sephardic Diaspora: Ancestral Birthplaces and Displacement, Diaspora Formation and Multiple Homelands. *Latin American Research Review* 54(4), (2019), pp. 1031–1038.
- [577] Ostrer H and Skorecki K. The population genetics of the Jewish people. *Human Genetics* 132(2), (2013), pp. 119–27.
- [578] Adams, S.M., Bosch, E., Balaesque, P.L., Ballereau, S.J., Lee, A.C., Arroyo, E., López-Parra, A.M., Aler, M., Grifo, M.S.G., Brion, M. and Carracedo, A. The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *The American Journal of Human Genetics*, 83(6) (2008) 725-736.
- [579] Fadhloui-Zid, K., Martinez-Cruz, B., Khodjet-el-khil, H., Mendizabal, I., Benammar-Elgaaied, A. and Comas, D. Genetic structure of Tunisian ethnic groups revealed by paternal lineages. *American Journal of Physical Anthropology*, 146(2), (2011), pp.271-280.
- [580] Fraser AM. *The gypsies*, Wiley-Blackwell, (1995).
- [581] Schurr TG. Reconstructing the origins and migrations of diasporic populations: the case of the European Gypsies. *American Anthropologist* 106, (2004), pp. 267–281.
- [582] Gusmão, A., Gusmão, L., Gomes, V., Alves, C., Calafell, F., Amorim, A. and Prata, M.J. A perspective on the history of the Iberian Gypsies provided by phylogeographic analysis of Y-chromosome lineages. *Annals of Human Genetics*, 72(2), (2008), pp.215-227.
- [583] Kenrick, D. *Historical dictionary of the Gypsies (Romanies)*. Scarecrow Press, (2007).
- [584] Kenrick, D. *Gypsies: from the Ganges to the Thames (Vol. 3)*. Univ of Hertfordshire Press, (2004).
- [585] Mendizabal, I., Valente, C., Gusmão, A., Alves, C., Gomes, V., Goios, A., Parson, W., Calafell, F., Alvarez, L., Amorim, A., Gusmão, L., Comas, D. and Prata, M.J.. Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. *PLOS One*, 6(1), (2011), p.e15988.
- [586] Moorjani, P., Patterson, N., Loh, P.R., Lipson, M., Kiszali, P., Melegh, B.I., Bonin, M., Kádaši, L., Rieβ, O., Berger, B., Reich, D. and Melegh, B. Reconstructing Roma history from genome-wide data. *PLOS One*, 8(3), (2013), p.e58633.
- [587] Melegh, B.I., Banfai, Z., Hadzsiev, K., Miseta, A. and Melegh, B. Refining the South Asian origin of the Romani people. *BMC Genetics*, 18(1), (2017), p.82.
- [588] Font-Porterias, N., Arauna, L.R., Poveda, A., Bianco, E., Rebato, E., Prata, M.J., Calafell, F. and Comas, D. European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. *PLOS Genetics*, 15(9), (2019), p.e1008417.
- [589] Rani, R., Fernandez-Vina, M.A. and Stastny, P. Associations between HLA class II alleles in a North Indian population. *Tissue Antigens*, 52(1), (1998), pp.37-43.
- [590] Rodriguez, J.P. *The historical encyclopedia of world slavery (Vol. 1-7)*. *Abc-Clio*, (1997)
- [591] Cressy, D.. *Gypsies: An English History*. *Oxford University Press*, (2018)
- [592] Marushiakova E, Popov V. *Gypsies in the Ottoman Empire: A Contribution to the History of the Balkans*. Vol. 22. *Hatfield: Univ of Hertfordshire Press*; (2001).
- [593] Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, Tournev I, de Pablo R, Kučinskis V, Perez-Lezaun A, Marushiakova E, Popov V, Kalaydjieva L. Origins and divergence of the Roma (gypsies). *The American Journal of Human Genetics*, 69(6), (2001), pp.1314-1331.

- [594] Kalaydjieva, L., Gresham, D. and Calafell, F. Genetic studies of the Roma (Gypsies): a review. *BMC Medical Genetics*, 2(1), (2001), p.5.
- [595] Anuario Estadístico de España 2019. [https://www.ine.es/prodyser/pubweb/anuarios\\_mnu.htm](https://www.ine.es/prodyser/pubweb/anuarios_mnu.htm) (last accessed December 15, 2019).
- [596] Hernández-Pedreño, M., García-Luque, O., and Gehrig, R. Situación social de la población gitana en España: balance tras la crisis. VIII Informe FOESSA Documento de Trabajo 3.12.
- [597] From India to Europe, <https://rm.coe.int/from-india-to-europe-factsheets-on-romani-history/16808b18ed>, last accessed December 15, 2019
- [598] Población Gitana, <https://www.msrebs.gob.es/ssi/familiasInfancia/PoblacionGitana/home.htm>, last accessed December 15, 2019
- [599] Sperling, J.. Spain: migration 1960s to present. *The Encyclopedia of Global Human Migration*, (2013).
- [600] Parga-Lozano, C., Rey-Medrano, D., Gomez-Prieto, P., Areces, C., Moscoso, J., Abd-El-Fatah-Khalil, S., Moreno, E. and Arnaiz-Villena, A., HLA genes in Amerindian immigrants to Madrid (Spain): epidemiology and a virtual transplantation waiting list. *Molecular Biology Reports*, 38(4), (2011), pp.2263-2271.
- [601] Brion, M., Salas, A., Gonzalez-Neira, A., Lareu, M. V. and Carracedo, A. Insights into Iberian population origins through the construction of highly informative Y-chromosome haplotypes using biallelic markers, STRs, and the MSY1 minisatellite. *Am. J. Phys. Anthropol.* 122(2), (2003), pp. 147–161.
- [602] Spínola, H., Middleton, D. and Brehm, A. HLA genes in Portugal inferred from sequence-based typing: in the crossroad between Europe and Africa. *HLA*, 66(1), (2005), pp.26-36.
- [603] São João, R., Papoila, A.L., Ligeiro, D. and Trindade, H. “HLA Allele and Haplotype Frequencies of the Portuguese Bone Marrow Donors Registry”. In: *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*. Springer, (2013), pp. 417-426.
- [604] Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, Á., Donnelly, P. and Myers, S. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nature Communications*, 10(1), (2019), pp.1-14.
- [605] Pimenta, J., Lopes, A.M., Carracedo, A., Arenas, M., Amorim, A. and Comas, D. Spatially explicit analysis reveals complex human genetic gradients in the Iberian Peninsula. *Scientific Reports*, 9(1), (2019), pp.1-9.
- [606] Gayán, J. , Galan, J.J., González-Pérez, A., Sáez, M.E., Martínez-Larrad, M.T., Zabena, C., Rivero, M.C., Salinas, A., Ramírez-Lorca, R., Morón, F.J., Royo, J.L., Moreno-Rey, C., Velasco, J., Carrasco, J.M., Molero, E., Ochoa, C., Ochoa, M.D., Gutiérrez, M., Reina, M., Pascual, R., Romo-Astorga, A., Susillo-González, J.L., Vázquez, E., Real, L.M., Ruiz, A., Serrano-Ríos, M. Genetic structure of the Spanish population. *BMC Genomics*, 11(1), (2010), pp.1-15.
- [607] Laayouni, H., Calafell, F. and Bertranpetit, J. A genome-wide survey does not show the genetic distinctiveness of Basques. *Human Genetics*, 127(4), (2010.), pp.455-458.
- [608] Martínez-Laso, J., Juan, D., Martínez-Quiles, N., Gomez-Casado, E., Cuadrado, E. and Arnaiz-Villena, A. The contribution of the HLA-A,-B,-C and-DR,-DQ DNA typing to the study of the origins of Spaniards and Basques. *HLA*, 45(4), (1995), 237-245.
- [609] Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, Bertranpetit J, Comas D, Bustamante CD. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences*, 110(29), (2013), pp.11791-11796

- [610] Torres-Galván MJ, Quiralte J, Blanco C, Castillo R, Carrillo T, Pérez-Aciego P and Sánchez-García F. Linkage of house dust mite allergy with the HLA region. *Annals of Allergy, Asthma & Immunology*, 82(2), (1999), pp.198-203.
- [611] Piancatelli, D., Canossi, A., Aureli, A., Oumhani, K., Del Beato, T., Di Rocco, M., Liberatore, G., Tessitore, A., Witter, K., El Aouad, R. and Adorno, D. Human leukocyte antigen-A,-B, and-Cw polymorphism in a Berber population from North Morocco using sequence-based typing. *HLA*, 63(2), (2004), pp.158-172.
- [612] Gómez-Casado E, del Moral P, Martínez-Laso J, García-Gómez A, Allende L, Silvera-Redondo C, Longas J, González-Hevilla M, Kandil M, Zamora J and Arnaiz-Villena A.. HLA gene in Arabic-Speaking Moroccans: close relatedness to Berbers and Iberians. *Tissue Antigens*. 55(3), (2000), pp. 239–49.
- [613] Canossi A, Piancatelli D, Aureli A, Oumhani K, Ozzella G, Del Beato T, Liberatore G, El Aouad R, Adorno D. Correlation between genetic HLA class I and II polymorphisms and anthropological aspects in the Chaouya population from Morocco (Arabic speaking). *Tissue Antigens*. 76(3), (2010), pp. 177-93.
- [614] Klitz, W., Gragert, L., Maiers, M., Fernandez-Viña, M., Ben-Naeh, Y., Benedek, G., Brautbar, C. and Israel, S.. Genetic differentiation of Jewish populations. *HLA*, 76(6), (2010), pp.442-458.
- [615] Manor, S., Halagan, M., Shriki, N., Yaniv, I., Zisser, B., Maiers, M., Madbouly, A. and Stein, J. High-resolution HLA A~B~DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel. *Human Immunology*, 77(12), (2016), pp.1114-1119.
- [616] Amar, A., Kwon, O.J., Motro, U., Witt, C.S., Bonne-Tamir, B., Gabison, R. and Brautbar, C.. Molecular analysis of HLA class II polymorphisms among different ethnic groups in Israel. *Human Immunology*, 60(8), (1999), pp.723-730.
- [617] Bonn -Tamir, B., Bodmer, J.G., Bodmer, W.F., Pickbourne, P., Brautbar, C., Gazit, E., Nevo, S. and Zamir, R. HLA polymorphism in Israel. 9. An overall comparative analysis. *Tissue Antigens*, 11(3), (1978), pp.235-250.
- [618] Martinez-Laso J, Gazit E, Gomez-Casado E, Morales P, Martinez-Quiles N, Alvarez M, Martin-Villa JM, Fernandez V, Arnaiz-Villena A. HLA DR and DQ polymorphism in Ashkenazi and non-Ashkenazi Jews: comparison with other Mediterraneans. *Tissue Antigens*, 47(1), (1996), pp.63-71.
- [619] Roitberg-Tambur, A., Witt, C.S., Friedmann, A., Safirman, C., Sherman, L., Battat, S., Nelken, D. and Brautbar, C.. Comparative analysis of HLA polymorphism at the serologic and molecular level in Moroccan and Ashkenazi Jews. *Tissue Antigens*, 46(2), (1995), pp.104-110.
- [620] Arnaiz-Villena, A., Gomez-Casado, E. and Martinez-Laso, J. Population genetic relationships between Mediterranean populations determined by HLA allele distribution and a historic perspective. *HLA*, 60(2), (2002), pp.111-121.
- [621] Cano, P., Testi, M., Andreani, M., Khoriaty, E., Monsef, J.B., Galluccio, T., Troiano, M., Fernandez-Vina, M. and Inati, A. HLA population genetics: a Lebanese population. *HLA*, 80(4), (2012), pp.341-355.
- [622] Massanet, M.F., Castro, J.A., Picornell, A. and Ramon, M.M.. Study of the populations of the Balearic Islands (Spain) using mtDNA RFLPs. *Human Biology*, (1997), pp.483-498.
- [623] Falchi, A., Giovannoni, L., Calo, C.M., Piras, I.S., Moral, P., Paoli, G., Vona, G. and Varesi, L. Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *Journal of Human Genetics*, 51(1), (2006), pp.9-14.
- [624] Balas, A., Garc a-S nchez, F. and Vicario, J.L. Allelic and haplotypic HLA frequency distribution in Spanish hematopoietic patients. Implications for unrelated donor searching. *HLA*, 77(1), (2011), 45-53.

- [625] Varela, T.A., Lodeiro, R. and Fariña, J. HLA gene and haplotype frequencies in Galicia (NW Spain). *International Journal of Anthropology*, 7(3), (1992), pp.19-24.
- [626] Arnaiz-Villena, A., Carballo, A., Juarez, I., Muñiz, E., Campos, C., Tejedor, B., Martín-Villa, M. and Palacio-Gruber, J.. HLA genes in Atlantic Celtic populations: are Celts Iberians?. *International Journal of Modern Anthropology*, 1(10), (2017), pp.50-72.
- [627] Comas D, Mateu E, Calafell F, Pérez-Lezaun A, Bosch E, Martínez-Arias R, and Bertranpetit, J. HLA class I and class II DNA typing and the origin of Basques. *Tissue Antigens*. 51(1), (1998), pp.30–40.
- [628] Alcoceba, M., Marin, L., Balanzategui, A., Sarasquete, M.E., Chillón, M.C., Martín-Jiménez, P., Puig, N., Santamaría, C., Corral, R., García-Sanz, R. and San Miguel, J.F. Frequency of HLA-A,-B and-DRB1 specificities and haplotypic associations in the population of Castilla y León (northwest-central Spain). *HLA*, 78(4), (2011), pp. 249-255.
- [629] Pérez-Miranda, A.M., Alfonso-Sánchez, M.A., Pena, J.A. and Calderón, R. HLA-DQA1 polymorphism in autochthonous Basques from Navarre (Spain): genetic position within European and Mediterranean scopes. *Tissue Antigens*, 61(6), (2003), pp.465-474
- [630] Pérez-Miranda, A.M., Alfonso-Sánchez, M.A., Vidales, M.C., Calderón, R. and Pena, J.A. Genetic polymorphism and linkage disequilibrium of the HLA-DP region in Basques from Navarre (Spain). *Tissue Antigens*, 64(3), (2004), pp.264-275.
- [631] López-Larrea, C., Alonso, J.T., Perez, A.R. and Coto, E.. HLA antigens in psoriatic arthritis subtypes of a Spanish population. *Annals of the Rheumatic Diseases*, 49(5), (1990), pp.318-319.
- [632] Valenzuela, A., Gonzalez-Escribano, M.F., Rodriguez, R., Moreno, I., Garcia, A. and Núñez-Roldan, A. Association of HLA shared epitope with joint damage progression in rheumatoid arthritis. *Human Immunology*, 60(3), (1999), pp.250-254.
- [633] González-Escribano, M.F., Rodríguez, M.R., Walter, K., Sanchez-Roman, J., Garcia-Lozano, J.R. and Núñez-Roldán, A.. Association of HLA-B51 subtypes and Behcet's disease in Spain. *Tissue Antigens*, 52(1), (1998), pp.78-80.
- [634] Medrano, L.M., Dema, B., López-Larios, A., Maluenda, C., Bodas, A., López-Palacios, N., Figueredo, M.Á., Fernández-Arquero, M. and Núñez, C.. HLA and celiac disease susceptibility: new genetic factors bring open questions about the HLA influence and gene-dosage effects. *PLOS One*, 7(10), (2012), p.e48403.
- [635] Villoslada P, Barcellos LF, Rio J, Begovich AB, Tintore M, Sastre-Garriga J, Baranzini SE, Casquero P, Hauser SL, Montalban X, Oksenberg JR. The HLA locus and multiple sclerosis in Spain. Role in disease susceptibility, clinical course and response to interferon- $\beta$ . *Journal of Neuroimmunology*, 130(1-2), (2002), pp.194-201.
- [636] González-Escribano, M.F., Jimenez, G., Walter, K., Montes, M., Perez-Bernal, A.M., Rodriguez, M.R., Conejo-Mir, J.S. and Núñez-Roldán, A. Distribution of HLA class II alleles among Spanish patients with pemphigus vulgaris. *Tissue Antigens*, 52(3), (1998), pp.275-278.
- [637] Balas, A., Planelles, D., Solves, P., Roig, R. and Vicario, J.L. Genomic full-length analysis of the B\*08:79 allele suggests exon shuffling involving the B\* 08:01:01 and B\*07:06 alleles. *Tissue Antigens*, 80(3), (2012), p.268.
- [638] Martínez-Laso, J., Herraiz, M.A., Vidart, J.A., Peñalosa, J., Barbolla, M.L., Jurado, M.L. and Cervera, I. Polymorphism of the HLA-B\*15 group of alleles is generated following 5 lineages of evolution. *Human Immunology*, 72(5), (2011), pp.412-421.
- [639] Montero-Martín, G., Cervera, I., Teniente-Serra, A., Fonolleda, M. and Martinez-Laso, J. Description of the novel HLA-DQB1\*02:02:01:02 allele in a Spanish individual. *HLA*, 87(2), (2016), p.113.

- [640] Balas, A., Santos, S., Aviles, M.J., García-Sánchez, F., Lillo, R., Alvarez, A., Villar-Guimerans, L.M. and Vicario, J.L. Elongation of the cytoplasmic domain, due to a point deletion at exon 7, results in an HLA-C null allele, Cw\*0409 N. *HLA*, 59(2), (2002), pp.95-100.
- [641] Suárez, M.B., Morales, P., Castro, M.J., Fernández, V., Varela, P., Alvarez, M., Martínez-Laso, J. and Arnaiz-Villena, A. A new HLA-G allele (HLA-G\*0105N) and its distribution in the Spanish population. *Immunogenetics*, 45(6), (1997), pp.464-465.
- [642] Matesanz, R., Domínguez-Gil, B., Coll, E., Mahillo, B., and Marazuela, R. How Spain reached 40 deceased organ donors per million population. *American Journal of Transplantation*, 17(6), (2017), pp.1447-1454.
- [643] Red Española de Donantes de Medula Osea -José Carreras Foundation: <https://www.fcarreras.org/en>, last accessed July 15, 2020.
- [644] Hauser SL, Goodin DS. Multiple Sclerosis and other demyelinating diseases. In: Harrison's Principles of Internal Medicine. *McGraw-Hill*, (2012), pp. 3395-3409.
- [645] Sospedra, M. and Martin, R. Immunology of multiple sclerosis. *Annu. Rev. Immunol.*, 23, (2005), pp.683-747.
- [646] Browne P, Chandraratna D, Angood C, Tremlett H, Baker C, Taylor BV, Thompson AJ. Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology*, 83(11), (2014), pp. 1022-1024.
- [647] Multiple Sclerosis International Federation. Atlas of MS, 2013. Mapping multiple sclerosis around the world. 2013. <http://www.msif.org/about-us/advocacy/atlas/>, last accessed July 15, 2020.
- [648] Koch-Henriksen, N. and Sørensen, P.S. The changing demographic pattern of multiple sclerosis epidemiology. *The Lancet Neurology*, 9(5), (2010), pp.520-532.
- [649] Belbasis, L., Bellou, V., Evangelou, E., Ioannidis, J.P. and Tzoulaki, I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. *The Lancet Neurology*, 14(3), (2015), pp.263-273.
- [650] Hollenbach, J.A. and Oksenberg, J.R.. The immunogenetics of multiple sclerosis: a comprehensive review. *Journal of Autoimmunity*, 64, (2015), pp.13-25.
- [651] Canto, E. and Oksenberg, J.R. Multiple sclerosis genetics. *Multiple Sclerosis Journal*, 24(1), (2018), pp.75-79.
- [652] International Multiple Sclerosis Genetics Consortium. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476, (2011), pp. 214–219.
- [653] International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*, 365(6460), (2019), p.eaav7188.
- [654] Bertrams, J., Kuwert, E., Liedtke, U., HLA antigens and multiple sclerosis, *Tissue Antigens*, 2, (1972), pp. 405-408.
- [655] Naito, S., Namerow, N., Mickey, M.R. and Terasaki, P.I. Multiple sclerosis: association with HL-A3. *Tissue Antigens*, 2(1), (1972), pp.1-4.
- [656] Jersild C, Fog T, Hansen GS, Thomsen M, Svejgaard A, Dupont B. Histocompatibility determinants in multiple sclerosis, with special reference to clinical course. *Lancet*, 2, (1973), pp. 1221-1225.
- [657] Sawcer, S., Franklin, R.J. and Ban, M. Multiple sclerosis genetics. *The Lancet Neurology*, 13(7), (2014), pp.700-709.

