



UNIVERSITY OF
LIVERPOOL

Department of Clinical Infection, Microbiology and Immunology
Institute of Infection and Global Health

**The development of a database
and bioinformatics applications for the
investigation of immune genes**

Thesis submitted in accordance with the requirements of
The University of Liverpool for the degree of Doctor in Philosophy

by

Faviel Francisco Gonzalez Galarza

September 2011

Abstract

The development of a database and bioinformatics applications for the investigation of immune genes

Faviel Francisco Gonzalez Galarza

The extensive allelic variability observed in several genes related to the immune response and its significance in transplantation, disease association studies and diversity in human populations has led the scientific community to analyse these variants among individuals.

This thesis is focussed on the development of a database and software applications for the investigation of several immune genes and the frequencies of their corresponding alleles in worldwide human populations. The approach presented in this thesis includes the design of a relational database, a web interface, the design of models for data exchange and the development of online searching mechanisms for the analysis of allele, haplotype and genotype frequencies.

At present, the database contains data from more than 1000 populations covering more than four million unrelated individuals. The repertory of datasets available in the database encompasses different polymorphic regions such as Human Leukocyte Antigens (HLA), Killer-cell Immunoglobulin-like Receptors (KIR), Major histocompatibility complex Class I chain-related (MIC) genes and a number of cytokine gene polymorphisms.

The work presented in this document has been shown to be a valuable resource for the medical and scientific societies. Acting as a primary source for the consultation of immune gene frequencies in worldwide populations, the database has been widely used in a variety of contexts by scientists, including histocompatibility, immunology, epidemiology, pharmacogenetics and population genetics among many others. In the last year (August 2010 to August 2011), the website was accessed by 15,784 distinct users from 2,758 cities in 136 countries and has been cited in 168 peer-reviewed publications demonstrating its wide international use.

Declaration

I hereby declare that the content of this thesis corresponds to my work, which has not been submitted for a degree at this University or any other institution, and that other sources of information used in the text have been clearly acknowledged.

The development of the database described on chapter 3 was initiated by Ralph Komerofsky in collaboration with Prof. Derek Middleton, and their contribution has been described on the 'Previous work' section of that chapter.

My contribution to the publications related to this research was as follows:

- **Gonzalez-Galarza F.F.**, Christmas S., Middleton D. and Jones A.R. Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acid Research* 2011; **39**:D913-D919.
*** Featured Article**
Corresponding and main author.
- Middleton D and **Gonzalez F.** Immunogenetic Databases. In: Mehra N, eds. *The HLA complex in Biology and Medicine: A resource book*. New Delhi: Jaypee 2010, 119-134.
Co-author with text, figures and tables included.
- Middleton D, **Gonzalez-Galarza F**, Meenagh A and Gourraud PA. Diversity of KIR genes, alleles and haplotypes In: J. Zimmer eds. *Natural Killer Cells: At the Forefront of Modern Immunology*. Berlin: Springer-Verlag 2010; 63-91.
Co-author with text, figures and tables included in the section "Website and KIR Genotypes".
- Middleton D and **Gonzalez F.** The extensive polymorphism of KIR genes. *Immunology* 2009; **129**:8-19.
Co-author with figures and tables included.

- Middleton D, **Gonzalez F**, Fernandez-Vina M, *et al.* A bioinformatic approach to ascertaining the rarity of HLA alleles. *Tissue antigens* 2009; **74**:480-485.
Co-author with figures and tables included, and leading the continuation of the project for the 16th International HLA and Immunogenetics Workshop in collaboration with Prof. Derek Middleton.
(http://www.16ibim.org/projects/gonzalez_rare.html)

.....

Signature

Faviel Francisco Gonzalez Galarza

Thesis overview

In recent years, technologies and computational techniques have been applied to the fields of medicine and biological sciences. A large number of algorithms and methods have been implemented in these areas to assist scientists in the analysis of large datasets generated after experiments. This information needs to be stored in electronic repositories to allow individuals to submit their own information, consult the data available and participate in the curation of other datasets where possible.

This thesis describes the design of a database, the Allele Frequency Net Database (AFND), containing information on several polymorphic regions, genes and their corresponding alleles involved in the immune response. A set of computational notations used in the construction of the database and formats for data exchange are also described in this work. The schemas presented in this thesis can be used by other database developers for the implementation of other related polymorphic regions or datasets with a similar structure. Additionally, the thesis describes a set of web-based applications which were designed for the consultation of information available.

AFND has assisted a number of different scientific groups from a diverse range of disciplines. The website has been extensively used in the investigation of the presence of alleles in worldwide human populations.

The content of this thesis is divided into eight related chapters which describe the design and implementation of the database and the analysis of the different frequencies of the polymorphic regions available.

Chapter 1 presents an introduction to the basis of immunology and immunogenetic databases. The chapter is intended to allow both informaticians and biologists to familiarise themselves with the basics of immunogenetics, the importance of public databases and the use of bioinformatics applications as a method to disseminate biological datasets. The polymorphisms covered in this work include the Human Leukocyte Antigens (HLA), Killer-cell Immunoglobulin-like Receptors (KIR), Major histocompatibility complex Class I chain-related (MIC) genes and a number of cytokine

gene polymorphisms. A review of other useful databases and bioinformatics applications commonly used for the analysis of immune genes is also presented in Chapter 1.

Chapters 2 and 3 describe the informatics approaches developed in this research. The computational methods and techniques are described in these two chapters. *Chapter 2* is focussed on the definition of the structure and content of the different datasets and computational techniques that were used in the construction of AFND. This chapter describes the implementation of a database schema to define the relationships among the existing datasets. *Chapter 3* describes the development of the AFND website as an electronic interface to explore the frequency data of the different populations. This chapter presents some examples of searching tools available on the AFND website including searches by allele, haplotype, genotype and amino acid frequencies.

Chapters 4, 5, 6 and 7 present a summary of the different analyses that were performed for each polymorphic region available in AFND. The aim of these chapters is to examine the allele, haplotype and genotype frequencies in different populations and illustrate their variability by presenting a collection of overlaid maps and breakdowns.

Chapter 4 focuses on the analysis of the frequency distribution of HLA genes at allele and haplotype level. The chapter includes the description of different software applications implemented for the analysis of frequencies. Additionally, several overlaid maps are presented to illustrate the variability of specific alleles in worldwide populations.

Taking into account that many of the existing HLA alleles have only been reported on one occasion, *Chapter 5* presents an extensive worldwide analysis of the rarity of HLA alleles. The analysis is based on the number of confirmations of these alleles by different scientific groups and international organisations.

Chapter 6 presents a summary of the distribution of KIR genes at allele and genotype level. The chapter includes a section to describe the content and structure of the KIR genotype database. The compilation of KIR genotypes presented in this chapter encompasses the largest KIR genotype collection in worldwide populations.

Chapter 7 comprises a summary of breakdowns of the frequencies of MIC genes and frequencies on associations with some of the HLA loci. Additionally, frequencies of several cytokine gene polymorphisms in worldwide populations are also presented in this chapter.

Finally, *Chapter 8* summarises the findings and includes a set of topics for discussion. Additionally, the section describes also a number of future projects which may be generated from the current work.

As complementary material, two appendices are included at the end of this thesis to further describe the methods and parameters used in the implementation of the database: Appendix A describes the schemas used in the construction of the database and Appendix B defines the metadata and data dictionary of the schemas.

Acknowledgments

I would like to thank my supervisors Prof. Derek Middleton, Dr. Andrew Jones and Dr. Stephen Christmas for their extraordinary support, time dedicated to this research, valuable criticism and the encouragement to participate in scientific and academic networks. I am deeply thankful to Prof. Middleton who provided me with guidance to understand the basics of immunogenetics and improve my research skills, and Dr. Jones and Dr. Christmas in topics related to the challenging and fascinating world of bioinformatics and immunology.

I also would like to express my gratitude to Prof. Steven G. E. Marsh, Prof. Alejandro Madrigal, Dr. Hazael Maldonado-Torres and James Robinson from the Anthony Nolan Research Institute, Dr. Asensio Gonzalez from the Northern Ireland Histocompatibility Laboratory and Dr. Rafael Arguello from the Autonomous University of Coahuila who provided me with aid and assistance during the early steps and throughout the duration of my research.

Likewise, I would like to thank the staff and PhD fellows of the Division of Immunology at the University of Liverpool: Mrs Helen Nelson, Mrs Patricia Fife, Dr. Deborah Howarth, Dr. Joanne Cummerson, Suhair Salman, Dr. Suliman Al-Omar, Ali Alejenef and Waleed Mahallawi.

Finally, I want to give my biggest thanks to my parents (Francisco and Manuela) and brothers (Fabricio and Favio) for their love, understanding and support during this time. They have always been the inspiration and guide to pursue my goals.

I dedicate this thesis to the memory of
my grandfather Julian Galarza and grandmother Maria de Jesus Hernandez

“Success is not final, failure is not fatal: it is the courage to continue that counts”.

Winston Churchill

Table of Contents

Abstract.....	i
Declaration.....	ii
Thesis overview.....	iv
List of figures.....	xv
List of tables.....	xvii
Commonly used abbreviations.....	xix
Chapter 1	
Investigation of immune genes	1
1.1 Introduction.....	1
1.2 The immune system.....	3
1.3 The Major Histocompatibility Complex.....	3
1.4 Human Leukocyte Antigens.....	4
1.4.1 Genes and alleles in the HLA system.....	6
1.4.2 HLA nomenclature.....	7
1.4.3 Genetics of the HLA system.....	11
1.4.4 Applications of the HLA system.....	13
1.5 Killer-cell Immunoglobulin-like Receptors.....	16
1.5.1 Genes and alleles in the KIR family.....	16
1.5.2 Organisation of the KIR genes.....	17
1.5.3 KIR nomenclature.....	18
1.5.4 KIR and HLA ligands.....	20
1.5.5 Applications of the KIR family.....	20
1.6 Other immune genes.....	21
1.6.1 Major histocompatibility Class I chain-related genes.....	21
1.6.2 Cytokine gene polymorphisms.....	22
1.7 Bioinformatics and immunogenetics databases.....	23
1.7.1 Evolution of biological databases.....	23
1.7.2 Architecture of databases.....	24
1.7.3 Database modelling.....	25
1.7.4 Information retrieval.....	27

1.7.5	Web applications and technologies	28
1.7.6	Programming languages in bioinformatics.....	29
1.7.7	Data exchange and formats	30
1.7.8	Curation of biological datasets.....	31
1.7.9	Immune gene databases and bioinformatics resources	32
1.8	Aim of the thesis	37
1.9	Summary.....	38

Chapter 2

Design of the Allele Frequency Net Database: datasets, schemas and methods 40

2.1	Introduction	40
2.2	Compilation of population datasets.....	41
2.2.1	Source of data.....	41
2.2.2	Submission protocol.....	41
2.2.3	Validation of population attributes and frequency data.....	44
2.2.4	Summary of population datasets	53
2.3	Design of the AFND schema.....	56
2.3.1	Database schemas	56
2.3.2	Metadata: data dictionaries and controlled vocabulary.....	58
2.4	Discussion	58
2.5	Conclusions.....	61

Chapter 3

Development of the AFND website and online tools 62

3.1	Introduction	62
3.1.1	Previous work in immune gene frequency websites.....	62
3.2	Systems and methods	64
3.2.1	Website organisation and dataflow.....	64
3.2.2	Software implementation.....	68
3.2.3	Data visualisation	69
3.2.4	Systems and hardware requirements.....	69
3.3	Navigation in population samples	70
3.4	Frequency Search Interfaces.....	71
3.4.1	The Allele Frequency Search (AFS)	71

3.4.2	The Haplotype Frequency Search (HFS)	75
3.4.3	The Genotype Frequency Search (GFS)	77
3.5	Amino acid frequency comparisons in populations.....	81
3.6	Accessibility of data	82
3.7	Database users and software metrics	82
3.8	Discussion	85
3.9	Conclusions.....	87

Chapter 4

HLA allele and haplotype frequency data 88

4.1	Introduction	88
4.2	Materials and methods.....	89
4.2.1	Population datasets	89
4.2.2	Frequency data and terminology.....	90
4.2.3	Data analysis and visualisation	92
4.3	Results.....	92
4.3.1	Summary of HLA frequency data in the AFND.....	92
4.3.2	Occurrence of high-resolution HLA alleles.....	98
4.3.3	Online applications for the analysis of HLA alleles.....	99
4.3.4	Software for analysis of HLA haplotype distribution.....	101
4.4	Discussion	104
4.5	Conclusions.....	107

Chapter 5

A bioinformatics approach to ascertain the rarity of HLA alleles 109

5.1	Introduction	109
5.2	Materials and methods.....	110
5.2.1	Datasets	110
5.2.2	Collection of confirmatory data from individual laboratories.....	111
5.3	Results.....	113
5.3.1	Summary of submissions	113
5.3.2	Classification of HLA alleles	116
5.3.3	Investigation of the rarity of HLA alleles in UK laboratories.....	119
5.3.4	The rare allele search (RAS)	121
5.3.5	AFND and confirmations in the US.....	124

5.3.6	Rare allele detector (RAD)	125
5.4	Discussion	126
5.5	Conclusions	128

Chapter 6

KIR genes and genotype frequencies 130

6.1	Introduction	130
6.2	Materials and methods.....	131
6.2.1	Population datasets	131
6.2.2	Frequency data and terminology.....	132
6.2.3	Construction of the KIR genotype database	132
6.2.4	Data analysis	134
6.2.5	Data visualisation	134
6.3	Results	134
6.3.1	Summary of data available in the AFND for KIR.....	134
6.3.2	Frequency distribution of KIR genes.....	136
6.3.3	The KIR genotype database	150
6.4	Discussion	153
6.5	Conclusions	156

Chapter 7

Other immune genes: MIC and Cytokine gene polymorphisms 157

7.1	Introduction	157
7.2	Materials and methods.....	158
7.2.1	Population datasets	158
7.2.2	MIC and cytokine gene frequency data	160
7.3	Results	161
7.3.1	Major histocompatibility Class I chain-related genes	161
7.3.2	Cytokine gene polymorphisms.....	167
7.4	Discussion	171
7.5	Conclusions	174

Chapter 8

Discussion, general conclusions and future work 175

8.1	Summary of thesis	175
8.2	Discussion	176

8.2.1	The Allele Frequency Net Database	177
8.2.2	Alternative approaches in the design of immune gene databases....	177
8.2.3	Use of relational models and DBMS	179
8.2.4	Archive of raw data	179
8.2.5	Standards in immune gene frequencies	180
8.3	Future work.....	180
8.3.1	Quality control.....	180
8.3.2	Software toolkit for data sharing and complementary analysis.....	181
8.3.3	KIR and HLA ligands	181
8.4	General conclusions.....	182
	Bibliography	183
	Appendix A	201
	Schemas of the Allele Frequency Net Database.....	201
	Appendix B	218
	Metadata and data dictionary.....	218

List of figures

Number	Page
Figure 1.1: Location and organisation of the HLA complex.....	4
Figure 1.2: Structure of the HLA Class I and Class II molecules.	5
Figure 1.3: Structure of the nomenclature for HLA alleles.....	8
Figure 1.4: Example of identical HLA alleles over exons 2 and 3.	11
Figure 1.5: Example of inheritance of HLA alleles.	12
Figure 1.6: Organisation of the KIR genes.	18
Figure 1.7: Structure of the nomenclature for KIR alleles.	19
Figure 1.8: Example of a relational model.....	26
Figure 1.9: Example of an object-oriented model.....	27
Figure 1.10: Example of Boolean algebra using AND, OR and NOT operators.	28
Figure 1.11: Example of XML document.	31
Figure 2.1: Dataflow diagram for submissions of populations in the AFND.	42
Figure 2.2: Example of demographic data capture in the AFND.....	43
Figure 2.3: Example of frequency data capture in the AFND	44
Figure 2.4: Classification of geographical regions in the AFND.	49
Figure 2.5: Conceptual schema of the AFND.	57
Figure 3.1: Number of populations reported on the AFND by year.	64
Figure 3.2: Workflow and system architecture of the AFND.	65
Figure 3.3: Screenshot of the AFND website.....	66
Figure 3.4: Organisation of the Allele Frequency Net Database website.	67
Figure 3.5: Program nomenclature in the AFND website.	68
Figure 3.6: Geographical location of population samples in the AFND.....	71
Figure 3.7: Example of the Allele Frequency Search (AFS).	72
Figure 3.8: Overall frequency distribution of the A*68:01 allele.....	74
Figure 3.9: Example of the grid view on the AFS.....	75
Figure 3.10: Example of the Haplotype Frequency Search (HFS).	76
Figure 3.11: Example of the Genotype Frequency Search (GFS).	78
Figure 3.12: KIR genotype searching options in the AFND.....	79

Figure 3.13: Example of the Amino Acid Frequency Analysis Tool (AAFAT).	81
Figure 3.14: Number of visits to the AFND by country.....	83
Figure 4.1: Occurrence of high-resolution alleles of HLA Class I loci by region.....	98
Figure 4.2: Occurrence of high-resolution alleles of HLA Class II loci by region.	99
Figure 4.3: Example of the Allele Frequency Calculator (AFC).	100
Figure 4.4: Frequency distribution of the HLA-A*02:01 and HLA-A*32:01 alleles.....	101
Figure 4.5: Example of the HLA haplotype frequency search.	102
Figure 4.6: Distribution of the HLA-DRB1*14:02-DQA1*05:01-DQB1*03:01 haplotype.	103
Figure 4.7: Distribution of the (A) DRB1*14:02, (B) DQA1*05:01 and (C) DQB1*03:01 alleles.	103
Figure 5.1: Screenshot of the online submission form for rare allele confirmations.	113
Figure 5.2: Summary of findings of HLA alleles.	116
Figure 5.3: Very rare alleles by year of submission.....	119
Figure 5.4: Example of the Rare Allele Search (RAS).	122
Figure 5.5: Example of identical HLA alleles over exons 2 and 3.	123
Figure 5.6: Information provided by individual laboratories on rare alleles.....	123
Figure 5.7: Screenshot of the Rare Allele Detector (RAD) module.	125
Figure 6.1: Screenshot of the KIR genotype submission form.....	133
Figure 6.2: Overall gene frequencies in KIR populations on AFND.	137
Figure 6.3: Frequency distribution of the KIR genes by geographical region.	141
Figure 6.4: Occurrence of KIR genes in worldwide populations.	147
Figure 6.5: Most common KIR genotypes in worldwide populations	150
Figure 6.6: Occurrence of distinct genotypes in populations and individuals.	151
Figure 6.7: Example of KIR AA genotypes with one gene missing.....	151
Figure 6.8: Linkage disequilibrium in the KIR genes.....	153
Figure 7.1: Frequency distribution of MICA microsatellites.	163
Figure 7.2: Frequency distribution of MICA microsatellites by geographical region. ...	164
Figure 7.3: Frequency distribution of top 10 MICA alleles.	165
Figure 7.4: MICA and HLA-B association frequency search.	167
Figure 7.5: Example of the frequency search for cytokine gene polymorphisms.....	170
Figure 7.6: Frequency distribution of the ‘TNFalpha/-308 AG’ polymorphism.	171

List of tables

Number	Page
Table 1.1: Genes and number of alleles in the HLA region	7
Table 1.2: HLA alleles with a special expression status	9
Table 1.3: Genes and number of alleles in the KIR family	17
Table 1.4: Useful databases for the investigation of immune genes	33
Table 1.5: Useful applications for immunogenetics and population genetics analyses	35
Table 2.1: Ethnic groups available in the AFND	50
Table 2.2: List of methods used in the estimation of frequencies	52
Table 2.3: Population frequency datasets by polymorphic region on the AFND	54
Table 2.4: Populations by sample size and polymorphic region	54
Table 2.5: Population samples by geographical region	55
Table 2.6: Population samples by ethnic group	55
Table 3.1: Top 10 searches accessed by users	83
Table 3.2: Number of visits by connection speed	84
Table 3.3: Number of visits by operating system	84
Table 3.4: Number of visits by web browser	84
Table 4.1: Population samples with allele frequency data by geographical region	90
Table 4.2: Populations with haplotype frequency data by geographical region	92
Table 4.3: Availability of frequency data of classical HLA loci by geographical region ..	93
Table 4.4: Availability of frequency data of non-classical HLA loci by region	95
Table 4.5: Number of HLA haplotype datasets available in the AFND	96
Table 4.6: HLA haplotype datasets by number of loci typed	97
Table 5.1: Rare alleles reported to AFND by individual laboratories	114
Table 5.2: Number of allele confirmations by times seen	115
Table 5.3: Summary of results of rare alleles	117
Table 5.4: Origin of rare alleles (1, 2 and 3 times)	118
Table 5.5: Very rare alleles by identical sequences	118
Table 5.6: Rare allele submissions by UK laboratories	120
Table 5.7: Summary of alleles by number of times seen	120

Table 5.8: Number of changes in allele status after submissions by UK labs.....	121
Table 5.9: Common and well-documented alleles and data in AFND	124
Table 6.1: KIR populations in the AFND by locus and level of resolution	135
Table 6.2: Confirmations of KIR alleles in the AFND	136
Table 6.3: Populations with the absence of a KIR framework gene	146
Table 6.4: Distribution of KIR genotypes by geographical region.....	152
Table 7.1: Cytokine names and references used in the AFND	160
Table 7.2: Availability of MIC data by level of resolution.....	162
Table 7.3: Availability of cytokine gene polymorphisms by geographical region	168

Commonly used abbreviations

AFC	Allele Frequency Calculator
AFND	Allele Frequency Net Database
AFS	Allele Frequency Search
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
ASHI	American Society for Histocompatibility and Immunogenetics
ASP	Active Server Pages
BMT	Bone Marrow Transplantation
CSS	Cascading Style Sheet
CSV	Comma-separated Value
CWD	Common and Well-Documented
DBMS	Database Management System
DTD	Document Type Definition
ERD	Entity-Relationship Diagram
FTP	File Transfer Protocol
GFS	Genotype Frequency Search
GNU GPL	GNU General Public License
HFS	Haplotype Frequency Search
HGP	Human Genome Project
HLA	Human Leukocyte Antigen
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HUGO	Human Genome Organisation
IHWS	International Histocompatibility Workshop
KIR	Killer-cell Immunoglobulin-like Receptor
LD	Linkage Disequilibrium
LRC	Leukocyte Receptor Complex
mHags	Minor Histocompatibility antigens
Mb	Mega base pairs
MHC	Major Histocompatibility Complex

MIC	Major histocompatibility complex Class I chain-related
NK	Natural Killer
NMDP	National Marrow Donor Program
OODB	Object Oriented Database
OOP	Object Oriented Programming
PCR	Polymerase Chain Reaction
RAD	Rare Allele Detector
RAM	Random Access Memory
RAS	Rare Allele Search
RDB	Relational Database
RM	Relational Model
RSS	Really Simple Syndication
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TSV	Tab-separated Value
UML	Unified Model Language
URL	Uniform Resource Locator
WWW	World Wide Web
W3C	World Wide Web Consortium
WHO	World Health Organisation
XHTML	Extensible Hypertext Markup Language
XML	Extensible Markup Language

Chapter 1

Investigation of immune genes

1.1 Introduction

In recent decades, the field of bioinformatics has played a significant role in the investigation of the genetics and genomics of human beings. The completion of the human genome project (HGP) (Lander et al. 2001; Venter et al. 2001) has led individuals to further extend the number of analyses and experiments with the aim of understanding the functioning of this biological complex.

One of the topics that has gained the attention of medical and biological researchers is to explore the relationship between the immune system and hereditary factors in individuals. Since 1936, when the term *immunogenetics* was initially introduced to describe this association, many studies have been performed by scientific groups to investigate the function of several genes which are involved in the immune system response in humans (Alper & Larsen 2004; Bontrop, Kasahara & Watkins 1999; Irwin 1976).

For many years, researchers have tended to combine experimental biology with analytical methods including algorithms and, more recently, computational models to assist scientific communities in the interpretation of their experiments and results. Thus, the role of programmers in the field of biology, known as *bioinformatics*, has become an important part of contemporary research by providing scientists with specialised software and databases to host and analyse an extensive number of biological datasets.

In the last fifty years, a vast number of studies on genes involved in the immune system response have been reported in the literature and in proceedings from *International Histocompatibility Workshops* (IHWSs) amassing an extremely large volume of information [For reviews of IHWSs see (Terasaki 2007; Thorsby 2009)]. One of the main outcomes from these IHWSs has been the massive number of DNA variants that have been found

in these immune genes among individuals in different worldwide populations. The variability present in the sequences of these genes has been demonstrated to have a significant impact in many fields of immunology and genetics research, i.e. explaining the basis of tissue rejection in transplantation, associations to infectious and autoimmune diseases, diversity in populations, etc. (See Section 1.4.4 for a wider review of applications of immune genes).

With the advances in technology and the use of novel molecular methods, data reported in journals becomes out of date rapidly. Therefore, the need to provide a public and up-to-date electronic resource to contain frequency datasets in populations has been an essential requirement for the scientific and medical societies.

In the past two decades, the introduction of *public databases* has fostered the interest of individuals from different research groups to interact with online repositories that can be freely accessed through the world wide web (WWW) (Varmus 2003). However, the design of biological databases encompasses a series of important factors that need to be considered during the implementation process such as (i) the standardisation of datasets, (ii) the definition of controlled vocabularies and (iii) integration with other databases (Stein 2003). Considering these essential factors, this research was focussed on the development of a database which can serve as a warehouse for the storage of immune gene frequencies along with an online repository to satisfy the demands of investigators who are interested in examining the occurrence of these genes.

The aim of this chapter is to present a concise review of the functioning of the immune system and some of the genes involved in the immune response, and introduce the reader to the basis of population genetics, immunogenetics databases and software applications which can be useful to perform population genetics analysis. The polymorphisms covered in this research include four polymorphic regions: the Human Leukocyte Antigens (HLA), Killer-cell Immunoglobulin-like Receptors (KIR), Major histocompatibility complex Class I chain-related (MIC) genes and a number of cytokine gene polymorphisms.

1.2 The immune system

The *immune system* in humans is a set of biological processes that protects individuals against infectious agents and tumour cells. These processes are based on the identification of non-self molecules in the body which are subsequently eliminated. The immune response is divided into two components, the innate and adaptive immune systems. The *innate immune system* provides individuals with an immediate response by recruiting immune cells, predominantly phagocytes (neutrophils, monocytes and macrophages), to the region infected, which have the capacity to distinguish host cells from pathogens (Delves & Roitt 2000a).

Usually, individuals acquire immunity during their lives due to the *adaptive immune system*, which can recognise an indefinite number of cells and act rapidly on re-exposure to the same infection.

Some aspects of the immune system have hereditary factors that vary among groups of individuals leading to the importance of investigating these genetic traits in different populations.

1.3 The Major Histocompatibility Complex

The *Major Histocompatibility Complex* (MHC) is a genomic region that encodes a set of proteins located on the surface of cells which permit T cells (known also as T lymphocytes) to identify and target infected cells (Delves & Roitt 2000b). The MHC is situated on the short arm of chromosome 6 at position 6p21.3 and is considered to be the most polymorphic region in the human genome. This site contains more than two hundred expressed genes of which approximately 40% are estimated to have immunological functions (Horton et al. 2004; The MHC sequencing consortium 1999).

The MHC region is approximately 7.6 mega base pairs (Mb) in length and many of the genes exhibit extensive polymorphism (Horton et al. 2004). The MHC is divided into three main regions referred as Class I, II and III (Figure 1.1). However, MHC Class III

presents different functions from those of Class I and II and only the first two classes are investigated in this thesis.