- [658] Patsopoulos NA, Barcellos LF, Hintzen RQ, Schaefer C, van Duijn CM, Noble JA, Raj T; IMSGC; ANZgene, Gourraud PA, Stranger BE, Oksenberg J, Olsson T, Taylor BV, Sawcer S, Hafler DA, Carrington M, De Jager PL, de Bakker PI. Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLOS Genetics*, 9(11), (2013), p.e1003926.
- [659] Stamatelos, P., Anagnostouli, M. HLA-Genotype in Multiple Sclerosis: The Role in Disease onset, Clinical Course, Cognitive Status and Response to Treatment: A Clear Step Towards Personalized Therapeutics. *Immunogenetics: Open Access* 2(1), (2017), pp. 1-12.
- [660] Isobe N, Keshavan A, Gourraud PA, Zhu AH, Datta E, Schlaeger R, Caillier SJ, Santaniello A, Lizée A, Himmelstein DS, Baranzini SE, Hollenbach J, Cree BA, Hauser SL, Oksenberg JR, Henry RG. Association of HLA genetic risk burden with disease phenotypes in multiple sclerosis. *JAMA Neurology*, 73(7), (2016) pp. 795–802.
- [661] Okuda DT, Srinivasan R, Oksenberg JR, Goodin DS, Baranzini SE, Beheshtian A, Waubant E, Zamvil SS, Leppert D, Qualley P, Lincoln R, Gomez R, Caillier S, George M, Wang J, Nelson SJ, Cree BA, Hauser SL, Pelletier D. Genotype-phenotype correlations in multiple sclerosis: HLA genes influence disease severity inferred by 1HMR spectroscopy and MRI measures. *Brain*, 132(1), (2009), pp.250-259.
- [662] Gourraud, P.A., Harbo, H.F., Hauser, S.L. and Baranzini, S.E. The genetics of multiple sclerosis: an up-to-date review. *Immunological Reviews*, 248(1), (2012), pp.87-103.
- [663] Ramagopalan, S.V. and Ebers, G.C. Multiple sclerosis: major histocompatibility complexity and antigen presentation. *Genome Medicine*, 1(11), (2009), pp.1-5.
- [664] Isobe, N. and Oksenberg, J.R. Genetic studies of multiple sclerosis and neuromyelitis optica: Current status in European, African American and Asian populations. *Clinical and Experimental Neuroimmunology*, 5(1), (2014), pp.61-68.
- [665] Barcellos LF, Oksenberg JR, Green AJ, Bucher P, Rimmler JB, Schmidt S, Garcia ME, Lincoln RR, Pericak-Vance MA, Haines JL, Hauser SL; Multiple Sclerosis Genetics Group. Genetic basis for clinical expression in multiple sclerosis. *Brain*, 125(1), (2002), pp.150-158.
- [666] Schmidt, H., Williamson, D. and Ashley-Koch, A., 2007. HLA-DR15 haplotype and multiple sclerosis: a HuGE review. *American journal of epidemiology*, 165(10), pp.1097-1109.
- [667] Fogdell, A., Hillert, J., Sachs, C. and Olerup, O. The multiple sclerosis-and narcolepsy-associated HLA class II haplotype includes the DRB5\*0101 allele. *Tissue Antigens*, 46(4), (1995), pp.333-336.
- [668] Caillier SJ, Briggs F, Cree BA, Baranzini SE, Fernandez-Viña M, Ramsay PP, Khan O, Royal W 3rd, Hauser SL, Barcellos LF, Oksenberg JR. Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. *The Journal of Immunology*, 181(8), (2008), pp.5473-5480.
- [669] Oksenberg JR, Barcellos LF, Cree BA, Baranzini SE, Bugawan TL, Khan O, Lincoln RR, Swerdlin A, Mignot E, Lin L, Goodin D, Erlich HA, Schmidt S, Thomson G, Reich DE, Pericak-Vance MA, Haines JL, Hauser SL. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. *The American Journal of Human Genetics*, 74(1), (2004), pp.160-167.
- [670] Marrosu MG, Murru R, Murru MR, Costa G, Zavattari P, Whalen M, Cocco E, Mancosu C, Schirru L, Solla E, Fadda E, Melis C, Porru I, Rolesu M, Cucca F. Dissection of the HLA association with multiple sclerosis in the founder isolated population of Sardinia. *Human Molecular Genetics*, 10(25), (2001), pp.2907-2916.
- [671] Haines JL, Terwedow HA, Burgess K, Pericak-Vance MA, Rimmler JB, Martin ER, Oksenberg JR, Lincoln R, Zhang DY, Banatao DR, Gatto N, Goodkin DE, Hauser SL. Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. *Human Molecular Genetics*, 7(8), (1998), pp.1229-1234.



- [672] Lincoln MR, Ramagopalan SV, Chao MJ, Herrera BM, Deluca GC, Orton SM, Dyment DA, Sadovnick AD, Ebers GC. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. *Proceedings of the National Academy of Sciences* 106(18), (2009), pp. 7542-7547.
- [673] Barcellos LF, Oksenberg JR, Begovich AB, Martin ER, Schmidt S, Vittinghoff E, Goodin DS, Pelletier D, Lincoln RR, Bucher P, Swerdlin A, Pericak-Vance MA, Haines JL, Hauser SL; Multiple Sclerosis Genetics Group. HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. *The American Journal of Human Genetics*, 72(3), (2003), pp.710-716.
- [674] Dyment, D.A., Herrera, B.M., Cader, M.Z., Willer, C.J., Lincoln, M.R., Sadovnick, A.D., Risch, N. and Ebers, G.C. Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance. *Human Molecular Genetics*, 14(14), (2005), pp.2019-2026.
- [675] Kwon, O.J., Karni, A., Israel, S., Brautbar, C., Amar, A., Meiner, Z., Abramsky, O. and Karussis, D. HLA class II susceptibility to multiple sclerosis among Ashkenazi and non-Ashkenazi Jews. *Archives of Neurology*, 56(5), (1999), pp.555-560.
- [676] Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, Briggs F, Cree BC, Begovich AB, Villoslada P, Montalban X, Uccelli A, Savettieri G, Lincoln RR, DeLoa C, Haines JL, Pericak-Vance MA, Compston A, Hauser SL, Oksenberg JR. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. *Human Molecular Genetics*, 15(18), (2006), pp.2813-2824.
- [677] Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, Shah TS, Patsopoulos NA, Alfredsson L, Anderson CA, Attfield KE, Baranzini SE, Barrett J, Binder TMC, Booth D, Buck D, Celius EG, Cotsapas C, D'Alfonso S, Dendrou CA, Donnelly P, Dubois B, Fontaine B, Fugger L, Goris A, Gourraud PA, Graetz C, Hemmer B, Hillert J; International IBD Genetics Consortium (IBDGC), Kockum I, Leslie S, Lill CM, Martinelli-Boneschi F, Oksenberg JR, Olsson T, Oturai A, Saarela J, Søndergaard HB, Spurkland A, Taylor B, Winkelmann J, Zipp F, Haines JL, Pericak-Vance MA, Spencer CCA, Stewart G, Hafler DA, Ivinson AJ, Harbo HF, Hauser SL, De Jager PL, Compston A, McCauley JL, Sawcer S, McVean G. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature Genetics*, 47(10), (2015), p.1107.
- [678] Yoshimura S, Isobe N, Yonekawa T, Matsushita T, Masaki K, Sato S, Kawano Y, Yamamoto K, Kira J; South Japan Multiple Sclerosis Genetics Consortium. Genetic and infectious profiles of Japanese multiple sclerosis patients. *PLOS One*, 7(11), (2012), p.e48592.
- [679] Matsuoka, T., Matsushita, T., Osoegawa, M., Kawano, Y., Minohara, M., Mihara, F., Nishimura, Y., Ohyagi, Y. and Kira, J. Association of the HLA-DRB1 alleles with characteristic MRI features of Asian multiple sclerosis. *Multiple Sclerosis Journal*, 14(9), (2008), pp.1181-1190.
- [680] McElroy, J.P., Isobe, N., Gourraud, P.A., Caillier, S.J., Matsushita, T., Kohriyama, T., Miyamoto, K., Nakatsuji, Y., Miki, T., Hauser, S.L., Oksenberg, J.R., and Kira, J. SNP-based analysis of the HLA locus in Japanese multiple sclerosis patients. *Genes & Immunity*, 12(7), (2011), pp.523-530.
- [681] Qiu, W., James, I., Carroll, W.M., Mastaglia, F.L. and Kermode, A.G. HLA-DR allele polymorphism and multiple sclerosis in Chinese populations: a meta-analysis. *Multiple Sclerosis Journal*, 17(4), (2011), pp.382-388.
- [682] Brassat, D., Salemi, G., Barcellos, L.F., McNeill, G., Proia, P., Hauser, S.L., Oksenberg, J.R. and Savettieri, G. The HLA locus and multiple sclerosis in Sicily. *Neurology*, 64(2), (2005), pp.361-363.
- [683] Isobe N, Gourraud PA, Harbo HF, Caillier SJ, Santaniello A, Khankhanian P, Maiers M, Spellman S, Cereb N, Yang S, Pando MJ, Piccio L, Cross AH, De Jager PL, Cree BA, Hauser SL, Oksenberg JR. Genetic risk variants in African Americans with multiple sclerosis. *Neurology*, 81(3), (2013), pp.219-227.
- [684] Dekker, J.W., Eastsal, S., Jakobson, I.B., Gao, X., Strwart, G.J., Buhler, M.M., Hawkins, B.R., Higgins, D.A., Yu, Y.L. and Serjeantson, S.W. HLA-DPB1 alleles correlate with risk for multiple sclerosis in

- Caucasoid and Cantonese patients lacking the high-risk DQB1\*0602 allele. *Tissue Antigens*, 41(1), (1993), pp.31-36.
- [685] Karni, A., Kohn, Y., Saftirminn, C., Abraimsky, O., Barcellos, L., Oksenberg, J.R., Kahana, E., Karussis, D., Chapman, J. and Brautbur, C. Evidence for the genetic role of human leukocyte antigens in low frequency DRB1\*1501 multiple sclerosis patients in Israel. *Multiple Sclerosis Journal*, 5(6), (1999), pp.410-415.
- [686] Marrosu, M.G., Cocco, E., Costa, G., Murru, M.R., Mancosu, C., Murru, R., Lai, M., Sardu, C. and Contu, P. Interaction of loci within the HLA region influences multiple sclerosis course in the Sardinian population. *Journal of Neurology*, 253(2), (2006), pp.208-213.
- [687] Cocco E, Sardu C, Pieroni E, Valentini M, Murru R, Costa G, Tranquilli S, Frau J, Coghe G, Carboni N, Floris M, Contu P, Marrosu MG. HLA-DRB1-DQB1 haplotypes confer susceptibility and resistance to multiple sclerosis in Sardinia. *PLOS One*, 7(4), (2012), p.e33972.
- [688] Lampis, R., Morelli, L., De Virgiliis, S., Congia, M. and Cucca, F. The distribution of HLA class II haplotypes reveals that the Sardinian population is genetically differentiated from the other Caucasian populations. *Tissue Antigens*, 56(6), (2000), pp.515-521.
- [689] Romero-Pinel, L., Pujal, J.M., Martínez-Yélamos, S., Gubieras, L., Matas, E., Bau, L., Torrabadella, M., Azqueta, C. and Arbizu, T. HLA-DRB1: genetic susceptibility and disability progression in a Spanish multiple sclerosis population. *European Journal of Neurology*, 18(2), (2011), pp.337-342.
- [690] Masterman, T., Ligers, A., Olsson, T., Andersson, M., Olerup, O. and Hillert, J. HLA-DR15 is associated with lower age at onset in multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 48(2), (2000), pp.211-219.
- [691] Stankovich J, Butzkueven H, Marriott M, Chapman C, Tubridy N, Tait BD, Varney MD, Taylor BV, Foote SJ; ANZgene Consortium, Kilpatrick TJ, Rubio JP. HLA-DRB1 associations with disease susceptibility and clinical course in Australians with multiple sclerosis. *Tissue Antigens*, 74(1), (2009), pp.17-21.
- [692] Smestad C, Brynedal B, Jonasdottir G, Lorentzen AR, Masterman T, Akesson E, Spurkland A, Lie BA, Palmgren J, Celius EG, Hillert J, Harbo HF. The impact of HLA-A and-DRB1 on age at onset, disease course and severity in Scandinavian multiple sclerosis patients. *European Journal of Neurology*, 14(8), (2007), pp.835-840.
- [693] Silva AM, Pereira C, Bettencourt A, Carvalho C, Couto AR, Leite MI, Marta M, Freijo M, Costa PP, Mendonça D, Monteiro L, Armas JB, Martins B. The role of HLA-DRB1 alleles on susceptibility and outcome of a Portuguese Multiple Sclerosis population. *Journal of the Neurological Sciences*, 258(1-2), (2007), pp.69-74.
- [694] Brum, D.G., Barreira, A.A., Louzada-Junior, P., Mendes-Junior, C.T. and Donadi, E.A. Association of the HLA-DRB1\* 15 allele group and the DRB1\*1501 and DRB1\*1503 alleles with multiple sclerosis in White and Mulatto samples from Brazil. *Journal of Neuroimmunology*, 189(1-2), (2007), pp.118-124.
- [695] Quelvennec, E., Bera, O., Cabre, P., Alizadeh, M., Smadja, D., Jugde, F., Edan, G. and Semana, G. Genetic and functional studies in multiple sclerosis patients from Martinique attest for a specific and direct role of the HLA-DR locus in the syndrome. *Tissue Antigens*, 61(2), (2003), pp.166-171.
- [696] Zipp, F., Windemuth, C., Pankow, H., Dichgans, J., Wienker, T., Martin, R. and Müller, C. Multiple sclerosis associated amino acids of polymorphic regions relevant for the HLA antigen binding are confined to HLA-DR2. *Human Immunology*, 61(10), (2000), pp.1021-1030.
- [697] Kira, J.I. Multiple sclerosis in the Japanese population. *The Lancet Neurology*, 2(2), (2003), pp.117-127.

- [698] Finn, T.P., Jones, R.E., Rich, C., Dahan, R., Link, J., David, C.S., Chou, Y.K., Offner, H. and Vandenberg, A.A. HLA-DRB1\* 1501 risk association in multiple sclerosis may not be related to presentation of myelin epitopes. *Journal of Neuroscience Research*, 78(1), (2004), pp.100-114.
- [699] Correa, E., Paredes, V. and Martínez, B. Prevalence of multiple sclerosis in Latin America and its relationship with European migration. *Multiple Sclerosis Journal—Experimental, Translational and Clinical*, 2, (2016), p.2055217316666407.
- [700] De Silvestri, A., Capittini, C., Mallucci, G., Bergamaschi, R., Rebuffi, C., Pasi, A., Martinetti, M. and Tinelli, C. The Involvement of HLA Class II Alleles in Multiple Sclerosis: A Systematic Review with Meta-analysis. *Disease Markers*, (2019).
- [701] Alvarado-de la Barrera, C., Zúñiga-Ramos, J., Ruíz-Morales, J.A., Estanol, B., Granados, J. and Llorente, L. HLA class II genotypes in Mexican Mestizos with familial and nonfamilial multiple sclerosis. *Neurology*, 55(12), (2000), pp.1897-1900.
- [702] Rivera, V.M.. Multiple sclerosis in Latin Americans: genetic aspects. *Current Neurology and Neuroscience Reports*, 17(8), (2017), p.57.
- [703] Werneck, L.C., Lorenzoni, P.J., Arndt, R.C., Kay, C.S.K. and Scola, R.H. The immunogenetics of multiple sclerosis. The frequency of HLA-alleles class 1 and 2 is lower in Southern Brazil than in the European population. *Arquivos de Neuro-Psiquiatria*, 74(8), (2016), pp.607-616.
- [704] Chi C, Shao X, Rhead B, Gonzales E, Smith JB, Xiang AH, Graves J, Waldman A, Lotze T, Schreiner T, Weinstock-Guttman B, Aaen G, Tillema JM, Ness J, Candee M, Krupp L, Gorman M, Benson L, Chitnis T, Mar S, Belman A, Casper TC, Rose J, Moodley M, Rensel M, Rodriguez M, Greenberg B, Kahn L, Rubin J, Schaefer C, Waubant E, Langer-Gould A, Barcellos LF. Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLOS Genetics*, 15(1), (2019), p.e1007808.
- [705] Romero-Hidalgo S, Flores-Rivera J, Rivas-Alonso V, Barquera R, Villarreal-Molina MT, Antuna-Puente B, Macias-Kauffer LR, Villalobos-Comparán M, Ortiz-Maldonado J, Yu N, Lebedeva TV, Alosco SM, García-Rodríguez JD, González-Torres C, Rosas-Madriral S, Ordoñez G, Guerrero-Camacho JL, Treviño-Frenk I, Escamilla-Tilch M, García-Lechuga M, Tovar-Méndez VH, Pacheco-Ubaldo H, Acuña-Alonzo V, Bortolini MC, Gallo C, Bedoya G, Rothhammer F, González-Jose R, Ruiz-Linares A, Canizales-Quinteros S, Yunis E, Granados J, Corona T. Native American ancestry significantly contributes to neuromyelitis optica susceptibility in the admixed Mexican population. *Scientific Reports*, 10(1), (2020), pp.1-11.
- [706] Maghbooli, Z., Sahraian, M.A. and Naser Moghadasi, A. Multiple sclerosis and human leukocyte antigen genotypes: Focus on the Middle East and North Africa region. *Multiple Sclerosis Journal—Experimental, Translational and Clinical*, 6(1), (2020), p.2055217319881775.
- [707] Mohajer, B., Abbasi, N., Pishgar, F., Abdolalizadeh, A., Ebrahimi, H., Razaviyoun, T., Mohebbi, F., Eskandarieh, S. and Sahraian, M.A. HLA-DRB1 polymorphism and susceptibility to multiple sclerosis in the Middle East North Africa region: A systematic review and meta-analysis. *Journal of Neuroimmunology*, 321, (2018), pp.117-124.
- [708] Zhang, Q., Lin, C.Y., Dong, Q., Wang, J. and Wang, W. Relationship between HLA-DRB1 polymorphism and susceptibility or resistance to multiple sclerosis in Caucasians: a meta-analysis of non-family-based studies. *Autoimmunity Reviews*, 10(8), (2011), pp.474-481.
- [709] Romero-Pinel, L., Pujal, J.M., Martínez-Yélamos, S., Gubieras, L., Matas, E., Bau, L., Torrabadella, M., Azqueta, C. and Arbizu, T. Epistasis between HLA-DRB1 parental alleles in a Spanish cohort with multiple sclerosis. *Journal of the Neurological Sciences*, 298(1-2), (2010), pp.96-100.

- [710] Tsai, S. and Santamaria, P. MHC class II polymorphisms, autoreactive T-cells, and autoimmunity. *Frontiers in Immunology*, 4, (2013), p.321.
- [711] Chao, M.J., Barnardo, M.C., Lincoln, M.R., Ramagopalan, S.V., Herrera, B.M., Dymont, D.A., Montpetit, A., Sadovnick, A.D., Knight, J.C. and Ebers, G.C. HLA class I alleles tag HLA-DRB1\*1501 haplotypes for differential risk in multiple sclerosis susceptibility. *Proceedings of the National Academy of Sciences*, 105(35), (2008), pp.13069-13074.
- [712] Harbo HF, Lie BA, Sawcer S, Celius EG, Dai KZ, Oturai A, Hillert J, Lorentzen AR, Laaksonen M, Myhr KM, Ryder LP, Fredrikson S, Nyland H, Sørensen PS, Sandberg-Wollheim M, Andersen O, Svejgaard A, Edland A, Mellgren SI, Compston A, Vartdal F, Spurkland A. Genes in the HLA class I region may contribute to the HLA class II-associated genetic susceptibility to multiple sclerosis. *Tissue Antigens*, 63(3), (2004), pp.237-247.
- [713] Chao MJ, Barnardo MC, Lui GZ, Lincoln MR, Ramagopalan SV, Herrera BM, Dymont DA, Sadovnick AD, Ebers GC. Transmission of class I/II multi-locus MHC haplotypes and multiple sclerosis susceptibility: accounting for linkage disequilibrium. *Human Molecular Genetics*, 16(16), (2007), pp.1951-1958.
- [714] Healy BC, Liguori M, Tran D, Chitnis T, Glanz B, Wolfish C, Gauthier S, Buckle G, Houtchens M, Stazzone L, Khoury S, Hartzmann R, Fernandez-Vina M, Hafler DA, Weiner HL, Guttman CR, De Jager PL. HLA B\*44: protective effects in MS susceptibility and MRI outcome measures. *Neurology*, 75(7), (2010), pp.634-640.
- [715] International MHC and Autoimmunity Genetics Network, Rioux JD, Goyette P, Vyse TJ, Hammarström L, Fernando MM, Green T, De Jager PL, Foisy S, Wang J, de Bakker PI, Leslie S, McVean G, Padyukov L, Alfredsson L, Annese V, Hafler DA, Pan-Hammarström Q, Matell R, Sawcer SJ, Compston AD, Cree BA, Mirel DB, Daly MJ, Behrens TW, Klareskog L, Gregersen PK, Oksenberg JR, Hauser SL. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of the National Academy of Sciences*, 106(44), (2009), pp.18680-18685.
- [716] Mayo, L., Quintana, F.J. and Weiner, H.L. The innate immune system in demyelinating disease. *Immunological reviews*, 248(1), (2012), pp.170-187.
- [717] Ugolotti, E., Vanni, I., Raso, A., Benzi, F., Malnati, M. and Biassoni, R. Human leukocyte antigen-B (-Bw6/-Bw4 I80, T80) and human leukocyte antigen-C (-C1/-C2) subgrouping using pyrosequencing analysis. *Human Immunology*, 72(10), (2011), pp.859-868.
- [718] Saunders PM, Vivian JP, Baschuk N, Beddoe T, Widjaja J, O'Connor GM, Hitchen C, Pymm P, Andrews DM, Gras S, McVicar DW, Rossjohn J, Brooks AG. The interaction of KIR3DL1\*001 with HLA class I molecules is dependent upon molecular microarchitecture within the Bw4 epitope. *The Journal of Immunology*, 194(2), (2015), pp.781-789.
- [719] Boudreau, J.E., Mulrooney, T.J., Le Luduec, J.B., Barker, E. and Hsu, K.C. KIR3DL1 and HLA-B density and binding calibrate NK education and response to HIV. *The Journal of Immunology*, 196(8), (2016), pp.3398-3410.
- [720] Biassoni, R. Human natural killer receptors, co-receptors, and their ligands. *Current Protocols in Immunology*, 84(1), (2009), pp.14-10.
- [721] Sanjanwala, B., Draghi, M., Norman, P.J., Guethlein, L.A. and Parham, P. Polymorphic sites away from the Bw4 epitope that affect interaction of Bw4+ HLA-B with KIR3DL1. *The Journal of Immunology*, 181(9), (2008), pp.6293-6300.
- [722] Mandelboim, O., Reyburn, H.T., Valés-Gómez, M., Pazmany, L., Colonna, M., Borsellino, G. and Strominger, J.L. Protection from lysis by natural killer cells of group 1 and 2 specificity is mediated by