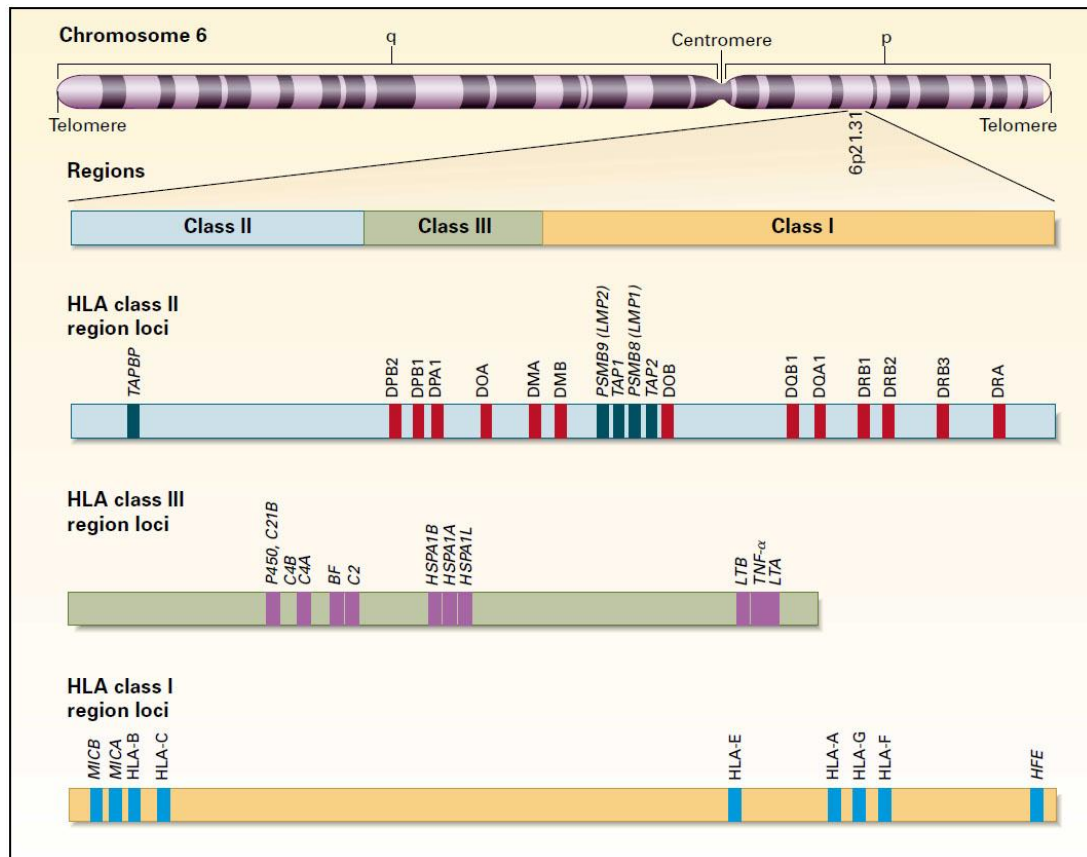


Figure 1.1: Location and organisation of the HLA complex.

(Klein & Sato 2000a)

1.4 Human Leukocyte Antigens

In humans, one of the most important components of the MHC is the *Human Leukocyte Antigen* (HLA) System. The principal role of the HLA system is in defence against microorganisms by presenting *peptides* (short sequences of amino acids) of the pathogen to T cells for recognition.

The HLA system is divided into two classes: Class I and Class II, which differ in molecular structure and function. *HLA Class I* molecules are found on almost all somatic cells, whereas *HLA Class II* molecules are specifically present on immune

system cells (e.g. B and T lymphocytes, macrophages and dendritic cells) (Klein & Sato 2000a).

The main function of the HLA Class I molecule is to present antigens to cytotoxic T lymphocytes (CD8+ T cells) which can eradicate cells typically infected by viruses. In contrast, HLA Class II molecules present peptides to T helper cells (CD4+ T cells) which, although they cannot directly eliminate infected cells, play an important role in the activation of other immune cells.

Based on the molecular structure, the HLA Class I molecule is formed by two polypeptides chains (α and β) containing a total of five domains: two peptide-binding domains ($\alpha 1$ and $\alpha 2$), one immunoglobulin-like domain ($\alpha 3$), a lateral β_2 -microglobulin gene (β_{2m}), the transmembrane region (TM) and the cytoplasmic tail (Figure 1.2). The structure of the class II is composed of two mirror peptides with four domains each: a peptide-binding domain ($\alpha 1$ or $\beta 1$), the immunoglobulin-like domain ($\alpha 2$ or $\beta 2$), the transmembrane region and the cytoplasmic tail in each of the chains.

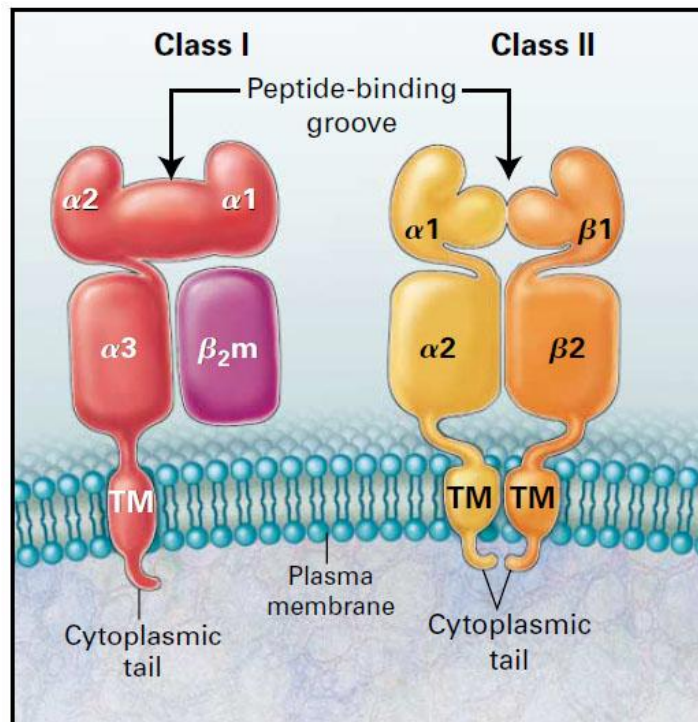


Figure 1.2: Structure of the HLA Class I and Class II molecules.

(Klein & Sato 2000a)

1.4.1 Genes and alleles in the HLA system

The HLA region is composed of more than twenty genes which are officially named by the World Health Organisation (WHO) Nomenclature Committee. Some of these genes are regularly tested for and thus called *classical HLA genes* (HLA-A, -B and -C for Class I; and HLA-DRB1, -DQA1, -DQB1, -DPA1 and -DPB1 for Class II). Other genes or *loci*¹ have also been identified in the HLA region; however, the majority of them probably do not function as peptide presenters and are not regularly tested for. Consequently, these loci have received the name of *non-classical HLA genes*.

Each locus of the HLA region presents a significant number of variants in their amino acid sequences. Variants in the DNA sequence of a gene are called *alleles*. The presence of HLA alleles differ considerably among populations and their corresponding frequencies are vastly supportive in investigating the function of HLA genes and genetic differences among individuals (See Section 1.4.4).

In the beginnings of the study of HLA, *antigens*² were identified using antisera (blood serum containing antibodies) in which the number of variants were encoded using a serological nomenclature [See review (McCluskey, Kanaan & Diviney 2003)]. However, with the advent of molecular techniques applied to the HLA typing since the mid 1980s, a massive number of alleles have been reported splitting the serological nomenclature into hundreds of alleles. At present, the officially recognised HLA loci and corresponding alleles are defined by the Nomenclature Committee for Factors of the HLA System (Marsh et al. 2010).

The number of alleles that have been recognised by molecular methods is constantly increasing. More than 6,000 HLA alleles have been reported as of January 2011 at release 3.3.0 on the IMGT/HLA Database (Robinson et al. 2003). The high number of alleles observed gives an indication of the extensive polymorphism that is found in this genomic region, except in non-classical genes in which the polymorphism is low (Table 1.1). Despite the elevated polymorphism observable in the HLA system, the variability is

¹ Locus (plural loci) is a term commonly used to specify the location of the gene on the chromosome.

² Antigens are molecules which activate the production of antibodies in the immune system response.

mainly presented in a short region of nucleotides located principally at exons 2 and 3 for HLA Class I molecules and at exon 2 for the HLA Class II molecule (Little & Parham 1999; Williams 2001). These two exons are known for encoding the peptide binding region.

Table 1.1: Genes and number of alleles in the HLA region

HLA locus	Alleles	HLA locus	Alleles
A	1,518	DRB4	14
B	2,068	DRB5	19
C	1,016	DRB6	3
DMA	4	DRB7	2
DMB	7	DRB8	1
DOA	12	DRB9	1
DOB	9	E	10
DPA1	28	F	22
DPB1	145	G	46
DQA1	35	H	12
DQB1	144	J	9
DRA	3	K	6
DRB1	873	L	5
DRB2	1	P	4
DRB3	52	V	3

Number of HLA alleles as of January 2011 from release 3.3.0 available in the IMGT/HLA database (Robinson et al. 2003). Classical loci are shown in bold.

1.4.2 HLA nomenclature

Since 1968, when the nomenclature to designate the name of HLA alleles was introduced initially, the Committee for factors of the HLA system has produced a series of nomenclature reports (http://hla.alleles.org/nomenclature/nomenc_reports.html) to meet the principles of HLA specificities [For a review of HLA specificities see (Park & Terasaki 2000)]. The current notation used to designate HLA alleles is composed of alphanumeric characters and symbols which are divided into two main components: (i) the name of the locus (e.g. HLA-A, -B, -C, etc.), followed by an asterisk (*), and (ii) the DNA sequence variant (Figure 1.3).

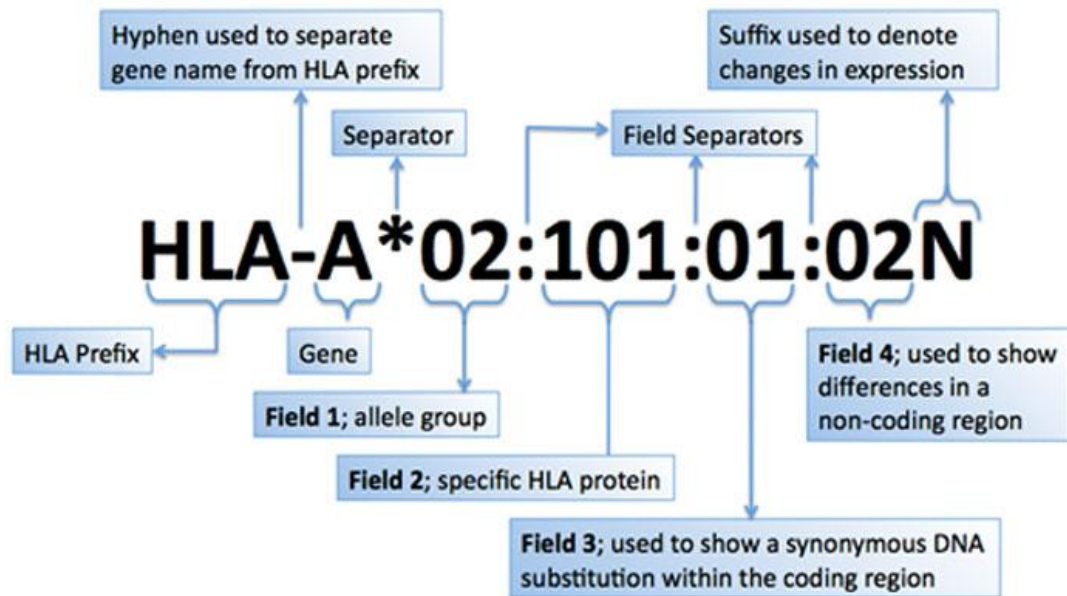


Figure 1.3: Structure of the nomenclature for HLA alleles.

(Reprinted with permission of S.G.E. Marsh)

Source: <http://hla.alleles.org/nomenclature/naming.html>

The notation used to define the DNA sequence variant consists of 2-4 levels (Fields) which are separated by a colon (:), as shown in Figure 1.3. *Level 1* (Field 1) encodes the *HLA allele family* which usually corresponds to the antigen group (e.g. A*02), *level 2* (Field 2) defines the DNA sequence variant at protein level which is typically given in a consecutive numerical order (e.g. A*02:101), *level 3* (Field 3) is used to describe a *synonymous DNA substitution*³ within the coding region (i.e. A*02:101:01 and A*02:101:02 alleles both encode for Arginine at codon 163 which is located in exon 3 in the sequence, however, nucleotide sequences of these two alleles correspond to CGG and AGG respectively) and *level 4* (Field 4) is used for those alleles that present a difference in the *non-coding region*⁴ of the sequence (e.g. A*02:101:01:02) [For more details on HLA allele nomenclature see (Marsh et al. 2010)].

Additionally, several suffixes are used in the HLA nomenclature to indicate special characteristics of the allele indicating whether the allele is not expressed (N), whether the allele has shown a *low expression* on the surface of the cell (L), whether the

³ A synonymous substitution is the replacement of one nucleotide in which the change does not affect the amino acid produced. E.g. CGG and AGG both encode the amino acid Arginine.

⁴ The non-coding region is the part of the DNA sequence that does not encode for protein sequences.

specification of proteins is expressed as *secreted molecule* (S), if the allele is present in the *cytoplasm* (C), if it corresponds to an *aberrant allele* (A) or if the allele is *questionable* regarding its level of expression (Q). These special suffixes are only present in a few alleles (~ 3.4% of the total of HLA alleles) and only in some loci (Table 1.2).

Table 1.2: HLA alleles with a special expression status

HLA locus	N	L	S	C	A	Q	Total
A	72	3				5	80
B	62	1	1			6	70
C	24					9	33
DOA	1						1
DPB1	3						3
DQA1	1						1
DQB1	1						1
DRB1	10						10
DRB4	3						3
DRB5	2						2
G	2						2
Total	181	4	1	0	0	20	206

Number of HLA alleles with suffixes as of January 2011 from release 3.3.0 available in the IMGT/HLA database (Robinson et al. 2003). N=Not expressed, L=Low expression, S=Secreted molecule, C=Cytoplasm, A=Aberrant, Q=Questionable.

Allele detection and groups

HLA alleles can be identified in laboratories using different typing methods based on the use of molecular techniques incorporating the *Polymerase Chain Reaction* (PCR) [See review of PCR in (Bartlett & Stirling 2003)]. The amplification of DNA sequence using PCR protocols has underpinned the improvements of allele detection which were originally identified by antisera [See reviews in (Erlich, Opelz & Hansen 2001; Middleton 2005)]. PCR methods performed in HLA typing include *Restriction fragment length polymorphism* (PCR-RFLP), *Sequence-Specific Primer* (PCR-SSP), *Sequence-Specific Oligonucleotide Probe* (PCR-SSOP), *Single-Strand Conformation Polymorphism* (PCR-SSCP), *Reference Strand Conformation Analysis* (PCR-RSCA) and *Sequence Based Typing* (PCR-SBT). The PCR-RFLP method is used to distinguish two alleles differing by the presence of a restriction enzyme site. The main problem of this method is that a failure to get complete cleavage may lead a homozygote to be considered as a heterozygote. Also, it may be difficult to find appropriate restriction sites in all of the alleles in certain loci. In

the *PCR-SSP* method, primers are designed to target only a specific set of alleles or a single allele. These unique sequences are usually located at the 3' end of the primer. However, this method is not suitable for large number of samples or for automation. The *PCR-SSOP* technique involves the use of single stranded DNA sequence which is used to complement a target sequence (~18 nucleotides). There are many alternatives of SSOP methods which mainly differ on the length and sequence of oligonucleotide probes. This method has been proved to be reliable, robust and accurate, although is not suitable for analysing individual alleles or small numbers. In the case of *PCR-SSCP*, PCR-amplified DNA is denatured and electrophoresed on a polyacrylamide gel. Each single strand moves at a position related to its conformation as determined by its sequence. This method could be useful in comparing the alleles of two individuals; however, this technique is not widely used due to the results may be difficult to interpret. The *PCR-RSCA* is a conformational technique which uses a fluorescein-labelled reference (FLR) in which pairs are formed between the PCR product of the locus of interest and a locus-specific FLR strand. RSCA is a rapid method for high resolution typing of large numbers of samples but limited to proprietary RSCA typing kits. Finally, *PCR-SBT* is based on the amplification of HLA alleles which are sequenced directly to identify the alleles carried by the individual. A sequence-specific sequencing primer is used to produce the sequence of a single allele from a PCR reaction containing both alleles at a locus. In this method, a software program is employed to identify the alleles based on their sequence. [For a review of typing methods see (Little 2007)].

Some of the HLA alleles share identical amino acid/nucleotide sequences over exons 2 and 3 for Class I and exon 2 for Class II. As mentioned in Section 1.4.1, these exons encompass the peptide binding domain. The WHO Nomenclature Committee classifies identical alleles at these exons in two main groups. One group corresponds to those alleles that *encode the same protein* (at exons 2 and 3 for HLA Class I and exon 2 for HLA Class II alleles) and which are grouped using the allele code designation at level 2 followed by the suffix **P** (e.g. A*02:03P) (Figure 1.4). The second group corresponds to those alleles that have *identical nucleotide sequences* (at exons 2 and 3 for HLA Class I and exon 2 for HLA Class II alleles) and which can be grouped using the allele code designation at level 3 followed by the suffix **G** (e.g. A*02:03:01G) (Figure 1.4). In the example shown in Figure 1.4, all six alleles (A*02:03:01, A*02:03:02, A*02:03:03, A*02:03:04, A*02:253 and A*02:264) have identical protein sequence over exons 2 and

3 (Group A*02:03P). However, only A*02:03:01, A*02:253 and A*02:264 have identical nucleotide sequences over these two exons (Group A*02:03:01G).

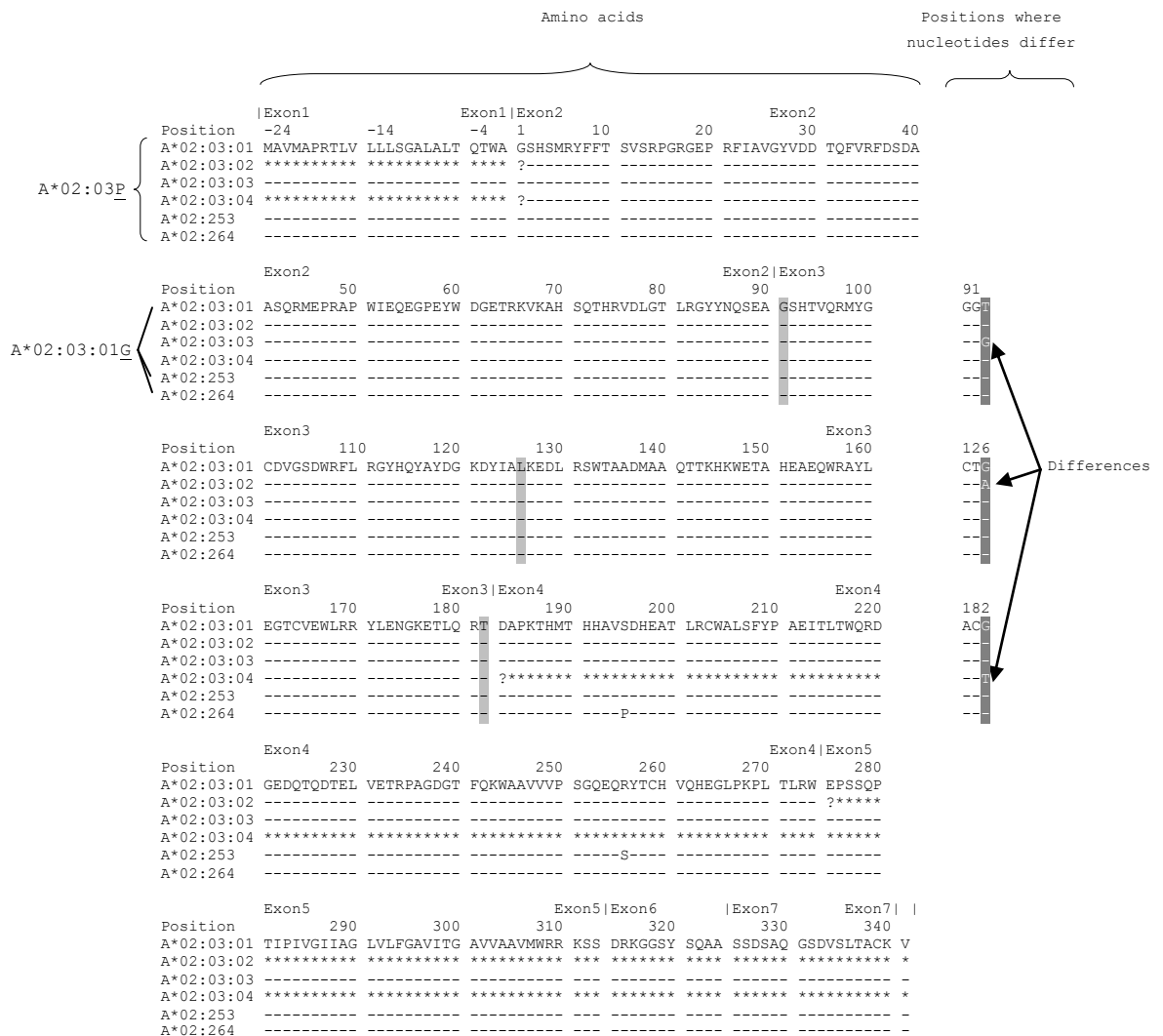


Figure 1.4: Example of identical HLA alleles over exons 2 and 3.

Symbols: (-) Identical sequence to first allele, (*) not sequenced, (?) incomplete sequence.

1.4.3 Genetics of the HLA system

In the HLA system, as elsewhere, individuals inherit two copies of DNA from their parents following the Mendelian law of segregation. Two blocks of alleles called *haplotypes*, are inherited together, one from the mother and one from the father. The set of alleles contained in the two haplotypes constitute the *genotype* of an individual. Figure 1.5 illustrates the total number of combinations of four different haplotypes (a, b, c and d) inherited from parents. Sibling I is shown to inherit haplotype ‘a’ (A*24:02-B*35:01-

C*03:02-DRB1*04:03-DQA1*03:01-DQB1*03:01) from the father and haplotype ‘c’ (A*02:01-B*39:02-C*03:04-DRB1*04:11-DQA1*04:01-DQB1*03:01) from the mother.

Individuals who present the same allele in each of the two copies in a particular locus are called *homozygous*, and *heterozygous* if the alleles are different. For example, in Figure 1.5, Sibling I is homozygous for the locus DQB1 by presenting the same allele (DQB1*03:01) and heterozygous for the rest of the loci whereas the other siblings are heterozygous at DQB1 locus. The inherited HLA alleles are co-dominant, which indicates that both alleles are equally expressed on the cell surface.

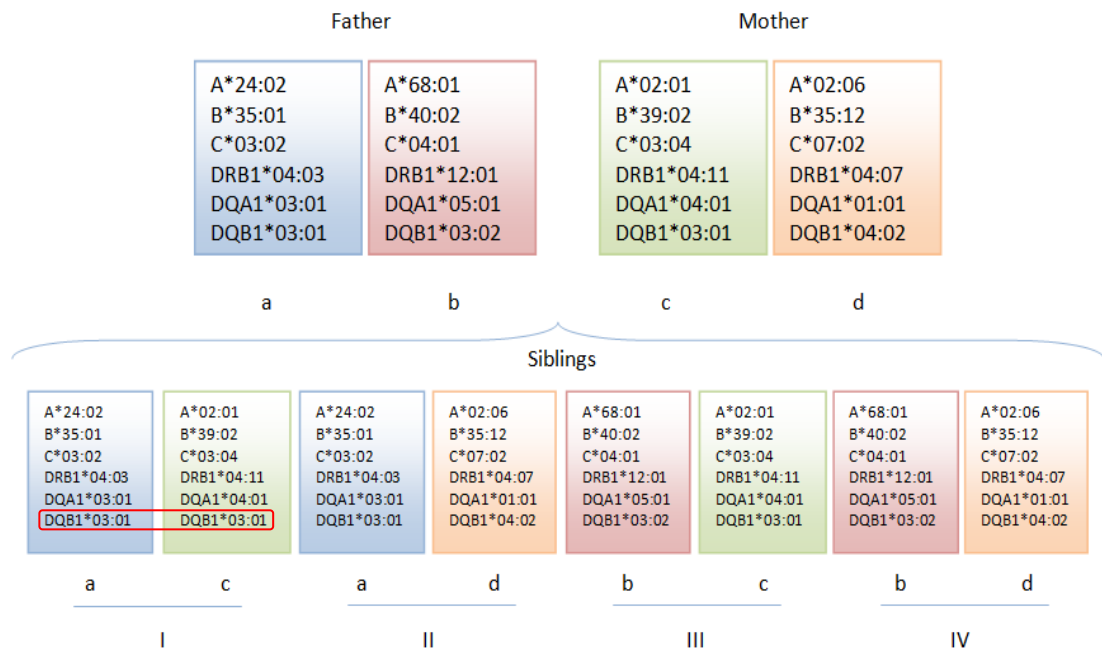


Figure 1.5: Example of inheritance of HLA alleles.

As shown in Figure 1.5, the chance of finding an identical genotype among siblings from the same family is one in four whereas the chance of sharing one of the haplotypes is 50%. Because this information is exceptionally significant for examining common diseases and potential donors in clinical transplantation (Crawford & Nickerson 2005), these data have been investigated by several groups such as the MHC Haplotype Consortium (Horton et al. 2008).

Taking into account the elevated number of existing alleles, the total number of haplotypic combinations may be immense. However, this number of combinations can be reduced as some loci and alleles are in *linkage disequilibrium* (LD) which, contrary to basic Mendelian genetics, states that those combinations have frequencies considerable higher than the frequencies theoretically expected knowing the individual allele frequency [See review of LD in (Slatkin 2008)]. The patterns of LD have been found to vary between populations and are of great interest for immune-related disease and population studies (Begovich et al. 1992; Evseeva et al. 2010; Sanchez-Mazas et al. 2000).

1.4.4 Applications of the HLA system

HLA and transplantation

The major applications of the HLA system have been in the field of tissue and organ transplantation. Originally the MHC was discovered as the region of the human genome which was the determinant for the *histocompatibility* between donor and recipient in tissue and organ graft. Since then, many studies and methods have been performed to examine the chances of survival after the transplant (Chinen & Buckley 2010; Sheldon & Poulton 2006; Terasaki & Cai 2008).

In some cases, a mismatch on the HLA genotype of the donor and recipient may cause the immune system of the donor to reject the transplanted organ or tissue. Recent studies have concluded that the matching of some loci is more important than others. For example, antigen matches of HLA-A, -B, -C and -DR loci are believed to be more crucial than mismatches in other loci (Sheldon & Poulton 2006). In the case of *Bone Marrow Transplantation* (BMT), the incompatibility not only affects the rejection but also may lead to a *Graft-Versus-Host-Disease* (GVHD) in which the recipient is attacked by the grafted bone marrow (Shlomchik 2007). In this case, complete HLA matching is normally needed.

As mentioned in the previous section, the chance of finding an identical HLA genotype among individuals from the same family is one in four. However, it is believed that only

less than a third part of the individuals who need a transplant have an HLA identical sibling (Petersdorf 2008). At present, computational matching algorithms and searching mechanisms are performed in Cord Blood and Bone Marrow Registries which contain large number of volunteer unrelated potential donors (Eberhard et al. 2010). One of the facts to consider is that if a match in two loci (e.g. HLA-DRB1, HLA-B) is found the chances for matching a third locus (e.g. HLA-C) are high, based on the LD that some of these loci present.

HLA in autoimmune and infectious diseases

Infectious diseases

The HLA system has also been recognised by its capability to offer some resistance to a wide number of infectious diseases (Blackwell, Jamieson & Burgner 2009). It is believed that human beings have acquired protection to diseases due to an evolutionary process called pathogen-driven balancing selection (Prugnolle et al. 2005). The HLA system has thus worked in conjunction with environmental factors to develop this defence through the adaptive and innate response (Traherne 2008).

One of the most common practices performed by researchers in infectious disease associations is to compare frequencies and/or genetic linkage of HLA alleles in patients against the corresponding frequencies in controls. For instance, carrying a particular combination of HLA variants may lead an individual to have a greater susceptibility to a given infection, e.g. malaria which has been associated with the absence of the HLA-B*53 allele, suggesting that the presence of this allele may reduce the risk of severe malaria by 40% (Hill et al. 1991). These association analyses are usually followed by a set of guidelines which involve the identification of an appropriate selection of individuals (controls) such as matching for ethnic background of individuals, selection of an optimal population/sample size, etc. (Balding 2006; Hattersley & McCarthy 2005). Thus, the selection of suitable parameters is essential in the quality control of information in disease associations.

Autoimmune diseases

On some occasions, individuals may present some alleles or haplotypes which may be linked to a particular disease (Klein & Sato 2000b). Although other factors such as environmental conditions and/or involvement of other non-HLA genes are clearly important, some HLA alleles have a significant correlation with some specific autoimmune diseases, for example, Type 1 Diabetes (T1D) in which some studies have shown a major predisposing factor in HLA genes and alleles implicated in the disorder [See reviews in (Bluestone, Herold & Eisenbarth 2010; Thorsby & Lie 2005)].

HLA and diversity

Despite the short length of the HLA system ($\sim 2.5 \times 10^{-3}$ percent of the total length of the human genome), this genomic region is the most polymorphic among individuals and thus has captured the attention of anthropologists due to the extensive diversity presented among populations (Sanchez-Mazas 2007; Shiina et al. 2009).

Anthropology studies have focussed on the analysis of allele frequencies and haplotypic variations in different geographic regions to predict and trace historical human migration. As allele and haplotype frequencies differ between groups of individuals, the HLA system has been extensively used in the characterisation of different ethnic groups [See examples in (Arnaiz-Villena et al. 2005; Begovich et al. 2001; Hollenbach et al. 2001; Yuliwulandari et al. 2009)]. The analyses performed in these populations include the comparison of genetic distance among populations, the representation of hierarchical structures by the creation of dendrograms⁵ and phylogenetic trees⁶, principal component analysis⁷ (PCA) for human genetic clustering⁸, among many others.

In recent years, the genetic variation of human beings has been analysed at different levels: within individuals from the same population, among populations within the same

⁵ Dendrograms are diagrams used to illustrate the genetic distance from populations.

⁶ Phylogenetic trees are diagrams to infer evolutionary relationships among species.

⁷ Principal component analysis is a statistical technique to identify correlations between different groups. For instance, in HLA, PCA are used to analyse correlations between populations.

⁸ Human genetic clustering refers to the similarity between individuals to infer population structures.

continent, and among different continents. The outcomes of HLA population analyses have assisted researchers in the confirmation of previous analyses performed using other DNA markers. In the case of HLA, approximately 84-92% of the variability (depending on the HLA locus examined) is presented by individuals from the same population, 4-6% among populations from the same continent and 3-9% among populations from different continents (Sanchez-Mazas 2007).

1.5 Killer-cell Immunoglobulin-like Receptors

The *Killer-cell Immunoglobulin-like Receptors* (KIR) constitute a group of proteins which are present on the cell surface of Natural Killer (NK) cells and in some T cells (Parham 2005a). The main task of these receptors is to participate in the regulation of the killing function against infections and malignancy by interacting with MHC Class I molecules (Kumar & McNerney 2005; Parham 2005b).

1.5.1 Genes and alleles in the KIR family

The KIR family consists of fifteen genes and two pseudogenes⁹ clustered on the *Leukocyte Receptor Complex* (LRC) residing on the long arm of chromosome 19 at position 19q13.4 (Wende et al. 1999). These receptors may be *activating* (*KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5* and *KIR3DS1*) or *inhibitory* (*KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL5A*, *KIR2DL5B*, *KIR3DL1*, *KIR3DL2* and *KIR3DL3*) based on the activating role on the function of the NK cell, except *KIR2DL4* which appear to have both functions (Middleton, Curran & Maxwell 2002).

KIR genes are also inherited as blocks of genes. According to recent studies, each individual may carry from seven to eleven different genes (Middleton, Meenagh & Gourraud 2007; Shilling et al. 2002; Uhrberg, Parham & Wernet 2002). Because of the high variability in the gene content and the allelic polymorphism found in this genomic region, individuals rarely present identical KIR genotypes (presence or absence of the KIR genes in an individual).

⁹ Pseudogenes are genes that lack expression in the cell.

The KIR family also presents a high level of polymorphism derived from presence or absence of genes, although not at the same level as the HLA region. A total of 601 KIR alleles have been reported as of release 2.4.0 April 2011 in the IPD-KIR database (Robinson et al. 2005). *KIR3DL1*, *KIR3DL2* and *KIR3DL3* are the most polymorphic genes with 70, 84 and 101 alleles respectively (Table 1.3).

Table 1.3: Genes and number of alleles in the KIR family

KIR locus	Alleles	KIR locus	Alleles
<i>KIR2DL1</i>	43	<i>KIR2DS1</i>	15
<i>KIR2DL2</i>	27	<i>KIR2DS2</i>	22
<i>KIR2DL3</i>	32	<i>KIR2DS3</i>	13
<i>KIR2DL4</i>	47	<i>KIR2DS4</i>	30
<i>KIR2DL5A</i>	15	<i>KIR2DS5</i>	16
<i>KIR2DL5B</i>	25	<i>KIR3DS1</i>	16
<i>KIR3DL1</i>	70	<i>KIR2DP1</i>	22
<i>KIR3DL2</i>	84	<i>KIR3DP1</i>	23
<i>KIR3DL3</i>	101		

Number of KIR alleles as of April 2011 from release 2.4.0 available in the IPD-KIR Database (Robinson et al. 2005).

1.5.2 Organisation of the KIR genes

In the last decade, many studies have been performed to analyse the organisation of the KIR genes. Researchers have classified genes in two main groups ‘A’ and ‘B’ according to the haplotype content, originally based on the presence or absence of a 24kb *HindIII* restriction enzyme fragment (Uhrberg et al. 1997). The basis of each haplotype group consists of four genes *KIR3DL3* and *KIR3DP1* located at the centromeric¹⁰ side and *KIR2DL4* and *KIR3DL2* at the telomeric¹¹ side, which are considered to be *framework genes* and always present (Figure 1.6). Group A consists of the presence of *KIR3DL1*, *KIR2DL1*, *KIR2DL3* and *KIR2DS4* genes whereas group B is composed of the presence of one or more of the following genes: *KIR2DL2*, *KIR2DL5*, *KIR3DS1*, *KIR2DS1*,

¹⁰ Centromere is the region of DNA situated near the middle of the chromosome and which is involved in cell division.

¹¹ Telomere is the region of DNA located at the end of the chromosome and which protects from deterioration or fusion with other chromosomes.

KIR2DS2, *KIR2DS3* and *KIR2DS5* [For KIR haplotypes see review in (Middleton & Gonzalez 2010)].

According to the A and B haplotype classification, the resulting haplotype groups can be categorised as AA or Bx (in which x can be either A or B) as defined by McQueen and colleagues due to the difficulty of distinguishing the other A or B haplotype without familial data (McQueen et al. 2007). The terms AA and Bx will be used throughout this thesis for referencing KIR haplotype groups.

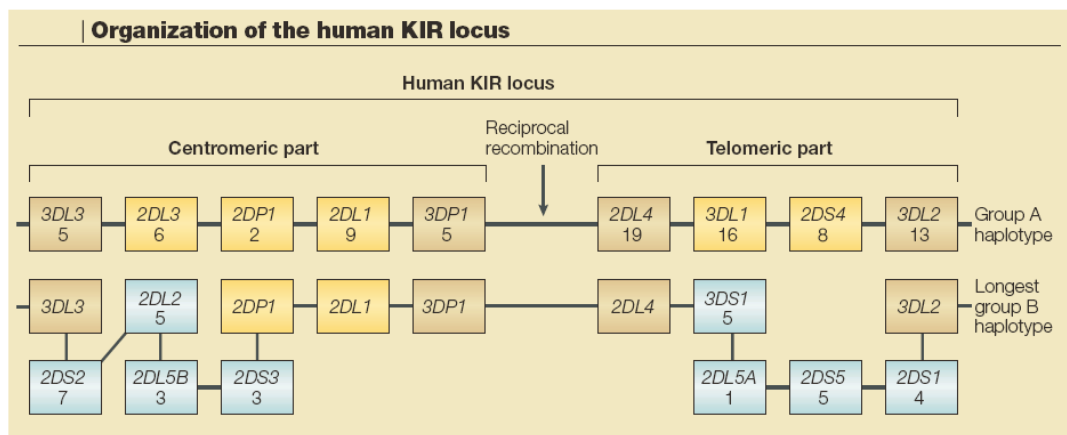


Figure 1.6: Organisation of the KIR genes.

In this figure, KIR genes are classified by centromeric and telomeric side. Two examples of haplotypes are shown in the figure. Framework genes are shown in brown. Genes specific to B haplotypes are shown in blue. Genes that are present in both groups are shown in yellow (Parham 2005b).

Based on the number of genes in each group (as shown in Figure 1.6), the number of haplotype combinations containing B genes is expected to be greater than A subsets (Hsu et al. 2002a). In recent studies, more than twenty different haplotypes have been identified by family studies. With these haplotypes, a high number of genotypes can be generated by haplotype permutations (Hsu et al. 2002b).

1.5.3 KIR nomenclature

The KIR nomenclature guidelines are defined by the Human Genome Organisation (HUGO) via the Genome Nomenclature Committee (Marsh et al. 2003). Following a similar notation as used for HLA allele designations, the KIR nomenclature is

composed of two main components: (i) the gene name and (ii) the code for the allele variant which are separated by an asterisk (*) (Figure 1.7).

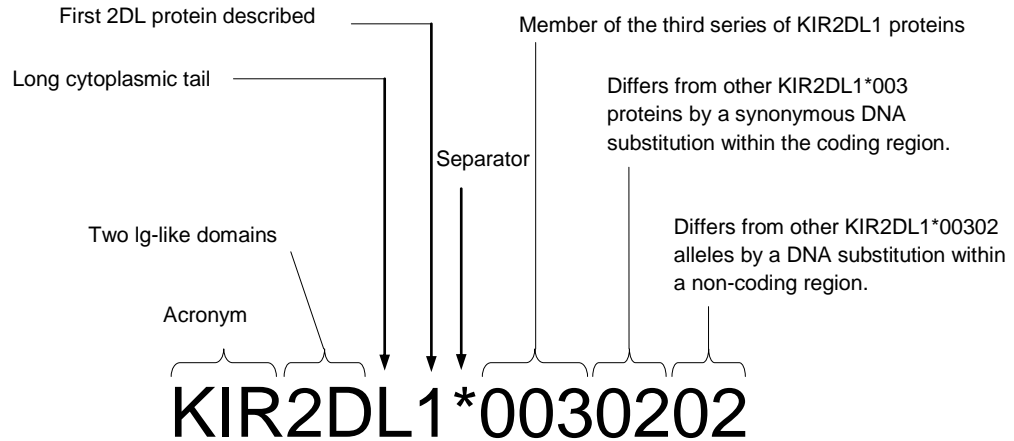


Figure 1.7: Structure of the nomenclature for KIR alleles.

(Reprinted with permission of S.G.E. Marsh)

Source: <http://www.ebi.ac.uk/ipd/kir/alleles.html>

The gene name is formed by the acronym of the polymorphic region (*KIR*) and the name of the corresponding gene (e.g. *2DL1*, *2DL2*, etc.). The name of the gene is composed of four to five alphanumeric characters which define the characteristics of the gene. The first character represents the number of immunoglobulin-like domains (2 or 3 domains), the second denotes the prefix of the word domain (D), the third character indicates whether the gene has a long (L) or short (S) cytoplasmic tail, or if it is a pseudogene (P). Long tail codes are commonly associated with inhibitory genes and short tail codes with activating genes, except *KIR2DLA* which appear to have both functions as mentioned in Section 1.5.1. The fourth character describes the number of the gene in consecutive numerical order which is assigned by the GenBank database (Benson et al. 2008) and a suffix 'A' or 'B' is added to distinguish genes that present similar structures and sequences, e.g. *2DL5A* and *2DL5B* (Marsh et al. 2003).

Finally, the KIR allele variant is composed of three to seven digits. The first three digits represent the sequence that encodes a specific protein (e.g. *KIR2DL1*003*), the next two digits are used to distinguished differences (synonymous nucleotide substitutions) within the coding region (e.g. *KIR2DL1*00302*), and the last two digits are used to denote differences which are located in the non-coding region (e.g. *KIR2DL1*0030202*).

1.5.4 KIR and HLA ligands

The products of certain HLA Class I loci (principally HLA-B and -C) and the products of some KIR genes present specific interactions and have been studied by immunologists in recent years.

The epitope Bw4 in the HLA-B locus is believed to function as a ligand with the *KIR3DL1* gene (Trowsdale 2001). In the case of the HLA-C locus, the subsets are divided into C1 and C2 epitopes whose differences are found at positions 77-80 in the amino acid sequence. *KIR2DL1* interacts with the C2 epitope (covering alleles under the C*02, *04, *05, *06 allele families) and *KIR2DL2/3* with the C1 epitope (for C*01, *03, *07, *08 allele families) (Trowsdale 2001).

These interactions are suggested to influence resistance to infections, susceptibility to autoimmune diseases, pregnancy and success in haematopoietic stem cell transplantation (Parham 2005b).

1.5.5 Applications of the KIR family

The investigation of the genes of this genomic region and their organisation has assisted scientists in the understanding of the functioning of these genes and their interaction with other molecules such as the HLA system.

From these studies a number of important topics have been under investigation by researchers. For example, the presence of activating KIR genes in individuals is believed to protect against infections, but has also been linked to the susceptibility to certain autoimmune diseases (Single et al. 2007a). In contrast, the presence of inhibitory KIR genes is believed to protect against inflammatory diseases (Single et al. 2007a).

In population genetics, the diversity present in the genes of the KIR family has been studied by anthropologists to understand theories of human migration (Rajalingam et al. 2008). The analysis of KIR gene/allele frequencies and the correlation of KIR and HLA ligands in different populations have provided evidence for coevolution of these

genomic regions (Norman et al. 2007; Single et al. 2007a). The number of different analyses has also been extended to include a range of populations among regions and continents (Middleton et al. 2008).

1.6 Other immune genes

1.6.1 Major histocompatibility Class I chain-related genes

The *Major histocompatibility complex Class I chain-related* (MIC) genes code for a set of proteins expressed on the surface of tissue cells (epithelial and endothelial) and fibroblasts (Bahram 2000). The MIC genomic region comprises seven loci located on chromosome 6 at position 6p21.33 in which MICA and MICB are the only genes that encode for proteins (Bahram 2000). MICA and MICB are composed of six exons encoding three extracellular domains ($\alpha 1$ - $\alpha 3$), a transmembrane region and a cytoplasmic tail (Bahram 2000). MIC proteins are known for presenting ligands to the Natural Killer activating receptor NKG2D and their interactions are believed to play an important role in the immune response to infectious agents [See review in (Collins 2004)]. MICA and MICB loci are located 46 and 141kb respectively from the HLA-B locus. Thus, MICA gene has been associated to some diseases and tissue rejection in transplantation possibly only as a result of its proximity and high linkage disequilibrium with the HLA-B locus (Collins 2004).

MICA and MICB loci also present a considerable number of allelic variants although much lower than those described in the HLA and KIR systems. Similarly to HLA, MICA and MICB alleles are officially named by the WHO Nomenclature Committee and sequences are deposited in the IMGT/HLA database (Robinson et al. 2003). As of release 3.3.0 (January 2011), 73 MICA and 31 MICB alleles had been reported in the IMGT/HLA database. The nomenclature used for allele designations of MIC genes follows a similar fashion as the HLA nomenclature. The code is divided into several components to describe the characteristics of the allele (e.g. MICA*002:01).

Prior to the official nomenclature of MICA, alleles were classified into eight variants (A4, A5, A5.1, A6, A7, A9, A9.1, A10) based on the number of Alanine repeats (GCT)

between codons 291-304 located at exon 5, which encodes for the transmembrane domain [See review in (Collins 2004)]. A5.1 and A9.1 alleles, which correspond to mutational variants, represent an insertion after the second Alanine repeat and a deletion at the start of exon 5 respectively. The presence of these eight allelic variants, also known as MICA microsatellites¹² or short tandem repeats (STRs), has been shown to vary among populations leading the interests of investigators to examine their frequencies among different ethnic groups.

The possibility that the diversity present in the MICA gene may be involved in the increase in graft survival after transplantation has been under investigation by researchers in recent years (Zou & Stastny 2009). The analysis of several diseases has also been extended to investigate the relationships between MICA and HLA-B due to the proximity of these two loci in the chromosome [See review in (Stephens 2001)].

1.6.2 Cytokine gene polymorphisms

Cytokines are a set of proteins which participate in the immune response against pathogens by modulating the activities of target cells. These proteins are released by several types of cells such as neutrophils, lymphocytes, monocytes, macrophages, NK cells, etc. by different means (Bidwell et al. 1999). Several cytokines have been demonstrated to present genetic polymorphism in specific regions of their DNA sequences [See review in (Bidwell et al. 1999)]. The variation, which is mainly presented in *single nucleotide polymorphisms* (SNPs) or in microsatellites, may affect the gene transcription and cause variations in cytokine production. For many years, investigators have studied the relationships between cytokine gene polymorphisms and their role in transplantation, autoimmune diseases and tumours [See review in (Dinarello 2007)].

Unlike HLA, KIR and MIC, the nomenclature of cytokines is not regulated by official bodies or online databases. In the last decade, a series of updates to list current cytokines were reported in an online database (Bidwell et al. 1999, 2001; Haukim et al. 2002; Hollegaard & Bidwell 2006). This online resource consisted of an extensive review

¹² Microsatellites are DNA repeat sequences of 1 to 6 base pairs which are used as biological markers in genetic studies.

of bibliographic references featuring the relationships between cytokine gene polymorphisms, cytokine gene expression and the susceptibility to several diseases. However, the online database which merely provided a list of references has been discontinued in recent years. The notation commonly used by individuals to describe cytokine gene polymorphisms is primarily composed of the gene name and the specific location of the gene that participates in the regulation. For example, 'IGF-1/-383' represents the nucleotide position on the Insulin-like Growth Factor 1 (IGF-1) gene that presents the polymorphism.

As the level of expression and frequencies of certain cytokines may lead to different diseases, allele and genotype frequencies in different populations have been under investigation by immunologists (Larcombe et al. 2005). The study of cytokine gene polymorphisms has been also extended to populations from different continents (Middleton et al. 2002).

1.7 Bioinformatics and immunogenetics databases

1.7.1 Evolution of biological databases

Since the introduction of the term *bioinformatics* by Hogeweg and colleagues to describe simulations of biological systems (Hogeweg 1978), the massive number of software applications in biological research has facilitated scientific analysis and retrieval of information (Ouzounis & Valencia 2003).

The branch of bioinformatics encompasses a wide number of different areas involving not only computational simulations and statistical methods but also the development of sophisticated databases and searching algorithms for the retrieval of information from biological datasets (Maier et al. 2009).

Biological databases are a set of electronic documents (e.g. raw text files, dictionaries, libraries, etc.) which contain information of experiments or computational simulations performed by scientists from a different range of areas such as genetics, genomics, proteomics, etc. The number of different biological repositories reported in different

sources is immense. For example, the Database issue of the Nucleic Acid Research journal contains more than 1,300 databases reported as of December 2010 and this only represents a small portion of all repositories available (Galperin & Cochrane 2011). The list of repositories includes data on a diverse number of types and formats. For instance, the GenBank (Benson et al. 2011), the EMBL Nucleotide Sequence database (Kulikova et al. 2007) and the DNA Data Bank of Japan (Tateno et al. 2002) were designed for the storage of sequences; UniProt (Apweiler & Consortium 2010) for the hosting of protein sequences and annotations; the Protein Data Bank (Rose et al. 2011) to examine three dimensional protein structures, among many others.

The creation of biological databases involves a series of processes including the design of an adequate structure for the storage of the data (Birney & Clamp 2004). Generally, these databases are developed with additional applications to execute specific analyses and utilities for querying data. The combination of databases and software applications are the fundamentals of the *biological repositories* which are becoming more popular and more frequently accessed by researchers.

The need to store biological datasets has encouraged bioinformaticians to create electronic repositories as a tool to assist researchers in the understanding of biological processes, gene structures and molecular interactions. However, in recent years, bioinformaticians have had to deal with large volumes of data, reaching in many cases dimensions of terabytes (TB = 2^{40}) and petabytes (PB = 2^{50}). As the growth in storage is constantly increasing, systems must be designed in such a way that new and heterogeneous data can be easily incorporated (Shah et al. 2008).

1.7.2 Architecture of databases

Based on its architecture, databases are generally composed of three main layers: internal, external and conceptual levels. These layers will help the reader in the understanding of the design of the database and the techniques and methods to access the information. In a database, the *external level* describes the organisation of the data and controls the visualisation of specific sections in a set of interfaces called *views*. Usually this information is presented to end users. The *conceptual level* depicts the relationships

and directions (flow of information) among datasets. This layer defines constraints and rules which are applied to the datasets. The *internal level* is used to define how the information is physically stored and processed on the computer. In this level, a description of type of data, path, directories, etc., are defined by the database designer.

1.7.3 Database modelling

In database development, informaticians are assisted by automated software applications which provide useful support in the design and maintenance of the database. The software used to perform the maintenance is called the *Database Management System* (DBMS). Nowadays, DBMS applications can support a variety of databases models of which the most commonly used by programmers are relational and object-oriented models. Some examples of DBMSs available are Microsoft Access, Microsoft SQL Server (MSSQL), Oracle, MySQL and PostgreSQL, the last two under the General Public License (GPL) and MIT-style license respectively.

Relational databases

Relational databases (RDBs) are models which consist of a collection of entities (*tables*) and their corresponding relationships that are used to define the organisation and flow of information among different tables based on a *relational model* (RM) (Figure 1.8). The RM is generally preceded by the design of a conceptual representation called the *Entity-Relationship Diagram* (ERD). ERDs are created to specify the relationship between different entities using an abstract model. Each table in an ERD contains a set of *attributes* (columns) which describe the characteristics of the entity. Tables represent the core of the RM structure and generally contain many *tuples* (rows) which share the same attributes. Tuples, commonly known by programmers as *records*, can be distinguished by one or several attributes denominated *keys*. Therefore, keys have the capability to uniquely identify a record within a table.

For example, in Figure 1.8, the table **locus** contains information of the loci available in the polymorphic region HLA (HLA-A, -B, -C, etc.). In the example, the attributes **loc_region** and **loc_name** are used to differentiate each record; therefore, these

constitute the key of the table. Chapter 2 will describe the list of tables, relationships and attributes that were generated to define the AFND.

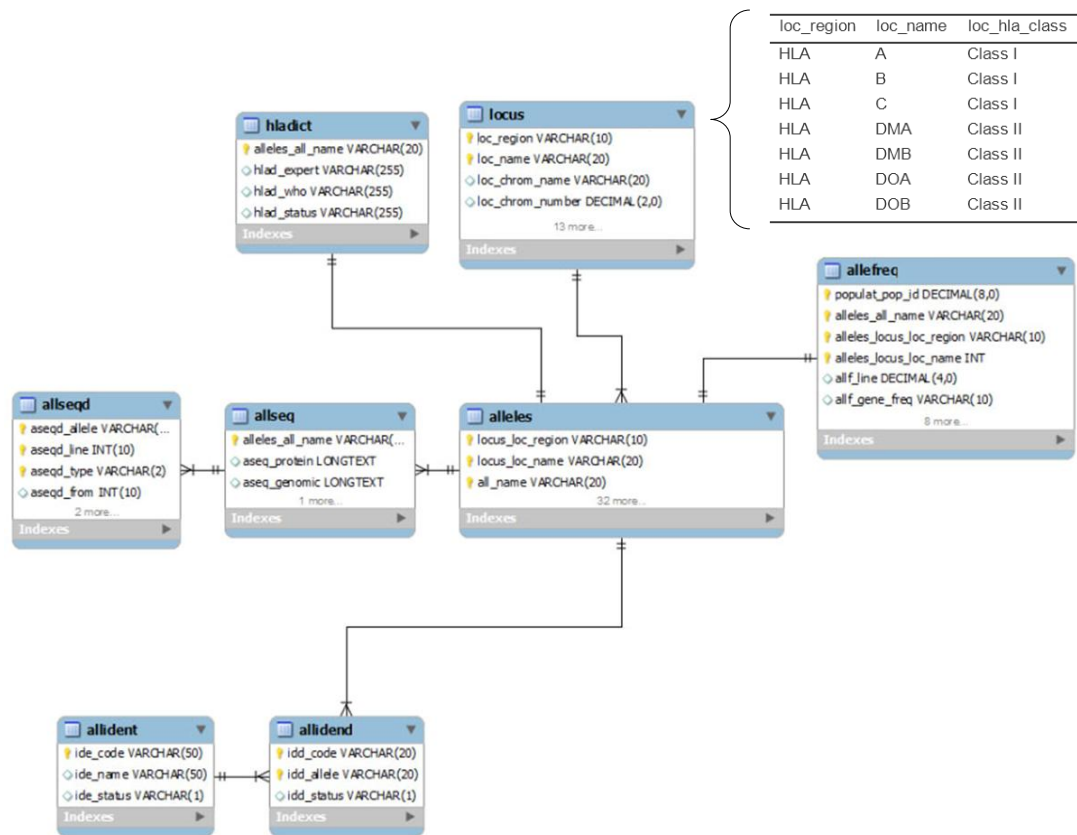


Figure 1.8: Example of a relational model.

Object oriented databases

Object oriented databases (OODBs) are models which follow a different concept of those used in RDB designs. An OODB consists of a collection of *objects* which, in the majority of cases, describe an object of the real world (Figure 1.9). OODBs, which were introduced since the early 1980s, are very well known by the use of concepts based on software reusability. These databases were originally created to facilitate the interaction with Object-Oriented Programming (OOP). In OOP, objects sharing similar characteristics are classified in *classes* and can contain a set of *attributes* and *methods*. Attributes listed in each class are used to describe different features of the object. Methods are subroutines which contain a set of instructions for the object. For instance, in Figure 1.9, **loc_name** and **loc_chrom_position** are attributes of the class **locus**. This class can contain a set of methods to access the information or perform specific

operations. In the example, `setLoc_name()` is used to allocate the name of the locus and `getLoc_name()` to access the value.

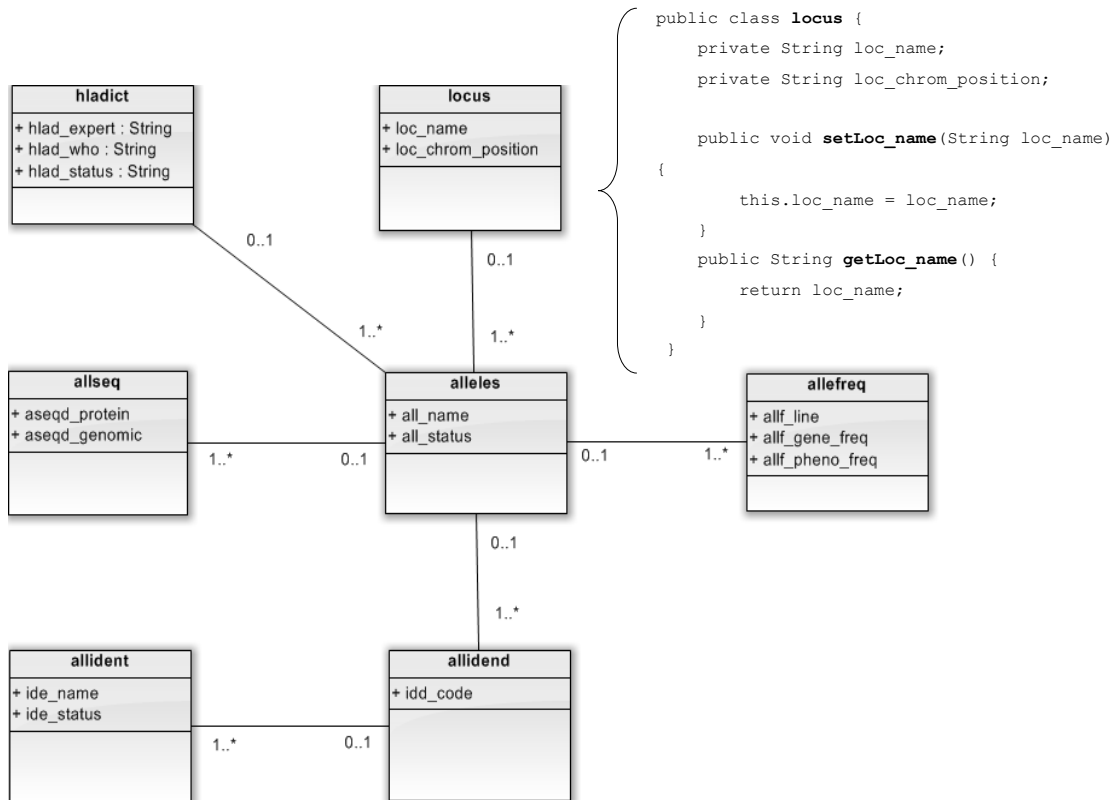


Figure 1.9: Example of an object-oriented model.

Since the inception of OODBs, a broad number of notations to define the relationships and flow of information among objects have been described in the literature. One of the notations most commonly used by software designers is the Unified Modelling Language (UML) which is widely used in industrial and scientific software development.

1.7.4 Information retrieval

The Structured Query Language (SQL) in biological databases

Nearly all DBMS systems provide utilities for information retrieval, giving the facility to programmers to perform both simple and complex queries by using a standard notation called the *Structured Query Language* (SQL). SQL is a syntactic-semantic computational notation which was designed to facilitate the maintenance of the information within a

database. The fundamentals of SQL are based on relational algebra using *Boolean logic* (logical operations) to allow individuals to manipulate specific sections of information by the use of predefined *reserved word* or instructions (e.g. SELECT, UPDATE, INSERT, DELETE, etc.).