- residue 80 in human histocompatibility leukocyte antigen C alleles and also occurs with empty major histocompatibility complex molecules. *The Journal of Experimental Medicine*, 184(3), (1996), pp.913-922.
- [723] Colonna, M., Brooks, E.G., Falco, M., Ferrara, G.B. and Strominger, J.L. Generation of allospecific natural killer cells by stimulation across a polymorphism of HLA-C. *Science*, 260(5111), (1993), pp.1121-1124.
- [724] Hollenbach, J.A., Pando, M.J., Caillier, S.J., Gourraud, P.A. and Oksenberg, J.R. The killer immunoglobulin-like receptor KIR3DL1 in combination with HLA-Bw4 is protective against multiple sclerosis in African Americans. *Genes & Immunity*, 17(3), (2016), pp.199-202.
- [725] Lorentzen AR, Karlsten TH, Olsson M, Smestad C, Mero IL, Woldseth B, Sun JY, Senitzer D, Celius EG, Thorsby E, Spurkland A, Lie BA, Harbo HF. Killer immunoglobulin-like receptor ligand HLA-Bw4 protects against multiple sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 65(6), (2009), pp.658-666.
- [726] García-León JA, Pinto-Medel MJ, García-Trujillo L, López-Gómez C, Oliver-Martos B, Prat-Arrojo I, Marín-Bañasco C, Suardíaz-García M, Maldonado-Sanchez R, Fernández-Fernández O, Leyva-Fernández L. Killer cell immunoglobulin-like receptor genes in Spanish multiple sclerosis patients. *Molecular Immunology*, 48(15-16), (2011), pp.1896-1902.
- [727] Mack SJ, Udell J, Cohen F, Osoegawa K, Hawbecker SK, Noonan DA, Ladner MB, Goodridge D, Trachtenberg EA, Oksenberg JR and Erlich HA. High resolution HLA analysis reveals independent class I haplotypes and amino-acid motifs protective for multiple sclerosis. *Genes & Immunity*, 20(4), (2019), pp.308-326. (Repeated in [292]).
- [728] Bw4 and Bw6 serological epitopes of HLA-B alleles, M.Tevfik Dorak: <http://www.dorak.info/hla/bw4bw6.html>, last accessed July 15, 2020.
- [729] Fernandez O, Fernandez V, Martinez-Cabrera V, Mayorga C, Alonso A, Leon A, Arnal C, Hens M, Luque G, de Ramon E, Caballero A, Leyva L. Multiple sclerosis in gypsies from southern Spain: prevalence, mitochondrial DNA haplogroups and HLA class II association. *Tissue Antigens*, 71, (2008), pp. 426–433.
- [730] Fernandez-Morera JL, Rodriguez-Rodero S, Tunon A, Martinez-Borra J, Vidal-Castineira JR, Lopez-Vazquez A, Rodrigo L, Rodrigo P, Gonzalez S, Lahoz CH, Lopez-Larrea C. Genetic influence of the nonclassical major histocompatibility complex class I molecule MICB in multiple sclerosis susceptibility. *Tissue Antigens*, 72, (2008), pp. 54–59
- [731] Clerici, N. and Fernández, M. Restriction fragment length polymorphism analysis of HLA-DR-and DQ-linked alleles in multiple sclerosis in Spain. *Journal of Neuroimmunology*, 41(2), (1992), pp.245-248.
- [732] De la Concha, E.G., Arroyo, R., Crusius, J.B.A., Campillo, J.A., Martín, C., De Seijas, E.V., Pena, A.S., Claveria, L.E. and Fernandez-Arquero, M. Combined effect of HLA-DRB1\*1501 and interleukin-1 receptor antagonist gene allele 2 in susceptibility to relapsing/remitting multiple sclerosis. *Journal of Neuroimmunology*, 80(1-2), (1997), pp.172-178.
- [733] De la Concha EG, Cavanillas, M.L., Cénit, M.C., Urcelay, E., Arroyo, R., Fernández, Ó., Álvarez-Cermeño, J.C., Leyva, L., Villar, L.M. and Núñez, C. DRB1\*03:01 haplotypes: differential contribution to multiple sclerosis risk and specific association with the presence of intrathecal IgM bands. *PLOS One*, 7(2), (2012), p.e31018.
- [734] Pina, M.A., Ara, J.R., Lasierra, P., Larrad, L. and Modrego, P.J. Major histocompatibility complex class II alleles and the course and outcome of MS. *Neurology*, 52(9), (1999), pp.1920-1920.
- [735] Pina, M.A., Ara, J.R., Lasierra, P., Modrego, P.J. and Larrad, L. Study of HLA as a predisposing factor and its possible influence on the outcome of multiple sclerosis in the sanitary district of Calatayud, northern Spain. *Neuroepidemiology*, 18(4), (1999), pp.203-209.

- [736] Uriá, D.F., Gutierrez, V., Menes, B.B., Arribas, J.M. and Lopez-Larrea, C. HLA class II susceptibility and resistance genes in patients with multiple sclerosis from northern Spain, by DNA-RFLP genotyping. *Journal of Neurology, Neurosurgery, and Psychiatry*, 56(6), (1993), p.722.
- [737] Villoslada P, Barcellos LF, Rio J, Begovich AB, Tintore M, Sastre-Garriga J, Baranzini SE, Casquero P, Hauser SL, Montalban X, Oksenberg JR. The HLA locus and multiple sclerosis in Spain. Role in disease susceptibility, clinical course and response to interferon- $\beta$ . *Journal of Neuroimmunology*, 130(1-2), (2002), pp.194-201. (Repeated in Reference [635]).
- [738] Coraddu, F., Reyes-Yanez, M.P., Parra, A., Gray, J., Smith, S.I., Taylor, C.J. and Compston, D.A.S. HLA associations with multiple sclerosis in the Canary Islands. *Journal of Neuroimmunology*, 87(1-2), (1998), pp.130-135.
- [739] Romero-Pinel, L., Martínez-Yélamos, S., Bau, L., Matas, E., Gubieras, L., Maria Pujal, J., Morandeira, F., Bas, J. and Arbizu, T. Association of HLA-DRB1\*15 allele and CSF oligoclonal bands in a Spanish multiple sclerosis cohort. *European Journal of Neurology*, 18(10), (2011), pp.1258-1262.
- [740] Uría, D.F. HLA y esclerosis múltiple. Estudios en la población española. *Rev Neurol*, 31(11), (2000), pp.1066-70.
- [741] Fernandez, O., Fernandez, V., Alonso, A., Caballero, A., Luque, G., Bravo, M., Leon, A., Mayorga, C., Leyva, L. and De Ramon, E. DQB1\*0602 allele shows a strong association with multiple sclerosis in patients in Malaga, Spain. *Journal of Neurology*, 251(4), (2004), pp.440-444.
- [742] Fernández O, R-Antigüedad A, Pinto-Medel MJ, Mendibe MM, Acosta N, Oliver B, Guerrero M, Papais-Alvarenga M, Fernández-Sánchez V, Leyva L. HLA class II alleles in patients with multiple sclerosis in the Biscay province (Basque Country, Spain). *Journal of Neurology*, 256(12), (2009), p.1977.
- [743] Scholz, E.M., Marcilla, M., Daura, X., Arribas-Layton, D., James, E.A. and Alvarez, I. Human leukocyte antigen (HLA)-DRB1\*15:01 and HLA-DRB5\*01:01 present complementary peptide repertoires. *Frontiers in Immunology*, 8, (2017), p.984.
- [744] DeLuca GC, Ramagopalan SV, Herrera BM, Dymment DA, Lincoln MR, Montpetit A, Pugliatti M, Barnardo MC, Risch NJ, Sadovnick AD, Chao M, Sotgiu S, Hudson TJ, Ebers GC. An extremes of outcome strategy provides evidence that multiple sclerosis severity is determined by alleles at the HLA-DRB1 locus. *Proceedings of the National Academy of Sciences*, 104(52), (2007), pp.20896-20901.
- [745] Fugger, L., Friese, M.A. and Bell, J.I. From genes to function: the next challenge to understanding multiple sclerosis. *Nature Reviews Immunology*, 9(6), (2009), pp.408-417.
- [746] Ramagopalan, S.V. and Ebers, G.C. Epistasis: multiple sclerosis and the major histocompatibility complex. *Neurology*, 72(6), (2009), pp.566-567.
- [747] Alcina A, Abad-Grau Mdel M, Fedetz M, Izquierdo G, Lucas M, Fernández O, Ndagire D, Catalá-Rabasa A, Ruiz A, Gayán J, Delgado C, Arnal C, Matesanz F. Multiple sclerosis risk variant HLA-DRB1\*1501 associates with high expression of DRB1 gene in different human populations. *PLOS One*, 7(1), (2012), p.e29819.
- [748] Dunne, C., McGuigan, C., Crowley, J., Hagan, R., Rooney, G., Kelleher, J., Hutchinson, M. and Lawlor, E. Human leucocyte antigen class II polymorphism in Irish patients with multiple sclerosis. *Tissue Antigens*, 68(3), (2006), pp.257-262.
- [749] Weinshenker, B.G., Santrach, P., Bissonet, A.S., McDonnell, S.K., Schaid, D., Moore, S.B. and Rodriguez, M. Major histocompatibility complex class II alleles and the course and outcome of MS: a population-based study. *Neurology*, 51(3), (1998), pp.742-747.

- [750] Saruhan-Direskeneli G, Esin S, Baykan-Kurt B, Ornek I, Vaughan R, Eraksoy M. HLA-DR and -DQ associations with multiple sclerosis in Turkey. *Human Immunology*, 55, (1997), pp. 59–65.
- [751] Mazdeh, M., Taheri, M., Sayad, A., Bahram, S., Omrani, M.D., Movafagh, A., Inoko, H., Akbari, M.T., Noroozi, R., Hajilooi, M. and Solgi, G. HLA genes as modifiers of response to IFN- $\beta$ -1a therapy in relapsing-remitting multiple sclerosis. *Pharmacogenomics*, 17(5), (2016), pp.489-498.
- [752] Ramagopalan, S.V., DeLuca, G.C., Degenhardt, A. and Ebers, G.C. The genetics of clinical outcome in multiple sclerosis. *Journal of Neuroimmunology*, 201, (2008), pp.183-199.
- [753] George MF, Briggs FB, Shao X, Gianfrancesco MA, Kockum I, Harbo HF, Celius EG, Bos SD, Hedström A, Shen L, Bernstein A, Alfredsson L, Hillert J, Olsson T, Patsopoulos NA, De Jager PL, Oturai AB, Søndergaard HB, Sellebjerg F, Sorensen PS, Gomez R, Caillier SJ, Cree BA, Oksenberg JR, Hauser SL, D'Alfonso S, Leone MA, Martinelli Boneschi F, Sorosina M, van der Mei I, Taylor BV, Zhou Y, Schaefer C, Barcellos LF. Multiple sclerosis risk loci and disease severity in 7,125 individuals from 10 studies. *Neurology Genetics*, 2(4), (2016).
- [754] Ramagopalan, S.V., Morris, A.P., Dymont, D.A., Herrera, B.M., DeLuca, G.C., Lincoln, M.R., Orton, S.M., Chao, M.J., Sadovnick, A.D. and Ebers, G.C. The inheritance of resistance alleles in multiple sclerosis. *PLoS Genetics*, 3(9), (2007), p.e150.
- [755] Brynedal, B., Duvefelt, K., Jonasdottir, G., Roos, I.M., Åkesson, E., Palmgren, J. and Hillert, J. HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. *PLoS One*, 2(7), (2007), p.e664.
- [756] Mamedov A, Vorobyeva N, Filimonova I, Zakharova M, Kiselev I, Bashinskaya V, Baulina N, Boyko A, Favorov A, Kulakova O, Ziganshin R, Smirnov I, Poroshina A, Shilovskiy I, Khaitov M, Sykulev Y, Favorova O, Vlassov V, Gabibov A, Belogurov A Jr. Protective Allele for Multiple Sclerosis *HLA-DRB1\*01:01* Provides Kinetic Discrimination of Myelin and Exogenous Antigenic Peptides. *Frontiers in Immunology*, 10, (2020), p.3088.
- [757] Ramal, L.M., Pablo, R.D., Guadix, M.J., Sánchez, J., Garrido, A., Garrido, F., Jiménez-Alonso, J. and López-Nevot, M.A. HLA class II allele distribution in the Gypsy community of Andalusia, southern Spain. *HLA*, 57(2) (2001) pp.138-143.
- [758] Ramal LM, López-Nevot MA, Sabio JM, Jáimez L, Paco L, Sánchez J, de Ramón E, Fernández-Nebro A, Ortego N, Ruiz-Cantero A, Rivera F, Martín J, Jiménez-Alonso J; Grupo Lupus Virgen de las Nieves. Systemic lupus erythematosus in southern Spain: a comparative clinical and genetic study between Caucasian and Gypsy patients. *Lupus*, 13(12), (2004), pp.934-940.
- [759] Milanov, I., Topalov, N. and Kmetski, T.S. Prevalence of multiple sclerosis in Gypsies and Bulgarians. *Neuroepidemiology*, 18(4), (1999), pp.218-222.
- [760] Kalman, B., Takacs, K., Gyodi, E., Kramer, J., Füst, G., Tauszik, T., Guseo, A., Kuntar, L., Komoly, S., Nagy, C. and Pálffy, G. Sclerosis multiplex in gypsies. *Acta Neurologica Scandinavica*, 84(3), (1991), pp.181-185.
- [761] López-Larrea, C., Uria, D.F. and Coto, E. HLA antigens in multiple sclerosis of northern Spanish population. *Journal of Neurology, Neurosurgery & Psychiatry*, 53(5), (1990), pp.434-435.
- [762] General Assembly of the World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *The Journal of the American Medical Association*, 81(3), (2014), p.14.
- [763] Wang, C., Mindrinos, M.N., Davis, M.M., Davis, R.W. and Krishnakumar, S., Leland Stanford Junior University. Haplotyping of HLA loci with ultra-deep shotgun sequencing. U.S. Patent 9,562,269 filed January 22, 2014, and issued Feb 7, 2017.

- [764] Wang, C., Mindrinos, M.N., Davis, M.M., Davis, R.W., Krishnakumar, S., Barsakis, K. and Fernandez-Vina, M.A., Leland Stanford Junior University. Software haplotyping of HLA loci. U.S. Patent Application 14/749,491 filed June 25, 2015, and issued December 30, 2015.
- [765] MIA FORA™ NGS FLEX HLA Typing Kit protocol, [http://www.immucor.com/LIFECODES%20Documents/SR-190-00525\\_MIA\\_FORA\\_NGS\\_FLEX\\_HLA\\_Typing\\_Package\\_Insert-RUO-A.pdf](http://www.immucor.com/LIFECODES%20Documents/SR-190-00525_MIA_FORA_NGS_FLEX_HLA_Typing_Package_Insert-RUO-A.pdf), last accessed July 15, 2020.
- [766] MIA FORA™ NGS FLEX HLA Typing Software v3.0, <http://www.immucor.com/global/Products/LIFECODES%20Software/MIA%20FORA%20NGS/SR-190-00523-EN-A%20MIA%20FORA%20FLEX%20Software%20User%20Guide.pdf> last accessed July 15, 2020.
- [767] Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, Correale J, Fazekas F, Filippi M, Freedman MS, Fujihara K, Galetta SL, Hartung HP, Kappos L, Lublin FD, Marrie RA, Miller AE, Miller DH, Montalban X, Mowry EM, Sorensen PS, Tintoré M, Traboulsee AL, Trojano M, Uitdehaag BMJ, Vukusic S, Waubant E, Weinshenker BG, Reingold SC and Cohen JA.. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2), (2018), pp.162-173.
- [768] Standards for Accredited Laboratories, American Society for Histocompatibility and Immunogenetics (ASHI), 2019 Revised Standards approved by the ASHI Board of Directors.
- [769] Standards for Histocompatibility & Immunogenetics testing, European Federation for Immunogenetics (EFI), 2019, approved by the EFI Quality Assurance Committee.
- [770] Gourraud, P.A., Hollenbach, J.A., Barnette, T., Single, R.M. and Mack, S.J. Standard methods for the management of immunogenetic data. In *Immunogenetics*, (2012), (pp. 197-213). *Humana Press*.
- [771] Mack, S.J., Gourraud, P.A., Single, R.M., Thomson, G. and Hollenbach, J.A. Analytical methods for immunogenetic population data. In: *Immunogenetics*, (2012), (pp. 215-244). *Humana Press*.
- [772] Mack, S.J., Tu, B., Yang, R., Masaberg, C., Ng, J. and Hurley, C.K. Human leukocyte antigen–A,-B,-C,-DRB1 allele and haplotype frequencies in Americans originating from southern Europe: Contrasting patterns of population differentiation between Italian and Spanish Americans. *Human Immunology*, 72(2) (2011) pp.144-149.
- [773] Hollenbach, J.A., Mack, S.J., Thomson, G. and Gourraud, P.A. Analytical methods for disease association studies with immunogenetic data. In *Immunogenetics* (2012), (pp. 245-266). *Humana Press*.
- [774] Caniatti, M.C.D.C.L., Borelli, S.D., Guilherme, A.L.F. and Tsuneto, L.T. Association between HLA genes and dust mite sensitivity in a Brazilian population. *Human Immunology*, 78(2), (2017), pp.88-94.
- [775] Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*. (48), (1992), pp.361–372.
- [776] Meyer D, Single R, Mack SJ, Lancaster A, Nelson MP, Erlich HA, Fernandez-Vina M, Thomson G. 13th IHWS anthropology/human genetic diversity joint report. Chapter 4. Single locus polymorphism of classical HLA genes. In: Hansen JA, editor. *Immunobiology of the human MHC. International Histocompatibility Working Group Press*, (2006).
- [777] Dorak, M. T. (2006). Basic population genetics. (<http://www.dorak.info/genetics/popgen.html>), last accessed July 15, 2020.
- [778] Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P. and Thomson, G. PyPop update--a software pipeline for large-scale multilocus population genomics. *Tissue Antigens*, 69, (2007), pp. 192–197.
- [779] Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. 12, (1995), pp.921–927.



- [780] Hedrick PW. Gametic disequilibrium measures: proceed with caution. *Genetics*. 117, (1987), pp.331–334. (*Repeated in Reference [98] and [787]*).
- [781] Lewontin RC. The interaction of selection and linkage. II. Optimum models. *Genetics*. 50, (1964), pp.757–782.
- [782] Lewontin RC. On measures of gametic disequilibrium. *Genetics*. 120, (1988), pp.849–852.
- [783] Klitz W, Stephen JC, Grote M, Carrington M. Discordant patterns of linkage disequilibrium of the peptide transporter loci within the HLA class II region. *Am J Hum Genetics*. 57, (1995), pp. 1436–1444.
- [784] Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet (Der Züchter)*, 38, (1968), pp. 226–231.
- [785] Cramer H. Mathematical methods of statistics. Princeton, NJ: *Princeton University Press*; (1946).
- [786] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: *Erlbaum*; (1988).
- [787] Hedrick PW. Gametic disequilibrium measures: proceed with caution. *Genetics*. 117(2), (1987), pp. 331–341. (*Repeated in Reference [98] and [780]*).
- [788] Thomson, G., Single, R.M.. Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic  $r^2$  measure. *Genetics*, 198 (2014), p. 321
- [789] Single, R.M., Strayer, N., Thomson, G., Paunic, V., Albrecht, M. and Maiers, M.. Asymmetric linkage disequilibrium: tools for assessing multiallelic LD. *Human Immunology*, 77(3), (2016), pp.288-294.
- [790] Ewens WJ. The sampling theory of selectively neutral alleles. *Theor Popul Biol*. 3(1), (1972), pp. 87–112.
- [791] Watterson G. The homozygosity test of neutrality. *Genetics*. 88, (1978), pp. 405–417.
- [792] Slatkin M. An exact test for neutrality based on the Ewens sampling distribution. *Genet Res*, 64, (1994), pp. 71–74.
- [793] Slatkin M. A correction to the exact test based on the Ewens sampling distribution. *Genet Res*. 68, (1996), pp. 259–260.
- [794] Salamon H, Klitz W, Eastel S, Gao X, Erlich HA, Fernandez-Vina M, Trachtenberg EA. Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics*. 152, (1999), pp. 393–400.
- [795] Conover W. Practical nonparametric statistics. New York: *Wiley*; (1980).
- [796] Takezaki, N., Nei, M. and Tamura, K. POPTREEW: Web version of POPTREE for constructing population trees from allele frequency data and computing some other quantities. *Molecular Biology and Evolution*, 31(6) (2014) pp.1622-1624.
- [797] Nei, M., Tajima, F. and Tateno, Y. Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, 19(2) (1983) pp.153-170.
- [798] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4) (1987) pp.406-425.
- [799] Pappas, D.J., Marin, W., Hollenbach, J.A. and Mack, S.J. Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): an integrated case-control analysis pipeline. *Human Immunology*, 77(3), (2016), pp.283-287.

- [800] R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria
- [801] GenBank ® (NIH genetic sequence database)  
<https://www.ncbi.nlm.nih.gov/Sequin/acc.html>, last accessed July 15, 2020.
- [802] Branden, C. I., and Tooze, J.. Introduction to protein structure. *Garland Science*.(2012)
- [803] Wang, J.H., Meijers, R., Xiong, Y., Liu, J.H., Sakihama, T., Zhang, R., Joachimiak, A. and Reinherz, E.L. Crystal structure of the human CD4 N-terminal two-domain fragment complexed to a class II MHC molecule. *Proceedings of the National Academy of Sciences*, 98(19), (2001), pp.10799-10804.
- [804] Li, X.L., Teng, M.K., Reinherz, E.L. and Wang, J.H. Strict major histocompatibility complex molecule class-specific binding by co-receptors enforces MHC-restricted  $\alpha\beta$  TCR recognition during T lineage subset commitment. *Frontiers in Immunology*, 4, (2013), p.383.
- [805] Brandt, D.Y., César, J., Goudet, J. and Meyer, D. The effect of balancing selection on population differentiation: a study with HLA genes. *G3: Genes, Genomes, Genetics*, 8(8), (2018), pp.2805-2815.
- [806] Hollenbach, J.A., Madbouly, A., Gragert, L., Vierra-Green, C., Flesch, S., Spellman, S., Begovich, A., Noreen, H., Trachtenberg, E., Williams, T., Yu, N., Shaw, B., Fleischhauer, K., Fernandez-Vina, M. and Maiers, M. A combined DPA1~ DPB1 amino acid epitope is the primary unit of selection on the HLA-DP heterodimer. *Immunogenetics*, 64(8), (2012), pp.559-569.
- [807] Cano, P. and Fernández-Viña, M.. Two sequence dimorphisms of DPB1 define the immunodominant serologic epitopes of HLA-DP. *Human Immunology*, 70(10), (2009), pp.836-843.
- [808] Hajjej, A., Almawi, W.Y., Arnaiz-Villena, A., Hattab, L. and Hmida, S. The genetic heterogeneity of Arab populations as inferred from HLA genes. *PLOS One*, 13(3), (2018), p.e0192269.
- [809] Alter, I., Gragert, L., Fingerson, S., Maiers, M. and Louzoun, Y. HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. *PLOS Computational Biology*, 13(8), (2017), p.e1005693.
- [810] Kwok, W.W., Kovats, S., Thurtle, P. and Nepom, G.T. HLA-DQ allelic polymorphisms constrain patterns of class II heterodimer formation. *The Journal of Immunology*, 150(6), (1993), pp.2263-2272.
- [811] Macdonald, W.A., Purcell, A.W., Mifsud, N.A., Ely, L.K., Williams, D.S., Chang, L., Gorman, J.J., Clements, C.S., Kjer-Nielsen, L., Koelle, D.M., Burrows, S.R., Tait, B.D., Holdsworth, R., Brooks, A.G., Lovrecz, G.O., Lu, L., Rossjohn, J. and McCluskey, J. A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire, and T cell recognition. *Journal of Experimental Medicine*, 198(5), (2003), pp.679-691.
- [812] Rufer, N., Breur-Vriesendorp, B.S., Tiercy, J.M., Slavcev, A.S., Lardy, N.M., Francis, P., Kressig, R., Speiser, D.E., Helg, C., Chapuis, B. and Gratwohl, A. High-resolution histocompatibility testing of a group of sixteen B44-positive, ABDR serologically matched unrelated donor-recipient pairs: analysis of serologically undisclosed incompatibilities by cellular techniques, isoelectrofocusing, and HLA oligotyping. *Human Immunology*, 38(3), (1993), pp.235-239.
- [813] Fleischhauer, K., Kernan, N.A., O'Reilly, R.J., Dupont, B. and Yang, S.Y. Bone marrow-allograft rejection by T lymphocytes recognizing a single amino acid difference in HLA-B44. *New England Journal of Medicine*, 323(26), (1990), pp.1818-1822.
- [814] Keever, C.A., Leong, N., Cunningham, I., Copelan, E.A., Avalos, B.R., Klein, J., Kapoor, N., Adams, P.W., Orosz, C.G. and Tutschka, P.J. HLA-B44-directed cytotoxic T cells associated with acute graft-versus-host disease following unrelated bone marrow transplantation. *Bone Marrow Transplantation*, 14(1), (1994), pp.137-145.