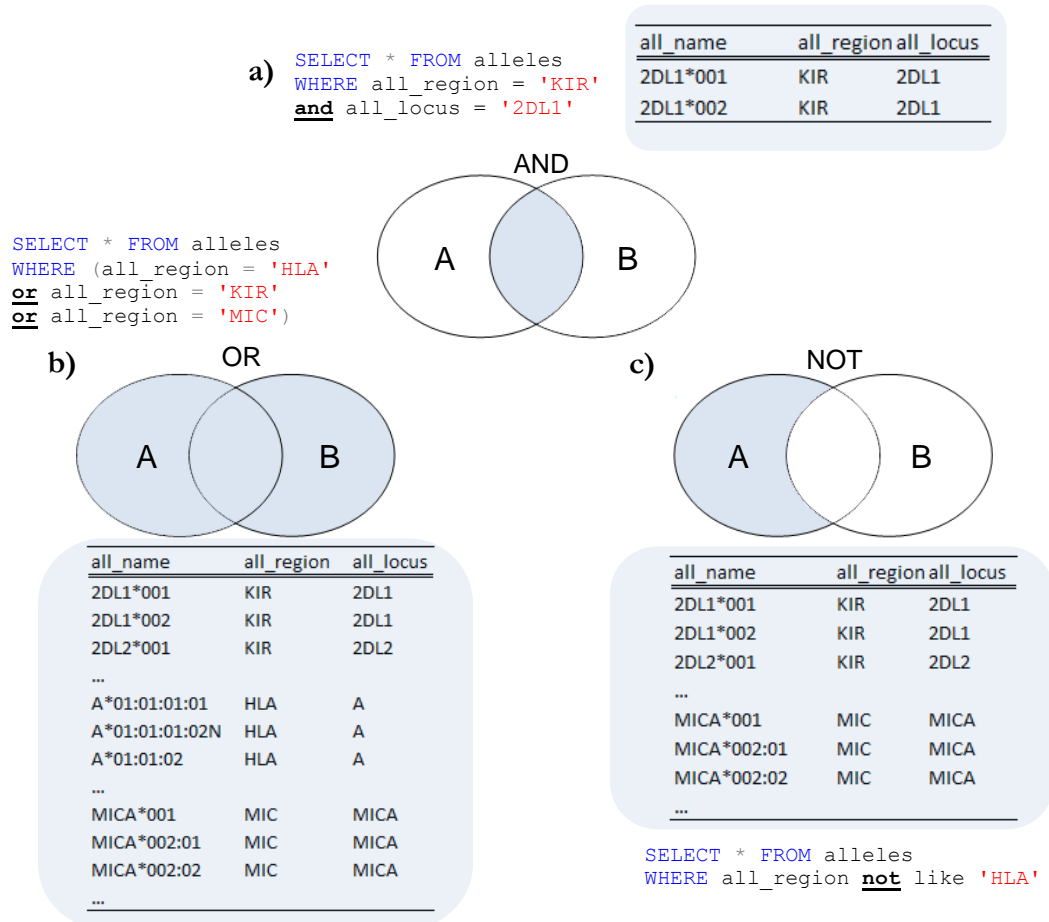


Figure 1.10: Example of Boolean algebra using AND, OR and NOT operators.

Figure 1.10 illustrates three different set of records that are retrieved after querying the table alleles using Boolean logic (AND, OR and NOT operators): (a) selects those alleles which belong to the KIR polymorphic region in locus 2DL1 only, (b) shows all alleles from the HLA, KIR and MIC polymorphic regions, and (c) lists all alleles of the existing polymorphisms except HLA alleles. With the use of relational algebra, programmers can easily add, remove or alter records in a database and retrieve a set of records that match the given criteria.

1.7.5 Web applications and technologies

In the last decade, bioinformaticians have been required to produce applications that can be accessed by a wider audience of researchers across diverse disciplines from different geographic locations. To do this, programmers have designed applications which can be accessed through the World Wide Web (WWW). *Web applications* can be accessed via a private (intranet) or public (internet) network with the use of a software interface installed on a client machine. Usually, this interface is provided by *web browsers* which have the capability of operating in a variety of different platforms, i.e. operating systems, system architecture, etc., without the need to install additional software for each application.

In biological research, websites can contain a number of different components including databases, electronic documents (e.g. text files), web pages, etc., which are accessed by different protocols such as the *Hypertext Transfer Protocol* (HTTP), the *Simple Object Access Protocol* (SOAP) and the *File Transfer Protocol* (FTP). The access to a particular website is given by entering the *Uniform Resource Locator* (URL) in a web browser which can locate the specific server in the worldwide network. Web pages can be classified according to their content in two categories: ‘dynamic’ and ‘static’. *Static websites* are typically used to display information whose content is considered to be fixed over time. On the contrary, *dynamic websites* provide users with the capabilities of performing queries and interacting with datasets on the hosting machine. Dynamic pages are generally supported by one or several programming languages running on the server.

1.7.6 Programming languages in bioinformatics

Since the beginnings of software development in biological research, bioinformaticians have used a massive number of different programming languages to produce their applications. Normally, programmers select a particular language based on the needs of the application, i.e. specific platform (operating system), time response, scalability, capabilities of software and hardware, interoperability, etc. (Dudley & Butte 2009). For example, applications written in C++ are considered to be the fastest programs due to an optimal use of Random-Access Memory (RAM) (Fourment & Gillings 2008). However, other languages such as Python, Perl and Java are commonly used in bioinformatics applications.

As of May 2011, the list of different programming languages encompassed more than 2,500 (Kinnersley 2011) covering a wide range of diverse methodologies (i.e. procedural, structured, object oriented, etc.). In web development some of the most common languages used are Perl, PHP, Java and *Active Server Pages* (ASP) which are generally compiled and executed on the hosting machine.

Web programming languages interact with other technologies and/or scripting languages such as *JavaScript* to provide end users with complementary tools to improve performance and user interaction. JavaScript is included in all modern browsers (e.g. Internet Explorer 8 ©, Firefox 3 ©, Google Chrome 7 ©, Opera 10 ©, Safari 5 ©, etc.) and is found in the majority of the current biological databases.

Since the introduction of dynamic websites, web developments have involved the combination of different technologies on both client and server sides. For instance, technologies such as the *Asynchronous JavaScript and XML language* (AJAX) have been used on the client-side to provide a dynamic interface between databases and the client machine without interfering in the display of the information of a given page. These technologies have also been used in conjunction with other electronic documents for standardisation, for example XML for data exchange and implementation of specific applications.

1.7.7 Data exchange and formats

One of the most important issues to consider in the development of biological repositories is the design of layers for data exchange which allow interoperability among different electronic resources. The majority of the databases are composed of diverse type of files for data transportation such as tab-separated values (TSV), comma-separated values (CSV) text files, Microsoft Excel spreadsheets (XLS), and in recent years, the *Extensible Markup Language* (XML). XML has been adopted by programmers as a standard for data exchange (Achard, Vaysseix & Barillot 2001) following the specifications provided by the *World Wide Web Consortium* (W3C). The XML syntax consists of a representation of a set of nodes called *elements* which are defined in a

hierarchical structure (Figure 1.11). Each node can contain a series of attributes and other nested elements.

Although the XML syntax can be considered as a verbose format, it is considered as the most optimal solution to validate the content of exchange documents. The validation is performed by the use of the *Document Type Definition* (DTD) and *XML Schemas*. In recent years, XML has been used in different software applications and technologies including SOAP, *Really Simple Syndication* (RSS) and web services.



Figure 1.11: Example of XML document.

The use of XML formats has facilitated the standardisation of complex datasets. One example of the application of these documents has been in the field of Proteomics. The Human Proteomics Organisation (HUPO) has adopted this syntax as a technique to implement their standards which have been developed by the Proteomics Standard Initiative (HUPO-PSI) (Orchard, Hermjakob & Apweiler 2003).

1.7.8 Curation of biological datasets

The development of biological databases involves not only the compilation and display of the information but also the inclusion of mechanisms to ensure that the quality of the

information provided can be continuously maintained. This is probably one of the most crucial steps in the designs of the biological databases. Thus, the importance of *data curation* has led developers to generate applications that can be easily validated by database administrators and end users.

Usually, data validation is performed by biocurators whose main roles are (i) to ensure that data is correctly extracted from different sources (e.g. peer-reviewed publications, external databases, etc.), (ii) validate the correct use of controlled vocabulary and (iii) provide facilities to interact with other researchers in the review process (Howe et al. 2008).

1.7.9 Immune gene databases and bioinformatics resources

In recent years, a considerable number of databases and bioinformatics applications for the investigation of the functions of immune genes have been published in the literature and set up online (Galperin & Cochrane 2011). The repositories available provide capabilities for a wide range of analyses as mentioned in Section 1.7.1. In the case of immunogenetics, the different resources include the IMGT/HLA (Robinson et al. 2009) and IPD-KIR databases (Robinson et al. 2010) for the definition of official nomenclatures for immune genes (HLA, KIR and MIC) and alleles in humans, the IPD-MHC database (Robinson et al. 2010) which encompasses the definition of sequences for non-humans species; NetMHCpan (Hoof et al. 2009; Nielsen et al. 2007), NetMHCIIpan (Nielsen et al. 2008) for peptide binding predictions of MHC Class I and Class II respectively; the SYFPEITHI database (Rammensee et al. 1999) for prediction of T-cell epitopes and MHC ligands; the dbMHC database (Helmberg, Dunivin & Feolo 2004) and ALFRED (Rajeevan et al. 2003) which encompass data of several polymorphisms in worldwide populations, among many others. Table 1.4 and Table 1.5 present a list of different useful resources available which can be used as a reference for the investigation of several immune genes including methods and techniques for peptide binding predictions and population genetic analyses.

Table 1.4: Useful databases for the investigation of immune genes

Database	Description	URL	References
ALFRED	Public database for the storage of Allele Frequency in different populations and DNA polymorphisms.	http://alfred.med.yale.edu/	(Cheung et al. 2000; Osier et al. 2001; Rajeevan et al. 2003)
dbMHC	Public database and online applications for the analysis of HLA and related genes including microsatellites.	http://www.ncbi.nlm.nih.gov/projects/mhc/	(Helmberg & Feolo 2007)
IMGT	Integrated repository specialised in immunoglobulins, T cell receptors, MHC and related proteins of the immune system in humans and other vertebrates.	http://www.imgt.cines.fr/	(Lefranc 2004)
IMGT/HLA	Public database for the storage of sequences of the HLA system including the official nomenclature for genes and alleles.	http://www.ebi.ac.uk/imgt/hla/	(Robinson et al. 2009; Robinson et al. 2003)

Table 1.4: Useful databases for the investigation of immune genes (*Continued*)

Database	Description	URL	References
IPD-KIR	Public repository for the storage of sequences of the KIR polymorphic region including the official nomenclature for genes and alleles.	http://www.ebi.ac.uk/ipd/kir/	(Robinson et al. 2010; Robinson et al. 2005)
SYFPEITHI	Public database for MHC ligands, peptide motifs and T-cell epitope prediction.	http://www.syfpeithi.de/	(Rammensee et al. 1999; Schuler, Nastke & Stevanovikc 2007)
The Immune Epitope Database (IEDB)	Public repository for the representation of the molecular structures recognised by adaptive immune receptors.	http://www.iedb.org/	(Beaver, Bourne & Ponomarenko 2007; Vita et al. 2010)

Table 1.5: Useful applications for immunogenetics and population genetics analyses

Application	Description	URL	References
Arlequin	Software for the analysis of population genetics including the estimation of haplotype and allele frequencies, linkage disequilibrium, test of neutrality, among others.	http://cmpg.unibe.ch/software/arlequin3/	(Excoffier, Laval & Schneider 2005)
ClustalW, ClustalX	Software for multiple alignments of nucleotide and amino acid sequences.	http://www.ebi.ac.uk/Tools/msa/clustalw2/ http://www.clustal.org/	(Chenna et al. 2003; Larkin et al. 2007; Thompson, Gibson & Higgins 2002)
Gene[RATE]	Online tools for the analysis of ambiguities in HLA datasets.	http://geneva.unige.ch/generate/	(Nunes 2007)
HLA Completion	Online tool for resolving HLA ambiguities in low and intermediate resolution types.	http://atom.research.microsoft.com/HLACompletion/	(Listgarten et al. 2008)
HLA Matchmaker	Computational algorithm for the analysis of HLA molecule and amino acids mismatches.	http://www.hlamatchmaker.net	(Duquesnoy 2002; Duquesnoy & Askar 2007; Duquesnoy, Howe & Takemoto 2003; Duquesnoy & Marrari 2002; Duquesnoy et al. 2003)

Table 1.5: Useful applications for immunogenetics and population genetics analyses (Continued)

Application	Description	URL	References
MEGA	Molecular Evolutionary Genetics Analysis Software for sequence alignments, inferring phylogenetic trees and molecular evolution.	http://www.megasoftware.net/	(Kumar et al. 2008; Tamura et al. 2007)
NetMHCpan NetMHCIIpan	Online software to analyse the function and interaction of HLA Class I and Class II molecules.	http://www.cbs.dtu.dk/services/NetMHCpan	(Buus et al. 2003; Hoof et al. 2009; Nielsen et al. 2007; Nielsen et al. 2008)
PHYLIP	Software package for inferring evolutionary trees.	http://evolution.genetics.washington.edu/phylip.html	(Retief 2000)
PyPop	Open-source software package for the analysis of populations at large scale and multiple loci.	http://pypop.org/	(Lancaster et al. 2003; Lancaster et al. 2007)
SKDM	Software for HLA and disease association analysis examining strong associations, amino acids and linkage disequilibrium between patients and controls.	http://sourceforge.net/projects/skdm/	(Kanterakis et al. 2008)

1.8 Aim of the thesis

The objective of this research was to develop a web-based repository containing data on frequencies of several immune genes and their corresponding alleles along with a set of online bioinformatics tools to ascertain the occurrence of these variants in worldwide human populations.

Giving the number of applications which could benefit by the use of frequency data of these immune genes and alleles, the research work was divided into three main cases of study:

a) **Compilation of immune gene frequency data and design of a database**

As mentioned in Section 1.1, the number of reports of allele, haplotype and genotype frequencies of different polymorphic regions in human populations has increased considerably in recent years. This information has been reported in different peer-reviewed journals in non-standardised formats leading to the difficulty to perform meta-analysis of these datasets. Thus, the first goal of the research was to provide a database to collect and curate publications and individual reports through an online submission system which could validate the different characteristic of the data, i.e. the official gene and allele names, demographic data of the population, etc., and the capability to incorporate other immune related genes. This work is summarised in Chapter 2 which comprises the description of population datasets and the schemas used in the construction of the database.

b) **Analysis of immune gene frequency data in worldwide populations**

The investigation of immune genes and their frequency distribution among worldwide human populations involves the analysis of data at different levels: allele, haplotype and genotype frequencies. Therefore, software applications must have the flexibility to summarise frequency data at different levels and provide searching mechanisms for the analysis of specific datasets, i.e. a specific population, geographic region, set of alleles, source of data (published or unpublished data), type

of populations (e.g. anthropological studies, individuals used as controls in disease association studies, etc.), level of resolution in allele typing (low and high-resolution), etc. Searching mechanisms implemented for the investigation of immune genes will be discussed in detail in Chapter 3.

As frequencies vary among populations, one interest was focussed in describing the variability presented by the different genes in specific geographic locations. For HLA, KIR, MIC and cytokine gene polymorphisms a scenario was formulated in which specific alleles would be clearly representative of a particular geographic region. In the case of the KIR polymorphism, the interest was extended to the investigation of the organisation of KIR genotypes by analysing which populations may present specific genes according to the genes being inhibitory or activating. The set of analysis carried out and software tools implemented for these four different polymorphic regions are presented in Chapters 4, 6 and 7.

c) Ascertaining of the rarity of HLA alleles

Regardless of the high number of alleles that have been reported in latest releases in the IMGT/HLA database, many HLA alleles have only been found once, suggesting that these alleles may have been wrongly sequenced or the methods used in the past were not optimal to distinguish alleles at high-resolution (levels 2-4). With this in mind, a worldwide analysis comprising the confirmation of alleles across international groups and laboratories was expected to be useful in demonstrating how relevant these alleles are, whether they only occurred in one individual, one particular family, a specific geographic region or ethnic group. The analysis of rare alleles is described in detail in Chapter 5.

1.9 Summary

This chapter presented an insight into the extensive variability that is found in several genes involved in the immune response. The polymorphic regions covered in this chapter include HLA, KIR, MIC and several cytokine genes. As reviewed in this chapter, allele frequencies of these genes vary significantly among individuals of

different populations. In the case of HLA, the variability is colossal, with more than two thousand alleles reported in a single locus (HLA-B) and more than 6,000 alleles in total. Knowing the distribution of alleles would lead to improved understanding of the functions of the human immune system and human evolution. Therefore, the need to develop a database to store such amount of data is a primarily need for scientists interested in the analysis of these immune genes.

Despite a very large amount of data has been in the public domain (mainly in the literature) over the last two decades, little effort has been carried out to create a generic database framework to store immune frequency data. Different methodologies, techniques and programming tools for the construction of online repositories have been described in this chapter. Due to the heterogeneity of data reported, a significant effort is required to extract information from different sources (data from journals, reports from IHWSs, etc.) for the design of a generic prototype. The following chapter presents the compilation of populations and the design of a database to contain immune gene frequencies in different worldwide human populations.

Chapter 2

Design of the Allele Frequency Net Database: datasets, schemas and methods

2.1 Introduction

The previous chapter presented a general introduction on the impact of several immune genes in different clinical and research areas and explained the importance of databases as tools to disseminate the distribution of these genes and their corresponding alleles in worldwide populations.

As reviewed in Chapter 1, the amount of data available in the public domain on frequencies of immune genes in recent years has been immense. The design of databases to contain immune frequencies requires the analysis of the data to identify special characteristics of the datasets.

Chapter 2 describes the compilation of datasets containing data on frequencies of HLA, KIR, MIC and cytokine genes from an extensive review in the literature. Data collection includes the definition of demographic and frequency attributes of the populations. Additionally, this chapter describes the criteria used in the validation of data along with the controlled vocabulary used in the assignment of population attributes values.

The second part of this chapter focuses on the construction of a generic database schema for the storage of the different immune gene frequencies. The design of the database includes the information workflow, metadata definition and data dictionaries used in the specification of population attributes and related entities.

2.2 Compilation of population datasets

For clarity and uniformity, the term ‘population’ used in this thesis refers to individuals who belong to the same geographical region, country, state, prefecture, city, town and/or ethnic background (See Section 2.2.3.1).

2.2.1 Source of data

The compilation of population samples was derived from two main sources: (i) from the literature (peer-reviewed publications and Proceedings from IHWSs) or (ii) from direct submissions to the AFND by users (i.e. submissions of unpublished results by individual laboratories). The aim of the literature review was to cover all previously published studies between January 1990 and December 2010 including publications from more than sixty five peer reviewed journals which can be consulted online via the <http://www.allelefrequencies.net/datasets.asp> link.

The initial criteria used for selecting the populations were as follows:

- Any peer-reviewed publication or direct submission to the AFND containing HLA, KIR, MIC or cytokine frequency data at allele, haplotype or genotype level at any level of resolution.
- Any population with a sample size over 10 individuals considering that studies in some ethnic groups contained a low number of members.

2.2.2 Submission protocol

Submissions of population datasets were performed by curators of the AFND (Derek Middleton and Faviel Gonzalez) or by users wishing to submit their data. For each submission, individuals were (and currently are) asked to complete a set of attributes to describe the characteristics of the population sample. The process is divided into three

different steps: (i) the capture of demographic data, (ii) the submission of frequency data and (iii) the validation of data which is usually performed by curators of the AFND (Figure 2.1).

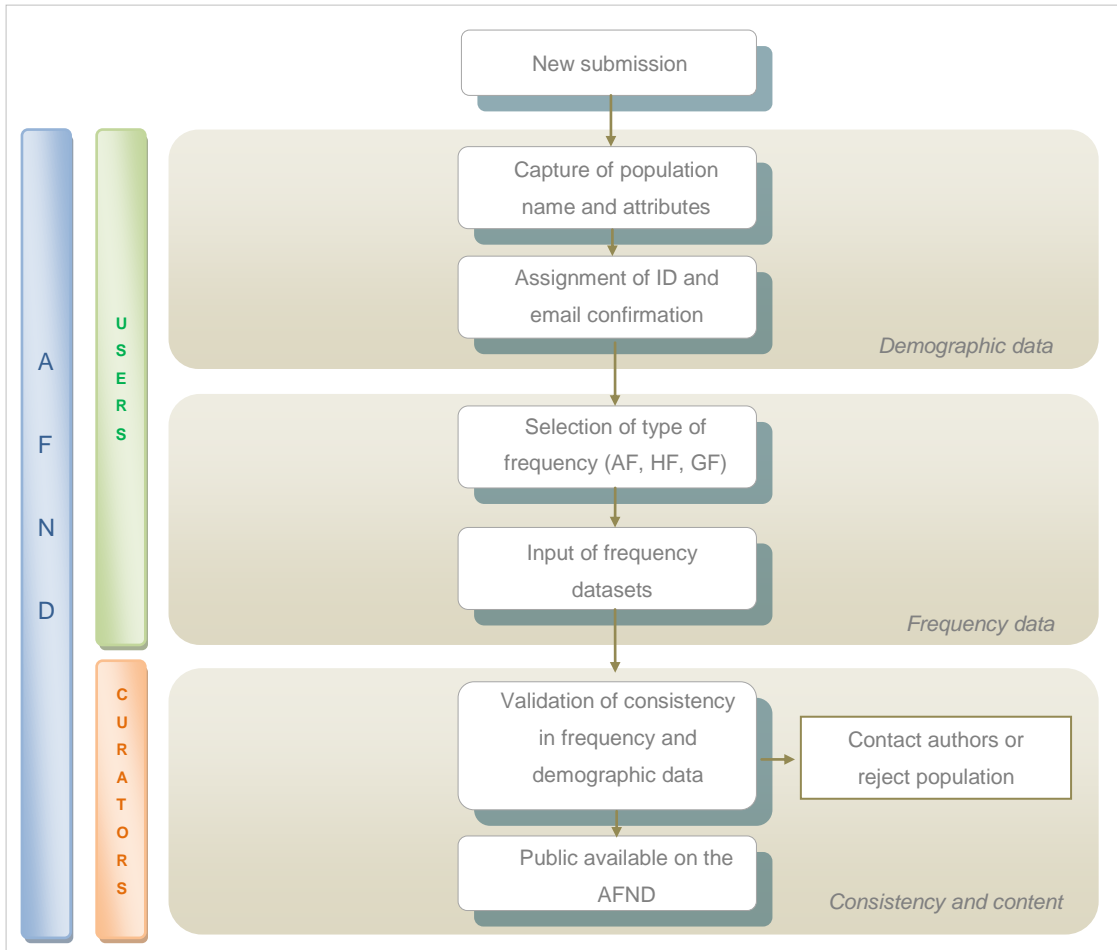


Figure 2.1: Dataflow diagram for submissions of populations in the AFND.

- i) **Demographic data.** For demographic data, authors were requested to input a set of minimal attributes such as the name of the population, polymorphic region, geographical region, country, ethnic origin, sample size, family background, year of sample, method(s) used for determination of frequencies, source of population, and the bibliographic reference if available. In this process, users were assisted with drop-down boxes to complete their data (Figure 2.2). When data for a given attribute was not available, values were set as ‘not known’.

The Allele Frequency Net Database [Add Population Study]

http://allelefrequencies.net/pop2001a.asp

Allele*Frequencies

in Worldwide Populations

Populations > Add New Population Study

Instructions

Please provide the demographic data of your population by completing the information below. Once you have completed your data click on "Submit" to add the population. You will receive an e-mail with a spreadsheet along with a set of instructions to enter your frequencies.

Note: Population names must be unique. You can check an appropriate name by consulting our database for a complete list of the existing populations.

* Population name: (Click [here](#) to check populations previously submitted)

Country + Region|Ethnic Group + Polymorphism (e.g. Pakistan Karachi KIR, France South East KIR, etc.
Note: HLA pops do not require to specify polymorphism e.g. Scotland Orkney)

* Polymorphic Region:

* Geographic Region:

* Country:

Latitude: ° '

Longitude: ° '

* Ethnic Origin:

* Sample Size:

* Urban|Rural:

* Family Background:

* Main author:

* Test date:

* Method(s) used: SSP SSOP SSCP SBT Sequencing RSCA RFLP Other

* Source of Population:

Publication Details:

Figure 2.2: Example of demographic data capture in the AFND.

After the submission of demographic data, a numeric identifier was assigned to each population to uniquely identify the record. Then, an e-mail confirmation was sent to the submitter to complete frequency data.

- ii) **Frequency data.** To complete frequency data, authors were requested to enter their frequencies via an online web form or by providing the pre-formatted spreadsheet containing the full list of alleles according to releases from the IMGT/HLA (for HLA and MIC) and IPD-KIR databases (for KIR) (Figure 2.3). In the case of cytokine gene polymorphisms, entries were based on a suggested nomenclature which will be described in detail in Chapter 7.

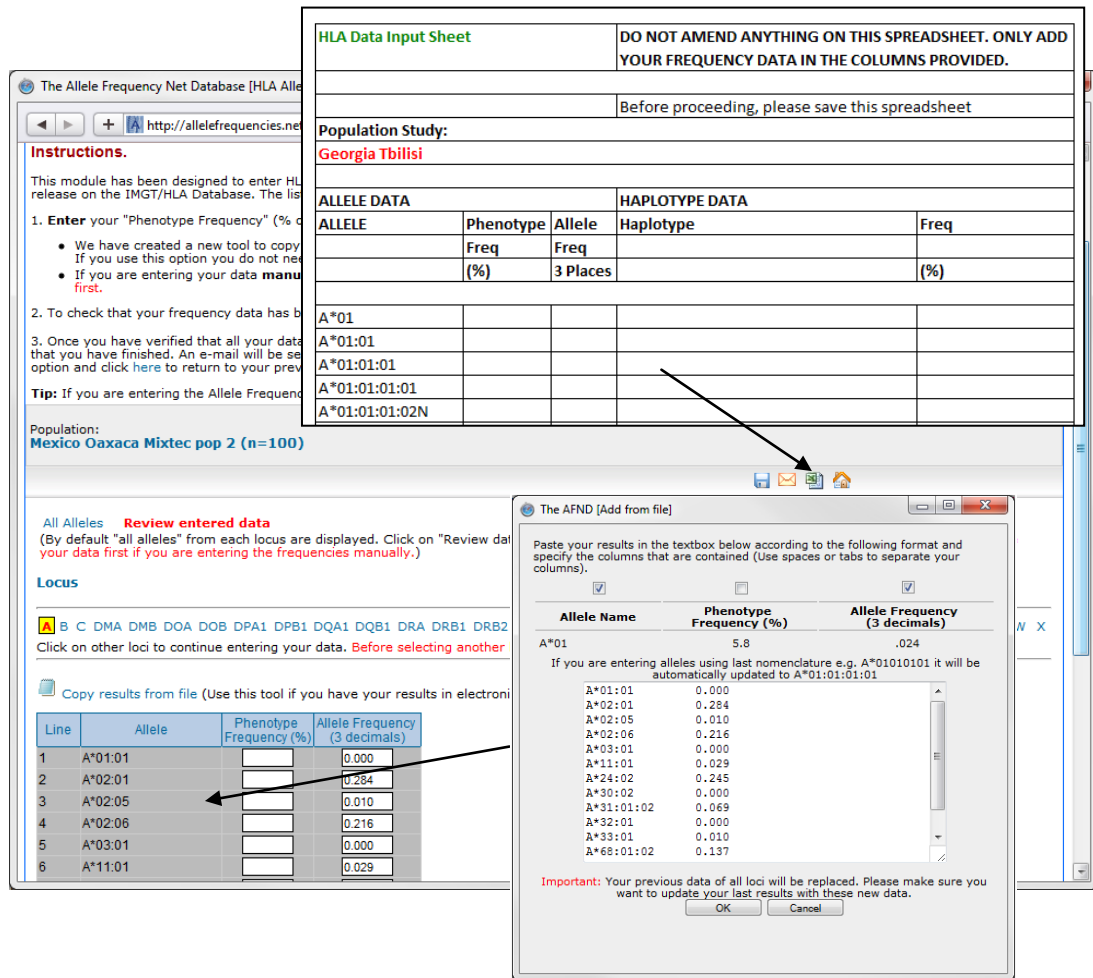


Figure 2.3: Example of frequency data capture in the AFND

- iii) **Data curation.** Each submission followed a data validation procedure for consistency of data, which is described in Section 2.2.3. This process is performed by curators of the AFND. If a particular problem in the content of datasets was detected, authors were contacted to verify the anomaly. Once the validation process was completed, datasets were set as publicly available in the AFND.

2.2.3 Validation of population attributes and frequency data

2.2.3.1 Validation of demographic data

For each attribute in a population, a set of drop-down boxes were provided to complete the data (Figure 2.2). When populations were submitted with different values in a particular field (e.g. ethnic group, source of population, etc.) which were not included in

the corresponding drop-down box, the value given by the submitter was added to the current list to extend the list of values. This list is under continuous maintenance by curators of the AFND to standardise reporting and has been subject of revision by International groups working in the standardisation of population attributes (See discussion).

a) Population attributes

- **Population identifier (ID).** A number in consecutive order was given to each population to uniquely identify the dataset within the database. The ID number was used as a primary code to retrieve the corresponding information from a given population.

E.g. **1202 Scotland Orkney**

Additionally, the AFND was designed to validate duplications in naming a population. For each new submission, demographic information of the population was also verified against the existing data which might have been entered under a similar name.

- **Population name.** Each population was named according to the combination of the country name, geographical location and ethnic group when available, to describe the population appropriately.

E.g. **South Africa Natal Zulu**

In order to identify the polymorphic region studied in a given population an additional term was incorporated at the end of the name, except for HLA populations which were the first populations entered in the database.

E.g. **South Africa Xhosa KIR**
Italy Milan Cytokine

If another set of individuals from a given population, which was geographically and ethnically similar to an existing population in the database, was typed by a second laboratory, a consecutive number was assigned to that population to differentiate the two samples.

E.g. **China Guangzhou Han,**
China Guangzhou Han pop 2

In some populations, individuals were typed in a different country from their original ethnic background (i.e., immigrants from a different country). In these cases, the name of the original ethnic background was included in the name of the population, and the country was defined as the current location in which individuals were typed.

E.g. **Canada Iranian KIR**
(Iranian descendant people living in Canada)
Singapore Thai
(Thai descendant people living in Singapore)

If individuals were typed within a region or country and were part of a Bone Marrow Donor Registry, Cord Blood Bank or any other centre, the name of the source of data was included in the population name.

E.g. **Taiwan Tzu Chi Cord Blood Bank**

Therefore, to determine the order and composition of a population name the following semantic notation was used:

**Population Name = Country + (Region | Sub-region) + (Locality)
 + (Ethnic group) + (Centre) + (Polymorphic Region) + (pop N)**

Where

Region = (East | West | North | South | any other region)

Sub-region = (State | Province | County | District)

Locality = (City | Town | Village)

Centre = (Bone Marrow Registry | any other centre)

Polymorphic Region = (KIR | Cytokine | MIC)

N = (2, 3, 4 ...)

- **Geographical region.** In order to classify all population datasets according to their geographical location, a total of ten geographical regions were used (Figure 2.4).
- **Country.** The name of the country was assigned to each population according to the location in which the individuals were sampled.
- **Sub-region (optional).** Different sub-regions or specific geographical locations were used to define a particular area within a country. E.g. State, province, county or district.
- **Ethnic group.** All populations were assigned to one of 19 major ethnic groups or additional categories (Mixed and Other) to identify the ethnic background (Table 2.1). The classification was based on the most frequent ethnicities reported in the literature. This classification was mainly used to differentiate different ethnic groups within a given country. For populations with a specific ethnic group such as **South African Zulu** (Williams et al. 2001), the major ethnic group was set as ‘Black’ and the specific ethnic group (Zulu) was included in the name of the population. It is important to highlight that Caucasian, Hispanic, and other ambiguous terms are being reviewed by the HLA-NET and IDAWG international groups (See discussion in this chapter).
- **Urban or Rural.** A field was included to define whether individuals were typed in an urban area or rural community or if they belonged to both sets.
- **Familial status.** An additional field was used to indicate whether individuals typed belonged to a) descendants from parents living at the same location, b) grandparents living at the same location, or c) if it was unknown.

- **Source.** All populations in the AFND were also classified according to the source of the population, e.g. anthropology study, blood donors, Bone Marrow Registry, controls for a disease association study, solid organ unrelated donors or another type of study.

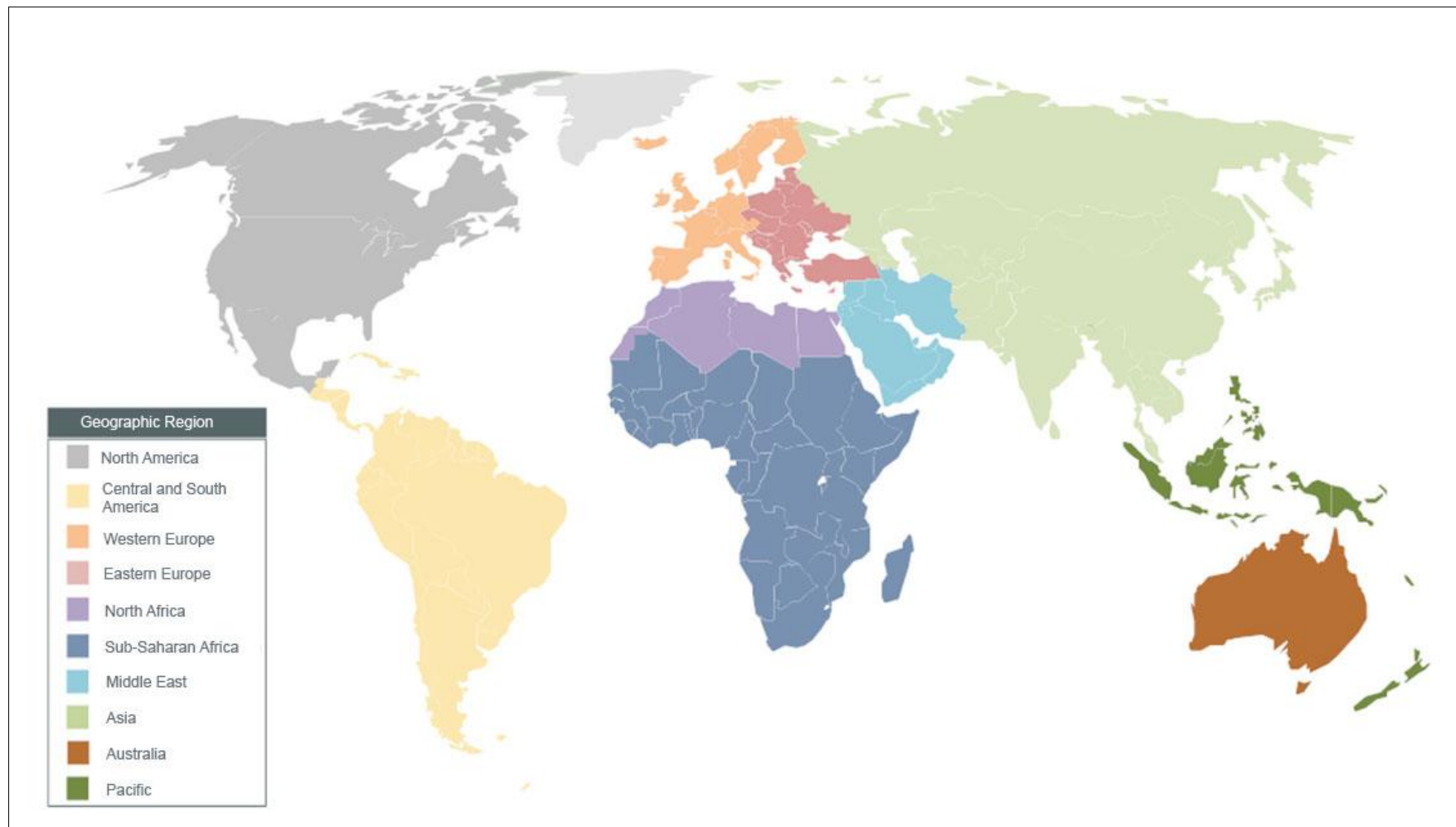


Figure 2.4: Classification of geographical regions in the AFND.

Table 2.1: Ethnic groups available in the AFND

Ethnic group	Description
Amerindian	Indigenous people from North, Central, South America and the Caribbean Islands.
Arab	Individuals from the Middle East or descendants of an Arabic speaking country.
Asian	Individuals from Central and Southern Asia (e.g. India, Pakistan, Sri Lanka, Uzbekistan and Kazakhstan).
Australian Aborigines	Individuals from indigenous communities in Australia.
Black	Individuals descending from Sub-Saharan Africa.
Caucasoid	Individuals descending from Western or Eastern Europe.
Hispanic	Individuals from Spanish or Latino American descents living in the United States.
Mestizo	Mixture of European and American indigenous people.
Oriental	Individuals from East and Southeast Asia.
Siberian	Individuals from the Siberian Region in Russia.
Melanesian	Individuals from Fiji, New Caledonia, Papua New Guinea and Solomon Islands.
Polynesian	Individuals from American Samoa, Cook Islands, New Zealand, Niue, Samoa, Tonga and Hawaii.
Austronesian	Individuals from Borneo, Canada, Indonesia, Malaysia, Philippines, Singapore and Timor-Lester.
Micronesian	Individuals from Kiribati and Nauru.
Berber	Indigenous individuals in North Africa from Tunisia.
Jew	Individuals of Jewish descent mainly from Israel.
Kurd	Individuals of Kurdish descending.
Mulatto	Individuals of black and white ancestry.
Persian	Individuals of Persian descending mainly from Iran.
Mixed	Populations which mixture was not able to be classified in another category.
Other	Individuals of other minor ethnicities or social group which could not be classified in other major group, e.g. Gypsies.

- **Geographical coordinates.** In addition, for a graphical representation of populations, latitude and longitude coordinates were assigned to all populations when information was provided.

b) Sample attributes

- **Sample size.** Sample size was used to identify the number of individuals which were considered in the study of the population. When the number of individuals typed for each locus was different, a note was included to specify sample sizes for each locus.
- **Sample year.** Year in which the individuals of a population were typed.
- **Typing method.** Method used for the typing. (E.g. SSP, SSOP, SSCP, SBT, RSCA, RFLP, Other – See Section 1.4.2 for description of typing methods).

c) Bibliographic reference

- **Reference.** The following data were used to cite the source from which the information was extracted.
 - Name of all authors, corresponding author, contact e-mail, journal name, year of publication, volume, issue and number of pages.

2.2.3.2 Validations of frequency data

All populations were validated for allele, haplotype and genotype frequency data depending on the availability of the information. To capture the information for each type of frequency, an online submission form or a preformatted spreadsheet were used (See Figure 2.3).

a) Estimation of allele, haplotype and genotype frequencies and terminology

The AFND contains a compilation of frequency data that was previously estimated by different methods by the corresponding authors. The bibliographic reference for each population was provided in order that a user may verify what type of analysis the author

used for calculating frequencies. A list of the different methods that were used to calculate the frequencies is shown in Table 2.2.

Table 2.2: List of methods used in the estimation of frequencies

Type of frequency	Type of submission			Description
	Extracted from literature	Direct submission		
		By the author	By AFND	
AF	Direct counting, Resampling [§]	Direct counting, Resampling	Direct counting	Proportion of allele copies in the gene in four-decimal format, e.g. 0.0010
HF	ML-EM	ML-EM	N/A	Percentage of individuals carrying the haplotype, e.g. 3.4%
GF	Direct counting	Direct counting	Direct counting	Percentage of individuals carrying the genotype e.g. 2.6%

AF=Allele Frequency; HF=Haplotype Frequency; GF=Genotype Frequency; [§] See (Dempster, Laird & Rubin 1977; Good 2005; Nunes 2007) for a review of resampling methods; ML-EM=Maximum Likelihood by Expectation-Maximization (Dempster, Laird & Rubin 1977; Excoffier & Slatkin 1995); N/A=Not available.

b) Validation of frequency data

- **Allele frequencies.** For each locus in a population sample, values were added and for any summation greater or less than 1.0000 authors were contacted to verify the difference. For frequencies that totalled greater than 1.0000 which could not be explained, the submission was rejected. Frequencies which added to less than 1.0000 were kept in the database. This occurred when a population which was published contained only the frequencies of the main alleles, the frequencies of the remaining alleles being added together to one frequency. Additionally, in each population the lowest expected frequency (observation of a single allele) was calculated and compared to possible multiples. For instance, in a population of 200 individuals, the minimum expected frequency was 0.0025 ($1/2n$). Therefore, all frequencies had to be multiples of 0.0025 considering truncation and rounding.
- **Haplotypes frequencies.** For haplotype data, only frequencies greater or equal to 1.0% were stored in the database, except for populations with a large number of individuals ($n \geq 1000$) in which no threshold was applied.

- **Genotype frequencies.** For genotype data, frequencies were added and for any summation greater or less than 100% the author was contacted. Frequencies in which the sum was less than 100% the information were also kept in the database.
- **Validation of allele names.** All populations on the AFND were validated to ensure that their corresponding frequencies were submitted under the official allele name for the HLA, KIR and MIC polymorphisms described on the IMGT/HLA and IPD-KIR databases (Robinson et al. 2003; Robinson et al. 2005). If necessary, the author of the data was contacted with any query or any change made. In the case of HLA and MIC, all frequency data submitted before April 2010 were updated according to the new nomenclature (Marsh et al. 2010). To do this, a script in the AFND was implemented to convert pre-2010 allele designations.
- **Levels of resolution of alleles.** In order to generate all possible alleles at different level of resolution (1-4 levels for HLA, MIC and KIR polymorphisms) a set of libraries were generated by splitting all possible subdivisions from a given allele.
E.g. $A*01:01:01:01 = (A*01, A*01:01, A*01:01:01)$
- **Ambiguities in typing.** When data, published or sent directly to the AFND, in which the author was not able to differentiate some alleles (i.e. ambiguous alleles), a note in the publication details was used to describe how the frequency data was entered for one of the ambiguous alleles. For example, ‘unable to differentiate alleles that are identical over exons 2 and 3 (Class I)’, in which frequencies were given under first allele.

2.2.4 Summary of population datasets

The collection of datasets available on the AFND as of May 2011 consisted of 1209 population samples covering 4,262,379 healthy unrelated individuals. Population

datasets comprised 844 populations for HLA, 194 for KIR, 113 for cytokine genes and 58 for MIC (Table 2.3). The compilation of frequency datasets included more than 100,000 records at allele, haplotype and genotype level in total. Some of the populations were only typed for one of the three levels. For instance, 829 out of the 844 population samples for HLA contained allele frequency data and 342 populations had haplotypes. As shown in Table 2.3, genotype frequencies were only available for 108 populations tested for KIR.

Table 2.3: Population frequency datasets by polymorphic region on the AFND

Region	Pops	Indiv	Allele			Haplotype			Genotype		
			Pops	Indiv	Freq	Pops	Indiv	Freq	Pops	Indiv	Freq
HLA	844	4,213,280	829	613,553	86,843	342	3,778,303	8,916	-	-	-
KIR	194	23,204	194	23,204	4,922	-	-	-	108	12,291	2,600
Cyt	113	18,246	113	18,246	3,603	-	-	-	-	-	-
MIC	58	7,649	58	7,649	723	20	2,426	257	-	-	-
Total	1,209	4,262,379	1,194	662,652	96,091	362	3,780,729	9,173	108	12,291	2,600

The vast majority of population samples ($\sim 97.7\%$) varied between 10 and 5,000 individuals with only 27 populations that exceeded 5,000 subjects (Table 2.4). Nearly half of the datasets contained populations with 100 or more subjects.

Table 2.4: Populations by sample size and polymorphic region

Sample Size	Cytokine		HLA		KIR		MIC	
	Pops	Indiv	Pops	Indiv	Pops	Indiv	Pops	Indiv
10-50	19	725	185	6,693	47	1,988	13	452
51-100	31	2,540	263	20,506	78	5,985	14	1,168
101-500	58	9,569	298	55,504	64	10,923	30	5,526
501-1000	3	1,835	30	20,454	4	2,843	1	503
1001-5000	2	3,577	41	79,700	1	1,465	-	-
> 5000	-	-	27	4,030,423	-	-	-	-
Totals	113	18,246	844	4,213,280	194	23,204	58	7,649

All populations were summarised by geographical region and ethnic group to illustrate the coverage of submissions. As shown in Table 2.5, there was a varied number of submissions in each polymorphic region. For instance, in HLA, KIR and MIC, most of the populations corresponded to individuals from Asia, Western Europe and South & Central America whereas for cytokines Western Europe and North America contained the highest number of submissions. Table 2.6 shows the number total of populations

classified by major ethnic groups in which the majority of submissions (51.7%) corresponded to individuals with Caucasian or Oriental background.

Table 2.5: Population samples by geographical region

Geographical Region	Cytokine		HLA		KIR		MIC	
	Pops	Indiv	Pops	Indiv	Pops	Indiv	Pops	Indiv
Asia	18	2,580	272	136,209	43	4,081	20	2,489
Australia	-	-	8	2,822	2	117	-	-
Eastern Europe	18	1,969	56	14,429	11	2,155	3	200
Middle East	9	781	35	33,486	14	1,580	2	36
North Africa	1	157	21	1,902	1	67	1	82
North America	26	4,630	82	3,684,249	26	4,619	5	645
Pacific	-	-	59	3,896	13	617	-	-
South and Central America	8	878	108	13,470	27	2,801	8	1,004
Sub-Saharan Africa	3	319	55	5,878	21	2,500	3	152
Western Europe	30	6,932	148	316,939	36	4,667	16	3,041
Total	113	18,246	844	4,213,280	194	23,204	58	7,649

Table 2.6: Population samples by ethnic group

Ethnic Group	Cytokine		HLA		KIR		MIC	
	Pops	Indiv	Pops	Indiv	Pops	Indiv	Pops	Indiv
Amerindian	3	231	84	6,138	15	931	7	804
Arab	3	355	29	3,691	7	745	3	118
Asian	3	733	54	7,096	15	1,312	1	38
Australian Aboriginal	-	-	5	587	1	67	-	-
Austronesian	1	26	20	2,811	3	143	-	-
Berber	-	-	2	136	-	-	-	-
Black	10	1,486	79	426,496	26	3,333	6	452
Caucasoid	59	11,545	237	2,782,067	67	10,845	21	3,586
Hispanic	5	726	7	455,386	4	475	-	-
Jew	1	48	12	24,572	2	74	-	-
Kurd	-	-	3	205	-	-	-	-
Melanesian	-	-	29	1,892	6	280	-	-
Mestizo	3	255	21	2,324	5	696	-	-
Micronesian	-	-	2	129	-	-	-	-
Mixed	3	359	24	365,822	7	801	1	200
Mulatto	1	100	3	270	1	42	-	-
Oriental	15	1,847	179	124,000	28	2,899	19	2,451
Other	-	-	7	412	-	-	-	-
Persian	6	535	8	6,918	2	316	-	-
Polynesian	-	-	17	1,054	4	194	-	-
Siberian	-	-	22	1,274	1	51	-	-
Total	113	18,246	844	4,213,280	194	23,204	58	7,649

2.3 Design of the AFND schema

To date various methods for the design of databases have been described in the literature. Some of the most commonly used techniques in database modelling were reviewed in Section 1.7.3. Each methodology has its advantages and drawbacks; for example, OODBs are suitable for the reusability of modules but have been found to be difficult to implement by the different public and commercial DBMS. To my knowledge only a limited number of DBMSs have been implemented following this methodology (e.g. MyOODB, Objectivity/DB, among few others.). In this research, the Relational Model was used in the design of the AFND schema since it is the most widely used approach for the implementation of scientific and commercial databases.

2.3.1 Database schemas

In order to specify relationships among tables, validate data integrity and constraints on database components, the conceptual, logical and physical model schemas were designed. The current schema of the AFND consists of 72 tables (listed in Table B.1 in Appendix B) of which 32 correspond to the core of the database (Figure 2.5). Several dynamic/virtual tables (views) were included in the schema to optimise time processing in the retrieval of information (Table B.1).

As shown in Figure 2.5, the database schema was divided into five main sections: (i) information on population demographic data (countries, ethnic groups, geographical regions, etc.), (ii) information on frequency data (allele, haplotype and genotype frequencies), (iii) information on DNA sequences of alleles compiled from external databases (IMGT/HLA and IPD-KIR), (iv) a section to analyse the rarity of alleles, and (v) a module to administrate the access to the information to users and/or curators of the AFND.

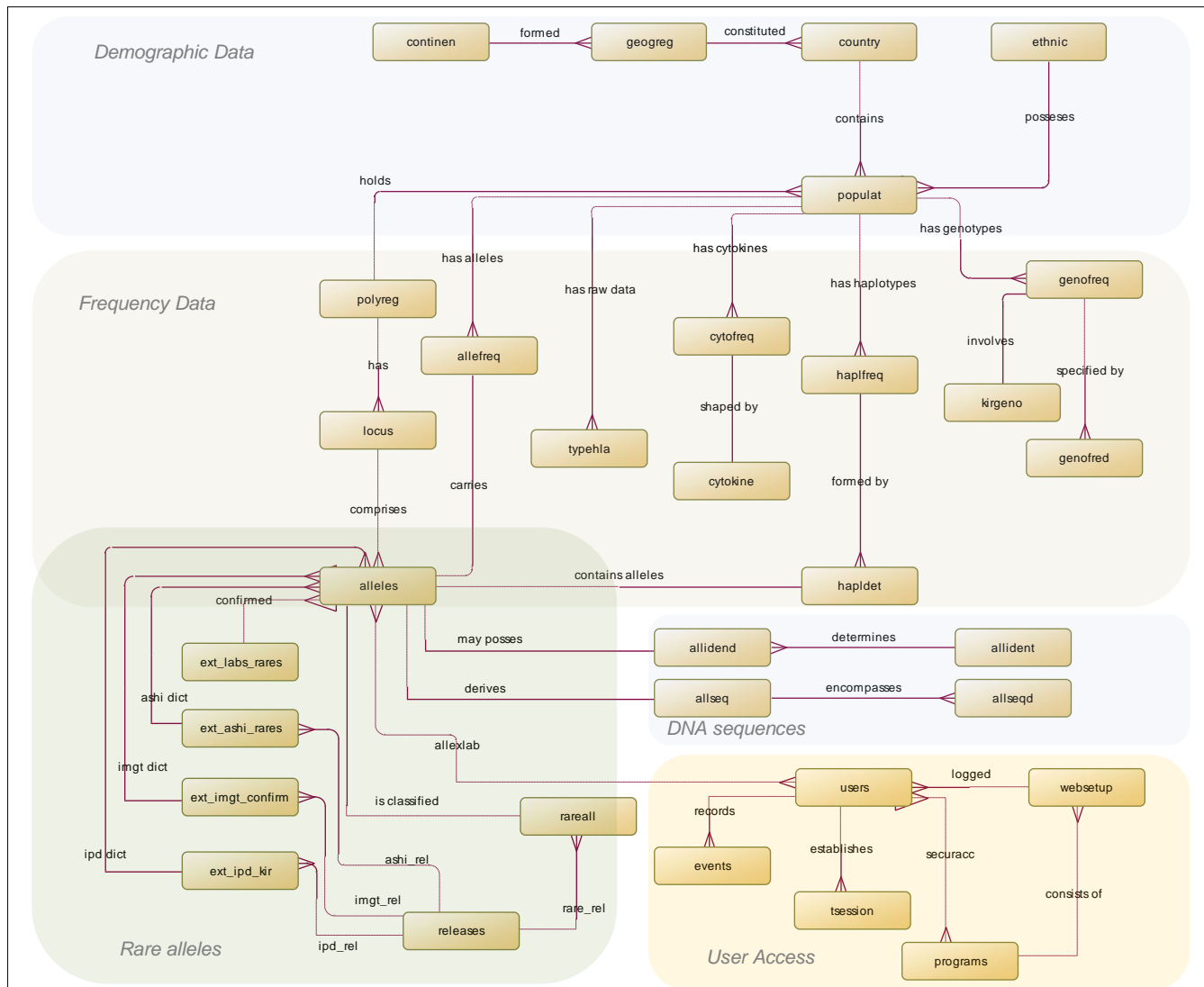


Figure 2.5: Conceptual schema of the AFND.

The conceptual, logical and physical model schemas were generated using the PowerDesigner Software Version 15.2.0 based on the Barker's notation¹³ to simplify readability. The schema was normalised to the fourth normal form (4NF) to minimise data redundancy. The logical and physical model schemas and the SQL schema for the generation of the database can be consulted in Appendix A (Figures A.1-A.10).

2.3.2 Metadata: data dictionaries and controlled vocabulary

In order to define the meaning and type of formats used in the different attributes included in the Logical and Physical Model schemas, a data dictionary was created containing the description of attributes and examples of accepted values (See Table B.2 in Appendix B).

The complete controlled vocabulary used in the definition of the values for ethnic groups, geographical regions, and other population attributes can be consulted on the website through the <http://www.allelefreqencies.net/datasets.asp> link.

2.4 Discussion

The compilation of datasets included in the AFND encompasses the largest collection of immune gene frequency data in the world. As mentioned in Section 1.1, the design of databases involves a number of challenges such as the standardisation of formats, the use of an appropriate controlled vocabulary and the interaction with other external databases for data validation and data exchange.

Due to the wide scope of data published on immune gene frequencies spanning the last two decades, the first objective in the design of the AFND was to provide a description of the population attributes to generate a controlled vocabulary based on the terminology found in the literature. In this process, several problems were found such as the standardisation of geographical regions and ethnic groups. During the

¹³ Barker's notation available in the PowerDesigner Software was developed by Richard Barker and collaborators as a method to represent Entity-Relationships Diagrams.

compilation of datasets it was decided to select the term which described the populations more appropriately. With the aim of formulating a standard terminology in reporting immune gene frequencies and population attributes, several groups from international organisations such as the Immunogenomics Data Analysis Working Group (IDAWG) (Mack et al. 2009) and the European network of the HLA diversity for histocompatibility, clinical transplantation, epidemiology and population genetics (HLA-NET) (Sanchez-Mazas et al. 2010) are actively participating in the definition of the controlled vocabulary. It is expected that in the near future the definition of terms used in these immune gene databases can be finalised. One of the advantages of the model described in this chapter is the flexibility in the modification of the physical schema which would allow the incorporation of new changes.

The second objective was focussed on the validation of the data submitted and the interaction with external databases such as the IMGT/HLA, IPD-KIR. As shown in Section 2.2.3.2, the list of valid alleles is automatically updated after a new release is available, assisting in the uniformity of data submitted by users.

Methodology used in the design of the AFND schema

The selection of a particular model in the design of databases has been a topic of controversy. Databases are generally constructed considering several factors such as the availability of data, the origins of the information, hardware requirements, etc. (Bornberg-Bauer & Paton 2002). Although it was believed that OODBs would replace RDBs, at present 90% of the biological datasets are still based on relational models (Hellerstein, Stonebraker & Hamilton 2007; Rob & Coronel 2008). Two of the main reasons in the delay of these implementations are the complexity of migrating existing relational models and the difficulty of performing new implementations in existing OODBs. Thus, the RDB schema was used as the model for the implementation of the AFND.

The first target in the design of the AFND schema was to generate a model which could encompass all four polymorphic regions presented in this research (HLA, KIR, MIC and several cytokine gene polymorphisms). After a preliminary analysis of the content of the different polymorphisms, it was concluded that the construction of general tables

for the storage of allele and haplotype frequency data was feasible (Figures A.2 and A.6). However, for genotype data, there was a need to generate a customised format to define the specific characteristics of the polymorphism of interest. This was the case of KIR genotypes. For cytokine frequencies it was opted for designing a separate table as the format in which cytokine polymorphisms were represented in the literature differed significantly from HLA, KIR and MIC.

The AFND as a generic framework for the collection of immune gene frequencies

The design of a generic model in the database schema permits the addition of other polymorphisms of interest with a minimum effort in the implementation. The use of preconfigured scripts and modules can speed up the incorporation of new polymorphisms such as minor Histocompatibility antigens (mHags), platelets, blood groups, etc. As such, an ongoing development includes a new section in the AFND to compile frequency data of minor Histocompatibility antigens (mHags) which are of relevance in haematopoietic stem cell transplantation (Simpson et al. 2002). Furthermore, other non-human species can be added into this schema as some of the genes share a similar structure, e.g. Bovine Human Leukocytes (BoLA), Dog Leukocyte Antigens (DLA), Feline Leukocyte Antigens (FLA), among others. This schema has assisted the implementation of a BoLA database prototype developed at the University of Liverpool which was used as supporting material for an ongoing grant application via the Biotechnology and Biological Sciences Research Council (BBSRC).

Accuracy of data

Unfortunately, on many occasions, data available in the literature are not always accurate. In many cases, Editors of journals were contacted to discuss these issues. Although the AFND cannot guarantee the accuracy of the tissue typing of the individuals, more than 90% of the data on the website has been peer-reviewed and published. Thus, the AFND relies on the accuracy of data being verified by the reviewers of the journals and acts mainly as a source for compiling data. Future developments in the AFND will include the collection the raw data in order that the

website can assist researchers in assessment of data quality. The module for the incorporation of raw data will be discussed in detail in Sections 8.2.4 and 8.3.1.

2.5 Conclusions

This chapter described the datasets, schemas and methods used for the construction of the AFND which encompasses the largest collection of immune gene frequency data in different worldwide populations. The information presented in this chapter included a full description of demographic and frequency attributes available for each population sample. Additionally, a set of minimal criteria to assess frequency and demographic data were also described in this section. The second part of the chapter comprised the definition of the schemas used in the implementation of the database and the metadata, data dictionary and controlled vocabulary employed in the standardisation of population attribute values.

The large number of datasets available in the AFND provides an extraordinary resource for the research community to compare allele, haplotype and genotype frequencies from populations of different geographical regions, countries or ethnic groups. To provide a rapid and user-friendly method for the consultation of data, the following chapter describes a set of computational approaches used for the examination of these immune genes.

Chapter 3

Development of the AFND website and online tools

3.1 Introduction

The previous chapter described the different methods employed in the development of the AFND including a complete description of the population datasets and the definition of the schemas utilised in the design of the back-end.

Based on the needs of researchers to examine the information available in AFND in an interactive manner, this chapter describes a bioinformatics web-based repository which was developed to assist researchers in the investigation of HLA, KIR, MIC and cytokines gene polymorphisms in worldwide populations. Furthermore, a section to evaluate the performance of the software implemented and examples of website users is included at the end of the chapter.