- [815] Archbold, J.K., Macdonald, W.A., Gras, S., Ely, L.K., Miles, J.J., Bell, M.J., Brennan, R.M., Beddoe, T., Wilce, M.C., Clements, C.S. and Purcell, A.W. Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. *Journal of Experimental Medicine*, 206(1), (2009), pp.209-219.
- [816] Herman, J., Jongeneel, V., Kuznetsov, D. and Coulie, P.G. Differences in the recognition by CTL of peptides presented by the HLA-B\*4402 and the HLA-B\*4403 molecules which differ by a single amino acid. *Tissue Antigens*, 53(2), (1999), pp.111-121.
- [817] Badrinath, S., Saunders, P., Huyton, T., Aufderbeck, S., Hiller, O., Blasczyk, R. and Bade-Doeding, C. Position 156 influences the peptide repertoire and tapasin dependency of human leukocyte antigen B\*44 allotypes. *Haematologica*, 97(1), (2012), pp.98-106.
- [818] Bailey, A., Dalchau, N., Carter, R., Emmott, S., Phillips, A., Werner, J.M. and Elliott, T. Selector function of MHC I molecules is determined by protein plasticity. *Scientific Reports*, 5, (2015), p.14928.
- [819] Tiercy, J.M. Analysis of 250 HLA-B44 genotypes in European Caucasoids: high diversity and preferential ABCDRB1 associations in B\*4402, B\*4403, and B\*4405 haplotypes. *HLA*, 65(5), (2005), pp.429-436.
- [820] Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M. and Fernández-Viña, M.A. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Human Immunology*, 62(9), (2001), pp.1009-1030.
- [821] Williams, F., Meenagh, A., Darke, C., Acosta, A., Daar, A.S., Gorodezky, C., Hammond, M., Nascimento, E. and Middleton, D. Analysis of the distribution of HLA-B alleles in populations from five continents. *Human Immunology*, 62(6), (2001), pp.645-650.
- [822] Vidan-Jeras, B., Breur-Vriesendorp, B., Bohinjec, M., Jeannet, M., Roosnek, E. and Tiercy, J.M.. HLA-B44 allele frequencies and haplotypic associations in three European populations. *European Journal of Immunogenetics*, 24(5), (1997), pp.335-343.
- [823] Vidan-Jeras B, Buhler S, Dubois V, Grubic Z, Ivanova M, Jaatinen T, Ligeiro D, Lokki ML, Papasteriades C, Poli F, Spyropoulou-Vlachou M, Tordai A, Viken MK, Wenda S, Nunes JM, Sanchez-Mazas A, Tiercy JM.. Resolution of HLA-B\*44:02:01G,-DRB1\*14:01:01G and-DQB1\*03:01:01G reveals a high allelic variability among 12 European populations. *Tissue Antigens*, 84(5), (2014), pp.459-464.
- [824] Mack, S.J., Tu, B., Lazaro, A., Yang, R., Lancaster, A.K., Cao, K., Ng, J. and Hurley, C.K. HLA-A,-B,-C, and-DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population. *HLA*, 73(1), (2009), pp.17-32.
- [825] Nerstheimer, S., Tauscher, P., Petek, E. and Schappacher-Tilp, G. HLA-frequencies of Austrian umbilical cord blood samples. *Human Immunology*, 76(11), (2015), pp.863-867.
- [826] Grubic, Z., Burek Kamenaric, M., Mikulic, M., Stingl Jankovic, K., Maskalan, M. and Zunec, R. HLA-A, HLA-B and HLA-DRB1 allele and haplotype diversity among volunteer bone marrow donors from Croatia. *International Journal of Immunogenetics*, 41(3), (2014), pp.211-221.
- [827] Matevosyan, L., Chattopadhyay, S., Madelian, V., Avagyan, S., Nazaretyan, M., Hyussian, A., Vardapetyan, E., Arutunyan, R. and Jordan, F. HLA-A, HLA-B, and HLA-DRB1 allele distribution in a large Armenian population sample. *HLA*, 78(1), (2011), pp.21-30.
- [828] Schmidt, A.H., Baier, D., Solloch, U.V., Stahr, A., Cereb, N., Wassmuth, R., Ehninger, G. and Rutt, C. Estimation of high-resolution HLA-A,-B,-C,-DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. *Human Immunology*, 70(11), (2009), pp.895-902.

- [829] Haimila, K., Peräsaari, J., Linjama, T., Koskela, S., Saarinen, T., Lauronen, J., Auvinen, M.K. and Jaatinen, T. HLA antigen, allele and haplotype frequencies and their use in virtual panel reactive antigen calculations in the Finnish population. *HLA*, 81(1), (2013), pp.35-43.
- [830] Fionnuala, W. and Derek, M. HLA-A,-B,-C, and-DRB1 genotyping of 1000 Northern Irish individuals from Belfast, Northern Ireland in the United Kingdom. *Human Immunology*, 76(6), (2015), pp.395-396.
- [831] Ashouri, E., Norman, P.J., Guethlein, L.A., Han, A.S., Nemat-Gorgani, N., Norberg, S.J., Ghaderi, A. and Parham, P. HLA class I variation in Iranian Lur and Kurd populations: high haplotype and allotype diversity with an abundance of KIR ligands. *HLA*, 88(3), (2016), pp.87-99.
- [832] Sánchez-Velasco, P., Karadsheh, N.S., García-Martín, A., de Alegría, C.R. and Leyva-Cobián, F. Molecular analysis of HLA allelic frequencies and haplotypes in Jordanians and comparison with other related populations. *Human Immunology*, 62(9), (2001), pp.901-909.
- [833] Mattiuz, P.L., Paolo, E., Fossombroni, V., Menicucci, A., Pradella, F., Porfiro, B. and Rombola, G. HLA-B44 subtypes and the chance of finding HLA compatible donor/recipient pairs for bone marrow transplantation: a haplotype study of 303 Italian families. *HLA*, 50(6), (1997), pp.602-609.
- [834] Santos, S., Vicario, J.L., Merino, J.L. and Balas, A. HLA-B44 subtyping in a Spanish population: further evidence of Caucasian population diversity. *HLA*, 49(2), (1997), pp.124-128.
- [835] Gallardo, D., Arostegui, J.I., Rodríguez-Luaces, M., Querol, S., González, J.R., García-López, J. and Grañena, A. HLA-B44 subtyping in the Catalan population using reference strand mediated conformation analysis. Implications for the selection of unrelated bone marrow donors. *HLA*, 56(2), (2000), pp.173-177.
- [836] Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R.J., Plowe, C.V., Kazura, J., Mann, D.L., Szein, M.B., Thomson, G. and Fernández-Viña M.A. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *HLA*, 63(4), (2004), pp.293-325. (Repeated in Reference [496]).
- [837] Spínola, H., Bruges-Armas, J., Middleton, D. and Brehm, A. HLA polymorphisms in Cabo Verde and Guiné-Bissau inferred from sequence-based typing. *Human Immunology*, 66(10), (2005), pp.1082-1092.
- [838] Ellis, J.M., Mack, S.J., Leke, R.F.G., Quakyi, I., Johnson, A.H. and Hurley, C.K. Diversity is demonstrated in class I HLA-A and HLA-B alleles in Cameroon, Africa: description of HLA-A\* 03012,\* 2612,\* 3006 and HLA-B\* 1403,\* 4016,\* 4703. *HLA*, 56(4), (2000), pp.291-302.
- [839] Zhou, X.Y., Zhu, F.M., Li, J.P., Mao, W., Zhang, D.M., Liu, M.L., Hei, A.L., Dai, D.P., Jiang, P., Shan, X.Y. and Zhang, B.W. High-resolution analyses of human leukocyte antigens allele and haplotype frequencies based on 169,995 volunteers from the China Bone Marrow Donor Registry Program. *PLOS One*, 10(9), (2015), p.e0139485.
- [840] Rani, R., Marcos, C., Lazaro, A.M., Zhang, Y. and Stastny, P. Molecular diversity of HLA-A,-B and-C alleles in a North Indian population as determined by PCR-SSOP. *International Journal of Immunogenetics*, 34(3), (2007), pp.201-208.
- [841] In, J.W., Roh, E.Y., Oh, S., Shin, S., Park, K.U. and Song, E.Y. Allele and haplotype frequencies of human leukocyte antigen-A,-B,-C,-DRB1, and-DQB1 from sequence-based DNA typing data in Koreans. *Annals of Laboratory Medicine*, 35(4), (2015), pp.429-435.
- [842] Park, H., Lee, Y.J., Song, E.Y. and Park, M.H. HLA-A, HLA-B and HLA-DRB1 allele and haplotype frequencies of 10918 Koreans from bone marrow donor registry in Korea. *International Journal of Immunogenetics*, 43(5), (2016), pp.287-296.

- [843] Bugawan, T.L., Klitz, W., Alejandrino, M., Ching, J., Panelo, A., Solfelix, C.M., Petrone, A., Buzzetti, R., Pozzilli, P. and Erlich, H.A. The association of specific HLA class I and II alleles with type 1 diabetes among Filipinos. *HLA*, 59(6), (2002), pp.452-469.
- [844] Yang, K.L. and Chen, H.B. (2017). Using high-resolution human leukocyte antigen typing of 11,423 randomized unrelated individuals to determine allelic varieties, deduce probable human leukocyte antigen haplotypes, and observe linkage disequilibria between human leukocyte antigen-B and-C and human leukocyte antigen-DRB1 and-DQB1 alleles in the Taiwanese Chinese population. *Tzu-Chi Medical Journal*, 29(2), p.84.
- [845] Kwok, J., Guo, M., Yang, W., Lee, C.K., Ho, J., Tang, W.H., Chan, Y.S., Middleton, D., Lu, L.W. and Chan, G.C. (2016). HLA-A,-B,-C, and-DRB1 genotyping and haplotype frequencies for a Hong Kong Chinese population of 7595 individuals. *Human Immunology*, 77(12), pp.1111-1112.
- [846] Gourraud, P.A., Pappas, D.J., Baouz, A., Balère, M.L., Garnier, F. and Marry, E. High-resolution HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies from the French Bone Marrow Donor Registry. *Human Immunology*, 76(5), (2015), pp.381-384.
- [847] Sacchi, N., Castagnetta, M., Miotti, V., Garbarino, L. and Gallina, A. High-resolution analysis of the HLA-A,-B,-C and-DRB1 alleles and national and regional haplotype frequencies based on 120,926 volunteers from the Italian Bone Marrow Donor Registry. *HLA*, 94(3), (2019), pp.285-295.
- [848] Fasano, M.E., Rendine, S., Pasi, A., Bontadini, A., Cosentini, E., Carcassi, C., Capittini, C., Cornacchini, G., Espadas de Arias, A., Garbarino, L. and Carella, G. The distribution of KIR-HLA functional blocks is different from North to South of Italy. *HLA*, 83(3), (2014), pp.168-173
- [849] Galgani, A., Mancino, G., Martínez-Labarga, C., Cicconi, R., Mattei, M., Amicosante, M., Bonanno, C.T., Di Sano, C., Gimil, G.S., Salerno, A. and Colizzi, V. HLA-A,-B and-DRB1 allele frequencies in Cyrenaica population (Libya) and genetic relationships with other populations. *Human Immunology*, 74(1), (2013), pp.52-59.
- [850] Hajjej, A., Almawi, W.Y., Hattab, L., El-Gaaied, A. and Hmida, S. The investigation of the origin of Southern Tunisians using HLA genes. *Journal of Human Genetics*, 62(3), (2017), p.419.
- [851] Hajjej, A., Almawi, W.Y., Hattab, L., El-Gaaied, A. and Hmida, S. HLA class I and class II alleles and haplotypes confirm the Berber Origin of the Present Day Tunisian Population. *PLOS One*, 10(8), (2015), p.e0136909.
- [852] Arrieta-Bolanos, E., Madrigal, J.A. and Shaw, B.E. Human Leukocyte Antigen profiles of Latin American populations: differential admixture and its potential impact on hematopoietic stem cell transplantation. *Bone Marrow Research*, (2012).
- [853] Catelli ML, Alvarez-Iglesias V, Gómez-Carballa A, Mosquera-Miguel A, Romanini C, Borosky A, Amigo J, Carracedo A, Vullo C and Salas A The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome. *BMC Genetics*. 12(1), (2011), pp. 77.
- [854] Arnaiz-Villena A, Benmamar D, Alvarez M, Diaz-Campos N, Varela P, Gomez-Casado E, Martinez-Laso J. HLA allele and haplotype frequencies in Algerians: Relatedness to Spaniards and Basques. *Human Immunology*. 43(4), (1995), pp. 259-68.
- [855] Ayed K, Ayed-Jendoubi S, Sfar I, Labonne MP and Gebuhrer L. HLA class-I and HLA class-II phenotypic, gene and haplotypic frequencies in Tunisians by using molecular typing data. *Tissue Antigens*, 64(4), (2004), pp. 520-532.
- [856] Hajjej, A., Hajjej, G., Almawi, W.Y., Kaabi, H., El-Gaaied, A. and Hmida, S. HLA class I and class II polymorphism in a population from south-eastern Tunisia (Gabes Area). *International Journal of Immunogenetics*, 38(3), (2011), pp.191-199.

- [857] Perkins, K. A history of modern Tunisia. *Cambridge University Press*, (2014).
- [858] Cherni L, Pakstis AJ, Boussetta S, Elkamel S, Frigi S, Khodjet-El-Khil H, Barton A, Haigh E, Speed WC, Ben Ammar Elgaaied A, Kidd JR. Genetic variation in Tunisia in the context of human diversity worldwide. *American Journal of Physical Anthropology*. 161(1), (2016), pp. 62-71.
- [859] Elloumi-Zghal H, Bouhamed HC. Genetics and genomic medicine in Tunisia. *Molecular Genetics & Genomic Medicine*. 6(2), (2018),134.
- [860] El Moncer W, Esteban E, Bahri R, Gayà-Vidal M, Carreras-Torres R, Athanasiadis G, Moral P, Chaabani H. Mixed origin of the current Tunisian population from the analysis of Alu and Alu/STR compound systems. *Journal of Human Genetics*. 55(12), (2010), pp. 827-33.
- [861] Arnaiz-Villena, A., Martínez-Laso, J., Gómez-Casado, E., Díaz-Campos, N., Santos, P., Martinho, A. and Breda-Coimbra, H. Relatedness among Basques, Portuguese, Spaniards, and Algerians studied by HLA allelic frequencies and haplotypes. *Immunogenetics*, 47(1), (1997), pp.37-43.
- [862] Adhikari, K., Chacón-Duque, JC, Mendoza-Revilla, J., Fuentes-Guajardo, M. and Ruiz- Linares, A. The genetic diversity of the Americas. *Annu. Rev. Genomics Hum. Genet.* 18, (2017), pp 277–296.
- [863] Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D. and Mountain, J.L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States, *Am. J. Hum. Genet.* 96 (2015) pp. 37–53.
- [864] Kruskall, M.S., Eynon, E.E., Awdeh, Z., Alper, C.A. and Yunis, E.J. Identification of HLA-B44 subtypes associated with extended MHC haplotypes. *Immunogenetics*, 26(4-5), (1987), pp.216-219.
- [865] Vilches, C., Pablo, R., Herrero, M.J., Moreno, M.E. and Kreisler, M. HLA-B73: an atypical HLA-B molecule carrying a Bw6-epitope motif variant and a B pocket identical to HLA-B27. *Immunogenetics*, 40(2), (1994), pp.166-166.
- [866] Yasukochi, Y. and Ohashi, J.. Elucidating the origin of HLA-B\*73 allelic lineage: Did modern humans benefit by archaic introgression?. *Immunogenetics*, 69(1), (2017), pp.63-67.
- [867] Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S.G., Maiers, M., Guethlein, L.A., Tavoularis, S., Little, A.M., Green, R.E., Norman, P.J. and Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, 334(6052), (2011), pp.89-94.
- [868] Kawashima, M., Ohashi, J., Nishida, N. and Tokunaga, K. Evolutionary Analysis of Classical HLA Class I and II Genes Suggests That Recent Positive Selection Acted on DPB1\*04:01 in Japanese Population. *PLOS One*, 7(10), (2012), p.e46806.
- [869] Smith, D.M., Baker, J.E., Gardner, W.B., Martens, G.W. and Agura, E.D. HLA class I null alleles and new alleles affect unrelated bone marrow donor searches. *HLA*, 66(2) (2005) pp.93-98.
- [870] Poli, F, Scalamogna, M and Sirchia, G. HLA Null Alleles: Implications in Stem-Cell Transplantation. *Cytotherapy*, 1(5), (1999), pp. 365-366
- [871] Henry, J., Kempenich, J., Bolon, Y.T., Hurley, C.K., Roers, B., Malmberg, C. and Jones, L. Frequency of Class I common or well documented null alleles in National Marrow Donor Program high resolution typing programs. *Human Immunology*, 76 (Supplement), (2015), p.133.
- [872] Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.* (215), (1990), pp.403–410.

- [873] Wang, Z.C., Smith, A.G., Yunis, E.J., Selvakumar, A., Ferrone, S., McKinney, S., Lee, J.H., Fernandez-Vina, M. and Hansen, J.A. Molecular characterization of the HLA-Cw\*0409N allele. *Human Immunology*, 63(4), (2002), pp.295-300.
- [874] Sutton, V.R., Kienzle, B.K. and Knowles, R.W. An altered splice site is found in the DRB4 gene that is not expressed in HLA-DR7, Dw11 individuals. *Immunogenetics*, 29(5), (1989), pp.317-322.
- [875] Sutton, V.R. and Knowles, R.W. An aberrant DRB4 null gene transcript is found that could encode a novel HLA-DR chain. *Immunogenetics*, 31(2), (1990), pp.112-117.
- [876] Naruse, T.K., Ando, R., Nose, Y., Kagiya, M., Ando, H., Kawata, H., Nabeya, N., Isshiki, G. and Inoko, H. HLA-DRB4 genotyping by PCR-RFLP: diversity in the associations between HLA-DRB4 and DRB1 alleles. *Tissue Antigens*, 49(2), (1997), pp.152-159.
- [877] Pinto, C., Smith, A.G., Larsen, C.E., Fernandez-Vina, M., Husain, Z., Clavijo, O.P., Wang, Z.C., Nisperos, B., Hansen, J.A., Alper, C.A. and Yunis, E.J. HLA-Cw\* 0409N is associated with HLA-A\*2301 and HLA-B\*4403-carrying haplotypes. *Human Immunology*, 65(2), (2004), pp.181-187.
- [878] Downing J, Street J, Hammond L and Darke C. Identification of HLA-Cw\*0409N and its distribution in blood donors resident in Wales. *Eur J Immunogenet* 29, (2002), pp. 351–370.
- [879] Bors, A., Inotai, D., Andrikovics, H., Benkő, S., Boros-Major, A., Illés, Z., Szilvási, A., Gelle-Hossó, A., Rajczy, K. and Tordai, A. Low occurrence of the HLA-C\*04:09N allele in a large Hungarian cohort. *HLA*, 86(1), (2015), pp.32-35.
- [880] Moore, B., Guerrero, E., Carmazzi, Y, Cao, K. OR19 FREQUENCY OF HLA-B\*44:03-C\*04:09N BEARING HAPLOTYPES AND PHENOTYPES IN LEUKEMIA PATIENTS. *Human Immunology*, 75(Supplement) (2014) pp.16.
- [881] Grubic, Z., Maskalan, M., Radmanic, L., Stingl Jankovic, K., Burek Kamenaric, M. and Zunec, R. The distribution of the DRB4\*01:03:01:02N null allele in HLA-DRB1~ DQB1 haplotypes in the Croatian population. *HLA*, 91(1), (2018), pp.23-28.
- [882] Reville, P., Thomas, D., Kawczak, P., Zhang, A., McMichael, J. and Askar, M. 163-P: Description of Allele Level HLA-DRB4 containing DRB1-DRB4-DQA1-DQB1 Haplotypes. *Human Immunology*, 73(Supplement), (2012), p.150.
- [883] Yarzabek, B., Zaitouna, A.J., Olson, E., Silva, G.N., Geng, J., Geretz, A., Thomas, R., Krishnakumar, S., Ramon, D.S., Raghavan, M. Variations in HLA-B cell surface expression, half-life and extracellular antigen receptivity. *Elife*, 7 (2018).
- [884] P. Chappell, K. Meziane el, M. Harrison, L. Magiera, C. Hermann, L. Mears, A.G. Wrobel, C. Durant, L.L. Nielsen, S. Buus, N. Ternette, W. Mwangi, C. Butter, V. Nair, T. Ahyee, R. Duggleby, A. Madrigal, P. Roversi, S.M. Lea, J. Kaufman. Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife*, 4 (2015), Article e05345
- [885] Balas, A., Pacho, A., Arrieta, A., García-Sánchez, F. and Vicario, J.L. Four new HLA class I alleles in Spaniards, HLA-A\*32:01:23, HLA-B\*18:01:24, HLA-B\*18:72:02 and HLA-C\*12:166. *HLA*, 88(1-2), (2016), pp.42-43.
- [886] Joannes, M.O.R.F., Voorter, C.E.M., Agis, F. and Tilanus, M.G.J. Full-length sequence of a novel HLA-B\*15:220 allele identified in an individual from Guadeloupe. *HLA*, 79(1), (2012), pp.75-76.
- [887] Schell, A., Leisenbach, R., Coman, C., Parissiadis, A. and Tourne, S. A new HLA-B\* 15 allele, B\*15:220, found in three individuals sharing the HLA-A\*66:01, HLA-C\*12:03 and HLA-DRB1\*07:01 alleles. *HLA*, 78(4), (2011), pp.287-288.

- [888] Begovich, A.B., Moonsamy, P.V., Mack, S.J., Barcellos, L.F., Steiner, L.L., Grams, S., Suraj-Baker, V., Hollenbach, J., Trachtenberg, E., Louie, L., Zimmerman, P., Hill, A.V., Stoneking, M., Sasazuki, T., Kononkov, V.I., Sartakova, M.L., Titanji, V.P., Rickards, O. and Klitz, W. Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. *HLA*, 57(5), (2001), pp.424-439.
- [889] Du, Z., Patel, J., Braun, C. and Norton, B. HLA-DPA1 and HLA-DPB1 Frequencies in the US Populations. *American Journal of Transplantation*, 17, (2017), pp. 528-528.
- [890] Rivas, F., Zhong, Y., Olivares, N., Cerda-Flores, R.M. and Chakraborty, R. Worldwide genetic diversity at the HLA-DQA1 locus. *American Journal of Human Biology*, 9(6), (1997), pp.735-749.
- [891] Huckenbeck, W., Alt, K.W., Zens, V., Vach, W., Stancu, V., Bonte, W. Population data on HLA-DQa system: Genotype and gene frequencies in Germany and a worldwide comparison. *Homo*, 46, (1996), pp. 239–52.
- [892] Bettens, F., Nicoloso de Faveri, G. and Tiercy, J.M.. HLA-B51 and haplotypic diversity of B-Cw associations: implications for matching in unrelated hematopoietic stem cell transplantation. *Tissue Antigens*, 73(4), (2009), pp.316-325.
- [893] Sinnock P. The Wahlund effect for the two-locus model. *American Naturalist*, (1975), pp.565–570.
- [894] Tarver, H.M. and Slape, E. eds. The Spanish Empire: A Historical Encyclopedia [2 volumes]. *ABC-CLIO*, (2016).
- [895] Harrison, J. An economic history of modern Spain. *Manchester University Press*. (1978).
- [896] Tiercy, J.M.. Unrelated hematopoietic stem cell donor matching probability and search algorithm. *Bone Marrow Research*, (2012).
- [897] Pedron B, Guérin-El Khourouj V, Dalle JH, Ouachée-Chardin M, Yakouben K, Corroyez F, Auvrignon A, Petit A, Landman-Parker J, Leverger G, Baruchel A and Sterkers G.. Contribution of HLA-A/B/C/DRB1/DQB1 common haplotypes to donor search outcome in unrelated hematopoietic stem cell transplantation. *Biology of Blood and Marrow Transplantation* 17(11), (2011), pp. 1612-1618.
- [898] Johansen, K.A., Schneider, J.F., McCaffree, M.A., Woods, G.L. and Council on Science and Public Health, American Medical Association. Efforts of the United States' National Marrow Donor Program and Registry to improve utilization and representation of minority donors. *Transfusion Medicine*, 18(4), (2008), pp.250-259.
- [899] Barker JN, Boughan K, Dahi PB, Devlin SM, Maloy MA, Naputo K, Mazis CM, Davis E, Nhaissi M, Wells D, Cooper C, Ponce DM, Kernan N, Scaradavou A, Giralt SA, Papadopoulos EB, Politikos I. Racial disparities in access to HLA-matched unrelated donor transplants: a prospective 1312-patient analysis. *Blood Advances*, 3(7), (2019), pp.939-944.
- [900] Robinson, M.A. Linkage Disequilibrium. In: Encyclopedia of Immunology (2nd Edition), *Academic Press*, (1998), pp. 1586-1588.
- [901] Crivello, P., Lauterbach, N., Zito, L., Sizzano, F., Toffalori, C., Marcon, J., Curci, L., Mulder, A., Wieten, L., Zino, E., Voorter, C.E., Tilanus, M.G.J. and Fleischhauer, K. Effects of transmembrane region variability on cell surface expression and allorecognition of HLA-DP3. *Human Immunology*, 74(8), (2013), pp.970-977.
- [902] King, G. and Dixon, A.M. Evidence for role of transmembrane helix–helix interactions in the assembly of the Class II major histocompatibility complex. *Molecular BioSystems*, 6(9), (2010), pp.1650-1661.



- [903] Dixon, A.M., Drake, L., Hughes, K.T., Sargent, E., Hunt, D., Harton, J.A. and Drake, J.R.. Differential transmembrane domain GXXXG motif pairing impacts major histocompatibility complex (MHC) class II structure. *Journal of Biological Chemistry*, 289(17), (2014), pp.11695-11703.
- [904] Dixon, A.M. and Roy, S. Role of membrane environment and membrane-spanning protein regions in assembly and function of the class II major histocompatibility complex. *Human Immunology*, 80(1), (2019), pp.5-14.
- [905] Harton, J., Jin, L., Hahn, A. and Drake, J. Immunological functions of the membrane proximal region of MHC class II molecules. *F1000Research*, 5, (2016).
- [906] Bodis, G., Toth, V. and Schwarting, A.. Role of human leukocyte antigens (HLA) in autoimmune diseases. *Rheumatology and Therapy*, 5(1), (2018), pp.5-20.
- [907] Ollila, H.M., Fernandez-Vina, M. and Mignot, E. HLA-DQ allele competition in narcolepsy: a comment on Tafti et al. DQB1 locus alone explains most of the risk and protection in narcolepsy with cataplexy in Europe. *Sleep*, 38(1), (2015), pp.147-151.
- [908] Witter, K., Mautner, J., Albert, T., Zahn, R. and Kauke, T. HLA-DQB1\*0319, a novel HLA-DQB1 allele, shows strong haplotype association to HLA-DRB1\*1102. *Tissue Antigens*, 70(1), (2007), pp.73-75.
- [909] Balgansuren, G., Lopes, K., Clark, A., Peel, L., Young, C., Deoliveira, A., Wegner, W. and Chen, D. P101: DQB1\*03:19 Association with DRB1 and DQA1 in North Carolina population. *Human Immunology*, 75(Supplement), (2014), p.121.
- [910] Al-Otaibi, A., Shawhatti, A., Zahrani, S., Al-Attas, R. and Liacini, A. Frequency and haplotype association of HLA-DQB1\*03:19 in Saudi Arabian population. *Human Immunology*, 76(Supplement), (2015), p.75.
- [911] Hurley CK, Fernandez-Vina M, Hildebrand WH, Noreen HJ, Trachtenberg E, Williams TM, Baxter-Lowe LA, Begovich AB, Petersdorf E, Selvakumar A, Stastny P, Hegland J, Hartzman RJ, Carston M, Gandham S, Kollman C, Nelson G, Spellman S and Setterholm M. A high degree of HLA disparity arises from limited allelic diversity: analysis of 1775 unrelated bone marrow transplant donor-recipient pairs. *Human Immunology*, 68(1), (2007), pp.30-40.
- [912] Biró P, Haase-Kromwijk B, Andersson T, Ásgeirsson EI, Baltessová T, Boletis I, Bolotinha C, Bond G, Böhmig G, Burnapp L, Cechlárová K, Di Ciaccio P, Fronck J, Hadaya K, Hemke A, Jacquelinet C, Johnson R, Kieszek R, Kuypers DR, Leishman R, Macher MA, Manlove D, Menoudakou G, Salonen M, Smeulders B, Sparacino V, Spieksma FCR, Valentín MO, Wilson N, van der Klundert J; ENCKEP COST Action. Building Kidney Exchange Programmes in Europe-An Overview of Exchange Practice and Activities. *Transplantation*. 103(7), (2019), pp.1514-1522.
- [913] Verity, D., Marr, J., Ohno, S., Wallace, G. and Stanford, M. Behçet's disease, the Silk Road and HLA-B51: historical and geographical perspectives. *Tissue Antigens*, 54, (1999), pp. 213-220.
- [914] Montes-Cano MA, Conde-Jaldón M, García-Lozano JR, Ortiz-Fernández L, Ortego-Centeno N, Castillo-Palma MJ, Espinosa G, Graña-Gil G, González-Gay MA, Barnosi-Marín AC, Solans R, Fanlo P, Camps T, Castañeda S, Sánchez-Bursón J, Núñez-Roldán A, Martín J, González-Escribano MF. HLA and non-HLA genes in Behçet's disease: a multicentric study in the Spanish population. *Arthritis Res Ther*. 15 (5), (2013), R145.
- [915] Burillo-Sanz, S., Montes-Cano, M.A., García-Lozano, J.R., Olivas-Martínez, I., Ortego-Centeno, N., García-Hernández, F.J., Espinosa, G., Graña-Gil, G., Sánchez-Bursón, J., Juliá, M.R., Solans, R., Blanco, R., Barnosi-Marín, A.C., Gómez de la Torre, R., Fanlo, P., Rodríguez-Carballeira, M, Rodríguez-Rodríguez, L., Camps, T., Castañeda, S., Alegre-Sancho, J.J., Martin, J. and González-Escribano, M.F. Behçet's disease and

genetic interactions between HLA-B\*51 and variants in genes of autoinflammatory syndromes. *Scientific Reports*, 9(1), (2019), pp.1-8.

[916] Abdennaji Guenounou, B., Loueslati, B.Y., Buhler, S., Hmida, S., Ennafaa, H., Khodjet-Elkhal, H., Moojat, N., Dridi, A., Boukef, K., Ben Ammar Elgaaied, A. and Sanchez-Mazas, A. HLA class II genetic diversity in southern Tunisia and the Mediterranean area. *International Journal of Immunogenetics*, 33, (2006), pp. 93-103.

[917] Tokić, S., Žižkova, V., Štefanić, M., Glavaš-Obrovac, L., Marczy, S., Samardžija, M., Sikorova, K. and Petrek, M. HLA-A,-B,-C,-DRB1,-DQA1, and-DQB1 allele and haplotype frequencies defined by next generation sequencing in a population of East Croatia blood donors. *Scientific Reports*, 10(1), (2020), pp.1-13.

[918] Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, Khusnutdinova EK, Balanovsky O, Semino O, Pereira L, Comas D, Gurwitz D, Bonne-Tamir B, Parfitt T, Hammer MF, Skorecki K, Villems R.. The genome-wide structure of the Jewish people. *Nature*. 466, (2010), pp. 238–242.

[919] Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E and Ostrer H.. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern ancestry. *Am J Hum Genet*, 86(6), (2010), pp.850-859.

[920] Cerná, M., Fernandez-Viña, M., Ivásková, E. and Stastny, P. Comparison of HLA class II alleles in Gypsy and Czech populations by DNA typing with oligonucleotide probes. *Tissue Antigens*, 39(3), (1992), pp.111-116.

[921] Inotai D, Szilvasi A, Benko S, Boros-Major A, Illes Z, Bors A, Kiss KP, Rajczy K, Gelle-Hossó A, Buhler S, Nunes JM, Sanchez-Mazas A, Tordai A. HLA genetic diversity in Hungarians and Hungarian Gypsies: complementary differentiation patterns and demographic signals revealed by HLA-A,-B and-DRB1 in Central Europe. *Tissue Antigens*, 86(2), (2015), pp.115-121.

[922] Fernández O, Fernández V, Martínez-Cabrera V, Mayorga C, Alonso A, León A, Arnal C, Hens M, Luque G, de Ramón E, Caballero A, Leyva L. Multiple sclerosis in Gypsies from southern Spain: prevalence, mitochondrial DNA haplogroups and HLA class II association. *Tissue Antigens*, 71(5), (2008), pp.426-433.

[923] de Pablo, R., Vilches, C., Moreno, M.E., Rementería, M.C., Solís, R. and Kreisler, M.. Distribution of HLA antigens in Spanish Gypsies: a comparative study. *Tissue Antigens*, 40(4), (1992), pp.187-196.

[924] Ogawa A, Tokunaga K, Lin L, Kashiwase K, Tanaka H, Herrero MJ, Vilches C, Park MH, Jia GJ, Chimge NO, Sideltseva EW, Ishikawa Y, Akaza T, Tadokoro K, Juji T.. Diversity of HLA-B61 alleles and haplotypes in East Asians and Spanish Gypsies. *Tissue Antigens*, 51(4), (1998), pp.356-366.

[925] Vilches, C., de Pablo, R., Herrero, M.J., Moreno, M.E. and Kreisler, M.. Molecular cloning and polymerase chain reaction-sequence-specific oligonucleotide detection of the allele encoding the novel allospecificity HLA-Cw6. 2 (Cw\* 1502) in Spanish gypsies. *Human Immunology*, 37(4), (1993), pp.259-263.

[926] Aboukaoud, M., Israel, S., Brautbar, C. and Eyal, S.. Genetic Basis of Delayed Hypersensitivity Reactions to Drugs in Jewish and Arab Populations. *Pharmaceutical Research*, 35(11), (2018), p.211.

[927] Kaback, M., Lopatequi, J., Portuges, A.R., Quindipan, C., Pariani, M., Salimpour-Davidov, N. and Rimoin, D.L. Genetic screening in the Persian Jewish community: A pilot study. *Genetics in Medicine*, 12(10), (2010), pp.628-633.

[928] Loewenthal R, Slomov Y, Gonzalez-Escribano MF, Goldberg I, Korostishevsky M, Brenner S, Nunez-Roldan A, Conejo-Mir JS, and Gazit E. Common ancestral origin of pemphigus vulgaris in Jews and Spaniards: a study using microsatellite markers. *Tissue Antigens*, 63(4), (2004), pp.326-334.

- [929] Goodman RM, Genetic disorders among the Jewish People. Baltimore, MD: *Johns Hopkins University Press*, (1979).
- [930] Boletín Oficial del Estado, Gobierno de España, 25 de Junio de 2015. <http://www.boe.es/boe/dias/2015/06/25/pdfs/BOE-A-2015-7045.pdf>, (last accessed July 2020).
- [931] Boletín Oficial del Estado, Gobierno de España, 12 de Noviembre de 1992. <https://www.boe.es/boe/dias/1992/11/12/pdfs/A38214-38217.pdf>, (last accessed July 2020).
- [932] Boletín Oficial de las Cortes Generales. Congreso de los Diputados, Gobierno de España, 22 de Junio de 2017. [http://www.congreso.es/public\\_oficiales/L12/CONG/BOCG/D/BOCG-12-D-179.PDF](http://www.congreso.es/public_oficiales/L12/CONG/BOCG/D/BOCG-12-D-179.PDF), (last accessed July 2020).
- [933] Gaig, P., Ferrer, M., Muñoz-Lejarazu, D., Lleó, R., García-Abujeta, J.L., Caballero, T., Rodríguez, A., Echechipia, S., Martínez-Cocera, C., Domínguez, F.J., Gonzalo, M.A. and Olona, M. Prevalencia de alergia en la población adulta española. *Alergología e Inmunología Clínica*, 19, (2004), 68-74.
- [934] Arnaiz-Villena, A., Parga-Lozano, C., Moreno, E., Areces, C., Rey, D. and Gomez-Prieto, P. The origin of Amerindians and the peopling of the Americas according to HLA genes: admixture with Asian and Pacific people. *Current Genomics*, 11(2), (2010), pp.103-114.
- [935] Arnaiz-Villena, A., Palacio-Grüber, J., Juárez, I., Lopez-Nares, A., Nieto, J., Campos, C. and Martín-Villa, J.M. HLA in Uros from Peru Titikaka Lake: Tiwanaku, Easter and Pacific Islanders. *Human Immunology*, 80(2), (2019), pp.91-92.
- [936] Arnaiz-Villena, A., Areces, C., Enríquez-de-Salamanca, M., Abd-El-Fatah-Khalil, S., Marco, J., Muñoz, E., Fernández-Honrado, M., Villa, M.M. and Rey, D. Pacific Islanders and Amerindian relatedness according to HLA autosomal genes. *International Journal of Modern Anthropology*, 1(7), (2014), pp.44-67.
- [937] Thorsby, E. The Polynesian gene pool: an early contribution by Amerindians to Easter Island. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1590), (2012), pp.812-819.
- [938] Cerna, M., Falco, M., Friedman, H., Raimondi, E., Maccagno, A., Fernandez-Viña, M. and Stastny, P. Differences in HLA class II alleles of isolated South American Indian populations from Brazil and Argentina. *Human Immunology*, 37(4), (1993), pp.213-220.
- [939] Galarza JM, Barquera R, Álvarez AMT, Hernández Zaragoza DI, Sevilla GP, Tamayo A, Pérez M, Dávila D, Birnberg L, Alonzo VA, Krause J and Grijalva M. Genetic diversity of the HLA system in human populations from the Sierra (Andean), Oriente (Amazonian) and Costa (Coastal) regions of Ecuador. *Human Immunology*, 79(9), (2018), pp.639-650.
- [940] Ramon, D., Scott, I., Cox, S.T., Pesa, S., Vullo, C., Little, A.M. and Madrigal, J.A. HLA-A\* 6817, identified in the Kolla Amerindians of North-West Argentina possesses a novel nucleotide substitution. *Tissue Antigens*, 55(5), (2000), pp.453-454.
- [941] Little, A.M., Scott, I., Pesa, S., Marsh, S.G., Argüello, R., Cox, S.T., Ramon, D., Vullo, C. and Madrigal, J.A. HLA class I diversity in Kolla Amerindians. *Human Immunology*, 62(2), (2001), pp.170-179.
- [942] Shyr DC, Zhang BM, Saini G, Madani ND, Schultz LM, Patel S, Kristovich K, Fernandez-Vina M, Bertaina A. HLA-haplotype loss after TCR $\alpha\beta$ /CD19-depleted haploidentical HSCT. *Bone Marrow Transplant*, (2020).
- [943] Leen G, Stein JE, Robinson J, Maldonado Torres H, Marsh SGE. The HLA diversity of the Anthony Nolan register. *HLA*, (2020).

- [944] Prat, E., Tomaru, U., Sabater, L., Park, D.M., Granger, R., Kruse, N., Ohayon, J.M., Bettinotti, M.P. and Martin, R. HLA-DRB5\*0101 and-DRB1\*1501 expression in the multiple sclerosis-associated HLA-DR15 haplotype. *Journal of Neuroimmunology*, 167(1-2), (2005), pp.108-119.
- [945] Field J, Browning SR, Johnson LJ, Danoy P, Varney MD, Tait BD, Gandhi KS, Charlesworth JC, Heard RN; Australia and New Zealand Multiple Sclerosis Genetics Consortium, Stewart GJ, Kilpatrick TJ, Foote SJ, Bahlo M, Butzkueven H, Wiley J, Booth DR, Taylor BV, Brown MA, Rubio JP, Stankovich J. A polymorphism in the HLA-DPB1 gene is associated with susceptibility to multiple sclerosis. *PLoS One*. 5(10), (2010), p.e13454.
- [946] Didonna, A., Damotte, V., Shams, H., Matsunaga, A., Caillier, S.J., Dandekar, R., Misra, M.K., Mofrad, M.R., Oksenberg, J.R. and Hollenbach, J.A. A splice acceptor variant in HLA-DRA affects the conformation and cellular localization of the class II DR alpha-chain. *Immunology*, (2020).
- [947] Kaur, G., Trowsdale, J. and Fugger, L. Natural killer cells and their receptors in multiple sclerosis. *Brain*, 136(9), (2013), pp.2657-2676.
- [948] Ødum, N., Hyldig-Nielsen, J.J., Morling, N., Sandberg-Wollheim, M., Platz, P. and Svejgaard, A. HLA-DP antigens are involved in the susceptibility to multiple sclerosis. *Tissue Antigens*, 31(5), (1988) ,pp.235-237.
- [949] Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9), (2010), pp.1564-1573.
- [950] Sullivan S, Fairchild PJ, Marsh SG, Müller C, Turner ML, Song J, Turner D. Haplobanking induced Pluripotent Stem Cells for clinical use. *Stem Cell Research*, 49, (2020), 102035.
- [951] Carey, B.S., Poulton, K.V. and Poles, A. Factors affecting HLA expression: A review. *International Journal of Immunogenetics*, 46(5), (2019), pp.307-320.
- [952] Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, Thomsen AR, Cardon LR, Bell JI, Fugger L. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, 443(7111), (2006), pp. 574-577.
- [953] Zhang, J., Zhan, W., Yang, B., Tian, A., Chen, L., Liao, Y., Wu, Y., Cai, B. and Wang, L. Genetic polymorphisms of rs3077 and rs9277535 in HLA-DP associated with systemic lupus erythematosus in a Chinese population. *Scientific Reports*, 7(1), (2017), pp.1-8.
- [954] Yu, M., Kinkel, R.P., Weinstock-Guttman, B., Cook, D.J. and Tuohy, V.K. HLA-DP: a class II restriction molecule involved in epitope spreading during the development of multiple sclerosis. *Human Immunology*, 59(1), (1998), pp.15-24.



***APPENDIXES***



**APPENDIX 1:****RÉSUMÉ/CURRICULUM VITAE, GONZALO MONTERO MARTIN****EDUCATION**

▪ **10/2014 - Present:** Programa de Doctorado en Investigación Biomédica R.D. 99/2011 (**Biomedical Research Doctorate Ph.D. Program R.D. 99/2011**), Facultad de Medicina (**School of Medicine**), Universidad Complutense de Madrid (UCM) (**Complutense University of Madrid**), Madrid (Spain).

Doctoral Thesis Title: *“Investigation of the distribution of HLA alleles in healthy and diseased populations by the application of novel sequencing methodologies”*. This Doctoral Thesis is included in the **research field #9** called **“Immunology and Immunopathology”** of this Biomedical Research Doctorate program.