3.1.1 Previous work in immune gene frequency websites

The collection of frequency datasets of immune genes, principally HLA, was implemented several years ago when a component for anthropology analysis was presented at the 11th IHWS [See review (Thorsby 2009)]. This component led to the implementation of the dbMHC database (Helmberg et al. 2004) which was used to store tissue typing data of several populations from different ethnic groups across the world.

In related work, researchers at Yale University created a database called ALFRED (Cheung et al. 2000) to store allele frequencies of different polymorphisms for anthropological analysis including the HLA region and many others.

To our knowledge, only these two publicly accessible databases for the storage of frequencies of immune genes have existed in the field of immunogenetics. The two databases have been described and referenced in Chapter 1 (See Section 1.7.9). Although both databases seem to target the same audience as the AFND, the aims of these two websites differ significantly from those described in this thesis. The AFND is a specialised database for the collection and analysis primarily of immune gene datasets whereas ALFRED focuses on the collection of frequencies across all polymorphisms in the human genome without considering any specific location in particular. To give an example, as of May 2011, ALFRED contained more than 600,000 polymorphisms in 710 populations in which only 70 populations comprised information on HLA allele frequencies. In contrast, AFND includes more than eight hundred populations for the HLA region. In the case of the dbMHC database, the polymorphisms covered are principally focussed on genes of the HLA region and correspond to only 90 populations compiled for anthropological analysis. Conversely, datasets in the AFND comprise data for HLA, KIR, MIC and cytokine genes from diverse sources including data from registries, blood donors and individuals selected as controls in disease studies.

AFND

The AFND website was initially introduced in 2002 when a first prototype was designed to collect information of several immune genes (Middleton et al. 2003). In its origins, the website was particularly designed to store frequencies for HLA Class I and Class II classical genes (HLA-A, -B, -C, -DRB1, -DPA1, -DPB1, -DQA1 and -DQB1). In its inception, the database consisted of only 59 HLA populations having a significant increase over the last 8 years (Figure 3.1).

In the early stages, the website was limited to the retrieval of information from a narrow number of dynamic web pages. One of the major concerns in the first release was the data validation process due to the null interaction with external databases. Moreover, the back-end of the database consisted of a Microsoft Access Database (.mdb) file with a small number of unrelated tables which did not fulfil the compliances of database integrity. In September 2007, the structure of the AFND was completely redesigned and new searching mechanisms and the new structure were set up online in October 2008, which are the subject of this thesis. The development of the new database version

included the addition of frequency data from non-classical HLA genes (HLA-DMA, -DMB, -E and -G), KIR genes, MHC Class I related genes A and B (MICA, MICB) and a number of cytokine gene polymorphisms.

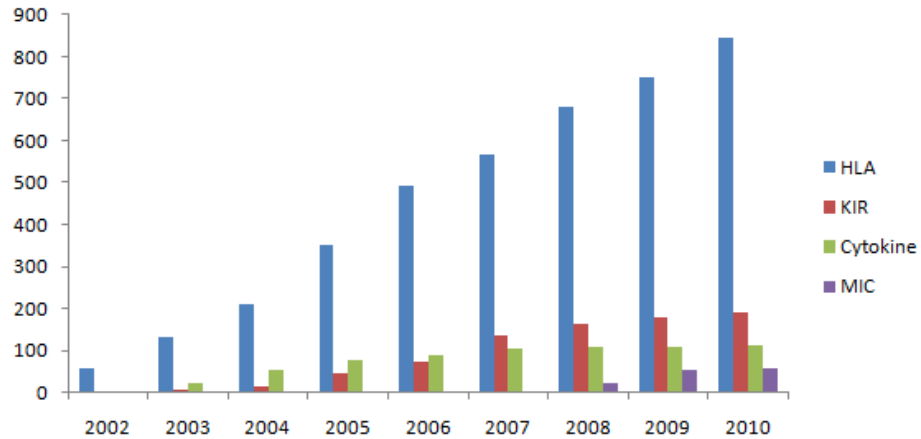


Figure 3.1: Number of populations reported on the AFND by year.

3.2 Systems and methods

3.2.1 Website organisation and dataflow

The main objective in the design of the AFND was to provide individuals with a software infrastructure to store immune gene frequency datasets, incorporate new submissions, validate existing data by the use of a controlled vocabulary and external libraries, and provide online tools for the consultation and analysis of data. Figure 3.2 shows the typical workflow performed in the submission of a new population. Each submission (normally provided in spreadsheets, tab-separated text files or XML formats) is sent to the AFND or entered online. Then, the website performs a validation process including the confirmation of the official nomenclatures from the IMGT/HLA and IPD-KIR databases. After that, data is classified according to the type of frequency (alleles, haplotype and genotypes). Finally, end users are provided with several views based on the different cases of study described in Section 1.8.

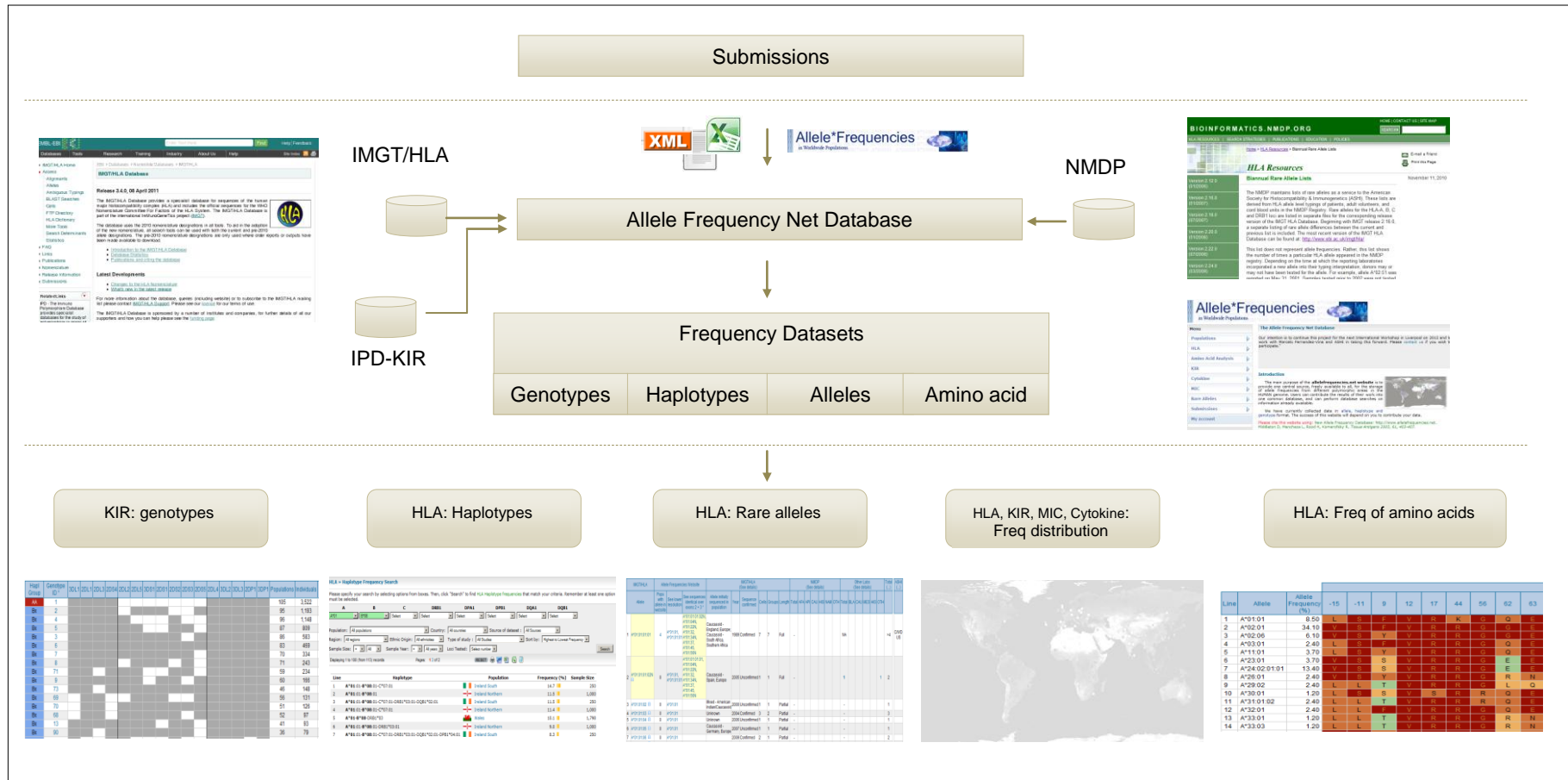


Figure 3.2: Workflow and system architecture of the AFND.

Based on the available polymorphisms, the AFND website was divided into four main sections containing information on HLA, KIR, MIC and cytokine gene polymorphisms. Each section consisted of one or more searches depending on the availability of data in each polymorphic region, e.g. allele, haplotype or genotype frequency, and a breakdown section to summarise the existing data (Figure 3.3). In the first release of the website, a registration process was mandatory for all users. The registration consisted of submission of basic information about the user. The aim of the registration was to identify the interest of the person using the website to consider possible improvements. To meet the requirements of public databases, user registration was excluded in August 2010 and was only required for submissions of new populations.

The Allele Frequency Net Database - Allele, haplotype and genotype frequencies in Worldwide Populations

http://allelefrequencies.net/

Allele*Frequencies

in Worldwide Populations

Menu

- Populations
- HLA**
- Amino Acid Analysis
- KIR
- Cytokine
- MIC
- Rare Alleles
- Submissions
- My account

The Allele Frequency Net Database

"We want to thank all individuals who contributed to the Rare Alleles Project presented at the 15th IHWS been published:

a M, Tiercy J-M, Marsh SGE, et al. (2009) A bioinformatics alleles. *Tissue Antigens* **74**:480-485.

for the next International Workshop in Liverpool on 2012 and to SHI in taking this forward. Please [contact us](#) if you wish to

Allele Frequency Net Database is able to all, for the storage orphic areas in the Human Genome. Users can contribute the results of their work into one common database, and can perform database searches on information already available.

We have currently collected data in **allele, haplotype and genotype** format. The success of this website will depend on you to contribute your data.

New: Please cite this website using our last publication: Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Gonzalez-Galarza FF, Christmas S, Middleton D and Jones AR *Nucleic Acid Research* 2011, **39**, D913-D919. [Full Text]

Database information

Polymorphic Region	Population Studies	Gene/Allele Data	Haplotype Data	Genotype Data
HLA	849	834	343	-
KIR	196	196	-	87
Cytokine	113	113	-	-
MIC	58	58	20	-
Totals	1,216	1,201	363	87

The current number of Frequencies stored in our database is: **87,314** (HLA), **4,959** (KIR), **3,603** (Cytokine) and **723** (MIC) from **4,265,576** individuals.

We have updated the website with the new IMGT/HLA nomenclature [guidelines](#).
 IMGT/HLA last Update: 3.5.0, 14 July, 2011. | IPD-KIR last Update: 2.4.0, 15 April, 2011.

Sponsors

- Abbott laboratories
- BAG Health Care
- BIO-RAD
- Invitrogen
- Innogenetics
- Olerup SSP AB
- One Lambda
- Gen-Probe

Figure 3.3: Screenshot of the AFND website.

Internally, the AFND website consists of two main layouts: *user searches* and a component for *data maintenance* (Figure 3.4). The first layer encompasses the different searching tools that can be accessed by any individual and the second a password restricted section for administrators/curators of the database.

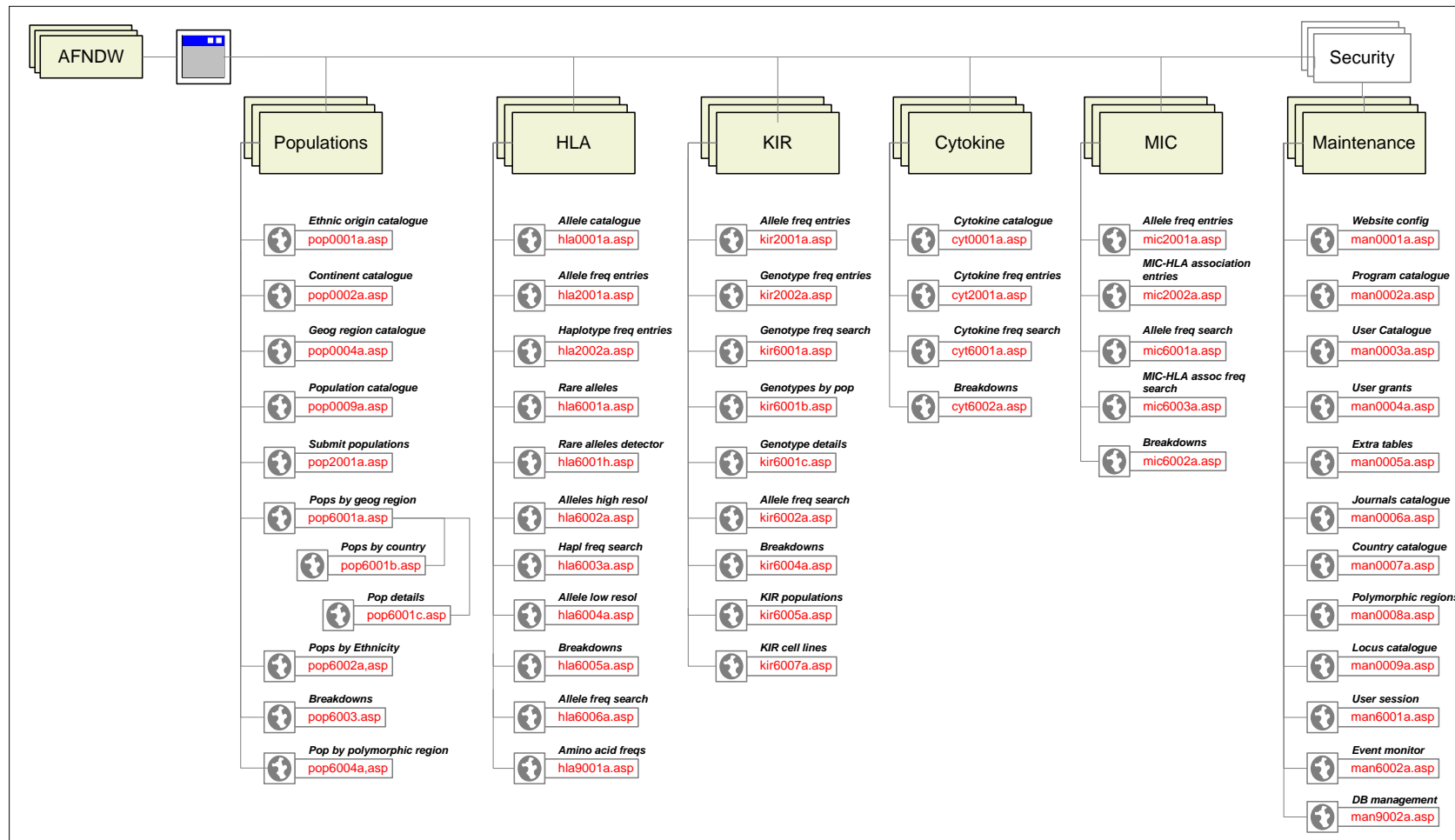


Figure 3.4: Organisation of the Allele Frequency Net Database website.

Nomenclature of programs

All searching interfaces available in the AFND website were classified using a code of 8 alphanumeric characters to uniquely identify the application and corresponding features (Figure 3.5). The use of these program codes provided an efficient approach to invoke different interrelated programs, i.e., allele and haplotype frequency searches for the HLA system.

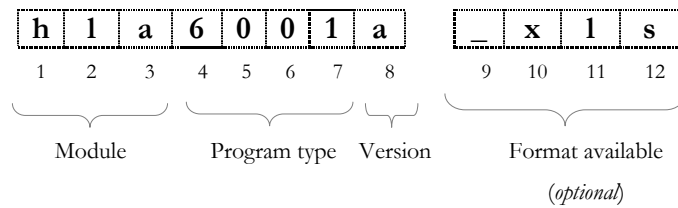


Figure 3.5: Program nomenclature in the AFND website.

The first three characters describe the type of module, which normally defines the polymorphic region (hla, kir, cyt and mic) or an alternative module, i.e. population samples (pop) and management of the catalogues (man). The next four digits (4-7) were used to identify the type of application. For instance, programs from 0001-1999 were assigned to catalogues, 2000-4999 for data processing, 5000-5999 for electronic data exchange, 6000-8999 for searches/reports and 9000-9999 for additional tools. The eighth character was used to identify the release version. Additionally, a suffix was used to define the type of format available for data reporting (xls, csv, txt, xml).

3.2.2 Software implementation

Web pages were implemented using the ASP scripting environment with the assistance of the JavaScript for data entry validation. Additionally, the AJAX technology was used in some of the searching mechanisms to allow simpler user interaction and improved visualisations. To facilitate the maintenance and retrieval of information, the back-end of the AFND is based on a relational database model utilizing MySQL 5.1 (<http://www.mysql.com/>) as the DBMS. For the curation of the datasets a set of stored procedures were executed in a mirror database using in the Microsoft SQL Server 2008

R2 Express Edition (<http://www.microsoft.com/sqlserver>) for rapid analysis of datasets. The analysis of information and summaries described in this thesis were performed by direct queries to the AFND using preconfigured stored procedures which were archived in the hosting server for future use. These stored procedures can be requested by demand via the website.

3.2.3 Data visualisation

To ensure the correct display of maps in web browsers, the graphical user interface was created using pure HTML and Cascading Style Sheets (CSS) to guarantee a standard visualisation in all browsers. All web pages in the AFND website were tested on Internet Explorer 7®, Firefox 3.5®, Safari 4®, Opera 9® and Google Chrome 3® web browsers. Additionally, the HTML code was validated using the Markup Validation Service (<http://validator.w3.org/>) to optimise the quality of the web pages based on the guidelines provided by the World Wide Web Consortium (W3C).

3.2.4 Systems and hardware requirements

- i) **Server requirements.** The AFND website was hosted on a remote server. The following minimal configuration is required for optimal processing.
 - *Hardware*
 - **Processing power.** Intel Core Quad CPU @ 2.83GHz.
 - **Memory.** 8GB RAM.
 - **Secondary storage.** 1GB for database files.
 - **Network access.** T1 or optical fibre.
 - *Software*
 - **Web server.** Internet Information Service 7 to support Classic ASP.
 - **Operating system.** Windows 2003/2008 Server (32-bits | 64-bits).
 - **Other utilities.** MailSender API for e-mail automatic notifications and ASPUpload for uploading files.

ii) Client machine requirements.

To guarantee the rapid access to the website from any modern computer, the website was tested on different platforms. However, the following configuration is recommended for optimal navigation performance.

- **Hardware**

- **Processing power.** 1 GHz (32-bit|64-bit) processor.
- **Memory.** 500MB RAM (1GB recommended).
- **Secondary storage.** 100MB for data buffering.
- **Network access.** The website was tested in different networks including Digital signal 1 (DS1) @ 1.4Mbps, Digital Subscriber Line (DSL) @ 256Kbs, Cable @ 512Kbs, Dialup @ 56Kbs, one Optical Carrier level 3 (OC3) @ 100Mbps, and Integrated Services Digital Network (ISDN) @ 128Kbs. A minimum connection of 512Kbs is recommended.

- **Software**

- **Operating system.** The website was tested in the following operating systems: Mac OS X v10.5, Windows XP Service Pack 2, Windows Vista, Windows 7 and Linux Fedora core 14.
- **Web browser.** Internet Explorer 7+, Safari 4+, Firefox 3+, Opera 9+ and Chrome 3+.
- **Other features.** Javascript should be enabled in web browsers for web pages with AJAX based implementations.

3.3 Navigation in population samples

As of May 2011, the AFND consisted of more than one thousand populations which were described in Section 2.2.4. To allow individuals to consult the populations available in an interactive manner, a graphical interface was included in the website based on the use of the Google maps API to display the population's geographical location (Figure

3.6). This tool was intended to serve as an index module to provide rapid access to the demographic details of the population sample. Figure 3.6 illustrates the wide coverage of HLA (blue), KIR (red), MIC (yellow) and cytokine genes (green) across the world.

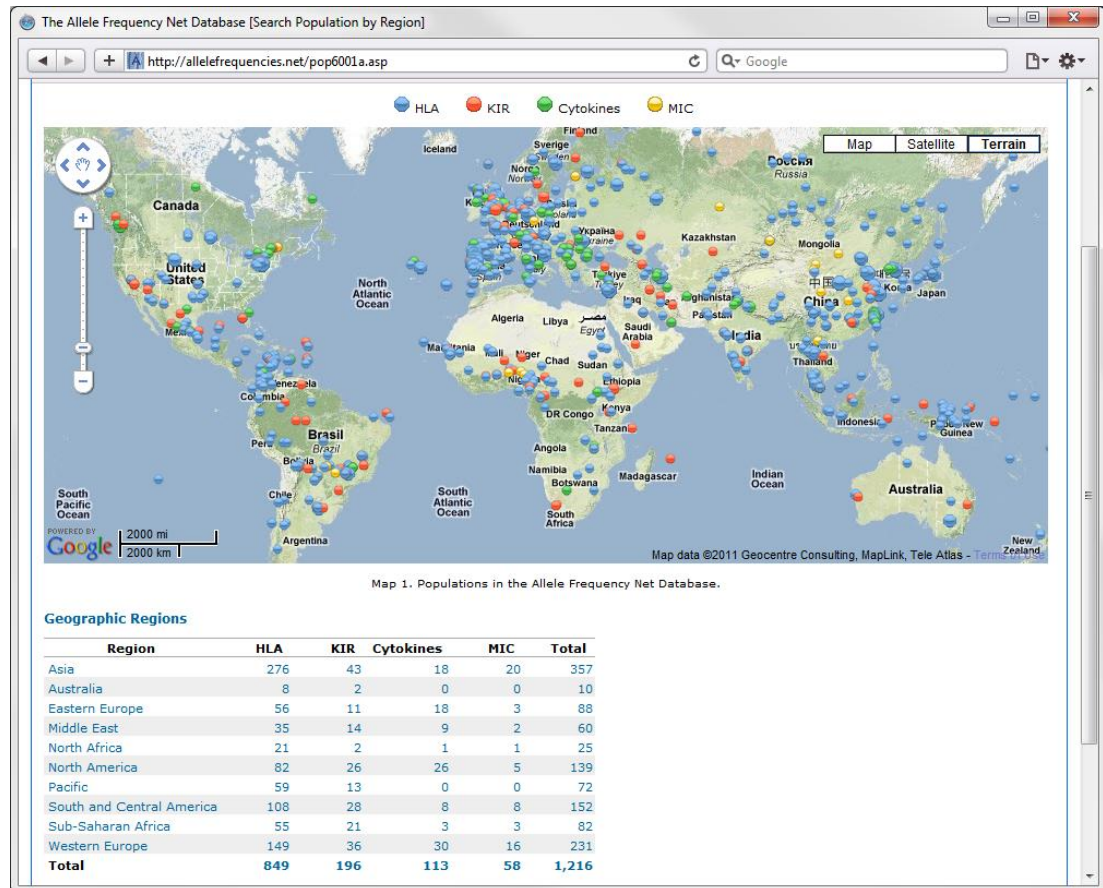


Figure 3.6: Geographical location of population samples in the AFND.

3.4 Frequency Search Interfaces

The core of the implementations in the AFND website involved principally the examination of frequency datasets by providing a set of tools to search and explore the different types of frequencies at allele, haplotype, genotype and amino acid level.

3.4.1 The Allele Frequency Search (AFS)

Since its inception, the most commonly used tool within AFND website has been the *Allele Frequency Search* (AFS). This search was designed to allow users to examine the

frequency of a particular allele in the existing population sample datasets, by filtering results using a set of criteria. The AFS was implemented for all polymorphisms available on the website (HLA, KIR, MIC and cytokine genes).

In this search, users normally start with the selection of a locus and a specific allele to identify which populations are more likely to present this allele. Figure 3.7 shows an example of the search by determining the A*68:01 allele in all populations available in the AFND. The example illustrated in Figure 3.7 includes two additional parameters (shown in green) to specify the source of datasets (i.e., populations extracted from the literature) and a minimum typing resolution level (≥ 2). By setting these two parameters users would guarantee that resulting frequencies followed a peer-review process performed by journals and would exclude alleles with lower resolution (e.g. A*68).

The screenshot displays the Allele Frequency Net Database search interface. The search criteria are as follows:

- Locus: A
- Starting Allele: A*68:01
- Ending Allele: A*68:01
- Source of dataset: Literature
- Level of resolution: ≥ 2

The search results are displayed in a table with the following columns: Line, Allele, Population, Phenotype Frequency (%), Allele Frequency (in_decimals), Sample Size, IMGT/HLA¹ Database, Distribution², Haplotype³ Association, and Notes⁴.

Line	Allele	Population	Phenotype Frequency (%)	Allele Frequency (in_decimals)	Sample Size	IMGT/HLA ¹ Database	Distribution ²	Haplotype ³ Association	Notes ⁴
1	A*68:01	India Delhi pop 2	18.9	0.0940	90	See			
2	A*68:01	Burkina Faso Mossi		0.0940	53	See			
3	A*68:01	Mexico Chihuahua Tarahumara		0.0910	44	See			
4	A*68:01	Pakistan Burusho		0.0820	92	See			
5	A*68:01	Senegal Niokholo Mandenka		0.0700	165	See			
6	A*68:01	Georgia Svaneti Region Svan		0.0630	80	See			
7	A*68:01	Portugal Center		0.0600	50	See			
8	A*68:01	Belgium	11.5	0.0570	99	See			
9	A*68:01	Cameroon Baka Pygmy		0.0500	10	See			

Figure 3.7: Example of the Allele Frequency Search (AFS).

To extend the searching parameters, users can select one, several or all populations, a set or range of alleles, or a particular country, geographical region and/or ethnicity. In HLA, MIC and KIR polymorphisms, alleles can be typed at different levels of resolution (i.e. allele group, specific HLA protein, synonymous allele with a nucleotide substitution within the coding region and alleles with differences in a non-coding

location in that order e.g. HLA-A*68:01:01:01). The official nomenclature available on the IMGT/HLA and IPD-KIR databases describes alleles only at the highest resolution. Therefore, to ensure that high-resolution data could be retrieved when a low level resolution allele was selected, the searching tool implements parsing methods to display information that may be relevant to the user. For instance, in the search of the HLA-A*68:01 allele, the results will include alleles at high-resolution that start with HLA-A*68:01 (e.g. A*68:01:01:01, A*68:01:01:02 and A*68:01:02 to A*68:01:10).

Additionally, users are able to optimise their queries to further refine datasets by selecting populations with a sample size from a range of values or the year in which the population was sampled. Populations from recent years are more likely to contain alleles with a high-resolution level and possibly, more accurate data. Furthermore, recent additions included filters to search information on a specific source of dataset (e.g. literature, unpublished data) and type of study (e.g. populations available in the literature oriented to anthropology studies, or individuals which were part of a Bone Marrow Registry, blood donors, solid organ unrelated donors or which were controls for a certain disease study).

Display of results

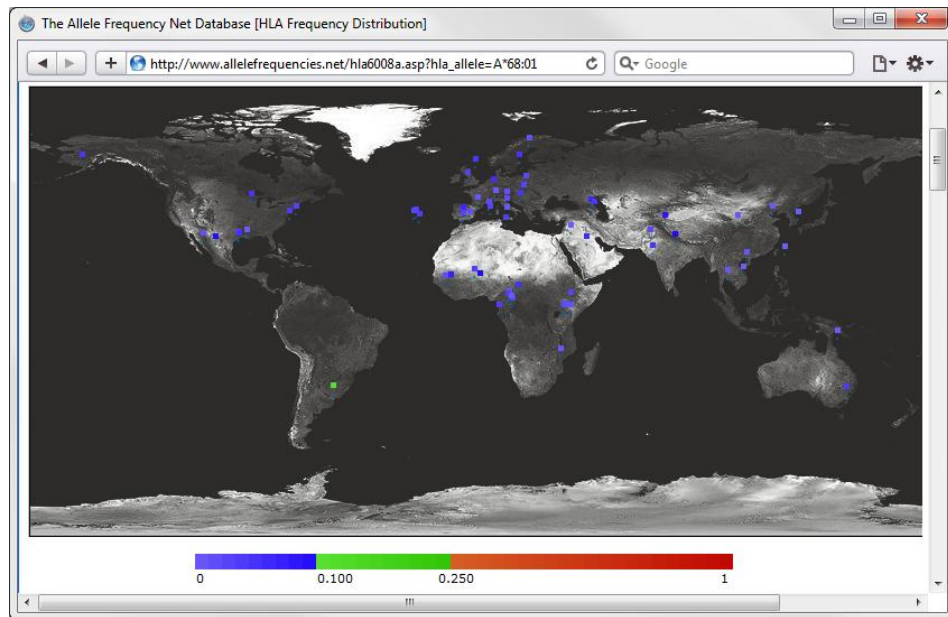
Results displayed in the AFS include the allele name, name of the population, allele frequency and/or the percentage of individuals which carry the allele, sample size of the population and additional notes included in the original submission. By clicking on the 'Population Name' hyperlink, users can access demographic details of the population in which the allele is present. The list of output records can be sorted by allele or population and the corresponding frequency from highest to lowest value. An additional column is included in the results to consult the original sequence of the allele in the IMGT/HLA or IPD-KIR databases depending on the selected search.

Frequency distribution maps

One of the most used tools in the examination of the distribution of allele frequencies is the *map overlay display*. This interface, which does not require a complex mechanism to depict allele frequencies, can clearly illustrate the frequency distribution of a particular

allele across the world (Figure 3.8A and B). The example shown in this figure illustrates the overall frequency distribution of the HLA-A*68:01 allele. If one examined the distribution at protein level (Figure 3.8A), the frequency distribution may look uniform, however, when the search includes a higher resolution, certain alleles may be more specific of certain regions. For example, HLA-A*68:01:02 has been found to be higher in several Amerindian populations (Figure 3.8B). The differences in frequencies in these two figures correspond to the availability of data for each different level of resolution.

A) A*68:01



B) A*68:01:02

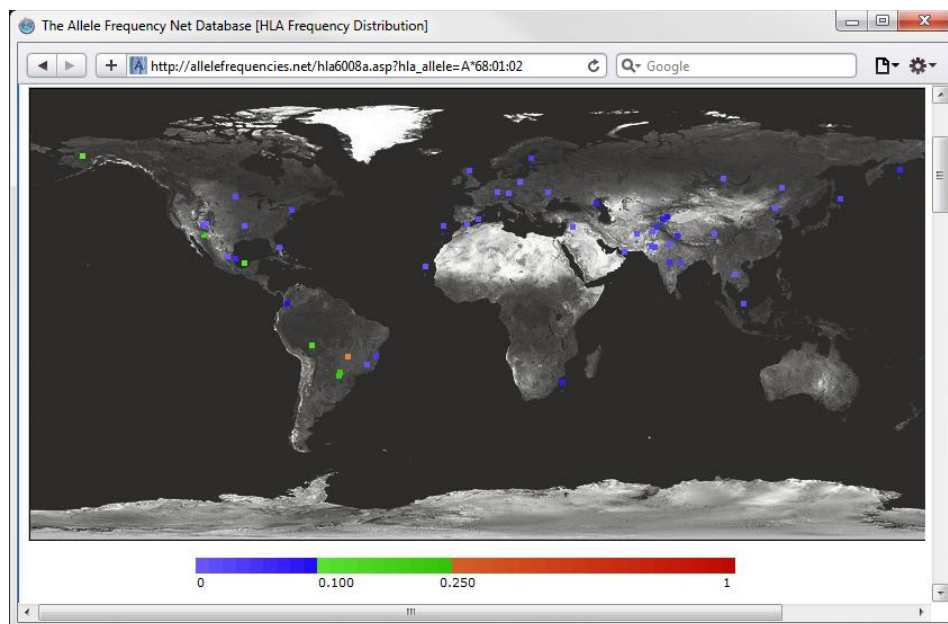


Figure 3.8: Overall frequency distribution of the A*68:01 allele.

In this figure, A) Frequency distribution of the A*68:01 allele and B) Frequency distribution of the A*68:01:02 allele.

Grid display

Another facility included in the AFS is the display of frequencies in a grid format. This option allows a set of alleles to be specified; e.g. alleles from an individual's tissue type. This option is mainly used by clinical scientists to compare the frequencies of a given tissue typing and rapidly identify those populations with the highest occurrences of the alleles. Figure 3.9 shows an example of the frequencies of two alleles HLA-A*68:01 and HLA-B*40:02 which belong to the same individual. To optimise the visualisation of the results, a maximum of 100 populations and twenty specific alleles are allowed in the grid display. The grid is only an approximation to use the frequencies of the allele in the populations when raw data is not available.

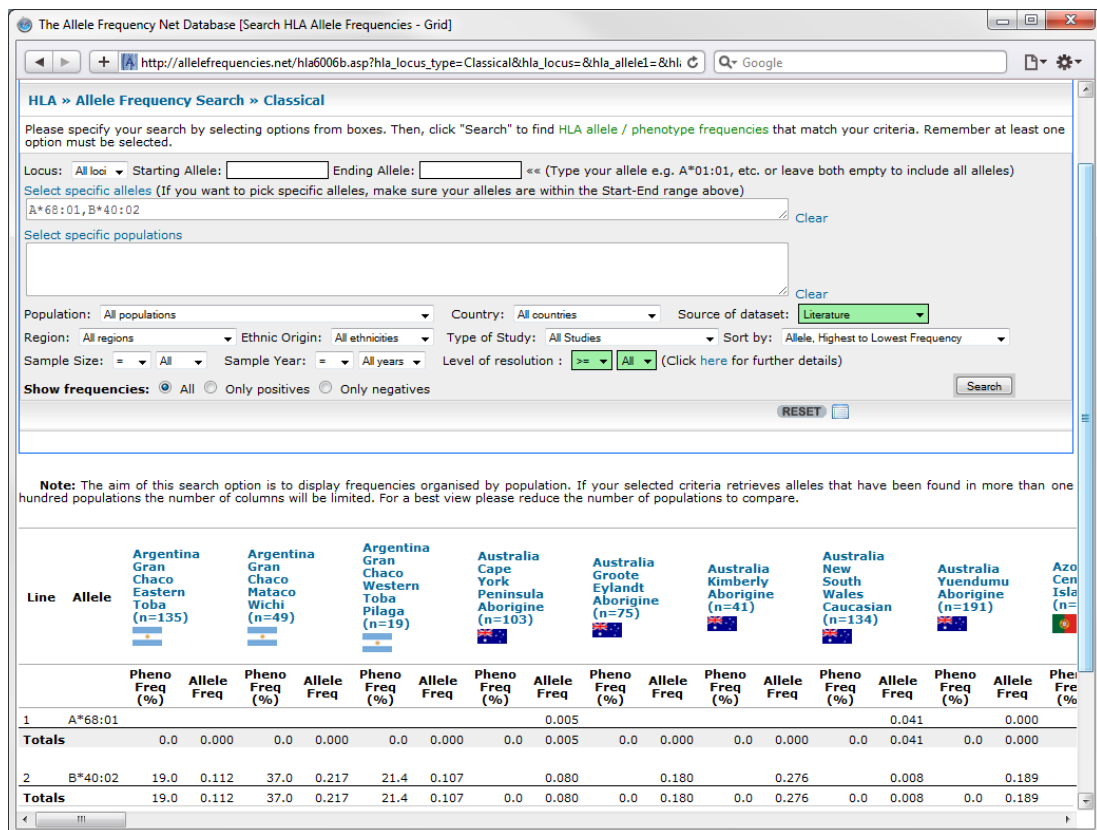


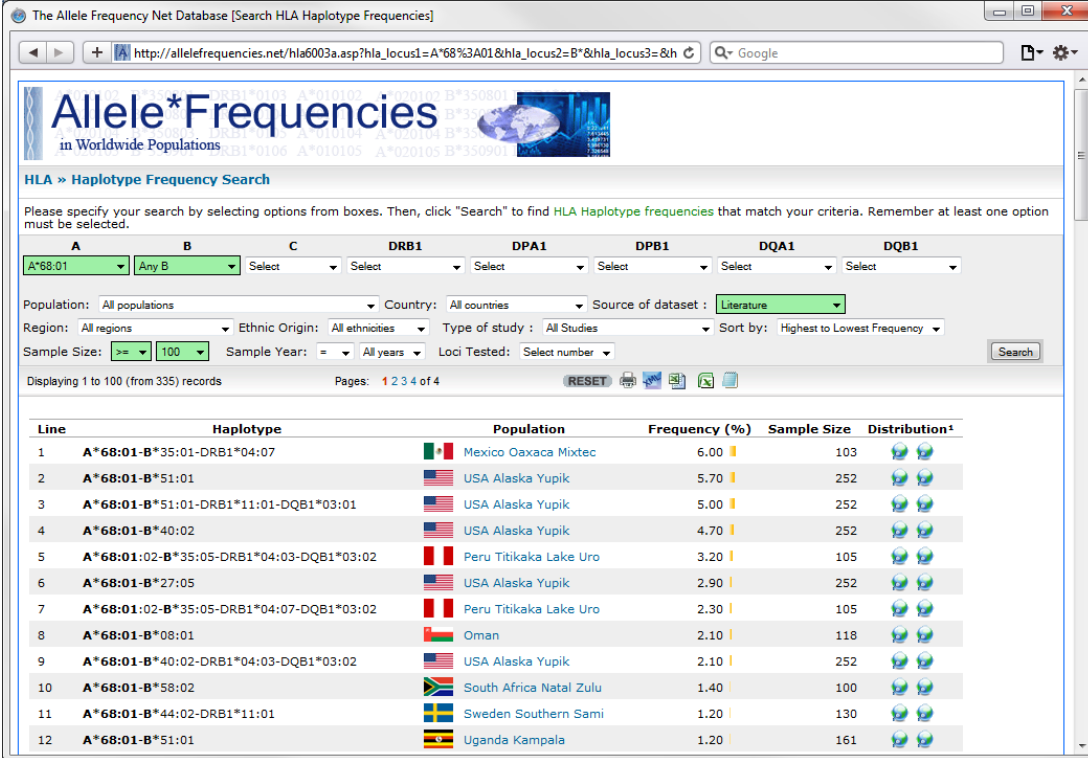
Figure 3.9: Example of the grid view on the AFS.

3.4.2 The Haplotype Frequency Search (HFS)

The AFND website also includes a tool for querying haplotype frequencies called the *Haplotype Frequency Search* (HFS) (Figure 3.10). As of May 2011, the data available for this search consisted of 8,916 HLA haplotypes and 257 MICA-HLA-B association records

from more than 3 million individuals, which are discussed in detail in Chapters 4 and 7. The origin of the haplotypes corresponded to 342 globally distributed populations from 81 countries.

The HFS allows users to customise a frequency search by inputting a given allele for one or several loci and search for associated haplotypes. Results can be filtered by a particular population, country, source of data, geographical region, ethnicity of the individual, and number of loci tested for the haplotype.



The screenshot shows the AFND HFS interface with the following search criteria: A*68:01, Any B, Population: All populations, Country: All countries, Source of dataset: Literature, Region: All regions, Ethnic Origin: All ethnicities, Type of study: All Studies, Sort by: Highest to Lowest Frequency, Sample Size: >= 100, Sample Year: All years, Loci Tested: Select number. The results table is as follows:

Line	Haplotype	Population	Frequency (%)	Sample Size	Distribution ¹
1	A*68:01-B*35:01-DRB1*04:07	Mexico Oaxaca Mixtec	6.00	103	
2	A*68:01-B*51:01	USA Alaska Yupik	5.70	252	
3	A*68:01-B*51:01-DRB1*11:01-DQB1*03:01	USA Alaska Yupik	5.00	252	
4	A*68:01-B*40:02	USA Alaska Yupik	4.70	252	
5	A*68:01:02-B*35:05-DRB1*04:03-DQB1*03:02	Peru Titikaka Lake Uro	3.20	105	
6	A*68:01-B*27:05	USA Alaska Yupik	2.90	252	
7	A*68:01:02-B*35:05-DRB1*04:07-DQB1*03:02	Peru Titikaka Lake Uro	2.30	105	
8	A*68:01-B*08:01	Oman	2.10	118	
9	A*68:01-B*40:02-DRB1*04:03-DQB1*03:02	USA Alaska Yupik	2.10	252	
10	A*68:01-B*58:02	South Africa Natal Zulu	1.40	100	
11	A*68:01-B*44:02-DRB1*11:01	Sweden Southern Sami	1.20	130	
12	A*68:01-B*51:01	Uganda Kampala	1.20	161	

Figure 3.10: Example of the Haplotype Frequency Search (HFS).

Figure 3.10 shows an example of the alleles associated to the HLA-A*68:01 allele containing at least one HLA-B allele. As shown in this search, the highest frequency was found in the 'Mexico Oaxaca Mixtec' population with a frequency of 6% in the A*68:01-B*35:01-DRB1*04:07 haplotype (Hollenbach et al. 2001). The haplotypic information can be more useful than information only on the allele, especially in clinical applications, i.e. in transplantation in which the higher number of matches will result in best chances of survival. Therefore, this search can be used as a complement of haplotype searches performed in Bone Marrow and Solid Organ Transplant Registries in which, on some occasions, the information about the ethnicity of the individual is not

known. The HFS is available for the eight routinely typed HLA loci (HLA-A, -B, -C, -DRB1, -DPA1, -DPB1, -DQA1 and -DQB1) and for MICA and HLA-B associations.

3.4.3 The Genotype Frequency Search (GFS)

One of the latest developments included in the AFND website was the compilation of an inventory of KIR genotype profiles (presence or absence of genes in and individual) published in the literature. This novel KIR genotype database comprises the most extensive archive of KIR genotypes and their corresponding frequencies in worldwide populations. As of May 2011, genotype data encompassed 2600 records of which 396 distinct KIR genotype profiles were identified. The content of the KIR genotype database will be reviewed in detail in Chapter 6.

In the *Genotype Frequency Search* (GFS), users can search for a particular genotype and examine its corresponding frequency from a list of 108 populations available with KIR genotype data. As mentioned in Chapter 1, the KIR gene cluster consists of fifteen genes and two pseudogenes; however, the availability of KIR genotype data from publications was limited to 16 genes due to *KIR2DL5A* and *KIR2DL5B* genes were reported as KIR2DL5. Therefore, the KIR genotype composition in the GFS consisted of sixteen genes which defined the presence or absence of the gene.

Genotype reference list

The GFS provides different approaches to find the occurrence of a specific KIR profile. A list of all genotypes and the number of populations and individuals in which the profile has been found is shown in the main screen (Figure 3.11).

The information displayed includes the haplotype group of the genotype (AA, Bx where x can be A or B), the genotype ID which is automatically assigned by the AFND as a consecutive number, the genes that constitutes the genotype, the number of populations and individuals in which the genotype was found, and the name of the population if the genotype was reported only once.

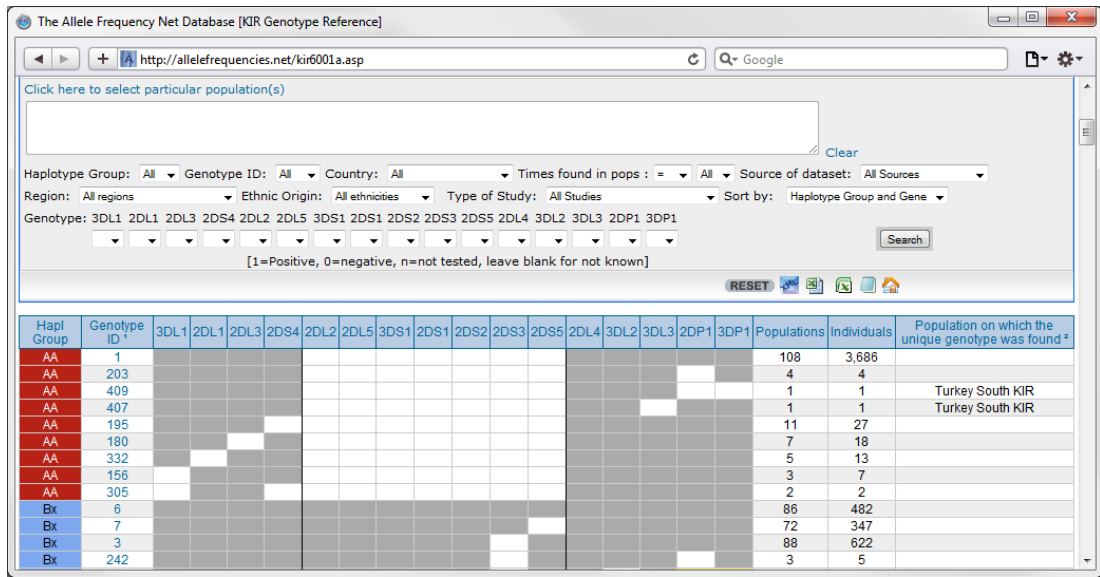


Figure 3.11: Example of the Genotype Frequency Search (GFS).

In this figure, KIR genotypes are illustrated as the presence (grey) and absence (white) of the gene.

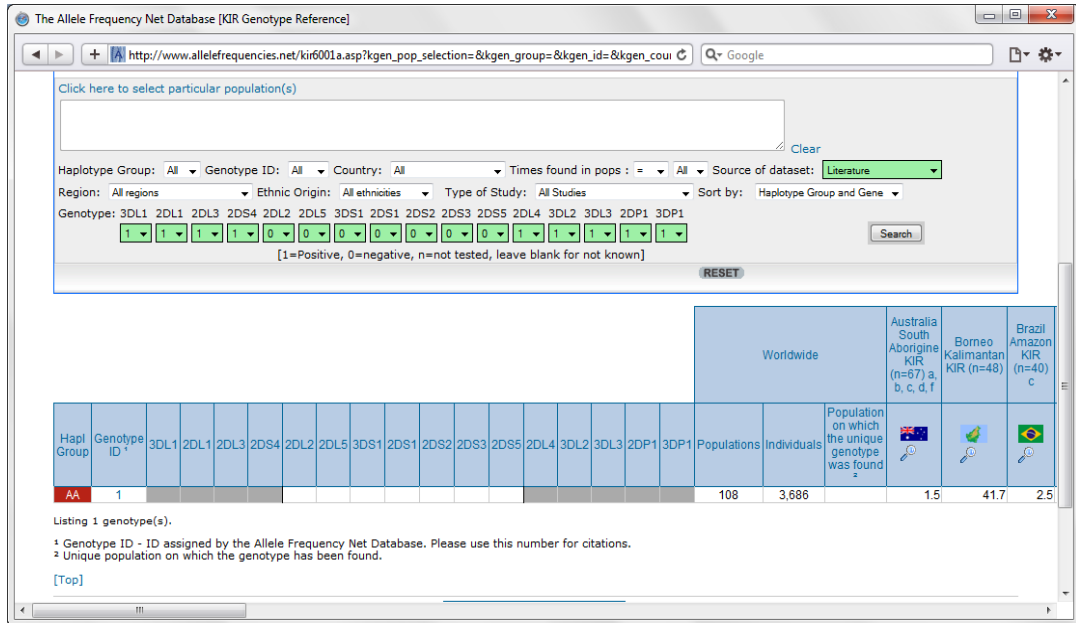
Searching a genotype

Users are provided with a range of options including the selection of one or many populations and one or many genes that constitute the genotype. Similarly to the AFS and HFS, users can select additional criteria to filter results by geographical region, ethnic origin, type of study and the source of datasets.

Figure 3.12A shows an example of the most common ‘A’ genotype (Genotype 1). This genotype has been reported to be present in all of the 108 population samples and in a total of 3,686 individuals at different frequencies. For instance, the ‘**Borneo Kalimantan KIR**’ population (Velickovic et al. 2009) presented a high frequency (41.7%), contrasting with ‘**Australia South Aborigine KIR**’ and ‘**Brazil Amazon KIR**’ populations with 1.5% and 2.5% respectively (Ewerton et al. 2007; Toneva et al. 2001). If users click on the genotype ID, a graphical summary of the frequency distribution across the world is presented as shown in Figure 3.12B. Additionally, the total number of genotypes of a given population can be accessed by clicking on the name of the population. The search presents a similar organisation of results by including the haplotype group, genotype ID, KIR genes in the genotype and the corresponding frequency (Figure 3.12C). Finally, if the genotype is not found from the initial search, users are provided with a list of the closest genotypes differing by one

gene (Figure 3.12D). This option allows individuals to identify possible mistakes in the determination of KIR genes or typographical errors in reporting genotypes.

A)



B)

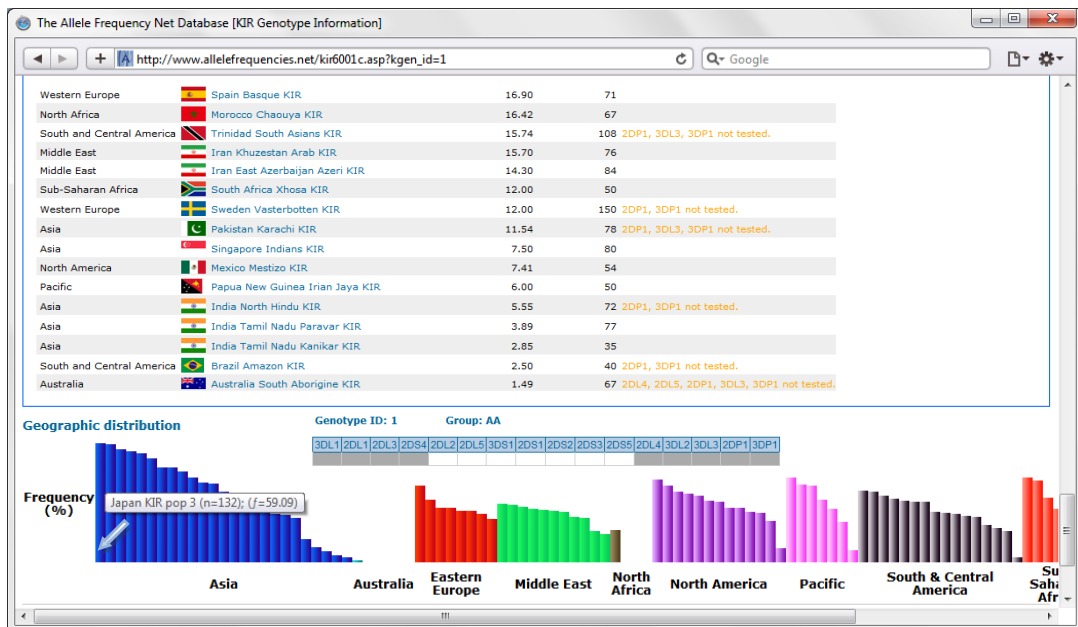
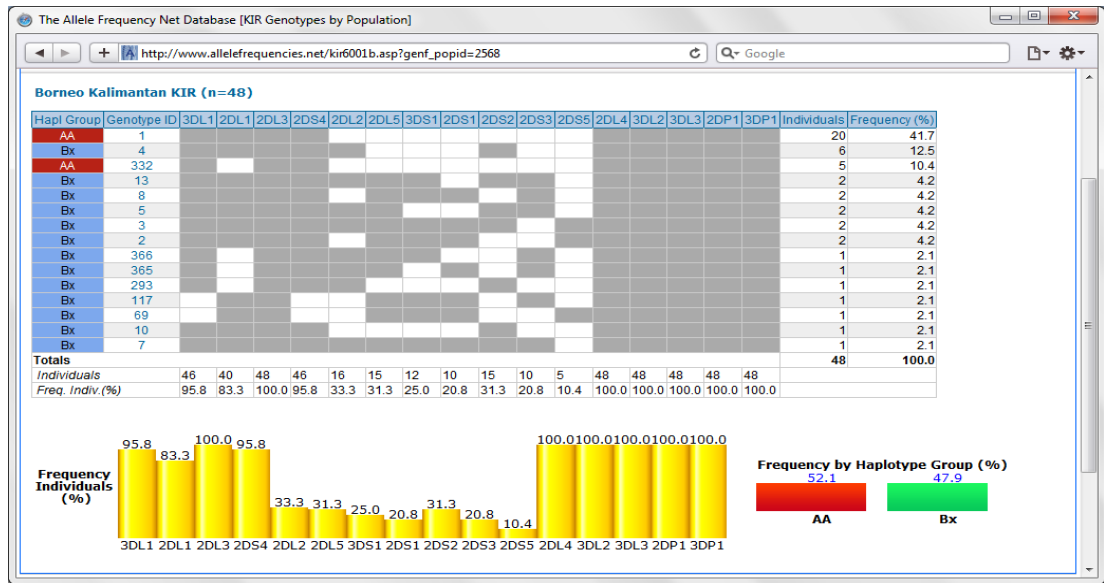


Figure 3.12: KIR genotype searching options in the AFND.

A) A search is performed to query a specific KIR genotype. B) Frequency distribution of the KIR genotype ‘1’ in worldwide populations.

C)



D)

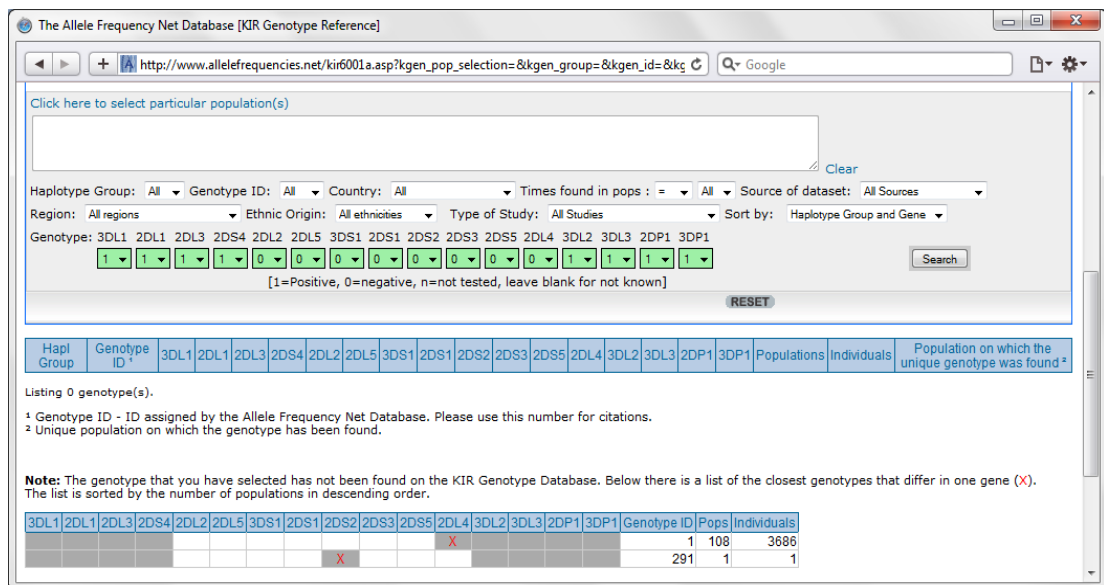


Figure 3.12: KIR genotype searching options in the AFND (Continued).

C) Total genotypes in the ‘Borneo Kalimantan KIR’ population (Velickovic et al. 2009). D) Search performing a query of a genotype which has not been reported to the AFND. Two closest genotypes differing by one gene are shown in the example.

3.5 Amino acid frequency comparisons in populations

One of the approaches commonly used by researchers in disease association studies is to compare frequencies of HLA alleles between patient and control groups. The development of the AFND website included a bioinformatics tool called the *Amino Acid Frequency Analysis Tool* (AAFAT). This tool allows users to investigate potential molecular mechanisms in disease associations by analysing the main differences in frequencies for a specific position of the allele at amino acid level. Figure 3.13 shows an example of the differences in amino acids at a given position, based on the original sequence submitted to the IMGT/HLA database.

A summary of frequencies for each differing amino acid is presented, allowing users to compare occurrences that may be implicated in the association. One additional feature is that users can also enter their own data in a tab-separated text file or use one of the existing populations available in the AFND. All populations available for this analysis need types at protein level or above (e.g. A*68:01) for the analysis to be performed.

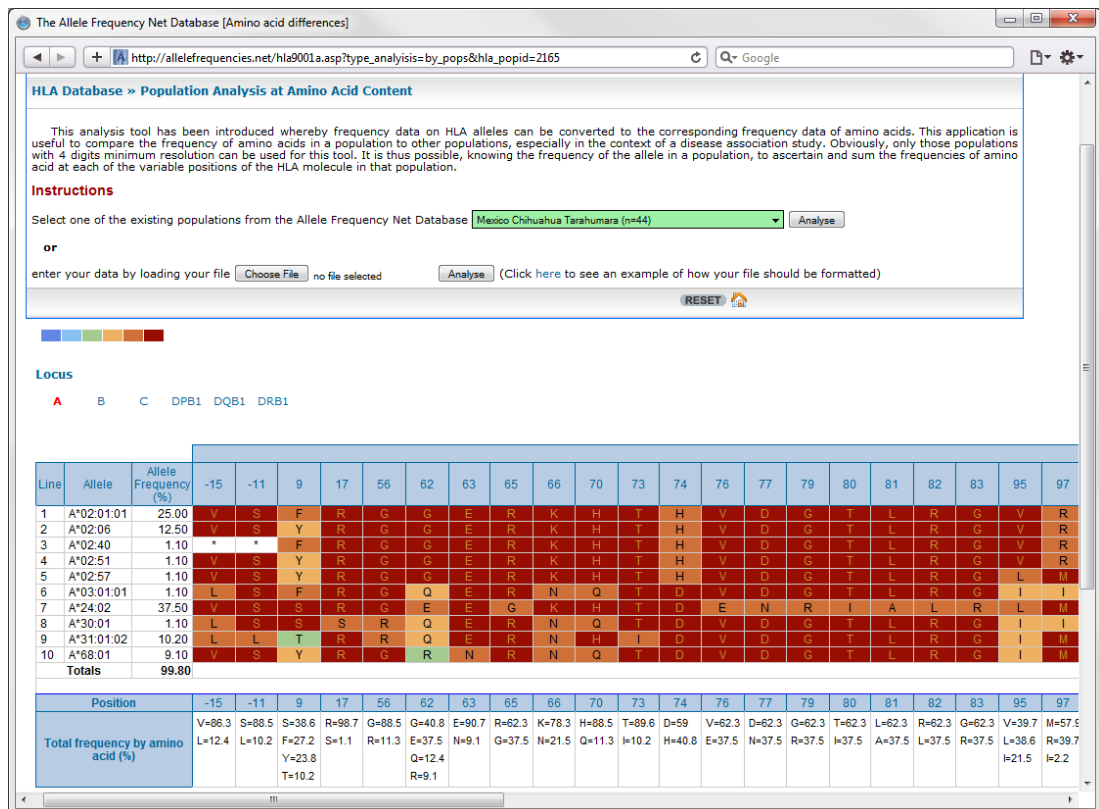


Figure 3.13: Example of the Amino Acid Frequency Analysis Tool (AAFAT).