Co-Directors of this Doctoral Thesis:

**Marcelo Fernández Viña**, Ph.D., D (ABHI), **Stanford Blood Center, Stanford University-School of Medicine**, Palo Alto, CA (USA). (*Email contact:* marcelof@stanford.edu)

**Jorge Martínez Laso**, Ph.D., **Immunogenetics Unit, Microbiology National Center-Instituto de Salud Carlos III**, Madrid (Spain) and **Immunology Unit, Complejo Hospitalario Universitario Insular Materno-Infantil**, Las Palmas de Gran Canaria (Spain). (*Email contact:* jmlaso12@gmail.com)

▪ **09/2013 - 07/2014:** Master en Investigación en Inmunología (**Immunology Research Master of Science Degree**), Facultad de Medicina (**School of Medicine**), Universidad Complutense de Madrid (UCM) (**Complutense University of Madrid**), Madrid (Spain). Spanish Universities grade system: 2.55 out of 4.00; and 9.09 out of 10.00.

▪ **09/2007 - 09/2013:** Licenciatura en Bioquímica (**Biochemistry Bachelor of Science Degree**), Facultad de Ciencias Químicas (**School of Chemistry**), Universidad Complutense de Madrid (UCM) (**Complutense University of Madrid**), Madrid (Spain). Spanish Universities grade system: 2.03 out of 4.00; and 7.48 out of 10.00.

**ADDITIONAL ACADEMIC-PROFESSIONAL ACHIEVEMENTS**

▪ **09/2017-Present:** **American Board of Histocompatibility and Immunogenetics (ABHI) Certified Histocompatibility Technologist license (CHT # 1955) (Total Score 99 out of 150; 72 scaled-score units)** (1.General Laboratory Skills: 6/12; 2.Histocompatibility and Immunogenetics Testing 35/58; 3.Test Interpretation and Reporting 30/45; 4.Histocompatibility and Immunogenetics Testing Principles and Theory 15/16, 5.Quality Systems 12/15; 6.Supervisory Functions and Management 01/04). This certificate permits to work as a Clinical Histocompatibility Technologist (CHT) in all areas of Clinical Histocompatibility Laboratories anywhere in United States of America as long as other additional specific state requirements are met.

▪ **06/2016-Present:** **California Clinical Histocompatibility Scientist Trainee license, Laboratory Field Services, California Department of Public Health (TRL01005232)**. This license permits to be trained in all areas of Clinical Histocompatibility in Department-approved training programs.



- **04/2014: GRE General Revised Test (ETS)**. Test Score: Verbal Reasoning (146) (28%), Quantitative Reasoning (149) (37%), Analytical Writing (3.0) (14%).
- **02/2013: iBT TOEFL Test (ETS)**. Test Score: 94 out of 120.

### PROFESSIONAL TRAINING

- **09/2020: XXXIX and XXXV Cursos Teóricos de Actualización en Inmunogenética y Genética Molecular 2020**. The period of time of this course was 38.10 hours in total: 38.10 hours of theory classes. This virtual course took place at the **Fundacion Comparte Vida, A.C., Mexico D.F. (Mexico)**. Course Organizer: **Fundacion Comparte Vida, A.C.** This course is ABHI (American Board of Histocompatibility and Immunogenetics) certified.
- **08/2020: ASHI Virtual Regional Education Workshop 2020**. The period of time of this course was 10.5 hours in total: 10.5 hours of theory classes. This virtual course took place online. Course Organizer: **American Society for Histocompatibility and Immunogenetics (ASHI)**. This course is ABHI (American Board of Histocompatibility and Immunogenetics) certified.
- **03/2017: Basic Histocompatibility Course 2017**. The period of time of this course was 15.5 hours in total: 15.5 hours of theory classes. This course took place at the **Orlando Marriot World Center Hotel**. Course Organizer: **AFDT (American Foundation for Donation and Transplantation)**. This course is ABHI (American Board of Histocompatibility and Immunogenetics) certified.
- **03/2015: IV Curso de Histocompatibilidad: del Laboratorio a la Clínica 2015 (4<sup>th</sup> Course of Histocompatibility: from the Lab to the Clinic 2015)**. The period of time of this course was 30 hours in total: 24 hours of theory classes and 6 hours of practical classes. Both theory and lab training sections of this course took place in the **Immunology Department (CDB), Hospital Clinic of Barcelona (Spain)**. Course Organizer: Aula Clinic (Immunology Department (CDB), Hospital Clinic of Barcelona). This course is officially recognized by the European Federation for Immunogenetics (EFI).
- **02/2015-09/2015: Curso de Postgrado en Técnicas de Diagnóstico Genético-2<sup>a</sup> Edición- (Course of Diagnostic Techniques in Genetics-2<sup>nd</sup> Edition-)**. The period of time of this course was 430 hours in total: 30 hours of theory classes and 400 hours of diagnostic techniques in genetics lab training. The theory section of this course took place in University-Enterprise Foundation (ADEIT), Valencia (Spain). The lab training section of this course was part of my thesis work at the Immunogenetics Unit in Microbiology National Center-Instituto de Salud Carlos III, Madrid (Spain). Course Organizer: Department of Genetics from University of Valencia (UV) and University-Enterprise Foundation (ADEIT). This course is accredited by **University of Valencia (UV) and University-Enterprise Foundation (ADEIT)**.
- **02/2014: Curso en Experimentación Animal (Categoría B) para Investigadores (Course of Laboratory Animal Science (Category B) for Researchers)**. The period of time of this course was 40 hours in total: 20 hours of theory classes and 20 hours of training with animals at the lab. This course took place in **Hospital**

**Universitario La Paz**, Madrid (Spain). Course Organizer: **Animalaria S.L.**. Test Score: 8.6 out of 10.0. This course is accredited by the respective Spanish Authorities according to the Directive 63/2010/EU and the Spanish Royal Decree 53/2013.

## **RESEARCH EXPERIENCE**

**03/2016-09/2017** Predoctoral Researcher. **Peter Parham Lab**, Department of Structural Biology, School of Medicine, Stanford University, Palo Alto, CA, USA.

▪Next Generation Sequencing platforms for high-throughput and high-resolution KIR molecular allele typing and gene content. KIR-disease association studies and KIR population studies.

Being part of United States National Institutes of Health (NIH) multi-center U19 project (NIH/NINDSU19NS095774) on behalf of the Immunogenetics of Neurological Diseases working GrOup (INDIGO) (led by Stanford Univeristy-Marcelo Fernández Viña, Ph.D., D (ABHI) and University California of San Francisco-Jorge Oksenberg, Ph.D.), which has as its primary goal to identify and characterize the repertoire of HLA and KIR genes and alleles that predispose to neurological diseases.

*\*Project responsible and Supervisor:* **Paul Norman**, Ph.D. (Senior Research Scientist).  
(Email contact: paul.norman@ucdenver.edu)

*\*Training responsible:* Neda Nemat-Gorgani, MSc. (Research Assistant).  
(Email contact: gorganin@stanford.edu)

**12/2015-Present** Predoctoral Researcher. **HLA Histocompatibility and Immunogenetics Laboratory, Stanford Blood Center**, Department of Pathology, Stanford University, Palo Alto, CA, USA.

▪Next Generation Sequencing platforms for high-throughput and high-resolution HLA molecular typing. HLA-disease association studies and HLA population studies.

Being part of United States National Institutes of Health (NIH) multi-center U19 project (NIH/NINDSU19NS095774) on behalf of the Immunogenetics of Neurological Diseases working GrOup (INDIGO) (led by Stanford Univeristy-Marcelo Fernández Viña, Ph.D., D (ABHI) and University California of San Francisco-Jorge Oksenberg, Ph.D.), which has as its primary goal to identify and characterize the repertoire of HLA and KIR genes and alleles that predispose to neurological diseases.

*\*Lab responsible and Supervisor:* **Marcelo Fernández Viña**, Ph.D., D (ABHI). (Co-Director of the Histocompatibility, Immunogenetics and Disease Profiling Laboratory).  
(Email contact: marcelof@stanford.edu)

*\*Training responsible:* Lisa Creary, Ph.D. (Senior Research Scientist) and Kazutoyo Osoegawa, Ph.D. (Senior Research Scientist) (Email contact: lcreary@stanford.edu) (Email contact: kazutoyo@stanford.edu)

**12/2014 - 09/2015** Predoctoral Researcher. **Immunogenetics Unit, Microbiology National Center-Instituto de Salud Carlos III**, Madrid (Spain).

- Refinement of a HLA Sequence Based Typing protocol and HLA typing analysis using IMGT/HLA database.
- Refinement of a Triplet Primed PCR protocol as a diagnostic tool for certain diseases with trinucleotide repeat expansion disorders as fragile X syndrome, Huntington's disease and Friederich's Ataxia.
- Refinement of a PCR protocol for sequencing all different 27 exons in *CFTR* gene to detect mutations as a diagnostic tool for cystic fibrosis.
- Refinement of a PCR protocol in order to develop a diagnostic tool for Duchenne muscular dystrophy. In which, we are developing specific amplifications with fluorescent-labeled primers and a later Fragment Analysis as well as sequencing the respective obtained DNA fragments. At the same time, this PCR protocol strategy is being optimized for sequencing KIR genes.
- Refinement of a PCR protocol for sequencing all different exons in beta globin (*HBB*) gene to detect mutations as a diagnostic tool for beta thalassemia.
- Refinement of a Real Time PCR protocol to determinate HLA allele association in autoimmune diseases as diabetes, celiac disease and ankylosing spondylitis.
- Development of a Real Time PCR protocol to detect the presence or lack of GSTT-1 locus in healthy female individuals versus pre-eclampsia patients. At the same time, related to this last, development of ELISA assay for detection of anti-GSTT1 antibodies in human sera comparing these same two groups of individuals.

*\*Lab responsible and Supervisor:* **Jorge Martinez Laso**, Ph.D., D (ABHI). (Director of Immunogenetics Unit). (Email contact: jmlaso12@gmail.com)

*\*Training responsible:* Isabel Cervera Hernández (Research Assistant). (Email contact: icervera@isciii.es)

**08/2014 - 10/2014** Predoctoral Researcher. **Vaccine Discovery-1 Lab, La Jolla Institute for Allergy and Immunology**, La Jolla, California (United States of America).

- Analysis of PBMCs responses to the major allergens from German cockroach: development of ELISPOT assays. This research project belongs to the Inner-City Asthma Consortium (ICAC).
- Development of HLA Restriction assays as part of the project: "T cell determinants of risk of TB in adolescents". This research project belongs to the Global Health Program against Tuberculosis (TB), which is funded and promoted by Bill & Melinda Gates Foundation.

*\*Lab responsible:* **Alessandro Sette**, Ph.D. (Head of the Division of Vaccine Discovery).

(Email contact: alex@lji.org)

*\*Training responsible and Supervisor:* Carla Oseroff, Ph.D. (Postdoctoral Researcher) (in the ICAC project) and Cecilia Lindestam Arlehamn, Ph.D. (Postdoctoral Researcher) (in the TB Gates project). (Emails contacts: coseroff@lji.org; cecilia@lji.org)

**09/2013 - 07/2014** Predoctoral Researcher. **Immunology Unit, Hospital Clínico San Carlos (IdISSC), Madrid (Spain).**

- Academic research project (as my Immunology Research Master of Science Degree project) about the extraction, purification and characterization of a murine tumor cell (Ehrlich tumor) surface carbohydrate (*called "A10"*).

*\*Lab responsible and Supervisor:* **José Luis Subiza**, Ph.D. (Director of Immunology Unit). (Email contact: jlsubiza@immunotek.com)

*\*Training responsible:* Carmen Diez Rivero, Ph.D. (Postdoctoral Researcher). (Email contact: cmdiezri@med.ucm.es)

**11/2011 - 07/2013** Student Internship. **Department of Cell Biology, School of Medicine, Complutense University of Madrid, Madrid (Spain):**

- Academic research project (as my Biochemistry Bachelor of Science Degree research project) about the *in vitro* effect of the morphogen Wnt5a in human naïve CD4<sup>+</sup>T cell activation.

*\*Lab responsible and Supervisor:* **Rosa Sacedón Ayuso**, Ph.D. (Professor of Department of Cell Biology). (Email contact: rosasacedon@med.ucm.es)

*\*Training responsible:* Jaris Valencia, Ph.D. (Postdoctoral Researcher). (Email contact: jarisval@ucm.es)

## **RESEARCH COMPETENCIES AND INTERESTS**

### **Molecular biology:**

RNA and DNA isolation, reverse transcription polymerase chain reaction (rt-PCR), polymerase chain reaction (PCR), quantitative polymerase chain reaction (qPCR), HLA typing PCR-SSOP-Luminex technique, HLA Sequence Based Typing (SBT) by Sanger method, Real Time PCR, and Triplet Primed PCR. Next-Generation Sequencing (NGS) platforms for KIR and HLA genotyping.

### **Cell biology:**

Processing blood, spleen and thymus from human tissue samples for cell isolation, Cell lines/primary cell culture, cell apoptosis assays, cell functionality assays, cell differentiation assays, cell proliferation assays, immunofluorescent double/triple staining, cell surface carbohydrate extraction and purification.

**Microscopy:**

Light microscopy, fluorescent microscopy.

**Immunology:**

Primary T cell isolation, PBMCs isolation, ELISA, ELISPOT, HLA restriction assays, Western blotting, immunocytochemistry, immunohistochemistry, flow cytometry.

**Tumor biology:**

Animal models (Mice strains: Swiss and C57BL/6J) of tumor metastasis (handling, restraint, anesthesia, analgesia, injections (intracardiac, subcutaneous, intraperitoneal), blood and ascites collection, animal dissection and euthanasia), primary tumor cell cultures (Ehrlich murine tumor cell line and lung, prostate and colon human tumor cell lines), purification (tangential flow filtration system with cassette) and characterization (Lectins binding assay and NMR) of carbohydrates.

**Automation/Robotics skills:**

Experience performing automated NGS DNA library preparation protocols with Beckman Coulter Biomek laboratory automation equipment (NxP, FxP, 4000 models). Maintain, troubleshoot, and improve automation and controls systems.

**Computer skills:**

Microsoft Word, Microsoft Excel, Power Point, FCS Express 3 (flow cytometry data analysis computer software), Derive, BLAST, Pymol, RasWin, IMGT/HLA database (sequence alignment tool), KIR (PING software, *Norman. P.J. et al. Am. J. Hum. Genet. 2016*) and HLA (MiaFora software, Immucor) genotyping analysis software programs, R language big data analysis, Pypop (Python for Population genomics, for performing HLA statistical analyses), Hapl-o-Mat (analysis software for haplotype inference via an expectation-maximization algorithm) and BIGDAWG (performs tests of Hardy-Weinberg equilibrium, and carries out case-control association analyses for haplotypes, individual loci, and HLA amino-acid positions on unambiguous genotype data).

**Research interests:**

T cell biology and vaccine discovery for treating cancer or allergic diseases are very promising fields. Apart from that, I consider NGS HLA/KIR genotyping studies a very high potential tool for clinical applications as it is in the transplantation clinical field or to contribute to decode the etiology of many complex human diseases such as narcolepsy, MS, NMO, MG, PD, SCZD, Diabetes, etc.

**Personal Statement:**

I have the knowledge, skills, training, experience and motivation necessary to successfully contribute in any immunogenetics related research project. I have a broad and rich background in cell biology, physiology, biochemistry and immunology with specific training and expertise in molecular genotyping methods, next-generation sequencing (NGS) technologies using automated platforms and analysis pipelines in the field of

immunogenetics, particularly for both HLA and KIR genes. My research background includes HLA- and KIR-disease association studies in neurological diseases. As a junior co-Investigator at Stanford University- and NIH-funded grants, I laid the groundwork for the proposed research by getting highly trained and later optimizing HLA and KIR NGS high-throughput genotyping protocols (including genotyping data analysis and application of statistical methods for case-control analyses), and by establishing significant collaborations with other international/national prestigious clinical and research groups on the field that will make possible to collect samples/clinical phenotypic data of different populations/ethnic groups/patient cohorts as documented in the following publications. In addition, I have produced several peer-reviewed publications from these different projects. As a result of these previous experiences, in addition to the increased level of knowledge and skills obtained in many aspects on the immunogenetics field, I am also aware of the importance of clear communication among project members and of designing a realistic and optimized research plan, defined timeline and sustainable budget.

### **SCIENTIFIC ASSOCIATIONS MEMBERSHIPS**

- **03/2017-Present:** Member of Sociedad Española de Inmunología-SEI (**Spanish Society for Immunology (SEI)**).
- **12/2016- Present:** Student/Fellow Member of **American Society for Histocompatibility and Immunogenetics (ASHI)**.
- **12/2018- Present:** Student/Fellow Member of **European Federation for Immunogenetics (EFI)**.
- **07/2014-07/2015:** Academic Trainee Member of **American Association of Immunologists (AAI)**.

### **RESEARCH-CLINICAL MEETINGS ATTENDANCE**

- **Federation of Clinical Immunology Societies (FOCIS) 2020 Virtual Annual Meeting.** Organizer: **Federation of Clinical Immunology Societies (FOCIS)**. October 28<sup>th</sup>-31<sup>st</sup>, 2020.
- **ASHI 2020 Virtual Annual Meeting.** Organizer: **American Society for Histocompatibility and Immunogenetics (ASHI)**. October 19<sup>th</sup>-21<sup>st</sup>, 2020.
- **Federation of Clinical Immunology Societies (FOCIS) 2018 Meeting.** Organizer: **Federation of Clinical Immunology Societies (FOCIS)**. San Francisco Marriott Marquis, San Francisco, California (USA). June 20<sup>th</sup>-23<sup>rd</sup>, 2018.
- **43<sup>rd</sup> American Society for Histocompatibility and Immunogenetics (ASHI).** Organizer: **American Society for Histocompatibility and Immunogenetics (ASHI)**. Hilton San Francisco Union Square, San Francisco, California (USA). September 11<sup>th</sup>-15<sup>th</sup>, 2017.

- **17<sup>th</sup> International HLA and Immunogenetics Workshop (IHIW)**. Organizer: **Stanford Blood Center (SBC)**. Asilomar Conference Grounds, Pacific Grove, California (USA). September 6<sup>th</sup>-10<sup>th</sup>, 2017.
- **International Day of Immunology at the ISCIII**. Organizer: **Immunology Department, Microbiology National Center (CNM)-Instituto de Salud Carlos III (ISCIII)**, Madrid (Spain). April 24<sup>th</sup>, 2015.
- **2014 AAI Introductory Course in Immunology**. Organizer: **American Association of Immunologists (AAI)**. Long Beach Convention Center, Long Beach, California (USA). July 12<sup>th</sup>-17<sup>th</sup>, 2014.
- **38<sup>th</sup> National Spanish Congress of Immunology**. Organizer: **Spanish Society of Immunology (SEI)**. Edificio Badajoz Siglo XXI, Badajoz (Spain). May 8<sup>th</sup>-10<sup>th</sup>, 2014.
- **1<sup>st</sup> Immunothercan Symposium. *From Inflammation to Cell Plasticity: The New Hallmarks of Cancer***. Centro Nacional de Biotecnología (CNB-CSIC), Madrid (Spain). November 27<sup>th</sup>-28<sup>th</sup>, 2013.

## **RESEARCH MEETINGS ORAL/POSTER PRESENTATIONS**

- **Federation of Clinical Immunology Societies (FOCIS) 2020 Virtual Annual Meeting**. Organizer: **Federation of Clinical Immunology Societies (FOCIS)**. October 28<sup>th</sup>-31<sup>st</sup>, 2020.

### **Poster Session – Autoimmunity**

#### **F185. The Killer Immunoglobulin-like Receptor KIR3DL1 in Combination with HLA-Bw4 is Associated with Pediatric Acute-onset Neuropsychiatric Syndrome (PANS)**

Gonzalo Montero-Martin<sup>1</sup>, Avis Chan<sup>2</sup>, Margo Thienemann<sup>3</sup>, Bahare Farhadian<sup>4</sup>, Theresa Willett<sup>2</sup>, Alicia Madden<sup>5</sup>, Elizabeth Mellins<sup>2</sup>, Tanya Murphy<sup>6</sup>, Susan Swedo<sup>7</sup>, Marcelo Fernández-Viña<sup>1</sup>, Jill A Hollenbach<sup>8</sup>, Jennifer Frankovich<sup>2</sup> and Kirsten M. Anderson<sup>8</sup>

<sup>1</sup>Department of Pathology, Stanford University School of Medicine, Stanford, California, USA., Palo Alto, CA, <sup>2</sup>Division of Allergy, Immunology and Rheumatology, Department of Pediatrics, Stanford University School of Medicine, California, USA, Stanford, CA, <sup>3</sup>Division of Child and Adolescent Psychiatry and Child Development, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, California, USA, Stanford, CA, <sup>4</sup>Stanford Immune Behavioral Health Clinic and PANS Research Program at Lucile Packard Children's Hospital, California, USA, Menlo Park, CA, <sup>5</sup>Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA., Palo Alto, CA, <sup>6</sup>Rothman Center for Pediatric Neuropsychiatry, Pediatrics and Psychiatry, University of South Florida, Florida, USA, St Petersburg, FL, <sup>7</sup>Pediatrics and Developmental Neuroscience Branch (PDN) in the Intramural Research Program (IRP) of the National Institute of Mental Health (NIMH), Bethesda, Maryland, USA, Bethesda, MD, <sup>8</sup>Department of Neurology, University of California San Francisco, San Francisco, California, USA, San Francisco, CA

### **TH38. Refining the HLA-Disease Association Landscape in Neuromyelitis Optica Spectrum Disorders (NMOSD)**

Gonzalo Montero-Martin<sup>1</sup>, Kalyan C. Mallemapati<sup>2</sup>, Sridevi Gangavarapu<sup>2</sup>, Stacy Caillier<sup>3</sup>, Lisa E. Creary<sup>1</sup>, Kazutoyo Osoegawa<sup>2</sup>, Danillo Augusto<sup>3</sup>, Kirsten M. Anderson<sup>3</sup>, Thais Armangue<sup>4</sup>, Maria Sepulveda<sup>4</sup>, Sara Llufrui<sup>4</sup>, Nathalie Dufay<sup>5</sup>, Guillaume Fiard<sup>5</sup>, Maria-Luiza Petzl-Erler<sup>6</sup>, Valérie Dubois<sup>7</sup>, Jose Luis Caro-Oleas<sup>8</sup>, Marius Ringelstein<sup>9</sup>, Romain Marignier<sup>10</sup>, Jun-ichi Kira<sup>11</sup>, Pablo Villoslada<sup>4</sup>, Albert Saiz<sup>4</sup>, Jill A Hollenbach<sup>3</sup>, Marcelo Fernández-Viña<sup>1</sup> and Jorge R Oksenberg<sup>3</sup>

<sup>1</sup>Department of Pathology, Stanford University School of Medicine, Stanford, California, USA., Palo Alto, CA, <sup>2</sup>Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA, Palo Alto, CA, <sup>3</sup>Department of Neurology, University of California San Francisco, San Francisco, California, USA, San Francisco, CA, <sup>4</sup>Center of Neuroimmunology, Service of Neurology, Hospital Clinic and Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain., Barcelona, Catalonia, Spain, <sup>5</sup>NeuroBioTec, Groupement Hospitalier Est, Hôpital Neurologique Pierre Wertheimer, Lyon, France., Lyon, Rhone-Alpes, France, <sup>6</sup>Laboratory of Human Molecular Genetics, Department of Genetics, Universidade Federal do Paraná, Curitiba, PR, Brazil., Curitiba, Parana, Brazil, <sup>7</sup>Etablissement Français du Sang, Lyon, France., Lyon, Rhone-Alpes, France, <sup>8</sup>Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain., Barcelona, Catalonia, Spain, <sup>9</sup>Department of Neurology, Center for Neurology and Neuropsychiatry, LVR-Klinikum Düsseldorf, Düsseldorf, Germany., Düsseldorf, Nordrhein-Westfalen, Germany, <sup>10</sup>Service de Neurologie, Sclérose en Plaques, Pathologies de la Myéline et Neuro-Inflammation, Hôpital Neurologique Pierre Wertheimer Hospices Civils de Lyon, Lyon, France, Lyon, Rhone-Alpes, France, <sup>11</sup>Department of Neurology, Neurological Institute, Graduate School of Medical Sciences, Kyushu University, 812-8582 Fukuoka, Japan., Fukuoka, Fukuoka, Japan.

▪ **ASHI 2020 Virtual Annual Meeting.** Organizer: **American Society for Histocompatibility and Immunogenetics (ASHI).** October 19<sup>th</sup>-21<sup>st</sup>, 2020.

#### **Workshop and Oral Abstract Session II: Immunogenetics**

##### **HIGH-RESOLUTION KIR ALLELIC CHARACTERIZATION IN AMERINDIANS AND BRAZILIAN URBAN POPULATIONS (Presented by Luciana B. Vargas)**

Luciana B. Vargas <sup>1</sup>, Brenda Ho <sup>2</sup>, Gonzalo Montero-Martin <sup>3</sup>, Wesley M. Marin <sup>2</sup>, Marcia H. Beltrame <sup>1</sup>, Marcelo Fernandez-Vina <sup>3</sup>, Maria Luiza Petzl-Erler <sup>1</sup>, Jill A. Hollenbach <sup>2</sup>, Danillo G. Augusto <sup>2</sup>,

<sup>1</sup> Universidade Federal do Paraná, Curitiba, BRAZIL, <sup>2</sup> University of California San Francisco, San Francisco, CA. <sup>3</sup> Stanford University Blood Center, Palo Alto, CA,

#### **Workshop and Oral Abstract Session VI: New Applications and Characterizations**

##### **THE UNPRECEDENTED HIGH-RESOLUTION CHARACTERIZATION OF ALL KIR GENES IN A LARGE NORTH AMERICAN COHORT IDENTIFIES A SUBSTANTIAL PROPORTION OF NEW VARIANTS (Presented by Leonardo M. Amorim)**

Leonardo M. Amorim <sup>1</sup>, Danillo G. Augusto <sup>2</sup>, Neda Nemat-Gorgani <sup>3</sup>, Gonzalo Montero-Martin <sup>4</sup>, Wesley M. Marin <sup>2</sup>, Ravi Dandekar <sup>2</sup>, Hengameh Shams <sup>2</sup>, Peter Parham <sup>5</sup>, Marcelo Fernandez-Vina <sup>4</sup>, Jorge R. Oksenberg <sup>2</sup>, Paul J. Norman <sup>6</sup>, Jill A. Hollenbach <sup>2</sup>,



1 Programa de Pós-Graduação em Genética, Universidade Federal do Paraná, Curitiba, BRAZIL, 2 Department of Neurology, University of California San Francisco, San Francisco, CA, 3 Department of Structural Biology, Stanford University, Stanford, CA, 4 Stanford University Blood Center, Palo Alto, CA, 5 Department of Structural Biology, Stanford University, Palo Alto, CA, 6 Division of Personalized Medicine, University of Colorado, Aurora, CO.

▪ **41<sup>st</sup> National Spanish Congress of Immunology.** Organizer: **Spanish Society of Immunology (SEI).** Hotel Meliá Sevilla, Sevilla (**Spain**). May 30<sup>th</sup> to June 1<sup>st</sup>, 2019.

### Oral Session – Immunogenetics

**Oral Communication # 38 “HIGH-RESOLUTION HLA ALLELIC AND HAPLOTYPIC ASSOCIATION WITH MULTIPLE SCLEROSIS IN SPANISH POPULATION USING NEXT-GENERATION SEQUENCING” (Presented by M. Fernandez-Vina)**

GONZALO MONTERO-MARTIN<sup>1</sup>, SARA LLUFRIU<sup>2</sup>, MARIA SEPULVEDA<sup>2</sup>, THAIS ARMANGUE<sup>2</sup>, JOSE LUIS CARO-OLEAS<sup>2</sup>, JORGE OKSENBERG<sup>2</sup>, PABLO VILLOSLADA<sup>2</sup>, ALBERT SAIZ<sup>2</sup>, MARCELO FERNANDEZ-VINA<sup>1</sup>.

1. STANFORD UNIVERSITY, SCHOOL OF MEDICINE DEPARTMENT OF PATHOLOGY.

2. HOSPITAL CLINIC, SERVICE OF NEUROLOGY.

▪ **33<sup>rd</sup> European Federation for Immunogenetics (EFI) Meeting.** Organizer: **Portuguese Society of Transplantation (SPT) and Portuguese Institute for Blood and Transplantation (IPST).** Centro Cultural de Belém (CCB), Lisbon (**Portugal**). May 8<sup>th</sup> -11<sup>th</sup>, 2019.

### Oral Presentations–Reproduction, Autoimmunity, Infection & Cancer

**O-18 “THE SHARED EPITOPE OF HLA-DRB1 MEDIATES RISK AND INTERACTS WITH SMOKING HISTORY IN PARKINSON'S DISEASE” (Presented by J.A. Hollenbach)**

Jill A. Hollenbach<sup>1</sup>, Paul J. Norman<sup>2</sup>, Lisa E. Creary<sup>3</sup>, Vincent Damotte<sup>1</sup>, Gonzalo Montero Martin<sup>4</sup>, Stacy Caillier<sup>5</sup>, Kirsten Anderson<sup>6</sup>, Maneesh K. Misra<sup>1</sup>, Neda Nemat- Gorgani<sup>7</sup>, Kazutoyo Osoegawa<sup>8</sup>, Wesley M. Marino<sup>9</sup>, Ravi Dandekar<sup>1</sup>, Marcelo A. Fernandez-Vina<sup>10</sup>, Jorge Oksenberg<sup>11</sup>.

1. University of California, San Francisco, San Francisco, United States of America.

2. University of Colorado Anschutz Medical Campus, Aurora, United States of America.

3. Stanford University School of Medicine, Department of Pathology, Palo Alto, United States of America.

4. Stanford University School of Medicine, Stanford, United States of America.

5. *Department of Neurology, University of California San Francisco, San Francisco, United States of America.*
6. *University of California San Francisco, San Francisco, United States of America.*
7. *Stanford University, Stanford, United States of America.*
8. *Stanford Blood Center, Palo Alto, United States of America.*
9. *University of California, San Francisco School of Medicine, San Francisco, United States of America.*
10. *Stanford University School of Medicine, Palo Alto, United States of America.*
11. *UCSF, San Francisco, United States of America.*

**O-21 “HIGH RESOLUTION HAPLOTYPE ANALYSES OF CLASSICAL HLA GENES IN FAMILIES WITH MULTIPLE SCLEROSIS” (Presented by K. Osoegawa)**

Kazutoyo Osoegawa<sup>1</sup>, Lisa E. Creary<sup>2</sup>, Kalyan C. Mallempati<sup>1</sup>, Sridevi Gangavarapu<sup>1</sup>, Gonzalo Montero-Martin<sup>3</sup>, Stacy J. Caillier<sup>4</sup>, Jill A. Hollenbach<sup>5</sup>, Jorge R. Oksenberg<sup>4</sup>, Marcelo A. Fernandez Vina<sup>3</sup>.

1. *Stanford Blood Center, Palo Alto, United States of America.*
2. *Stanford University School of Medicine, Department of Pathology, Palo Alto, United States of America.*
3. *Stanford University School of Medicine, Palo Alto, United States of America.*
4. *University of California, San Francisco, San Francisco, United States of America.*
5. *University of California, San Francisco School of Medicine, San Francisco, United States of America.*

**Oral Presentations–New Technologies**

**O-63 “DEEP ANALYSIS OF KIR2DL1 AND KIR3DL1S1 BY NEXT GENERATION SEQUENCING IN 3,695 INDIVIDUALS IDENTIFIES NOVEL VARIANTS WITH POSSIBLE FUNCTIONAL RELEVANCE”. (Presented by D.G. Augusto)**

Danillo G. Augusto<sup>1</sup>, Neda Nemat-Gorgani<sup>2</sup>, Gonzalo Montero-Martin<sup>3</sup>, Wesley Marin<sup>1</sup>, Ravi Dendekar<sup>4</sup>, Peter Parham<sup>2</sup>, Marcelo A. Fernandez-Vina<sup>3</sup>, Jorge R. Oksenberg<sup>1</sup>, Paul J. Norman<sup>2</sup>, Jill A. Hollenbach<sup>5</sup>.

1. *University of California San Francisco, San Francisco, United States of America.*
2. *Stanford University, Stanford, United States of America.*
3. *Stanford University School of Medicine, Stanford, United States of America.*
4. *Department of Structural Biology, San Francisco, United States of America.*
5. *University of California, San Francisco, San Francisco, United States of America.*

**Poster Session–Evolution, Anthropology & Population Genetics**

**P-106 “HIGH-RESOLUTION HLA ALLELIC AND HAPLOTYPIC ASSOCIATION WITH MULTIPLE SCLEROSIS IN SPANISH POPULATION USING NEXT-GENERATION SEQUENCING”**

Gonzalo Montero-Martin<sup>1</sup>, Sara Llufrui<sup>2</sup>, Maria Sepulveda<sup>2</sup>, Thais Armangue<sup>2</sup>, Kazutoyo Osoegawa<sup>3</sup>, Kalyan C. Mallempati<sup>3</sup>, Sridevi Gangavarapu<sup>3</sup>, Lisa E. Creary<sup>1</sup>, Stacy Caillier<sup>4</sup>, Jorge R. Oksenberg<sup>4</sup>, Pablo Villoslada<sup>2</sup>, Albert Saiz<sup>2</sup>, Marcelo A. Fernandez-Vina<sup>1</sup>.

1. *Stanford University School of Medicine, Department of Pathology, Palo Alto, United States of America.*
2. *Service of Neurology, Hospital Clinic, University of Barcelona, Spain Neuroimmunology Program, Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), Barcelona, Spain,*

3. *Stanford Blood Center, Palo Alto, United States of America.*

4. *Department of Neurology, University of California San Francisco, San Francisco, United States of America.*

▪ **Childhood Arthritis and Rheumatology Research Alliance (CARRA) 15<sup>th</sup> Annual Scientific Meeting.** Organizer: **Childhood Arthritis and Rheumatology Research Alliance.** Louisville, Kentucky, (USA). April 10<sup>th</sup>-14<sup>th</sup>, 2019.

## Poster Session

**Abstract # 614399 “HLA Findings in Youth with Pediatric Acute-onset Neuropsychiatric Syndrome (PANS)”.**

Jennifer Frankovich<sup>1</sup>, Jill Hollenbach<sup>2</sup>, Gonzalo Montero-Martin<sup>1\*</sup>, Avis Chan<sup>1</sup>, Margo Thienemann<sup>1</sup>, Bahare Farhadian<sup>1</sup>, Theresa Willett<sup>1</sup>, David Lewis<sup>1</sup>, Elizabeth Mellins<sup>1</sup>, Tanya Murphy<sup>3</sup>, Marcelo Fernandez-Vina<sup>1</sup>.

1. *Stanford University, California, USA.*

2. *University of California San Francisco, California, USA.*

3. *University of South Florida, Tampa, Florida, USA.*

▪ **The Guthy-Jackson Charitable Foundation’s (GJCF) 2019 11th International NMO Roundtable Conference.** UCLA Luskin Conference Center, UCLA, Los Angeles, California (USA). March 31<sup>st</sup>, 2019.

## Oral Session III – Uniting the World to Solve NMOSD

**O-3 “Next-Gen HLA Class I & II Genotyping in NMOSD”**

Montero-Martín, G.<sup>1</sup>, Mallempati, K.<sup>2</sup>, Gangavarapu, S.<sup>2</sup>, Caillier, S.<sup>3</sup>, Creary, L.E.<sup>1</sup>, Osoegawa K.<sup>2</sup>, Augusto, D.<sup>3,4</sup>, Anderson, K.<sup>3</sup>, Armangue, T.<sup>5</sup>, Sepulveda, M.<sup>5</sup>, Llufríu, S.<sup>5</sup>, Dufay, N.<sup>6</sup>, Fiard, G.<sup>6</sup>, Petzl-Erler, M.L.<sup>4</sup>, Dubois, V.<sup>7</sup>, Caro-Oleas, J.L.<sup>8</sup>, Marignier, R.<sup>9</sup>, Kira, J.I.<sup>10</sup>, Villoslada, P.<sup>5</sup>, Saiz, A.<sup>5</sup>, Hollenbach, J.A.<sup>3</sup>, Fernández-Viña, M.A.<sup>1</sup> and Oksenberg, J.R.<sup>3</sup>

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*

2. *Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.*

3. *Department of Neurology, University of California San Francisco, San Francisco, California, USA*

4. *Laboratory of Human Molecular Genetics, Department of Genetics, Universidade Federal do Paraná, Curitiba, PR, Brazil.*

5. *Center of Neuroimmunology, Service of Neurology, Hospital Clinic and Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain.*

6. *NeuroBioTec, Groupement Hospitalier Est, Hôpital Neurologique Pierre Wertheimer, Lyon, France.*

7. *Etablissement Français du Sang, Lyon, France.*

8. *Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain.*
9. *Service de Neurologie, Sclérose en Plaques, Pathologies de la Myéline et Neuro-Inflammation, Hôpital Neurologique Pierre Wertheimer Hospices Civils de Lyon, Lyon, France.*
10. *Department of Neurology, Neurological Institute, Graduate School of Medical Sciences, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, 812-8582, Japan.*

▪ **HLA and KIR Region Genomics in immune-mediated diseases Consortium (HLARGC), National Institute of Health (NIH) Steering Committee (National Institute of Allergy and Infectious Diseases (NIAID) and National Institute of Neurological Disorders and Stroke (NINDS)).** Li Ka Shing Building, Stanford University, Stanford, California (USA). March 20<sup>th</sup>, 2019.

## **Oral Session 4 – Immunogenetic Determinants of Disease in Neurological Disease (U19), HLA Class I and Class II in neurological diseases (U19)**

### **O-1 “Refining the HLA-disease association landscape in Neuromyelitis Optica”**

Montero-Martín, G.<sup>1</sup>, Mallempati, K.<sup>2</sup>, Gangavarapu, S.<sup>2</sup>, Caillier, S.<sup>3</sup>, Creary, L.E.<sup>1</sup>, Osoegawa K.<sup>2</sup>, Augusto, D.<sup>3,4</sup>, Anderson, K.<sup>3</sup>, Armangue, T.<sup>5</sup>, Sepulveda, M.<sup>5</sup>, Llufríu, S.<sup>5</sup>, Dufay, N.<sup>6</sup>, Fiard, G.<sup>6</sup>, Petzl-Erler, M.L.<sup>4</sup>, Dubois, V.<sup>7</sup>, Caro-Oleas, J.L.<sup>8</sup>, Marignier, R.<sup>9</sup>, Kira, J.I.<sup>10</sup>, Villoslada, P.<sup>5</sup>, Saiz, A.<sup>5</sup>, Hollenbach, J.A.<sup>3</sup>, Fernández-Viña, M.A.<sup>1</sup> and Oksenberg, J.R.<sup>3</sup>

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*
2. *Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.*
3. *Department of Neurology, University of California San Francisco, San Francisco, California, USA*
4. *Laboratory of Human Molecular Genetics, Department of Genetics, Universidade Federal do Paraná, Curitiba, PR, Brazil.*
5. *Center of Neuroimmunology, Service of Neurology, Hospital Clinic and Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain.*
6. *NeuroBioTec, Groupement Hospitalier Est, Hôpital Neurologique Pierre Wertheimer, Lyon, France.*
7. *Etablissement Français du Sang, Lyon, France.*
8. *Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain.*
9. *Service de Neurologie, Sclérose en Plaques, Pathologies de la Myéline et Neuro-Inflammation, Hôpital Neurologique Pierre Wertheimer Hospices Civils de Lyon, Lyon, France.*
10. *Department of Neurology, Neurological Institute, Graduate School of Medical Sciences, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, 812-8582, Japan.*

▪ **11<sup>th</sup> KIR WORKSHOP 2018.** Organizer: **Italian Society for Immunogenetics and Transplantation Biology (AIBT).** Camogli, Genoa (**Italy**). October 25<sup>th</sup>-27<sup>th</sup>, 2018.

### **Oral Session - KIR sequencing and typing**

**“Novel variants of KIR2DL1 and KIR3DL1/S1 identified in 3,695 individuals analyzed by next generation sequencing”. (Presented by D.G. Augusto)**

Danillo G. Augusto<sup>1</sup>, Neda Nemat-Gorgani<sup>2</sup>, Gonzalo Montero-Martin<sup>3</sup>, Wesley Marin<sup>1</sup>, Ravi Dendekar<sup>4</sup>, Peter Parham<sup>2</sup>, Marcelo A. Fernandez-Vina<sup>3</sup>, Jorge R. Oksenberg<sup>1</sup>, Paul J. Norman<sup>2</sup>, Jill A. Hollenbach<sup>5</sup>.

1. *University of California San Francisco, San Francisco, United States of America.*
2. *Stanford University, Stanford, United States of America.*
3. *Stanford University School of Medicine, Stanford, United States of America.*
4. *Department of Structural Biology, San Francisco, United States of America.*
5. *University of California, San Francisco, San Francisco, United States of America.*

▪ **2018 American College of Rheumatology (ACR/ARHP) Annual Meeting.**  
Organizer: **American College of Rheumatology.** Chicago, Illinois (USA). October 19<sup>th</sup>-24<sup>th</sup>, 2018.

**Poster Session - Genetics, Genomics and Proteomics**

**Abstract # 1975 “Behcet’s Disease Lies in the “B” Holder. New Associations in Disease Susceptibility and Manifestations”.**

Mohanad Elfishawi<sup>1,2</sup>, Sally Elfishawi<sup>3</sup>, Ghada Mossallam<sup>3</sup>, Paul Norman<sup>4</sup>, Jill Hollenbach<sup>5</sup>, Maneesh Misra<sup>5</sup>, Gonzalo Montero Martin<sup>6</sup>, Helma de Bruin<sup>7</sup>, Leos Van de Pasch<sup>7</sup>, Erik Rozemuller<sup>7</sup>, Marcelo Fernandez-Vina<sup>6</sup>, Adriana Abrudescu<sup>8</sup> and Khaled Zaky<sup>9</sup>.

1. *Internal Medicine, Icahn School of Medicine at Mount Sinai, Queens Hospital Center, New York, NY, USA.*
2. *Rheumatology, Kasr Alainy Hospital, Cairo University, Cairo, Egypt.*
3. *Clinical Pathology and Immunology Laboratory, National Cancer Institute, Cairo University, Cairo, Egypt.*
4. *Division of Personalized Medicine and Department of Immunology, University of Colorado School of Medicine, Denver, Colorado, USA.*
5. *Department of Neurology, University of California San Francisco, San Francisco, California, USA.*
6. *Department of Pathology, Stanford University, School of Medicine, Stanford, California, USA.*
7. *GenDx, Utrecht, Netherlands.*
8. *Internal Medicine and Rheumatology, Icahn School of Medicine at Mount Sinai, Queens Hospital Center, NYC, NY, USA.*
9. *Rheumatology and Rehabilitation, Faculty of medicine, Al-Azhar University, Cairo, Egypt.*

▪ **44<sup>th</sup> American Society for Histocompatibility and Immunogenetics (ASHI).**  
Organizer: **American Society for Histocompatibility and Immunogenetics (ASHI).** Baltimore Marriott Waterfront, Baltimore, Maryland (USA). October 1<sup>st</sup>-5<sup>th</sup>, 2018.

**Oral Presentations**

**OR24 “HLA allele and haplotype frequencies characterized using next-generation sequencing methods in unrelated world-wide populations: Summary from the 17th International HLA and Immunogenetics Workshop” (Presented by L.E. Creary)**

Lisa E. Creary<sup>1</sup>, Chia-Jung Chang<sup>2</sup>, Gonzalo Montero-Martin<sup>1</sup>, Kalyan C. Mallempati<sup>3</sup>, Sridevi Gangavarapu<sup>3</sup>, Kazutoyo Osoegawa<sup>3</sup>, Tamara Vayntrub<sup>3</sup>, Marcelo A. Fernandez-Vina<sup>1</sup>.

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*

2. *Stanford Genome Technology Center, Palo Alto, California, USA.*

3. *Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.*

▪ **Federation of Clinical Immunology Societies (FOCIS) 2018 Meeting.**

Organizer: **Federation of Clinical Immunology Societies (FOCIS)**. San Francisco Marriott Marquis, San Francisco, California (USA). June 20<sup>th</sup>-23<sup>rd</sup>, 2018.

**Poster Session – Immunogenetics, Immunology of the eye**

**P-T.56 “Association Study between HLA genes and Climatic Droplet Keratopathy (CDK) in a cohort from the Patagonian region of Argentina”**

Montero-Martin, G.<sup>1</sup>, Suárez, M.F.<sup>2</sup>, Mallempati, K.<sup>3</sup>, Fernández-Viña, M.A.<sup>1</sup>, Urrets-Zavalía, J.A.<sup>4</sup>, and Serra, H.M.<sup>2</sup>.

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*

2. *CIBICI-CONICET, Faculty of Chemical Sciences, Department of Clinical Biochemistry, Universidad Nacional de Córdoba, Córdoba, Argentina.*

3. *Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.*

4. *Department of Ophthalmology, University Clinic Reina Fabiola, Universidad Católica de Córdoba, Córdoba, Argentina.*

▪ **XII Congress of the Latin American Association of Immunology (ALAI) and XXIII Congress of the Mexican Society of Immunology (SMI).** Organizer: **Mexican Society of Immunology (SMI)**., Cancun, Quintana Roo (Mexico). May 14<sup>th</sup>-18<sup>th</sup>, 2018.

**Poster Session I – Clinical Immunology**

**P-125 “Association Study between HLA genes and Climatic Droplet Keratopathy (CDK) in a cohort from the Patagonian region of Argentina”**

Montero-Martin, G.<sup>1</sup>, Suárez, M.F.<sup>2</sup>, Mallempati, K.<sup>3</sup>, Fernández-Viña, M.A.<sup>1</sup>, Urrets-Zavalía, J.A.<sup>4</sup>, and Serra, H.M.<sup>2</sup>.

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*

2. CIBICI-CONICET, Faculty of Chemical Sciences, Department of Clinical Biochemistry, Universidad Nacional de Córdoba, Córdoba, Argentina.

3. Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.

4. Department of Ophthalmology, University Clinic Reina Fabiola, Universidad Católica de Córdoba, Córdoba, Argentina.

▪ **32<sup>nd</sup> European Federation for Immunogenetics (EFI) and 25<sup>th</sup> Italian Society for Immunogenetics and Transplantation Biology (AIBT) Joint Meeting.** Organizer: **Italian Society for Immunogenetics and Transplantation Biology (AIBT).** Palazzo del Cinema and Palazzo del Casinò, Venice (Italy). May 9<sup>th</sup>-12<sup>th</sup>, 2018.

### **Oral Presentations—MHC Evolution, Anthropology & Population Genetics**

**O-11 “HLA allele and haplotype frequencies characterized using next-generation sequencing methods in unrelated world-wide populations: Summary from the 17th International HLA and Immunogenetics Workshop” (Presented by L.E. Creary)**

Lisa E. Creary<sup>1</sup>, Chia-Jung Chang<sup>2</sup>, Gonzalo Montero-Martin<sup>1</sup>, Kalyan C. Mallemapati<sup>3</sup>, Sridevi Gangavarapu<sup>3</sup>, Kazutoyo Osoegawa<sup>3</sup>, Tamara Vayntrub<sup>3</sup>, Marcelo A. Fernandez-Vina<sup>1</sup>.

1. Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.

2. Stanford Genome Technology Center, Palo Alto, California, USA.

3. Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.

### **Poster Group No. 6 – MHC Evolution, Anthropology & Population Genetics**

**P-155 “High-Resolution Characterization of Allelic and Haplotypic HLA Frequencies Distribution in a Spanish Population using High-Throughput Next-Generation Sequencing”**

Montero-Martín, G.<sup>1</sup>, Creary, L.E.<sup>1</sup>, Mallemapati, K.<sup>2</sup>, Gangavarapu, S.<sup>2</sup>, Vayntrub, T.<sup>2</sup>, Planelles, D.<sup>3</sup>, Vilches, C.<sup>4</sup>, Caro-Oleas, J.L.<sup>5</sup>, Herrero-Mata, M. J.<sup>5</sup>, Sánchez-Gordo, F.<sup>6</sup>, González-Escribano, F.<sup>7</sup>, Muro, M.<sup>8</sup>, Moya-Quiles, M.R.<sup>8</sup>, González-Fernández, R.<sup>9</sup>, Sánchez-García, F.<sup>10</sup>, Ocejo-Vinyals, J.G.<sup>11</sup>, Balas, A.<sup>12</sup>, Vicario, J.L.<sup>12</sup>, Marín, L.<sup>13</sup> and Fernández-Viña, M.A.<sup>1</sup>.

1. Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.

2. Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.

3. Histocompatibility, Centro de Transfusión de la Comunidad Valenciana, Valencia, Spain.

4. Immunogenetics and Histocompatibility, Instituto de Investigación Sanitaria Puerta de Hierro, Madrid, Spain.

5. Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain.

6. Histocompatibility, Centro de Transfusión de Málaga, Málaga, Spain.

7. Immunology, Hospital Universitario Virgen del Rocío, Sevilla, Spain.

8. Immunology, Hospital Clínico Universitario Virgen de la Arrixaca, Murcia, Spain.

9. Immunology, Hospital Universitario Reina Sofía, Córdoba, Spain.

10. Immunology, Hospital Universitario de Gran Canaria Dr Negrín, Las Palmas de Gran Canaria, Spain.

11. Immunology, Hospital Universitario Marqués de Valdecilla, Santander, Spain.

12. Histocompatibility, Centro de Transfusión de la Comunidad de Madrid, Madrid, Spain.

13. Molecular Biology-Hematology, Hospital Clínico Universitario, Salamanca, Spain.

▪ **17<sup>th</sup> International HLA and Immunogenetics Workshop (IHIW).** Organizer: **Stanford Blood Center (SBC).** Asilomar Conference Grounds, Pacific Grove, California (USA). September 6<sup>th</sup>-10<sup>th</sup>, 2017.

### **Oral Session – Workshop-Activities: NGS HLA Projects, Population Genetics, IHIWS Unrelated project contributors, individual brief reports**

#### **O-1 “HLA-NGS Spain Population Study”**

Montero-Martín, G.<sup>1</sup>, Creary, L.E.<sup>1</sup>, Mallemapati, K.<sup>2</sup>, Gangavarapu, S.<sup>2</sup>, Vayntrub, T.<sup>2</sup>, Planelles, D.<sup>3</sup>, Vilches, C.<sup>4</sup>, Caro-Oleas, J.L.<sup>5</sup>, Herrero-Mata, M. J.<sup>5</sup>, Sánchez-Gordo, F.<sup>6</sup>, González-Escribano, F.<sup>7</sup>, Muro, M.<sup>8</sup>, Moya-Quiles, M.R.<sup>8</sup>, González-Fernández, R.<sup>9</sup>, Sánchez-García, F.<sup>10</sup>, Oejo-Vinyals, J.G.<sup>11</sup>, Balas, A.<sup>12</sup>, Vicario, J.L.<sup>12</sup>, Marín, L.<sup>13</sup> and Fernández-Viña, M.A.<sup>1</sup>.

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*
2. *Stanford Blood Center, Stanford University School of Medicine, Palo Alto, California, USA.*
3. *Histocompatibility, Centro de Transfusión de la Comunidad Valenciana, Valencia, Spain.*
4. *Immunogenetics and Histocompatibility, Instituto de Investigación Sanitaria Puerta de Hierro, Madrid, Spain.*
5. *Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain.*
6. *Histocompatibility, Centro de Transfusión de Málaga, Málaga, Spain.*
7. *Immunology, Hospital Universitario Virgen del Rocío, Sevilla, Spain.*
8. *Immunology, Hospital Clínico Universitario Virgen de la Arrixaca, Murcia, Spain.*
9. *Immunology, Hospital Universitario Reina Sofía, Córdoba, Spain.*
10. *Immunology, Hospital Universitario de Gran Canaria Dr Negrín, Las Palmas de Gran Canaria, Spain.*
11. *Immunology, Hospital Universitario Marqués de Valdecilla, Santander, Spain.*
12. *Histocompatibility, Centro de Transfusión de la Comunidad de Madrid, Madrid, Spain.*
13. *Molecular Biology-Hematology, Hospital Clínico Universitario, Salamanca, Spain.*

### **Oral Session – Workshop-Activities: 17th Workshop, KIR component population reports**

#### **O-3 “KIR diversity in a Spanish Population”**

Montero-Martín, G.<sup>1</sup>, Misra, M.K.<sup>2</sup>, Nemat-Gorgani, N.<sup>3</sup>, Balas, A.<sup>4</sup>, Cisneros, E.<sup>5</sup>, González-Fernández, R.<sup>6</sup>, Herrero-Mata, M.J.<sup>7</sup>, Moreno-Hidalgo, M.A.<sup>4</sup>, Sánchez-Gordo, F.<sup>8</sup>, Vicario, J.L.<sup>4</sup>, Sánchez-García, F.<sup>9</sup>, Fernández-Viña, M.A.<sup>1</sup>, Oksenberg, J.R.<sup>2</sup>, Parham, P.<sup>3</sup>, Caro-Oleas, J.L.<sup>7</sup>, Planelles, D.<sup>10</sup>, Vilches, C.<sup>5</sup>, Norman, P.J.<sup>11</sup> and Hollenbach, J.A.<sup>2</sup>.

1. *Department of Pathology, Stanford University School of Medicine, Stanford, California, USA.*
2. *Department of Neurology, University of California San Francisco, San Francisco, California, USA.*
3. *Department of Structural Biology, Stanford University School of Medicine, Stanford, California, USA.*
4. *Histocompatibility, Centro de Transfusión de la Comunidad de Madrid, Madrid, Spain.*
5. *Immunogenetics and Histocompatibility, Instituto de Investigación Sanitaria Puerta de Hierro, Madrid, Spain.*
6. *Immunology, Hospital Universitario Reina Sofía, Córdoba, Spain.*
7. *Histocompatibility and Immunogenetics, Banc de Sang i Teixits, Barcelona, Spain.*
8. *Histocompatibility, Centro de Transfusión de Málaga, Málaga, Spain.*
9. *Immunology, Hospital Universitario de Gran Canaria Dr Negrín, Las Palmas de Gran Canaria, Spain.*
10. *Histocompatibility, Centro de Transfusión de la Comunidad Valenciana, Valencia, Spain.*
11. *Department of Structural Biology, Stanford University School of Medicine, Stanford, California, USA.; Division of Biomedical Informatics and Personalized Medicine, and Department of Immunology, University of Colorado, Denver, Colorado, USA.*



▪ **31<sup>st</sup> European Federation for Immunogenetics (EFI) and 25<sup>th</sup> German Society for Immunogenetics (DGI) Joint Meeting.** Organizer: **German Society for Immunogenetics (DGI).** Mannheim/Heidelberg (**Germany**). May 30<sup>th</sup>-June 2<sup>nd</sup>, 2017.

### **Oral Presentations– Best Abstracts**

**O-6 “HLA and KIR Mediate Susceptibility to Parkinson’s Disease” (Presented by J.A. Hollenbach)**

Jill A. Hollenbach<sup>1</sup>, Paul J. Norman<sup>2</sup>, Gonzalo Montero Martin<sup>3</sup>, Neda Nemat-Gorgani<sup>2</sup>, Lisa Creary<sup>3</sup>, Maneesh K. Misra<sup>1</sup>, Vincent Damotte<sup>1</sup>, Stacy Caillier<sup>1</sup>, Jorge R. Oksenberg<sup>1</sup>, Marcelo Fernandez Viña<sup>3</sup>

1. Department of Neurology, University of California San Francisco School of Medicine, CA, USA.

2. Department of Structural Biology, Stanford University School of Medicine, CA, USA.

3. Department of Pathology, Stanford University School of Medicine, CA, USA.

▪ **40<sup>th</sup> National Spanish Congress of Immunology.** Organizer: **Spanish Society of Immunology (SEI).** Edificio Paraninfo Universidad de Zaragoza, Zaragoza (**Spain**). May 25<sup>th</sup>-27<sup>th</sup>, 2017.

### **Poster Session – Immunogenetics and Transplantation**

**P-078 “HLA and KIR Mediate Susceptibility to Parkinson’s Disease”**

Jill A. Hollenbach<sup>1</sup>, Paul J. Norman<sup>2</sup>, Gonzalo Montero Martin<sup>3</sup>, Neda Nemat-Gorgani<sup>2</sup>, Lisa Creary<sup>3</sup>, Maneesh K. Misra<sup>1</sup>, Vincent Damotte<sup>1</sup>, Stacy Caillier<sup>1</sup>, Jorge R. Oksenberg<sup>1</sup>, Marcelo Fernandez Viña<sup>3</sup>

1. Department of Neurology, University of California San Francisco School of Medicine, CA, USA.

2. Department of Structural Biology, Stanford University School of Medicine, CA, USA.

3. Department of Pathology, Stanford University School of Medicine, CA, USA.

▪ **16<sup>th</sup> Annual Meeting of the Society for Natural Immunity. NK2016.** Organizer: **Society for Natural Immunity (SNI).** Atahotel Capotaormina, Taormina (ME) (Italy). October 2<sup>nd</sup>-5<sup>th</sup>, 2016.

### **Poster Session - NK cell receptors and signaling**

**P-032 - KIR3DL1\*004 in the presence of HLA Bw4 is protective in Parkinson’s Disease**

Jill A. Hollenbach<sup>1</sup>, Paul J. Norman<sup>2</sup>, Neda Nemat-Gorgani<sup>2</sup>, Gonzalo M. Martin<sup>3</sup>, Maneesh Misra<sup>1</sup>, Vincente Damotte<sup>1</sup>, Lisa Creary<sup>3</sup>, Marcelo Fernandez-Vina<sup>3</sup>, Jorge R. Oksenberg<sup>1</sup>

1. Department of Neurology, University of California, San Francisco, CA, USA.

2. Department of Structural Biology, Stanford University, CA, USA.
3. Department of Pathology, Stanford University, CA, USA.

▪ **39<sup>th</sup> National Spanish Congress of Immunology.** Organizer: **Spanish Society of Immunology (SEI).** Auditorio de la Diputación de Alicante (ADDA), Alicante (Spain). May 5<sup>th</sup>-7<sup>th</sup>, 2016.

## Poster Session – Immunogenetics

### P-094 “Description of the novel HLA-DQB1\*02:02:01:02 allele in a Spanish individual”

G. Montero Martín<sup>1</sup>, E. Rivas García<sup>4</sup>, I. Cervera<sup>1</sup>, A. Teniente Serra<sup>6</sup>, M. Fonolleda<sup>6</sup>, M. C. Trejo Benítez<sup>4</sup>, M. I. González Henríquez<sup>4</sup>, J. Martínez Laso<sup>5</sup>.

1. Centro Nacional de Microbiología-ISCIII.
2. German Trias i Pujol University IGTP-Campus Can Ruti.
3. Universitat Autònoma de Barcelona.
4. Complejo Hospitalario Universitario Insular Materno-Infantil.
5. Centro Nacional de Microbiología- ISCIII, Complejo Hospitalario Universitario Insular Materno-Infantil.
6. German Trias i Pujol University IGTP-Campus Can Ruti, Universitat Autònoma de Barcelona.

**APPENDIX 2:****PUBLISHED ARTICLES AND SUBMITTED MANUSCRIPTS UNDER REVIEW****PUBLICATIONS**

- 1) Anderson, K.M., Augusto, D.G., Dandekar, R., Shams, H., Zhao, C., Yusufali, T., Montero-Martín, G., Marin, W.M., Nemat-Gorgani, N., Creary, L.E., Caillier, S., Mofrad, M.R.K., Parham, P., Fernández-Viña, M., Oksenberg, J.R., Norman, P.J. and Hollenbach, J.A. (2020). Killer-cell Immunoglobulin-like Receptor Variants Are Associated with Protection from Symptoms Associated with More Severe Course in Parkinson's Disease. *The Journal of Immunology*, 205(5), 1323-1330.
- 2) Creary LE, Gangavarapu S, Mallempati KC, Montero-Martín G, Caillier SJ, Santaniello A, Hollenbach JA, Oksenberg JR and Fernández-Viña MA (2019). Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population. *Human Immunology*. (PMCID: PMC6778037 [Available on 2020-10-01]).
- 3) Osoegawa K, Mallempati KC, Gangavarapu S, Oki A, Gendzekhadze K, Marino SR, Brown NK, Bettinotti MP, Weimer ET, Montero-Martin G, Creary LE, Vayntrub TA, Chang CJ, Askar M, Mack SJ and Fernández-Viña MA (2019). HLA alleles and haplotypes observed in 263 US families. *Human Immunology*. (PMCID: PMC6773484 [Available on 2020-09-01]).
- 4) Hollenbach, J.A., Norman, P.J., Creary, L.E., Damotte, V., Montero-Martin, G., Caillier, S., Anderson, K.M., Misra, M.K., Nemat-Gorgani, N., Osoegawa, K., Santaniello, A., Renschen, A., Marin, W.M., Dandekar, R., Parham, P., Tanner, C.M., Hauser, S.L., Fernandez-Viña, M. and Oksenberg, J.R. (2019). A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proceedings of the National Academy of Sciences*, 116(15), 7419-7424. (PMCID: PMC6462083).
- 5) Montero-Martín, G., Mallempati, K.C., Gangavarapu, S., Sánchez-Gordo, F., Herrero-Mata, M.J., Balas, A., Vicario, J.L., Sánchez-García, F., González-Escribano, M.F., Muro, M., Moya-Quiles, M.R., González-Fernández, R., Oejo-Vinyals, J.G., Marín, L., Creary, L.E., Osoegawa, K., Vayntrub, T., Caro-Oleas, J.L., Vilches, C., Planelles, D. and Fernández-Viña, M.A. (2019). High-resolution characterization of allelic and haplotypic HLA frequency distribution in a Spanish population using high-throughput next-generation sequencing. *Human Immunology*, 80(7), 429-436. (NIHMS1523909, Publ.ID: HIM10172; PMCID: PMC6599556 [Available on 2020-07-01]).
- 6) Misra, M.K., Augusto, D.G., Martin, G.M., Nemat-Gorgani, N., Sauter, J., Hofmann, J.A., Traherne, J.A., González-Quezada, B., Gorodezky, C., Bultitude, W.P., Marin, W., Vierra-Green, C., Anderson, K.M., Balas, A., Caro-Oleas, J.L., Cisneros, E., Colucci, F., Dandekar, R., Elfishawi, S.M., Fernández-Viña, M.A., Fouda, M., González-Fernández, R., Große, A., Herrero-Mata, M.J., Hollenbach, S.Q., Marsh, S.G.E., Mentzer, A., Middleton, D., Moffett, A., Moreno-Hidalgo, M.A., Mossallam, G.I., Nakimuli, A., Oksenberg, J.R., Oppenheimer, S.J., Parham, P., Petzl-Erler, M.L., Planelles, D., Sánchez-García, F., Sánchez-Gordo, F., Schmidt, A.H., Trowsdale, J., Vargas, L.B., Vicario, J.L., Vilches, C., Norman, P.J. and Hollenbach, J.A. (2018). Report from the killer-cell immunoglobulin-like receptors (KIR) component of the 17th international HLA and immunogenetics workshop. *Human Immunology*, 79(12):825-833. (PMCID: PMC6322681 [Available on 2019-12-01]).

7) Belbezier, A., Joubert, B., Montero-Martin, G., Fernandez-Vina, M., Fabien, N., Rogemond, V., Mignot, E. and Honnorat, J. (2018). Multiplex family with GAD65-Abs neurologic syndromes. *Neurology-Neuroimmunology Neuroinflammation*, 5(1), e416. (PMCID: PMC5778747).

8) Montero-Martin, G., Cervera, I., Teniente-Serra, A., Folloneda, M., and Martinez-Laso, J. (2016). Description of the novel HLA-DQB1\*02:02:01:02 allele in a Spanish Individual. *HLA*, 87(2), 113.

### **MANUSCRIPTS SUBMITTED UNDER REVIEW**

1) Elfishawi, M., Mossallam, G., Augusto, D., Montero-Martin, G., de Bruin, H, Van de Pasch, L, Norman, PJ, Rozemuller, E, Fernandez-Vina, M., Abrudescu, A., Hollenbach, JA, Zaky, K. and Elfishawi, S. (2020). Behçet disease, New insights in disease associations and manifestations. A next generation sequencing study. (*Manuscript under review*).

2) Hollenbach, J.A., Ombrello, M., Tremoulet, A., Rosa Duque, J.S., Chua, G., Montero-Martin, G., Burns, J., Shimizu, C., Deutsch, G., Monos, D., Mallajosyu, V., Xu, J., Ghosh, D., Tan, S., Remmers, E., Fernandez-Viña, M.A., Canna, S., Szymanski, A.M., Rubin, D., Saper, V. and Mellins, E. (2020). Hypersensitivity reactions to IL-1 and IL-6 inhibitors are linked to certain HLA-DRB1\*15 alleles. (*Manuscript under review*).

3) Gorzynski, J.E., De Jong, H.N., Amar, D., Hughes, C., Tanigawa, Y., Kistler, A., Kamm, J., Neff, N., Rubinacci, S., Delaneau, O., Shoura, M. J., Seo, K., Kirillova, A., Raja, A., Sutton, S., Huang, C., Sahoo, M. K, Mallempati, K.C., Montero-Martin, G., Osoegawa, K., Watson, N., Hammond, N., Joshi, R., Fire, A., Fernandez-Vina, M.A., Christle, J.W., Wheeler, M.T., Pinsky, B.A., Rivas, M.A., Bustamante, C., Ashley, E.A., Parikh, V.N. (2020). High-throughput SARS-CoV-2 and host genome sequencing from single nasopharyngeal swabs. (*Manuscript under review*).

4) Creary, L.E., Sacchi, N., Mazzocco, M., Morris, G.P., Montero-Martin, G., Chong, W., Brown, C.J., Dinou, A., Stavropoulos-Giokas, C., Gorodezky, C., Narayan, S., Periathiruvadi, S., Thomas, R., De Santis, D., Pepperall, J., El Ghazali, G.E., Al Yafei, Z., Askar, M., Tyagi, S., Kanga, U., Marino, S.R., Planelles, D., Chang, C.J. and Fernández-Viña, M.A. (2020). High-resolution HLA allele and haplotype frequencies characterized using next-generation sequencing in worldwide unrelated populations: 17th International HLA and Immunogenetics Workshop joint report. (*Manuscript under review*).

5) Osoegawa, K., Creary, L.E, Montero-Martin, G., Mallempati, K.C., Gangavarapu, S., Caillier, S., Santaniello, A, Isobe, N., Hollenbach, JA, Hauser, S.L., Oksenberg, J.R. and Fernández-Viña, M. (2020). High resolution haplotype analyses of classical HLA genes in families with multiple sclerosis highlights the role HLA-DP alleles in disease susceptibility. (*Manuscript under review*).