3.6 Accessibility of data

An important feature included in the AFND website was the ability to interact with external websites by the use of bidirectional links. For instance, in the AFS, a complete list of all population samples carrying the A*68:01 allele can be accessed using the code **hla_selection=A*68:01** at the end of the application's URL.

E.g. http://www.allelefrequencies.net/hla6006a.asp?hla_selection=A*68:01

The complete list of reference links is available on the 'External access' section of the website (<http://www.allelefrequencies.net/extaccess.asp>).

In addition, the site provides the option to export data to different format files including XML, tab-separated and comma-separated text files for the 'HLA Rare Allele' section (described in Chapter 5), allowing users to integrate the information available in AFND with alternative bioinformatic packages. At present, users can print results from all searches using the printer-friendly version available for each search which can be used to export datasets in tabular format. To complement frequency data in searches for further analyses, the printer-friendly option includes information of latitude and longitude if users wish to plot frequencies on maps.

3.7 Database users and software metrics

The use of software metrics, especially in research, is an essential requirement. Thus, quality assurance tests were performed by optimising the number of line codes and program size for each module available in the website. To complement the software metrics, the AFND website included the Google Analytics API (available at <http://www.google.com/analytics/>) to examine the different users who have accessed the website along with the system configuration used for querying data. From 4 August 2010 to 4 August 2011, a total of 37,077 visits were recorded comprising 15,784 unique visitors from 2,758 cities in 136 countries (Figure 3.14).

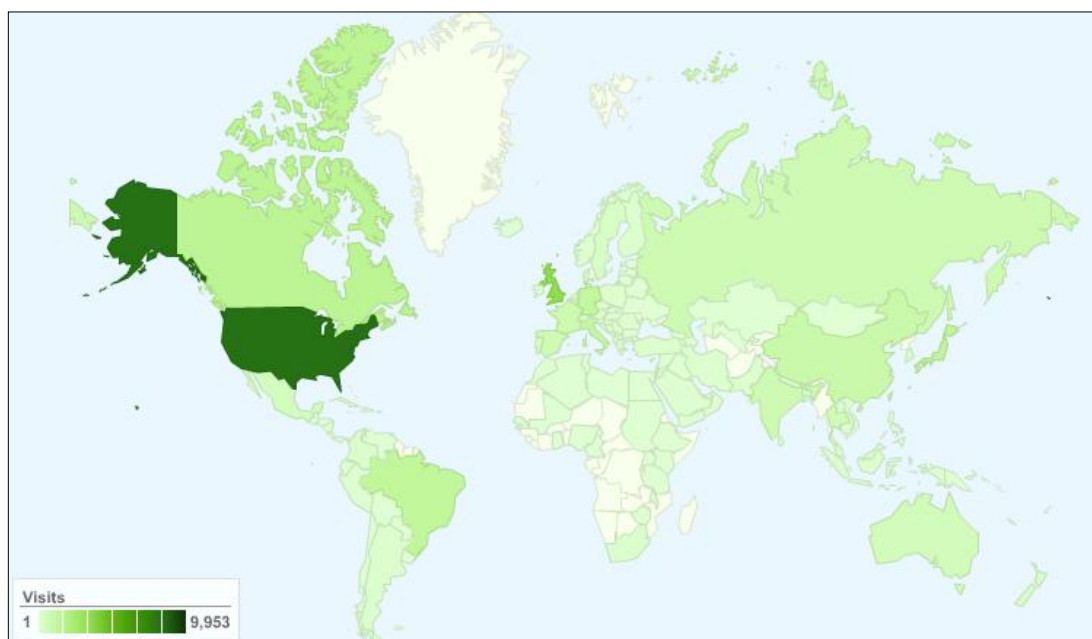


Figure 3.14: Number of visits to the AFND by country.

To identify user preferences, a page count was implemented in all searches. Table 3.1 summarises the top 10 searching tools accessed in the AFND website. The majority of visits corresponded to users interested in consulting the AFS and HFS in the HLA region. Interestingly, the Rare Allele Search, which will be discussed in detail in Chapter 5, was the third most popular search in the AFND despite its recent implementation.

Table 3.1: Top 10 searches accessed by users

Searching tool	Page views	%
HLA Allele Frequency Search	95,177	40.0
HLA Haplotype Frequency Search	35,876	15.1
Publication Details Report	16,222	6.8
Rare Allele Search	12,698	5.3
KIR Genotype Search	7,633	3.2
KIR Allele Frequency Search	4,892	2.1
Population by Geographical Region Search	3,070	1.3
HLA Frequency Distribution Map	2,911	1.2
HLA Allele Frequency Grid Search	2,111	0.9
Geographical Regions Search	1,984	0.8
Other searches	55,185	23.2

The analysis of software metrics was extended to examine the different network connections (Table 3.2), platforms (Table 3.3) and browser capabilities (Table 3.4) to provide users with an optimal performance for the retrieval and visualisation of data. As

expected, the operating systems and web browsers in which the website was tested comprised 99.3% and 99.6% of the visits respectively, ensuring the correct retrieval and visualisation of data.

Table 3.2: Number of visits by connection speed

Connection Speed	Visits	%	Pages/ Visit	Average time on site in seconds
Unknown	24,154	65.1	6.4	7.4
T1@ 1.4Mbps	7,528	20.3	6.8	7.5
DSL @ 256Kbs	2,892	7.8	6.6	8.0
Cable @ 512Kbs	1,915	5.2	5.9	5.6
Dialup @ 56Kbs	510	1.4	6.6	7.1
OC3 @ 100Mbps	75	0.2	5.0	6.3
ISDN @ 128Mbps	3	0.0	1.7	1.1

Table 3.3: Number of visits by operating system

Operating System	Visits	%
Windows	31,845	85.9
Macintosh	4,465	12.0
Linux	521	1.4
iPhone	78	0.2
iPad	57	0.2
Android	23	0.1
iPod	13	0.0
Others	18	0.0
Not identified	57	0.2

Table 3.4: Number of visits by web browser

Browser	Visits	%
Internet Explorer	20,686	55.8
Firefox	9,551	25.8
Safari	3,379	9.1
Chrome	2,986	8.1
Opera	306	0.8
SeaMonkey	48	0.1
Others	121	0.3

3.8 Discussion

The implementation of searching mechanisms in the AFND website has provided the medical and scientific community with a practical approach to investigate the occurrence of genes and their corresponding alleles among populations. The multiple filter schemes performed in each of the frequency searches allows users to optimise their searches and obtain the closest matching results. For instance, by setting customised criteria, individuals can look into only specific datasets which followed a previous peer reviewed process and/or only alleles typed at high-resolution.

Comparison of AFND and other immune gene frequency databases

To my knowledge, two databases also contain frequency data of alleles of the HLA system: the dbMHC database (Single et al. 2007b) and ALFRED (Cheung et al. 2000). However, the aims of these two databases differ considerably from the objectives of AFND. The dbMHC database, which contains more than 90 populations, was mainly designed for the storage of data from anthropological analysis components presented at the 13th and 14th IHWS (Single et al. 2007b). Thus, the aim of the database was to serve as an electronic warehouse to display results analysed in these components without the option to add more data. In contrast, ALFRED contains a large amount of datasets (~700 populations from >3,500 samples). However, the aim of ALFRED was to cover any type of polymorphism (>650,000 as of August 2011). Consequently, only few populations containing data on immune gene frequencies are stored in ALFRED. For instance, HLA-A frequency data is limited to only 51 populations compared to >800 population samples listed in the AFND. Additionally, AFND includes data from a wide number of sources such as anthropological studies, healthy individuals selected as controls in disease association studies, and large datasets from registries (e.g. bone marrow registries, cord blood banks, etc.). Lastly, none of the two databases are active in trying to obtain new data or possess specific tools for the analysis of the frequencies in different regions as those presented in this research, making the AFND as a unique resource for the analysis not only of HLA but also KIR, MIC and cytokine gene polymorphisms.

The AFND website aiding the research

The display of frequencies using distribution maps was shown to be a useful mechanism for illustrating differences in the frequencies of HLA, KIR or MIC alleles. With the use of this tool, researchers have included in their publications specific regions in which the allele has been found to be more frequent. For example, Chen and colleagues included two figures which were generated with this interface to exemplify the distribution of HLA-DPB1*05:01 and HLA-DRB1*03:01 which were associated to Graves' disease (Chen et al. 2011). The current overlaid maps could be optimised by setting gradients layers to display possible pathways of human migration, however, data with a strictly defined sample area would be needed.

In the haplotype frequency display, perhaps one of the limits is the narrow number of haplotypes submitted to the AFND. Haplotype data is not often disclosed in publications. On some occasions, only two loci haplotypes were reported, whilst in others, only the most frequent haplotypes were included in the report. It is known that long haplotypes (six to eight routinely typed alleles) would be more useful in defining the possible origin of the haplotype in populations. Conversely, two loci haplotypes could be useful in the examination of strongest relationships between two loci. Due to the limited amount of data for haplotypes, at present, the graphical display of the haplotypes is not as globally representative as for the AFS, especially in the selection of haplotypes with more than three loci.

Future implementations in the AFND website

Future developments in the website include the incorporation of HLA haplotypes using raw data (genotypes of individuals). This is expected to be a useful tool for a wide range of areas including population genetics and pharmacogenetics analyses. For instance, having genotype data would assist scientists in the examination of the strongest relationships between two loci and generate all possible haplotype combinations based on the number of loci typed. In addition, the amino acid frequency tool will be expanded to include eplets (polymorphic residues for epitopes) and triplets as examined by Duquesnoy and colleagues in which the analysis of specific group of amino acids at

certain positions may be more significant than presenting frequencies for each single position (Duquesnoy & Askar 2007).

3.9 Conclusions

This chapter described several searching mechanisms and bioinformatics tools used for the examination of immune genes in worldwide populations. Being accessed more than 100 times per day on average, the AFND website has been shown to be an important resource for a diverse number of fields and disciplines including histocompatibility, immunogenetics, pharmacogenetics, population genetics, among many others. The development of novel mechanisms for querying and the incorporation of new polymorphisms have enriched the functionality of the AFND. At present, there are no similar online mechanisms publicly available for consultation of allele, haplotype and genotype frequency of immune genes converting the AFND website as the central source for the investigation of frequencies of these genes at different level.

The implementation of the AFS, HFS, GFS, AFAAT and the visualisation of frequency distribution in overlaid maps has assisted researchers in the analysis of immune genes. The capabilities of these tools will be exemplified in the analyses carried out in each of the polymorphic regions which will be discussed in the following four chapters.

Chapter 4

HLA allele and haplotype frequency data

4.1 Introduction

The previous two chapters described the design and implementation of the AFND as an electronic resource for the collection of frequency data of several immune genes. To examine the genes and the corresponding alleles in human populations, a set of online software tools were presented in Chapter 3, including several examples of queries performed in different polymorphic regions.

This chapter focuses on the analysis of the frequencies of HLA alleles. As mentioned in Section 1.4, the HLA region contains more than twenty genes located on chromosome 6 at position 6p21.3 and encompasses the most polymorphic region in the human genome. The diversity is mainly presented in several loci commonly known as *classical loci* (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPBA1 and -DPB1) and less observed in *non-classical loci*. As of January 2011, 6,072 alleles had been described in release 3.3.0 available in the IMGT/HLA database (Robinson et al. 2003) of which 5,827 corresponded to alleles from classical loci and 245 alleles from non-classical loci.

For many years, researchers have investigated the genetic factors in this genomic region. Due to the observation that HLA alleles were shown to present different frequencies in various populations, several studies have been carried out in anthropological components of the IHWSs to investigate the polymorphism in HLA genes [See reviews in (Terasaki 2007; Thorsby 2009)]. To examine the global distribution of HLA alleles, different human populations across the world were analysed and data was deposited in the dbMHC database (Single et al. 2007b). However, more information on HLA frequencies has become available in the public domain (mainly in the literature), comprising not only data from anthropological studies but also from healthy individuals

used as controls in disease association studies or from individuals from large registries (e.g. Bone Marrow Donor Registries, Cord blood banks, etc.). Therefore, the design of a database and software tools to analyse such extensive amount of data was a primary need for the scientific community.

This chapter provides an insight into the frequencies of HLA alleles in worldwide populations. The information presented in this chapter includes the compilation of more than 800 populations in more than four million individuals comprising the most extensive electronic resource of HLA frequency data. To allow individuals to analyse frequency data, a set of software tools were implemented in the AFND and are described in this chapter. The implementation of searching mechanisms was aimed to assist individuals in the investigation of the distribution of HLA alleles.

4.2 Materials and methods

4.2.1 Population datasets

A total of 844 population samples containing data on HLA frequencies from 4,213,280 healthy and unrelated individuals were submitted to the AFND and included in the analysis. The compilation of data comprised three main sources: 639 population samples reported in the literature (forty-nine peer-reviewed Journals from January 1990 to December 2010), 76 population samples from Proceedings of IHWSs and 129 unpublished datasets which were submitted directly to the AFND. The criteria applied for the selection of populations are described in Section 2.2.3.

From the 844 population datasets, only 327 covering 178,576 individuals contained both allele and haplotype frequencies, 502 populations including 434,977 individuals contained allele frequencies only and 15 populations comprising 3,599,727 individuals contained haplotype frequencies only. The contrasting difference between the number of individuals containing allele and haplotype frequencies corresponded to 3,583,336 volunteers listed in the National Marrow Donor Program Registry in the US in which only haplotype frequencies were disclosed (Kollman et al. 2007).

In order to compare the distribution and availability of frequency data, samples were classified in ten geographical regions: Asia (ASIA), Australia (AUST), Eastern Europe (EEUR), Middle East (MIDE), North Africa (NAFR), North America (NAME), Pacific (PACI), Sub-Saharan Africa (SAFR), South and Central America (SCAM) and Western Europe (WEUR).

4.2.2 Frequency data and terminology

Frequency datasets used for the analysis of HLA genes were available in two formats: allele and haplotype frequencies.

Allele frequencies

Frequency information for each HLA allele was given in two types: (i) *allele frequency* usually at four-decimal format and (ii) as the percentage of individuals carrying the allele with two decimals of precision. Although the schema in the AFND was designed to store up to eight decimals of precision, only few large datasets (>10000 individuals) needed more than four decimals. The compilation of allele frequencies consisted of 16,718 records containing alleles at low resolution (level 1) and 69,725 records for high resolution (levels 2-4) (For levels of resolution see Section 1.4.2).

Table 4.1: Population samples with allele frequency data by geographical region

Geographical region	Populations	Individuals	Populations by country
Asia	272	136,209	China (69), India (43), Taiwan (37), Russia (35), Japan (26), Others (62)
Australia	7	2,645	Australia (7)
Eastern Europe	56	14,429	Croatia (7), Greece (7), Turkey (7), Others (35)
Middle East	34	33,188	Israel (13), Iran (9), Others (12)
North Africa	21	1,902	Tunisia (7), Morocco (6), Others (8)
North America	77	99,014	US (52), Mexico (23), Canada (2)
Pacific	59	3,896	Papua New Guinea (27), Indonesia (8), Others (24)
South and Central America	108	13,470	Colombia (29), Brazil (25), Argentina (16), Venezuela (13), Others (25)
Sub-Saharan Africa	55	5,878	Cameroon (8), Guinea-Bissau (5), Others (42)
Western Europe	140	302,922	Portugal (30), Spain (27), Italy (18), England (15), France (12), Others (38)
Total	829	613,553	

Nearly half of the allele frequency datasets submitted to the AFND corresponded to population samples from Asia and Western Europe as shown in Table 4.1. Based on the number of individuals typed, the highest percentage (49.3%) belonged to individuals from Western Europe.

Validation of HLA alleles and ambiguities

All alleles were validated against release 3.3.0 available in the IMGT/HLA database using the online submission form shown in Section 2.2.2. For entries in which authors were unable to distinguish alleles, i.e. alleles with identical sequences over exons 2 and 3 for Class I and exon 2 for Class II, frequencies were inputted under the first allele and the corresponding notes were included in the demographic details of the population.

Haplotype frequencies

To analyse the combinations of HLA alleles, a total of 8,916 *haplotype frequencies* (proportion of haplotype copies in the population) were compiled in the AFND. HLA haplotypes consisted of 2 to 8 classical loci (HLA-A, -B, -C, -DRB1, -DPA1, -DPB1, -DQA1 and -DQB1). To cover populations with large numbers of individuals typed, frequencies were collected in percentages with two decimals of precision for optimal representation of low frequencies.

Table 4.2 summarises all 342 HLA populations with haplotype data divided by geographical region. The majority of the submissions (38.8%) belonged to population samples reported in the Asian continent. In terms of individuals typed, the vast majority comprised data from individuals in North America in which more than three million individuals belonged to NMDP registered donors as mentioned previously in this section.

Table 4.2: Populations with haplotype frequency data by geographical region

Geographical Region	Populations	Individuals	Populations by country
Asia	133	72,050	Taiwan (28), China (27), Russia (26), India (15), Others (37)
Australia	1	177	Australia (1)
Eastern Europe	19	3,172	Belarus (3), Greece (3), Others (13)
Middle East	19	2,155	Israel (10), Iran (6), Others (3)
North Africa	11	1,027	Tunisia (1), Morocco (3)
North America	32	3,596,559	US (16), Mexico (15), Canada (1)
Pacific	14	682	Papua New Guinea (5), Others (9)
South and Central America	25	6,138	Brazil (9), Others (16)
Sub-Saharan Africa	23	2,778	Uganda (3), Kenya (2), Others (18)
Western Europe	65	93,565	Portugal (28), Spain (14), Italy (10), Others (13)
Total	342	3,778,303	

4.2.3 Data analysis and visualisation

Data analysis described in this chapter was performed by direct queries to the AFND using preconfigured SQL stored procedures which were archived in the hosting server for future use.

To illustrate the geographical location of HLA allele and haplotypes across the world, overlaid maps were automatically generated using the frequency distribution map tool described in Section 3.4.1. The automated maps were based on geographical coordinates for each of the populations with HLA data. For those populations in which geographical coordinates were not described by authors, the coordinates were assigned to the closest location (city, town, region) using the Google maps API (<http://code.google.com/apis/maps/index.html>).

4.3 Results

4.3.1 Summary of HLA frequency data in the AFND

4.3.1.1 HLA allele frequencies

The majority of population samples submitted to the AFND with allele frequency data were typed at a different number of loci (1-8 loci). To show the availability of data,

samples were organised by classical (Table 4.3) and non-classical loci (Table 4.4). Table 4.3 shows the total of submissions containing classical loci classified by geographical region and level of resolution. As shown in this table, all classical loci bar HLA-DPA1 contained allele frequencies in at least one population in each geographical region, indicating the wide coverage of populations across all regions. HLA-A, -B and -DRB1 were the loci with more submissions containing more than half million individuals typed in each locus. In terms of frequencies, the vast majority of the inputs (~80%) were samples of individuals typed at high-resolution. A similar analysis was performed in non-classical HLA loci as shown in Table 4.4. Contrary to classical HLA loci, very few populations were available for each geographical region. The HLA-G locus, which is believed to play an important role in pregnancy (Hviid 2006), was the locus with the most submissions in the AFND.

Table 4.3: Availability of frequency data of classical HLA loci by geographical region

Locus	Geographical Region	Countries	Pops	Individuals	Frequency data records by level of resolution			
					1	2	3	4
A	ASIA	15	138	117,866	1,384	4,375	278	34
	AUST	1	6	1,435	51	78	6	0
	EEUR	12	23	10,567	291	688	46	7
	MIDE	6	14	31,308	168	220	19	4
	NAFR	4	12	1,068	138	1,031	197	6
	NAME	2	41	88,627	237	2,330	109	18
	PACI	6	14	1,259	22	265	24	10
	SAFR	14	32	3,049	89	894	55	4
	SCAM	13	47	9,901	654	438	27	8
	WEUR	16	81	278,294	1,234	737	40	3
B	ASIA	15	147	119,007	2,509	6,584	418	18
	AUST	1	6	1,435	75	164	0	0
	EEUR	12	24	10,669	476	1,285	98	8
	MIDE	6	14	31,308	264	450	25	1
	NAFR	4	12	1,068	196	1,771	275	5
	NAME	2	43	90,378	450	4,564	99	15
	PACI	6	14	1,259	36	495	45	6
	SAFR	16	31	3,290	161	1,291	52	1
	SCAM	12	46	9,718	1,052	961	17	3
	WEUR	16	81	277,103	2,039	1,240	26	1
C	ASIA	13	97	12,851	663	2,038	309	12
	AUST	1	5	544	28	71	2	0

Table 4.3: Availability of frequency data of classical HLA loci by geographical region (Continued)

Locus	Geographical Region	Countries	Pops	Individuals	Frequency data records by level of resolution			
					1	2	3	4
C	EEUR	10	13	8,748	109	133	9	0
	MIDE	5	7	923	4	201	28	0
	NAFR	4	8	686	63	598	133	4
	NAME	2	25	14,120	32	1,253	56	4
	PACI	5	11	786	14	181	13	0
	SAFR	12	21	2,502	46	413	13	0
	SCAM	9	21	2,100	196	75	10	0
	WEUR	14	43	39,497	365	610	15	0
DRB1	ASIA	14	199	126,444	874	3,964	884	0
	AUST	1	5	2,436	38	45	2	0
	EEUR	16	41	12,674	269	1,278	112	0
	MIDE	6	28	32,504	69	668	35	0
	NAFR	4	18	1,632	63	719	81	0
	NAME	3	55	94,427	177	3,949	136	0
	PACI	14	49	3,199	30	913	130	0
	SAFR	15	22	2,500	84	668	86	0
	SCAM	15	77	11,725	497	3,785	127	0
	WEUR	18	103	293,102	890	2,596	93	0
DPA1	ASIA	6	15	1,835	10	39	41	0
	EEUR	1	1	100	0	3	4	0
	NAME	2	3	152	0	5	4	0
	PACI	6	9	436	0	37	27	0
	SAFR	7	8	969	0	30	35	0
	SCAM	4	9	741	8	14	7	0
	WEUR	5	6	908	2	24	9	0
DPB1	ASIA	10	69	8,084	1	1,299	125	0
	AUST	1	2	144	7	15	0	0
	EEUR	5	9	957	0	213	13	0
	MIDE	2	5	305	0	75	7	0
	NAFR	2	3	397	0	64	0	0
	NAME	2	15	1,638	0	182	2	0
	PACI	9	22	1,174	0	431	28	0
	SAFR	10	12	1,549	1	465	25	0
	SCAM	6	39	2,192	4	341	4	0
	WEUR	14	35	9,792	10	1,017	15	0
DQA1	ASIA	8	68	9,284	28	622	29	0
	AUST	1	2	144	8	12	2	0
	EEUR	8	23	2,715	13	168	9	0
	MIDE	3	16	1,574	16	98	4	0
	NAFR	3	5	498	5	54	0	0

Table 4.3: Availability of frequency data of classical HLA loci by geographical region (Continued)

Locus	Geographical Region	Countries	Pops	Individuals	Frequency data records by level of resolution			
					1	2	3	4
	NAME	3	19	2,165	15	136	7	0
	PACI	10	13	761	2	96	0	0
	SAFR	9	11	1,403	7	88	5	0
	SCAM	12	42	3,544	78	215	11	0
	WEUR	13	36	6,587	42	320	29	0
DQB1	ASIA	13	116	13,752	99	1,337	181	0
	AUST	1	2	144	9	18	0	0
	EEUR	11	31	10,737	37	324	8	0
	MIDE	5	25	2,831	15	268	31	0
	NAFR	4	13	1,327	2	244	6	0
	NAME	3	40	23,771	38	503	3	0
	PACI	12	24	1,483	7	280	39	0
	SAFR	14	19	2,535	31	234	34	0
	SCAM	15	59	5,001	94	580	47	0
	WEUR	16	55	16,122	92	743	19	0
Totals					16,718	64,613	4,940	172

ASIA=Asia; AUST=Australia; EEUR=Eastern Europe; MIDE=Middle East; NAFR=North Africa; NAME=North America; PACI=Pacific; SAFR=Sub-Saharan Africa; SCAM=South and Central America; WEUR=Western Europe.

Table 4.4: Availability of frequency data of non-classical HLA loci by region

Locus	Geographical Region	Countries	Pops	Individuals	Frequency data records by level of resolution			
					1	2	3	4
DMA	EEUR	1	1	202	0	4	0	0
	NAME	1	1	263	0	4	0	0
	WEUR	1	1	90	0	4	0	0
DMB	EEUR	1	1	202	0	4	0	0
	NAME	1	1	263	0	4	0	0
DMB	WEUR	1	1	90	0	5	0	0
DRB3,4,5	SCAM	1	1	130	8	0	0	0
E	ASIA	4	4	756	0	8	6	0
	NAME	2	3	194	0	4	6	0
	SCAM	2	4	215	0	4	8	0
	WEUR	2	2	142	0	5	0	2
G	ASIA	4	5	847	0	11	27	0
	EEUR	1	1	58	0	2	6	0
	MIDE	1	1	102	0	1	8	0

Table 4.4: Availability of frequency data of non-classical HLA loci by region (Continued)

Locus	Geographical Region	Countries	Pops	Individuals	Frequency data records by level of resolution			
					1	2	3	4
	NAME	1	2	122	0	5	3	0
	SAFR	3	3	406	0	5	16	0
	SCAM	1	2	155	0	7	19	0
	WEUR	3	3	396	0	11	3	0
Totals					8	88	102	2

ASIA=Asia; AUST=Australia; EEUR=Eastern Europe; MIDE=Middle East; NAFR=North Africa; NAME=North America; PACI=Pacific; SAFR=Sub-Saharan Africa; SCAM=South and Central America; WEUR=Western Europe.

4.3.1.2 HLA haplotype frequencies

At the beginning of data collection, only haplotypes with the highest number of loci typed were entered. In later submissions, haplotype data with different numbers of loci combinations were included for the same population. Therefore, several submissions in the AFND contained HLA haplotype frequencies with a different number of loci from the same population.

Table 4.5 shows a summary of populations containing 1-4 different HLA haplotype datasets. For instance, one population stored in the AFND contained data in four different formats (e.g. A-B, A-B-C, DRB1-DQB1 and A-B-C-DRB1-DQB1). Thus, haplotypes in the AFND corresponded to 403 datasets from 342 population samples.

Table 4.5: Number of HLA haplotype datasets available in the AFND

Populations	Formats reported	Total datasets
1	4	4
3	3	9
52	2	104
286	1	286
Total	342	403

A further analysis was carried out to explore the availability of haplotypes according to the number of loci typed (2 to 8 loci). Table 4.6 shows the number of haplotype submissions for each haplotypic combination. As an example, one population contained 15 haplotype frequencies in which 8 loci were typed. In terms of datasets, haplotypes containing two and three loci were those with most entries (128 and 209 datasets

respectively). The main reason was that in many publications few loci were considered for haplotype analysis. Submissions for A-B-DRB1 and DRB1-DQA1-DQB1 were the haplotypes with more haplotype frequencies reported comprising 44% of the haplotypes available in the AFND.

Table 4.6: HLA haplotype datasets by number of loci typed

Number of loci typed	Number of datasets	Loci typed	Haplotype Frequency Records
8	1	A-B-C-DRB1-DPA1-DPB1-DQA1-DQB1	15
6	6	A-B-C-DRB1-DPB1-DQB1	58
		A-B-DRB1-DRB3-DQA1-DQB1	3
		A-B-DRB1-DRB4-DQA1-DQB1	3
		A-B-DRB1-DRB5-DQA1-DQB1	1
5	23	A-B-C-DRB1-DQB1	131
		A-B-DRB1-DQA1-DQB1	52
4	36	A-B-C-DRB1	76
		A-B-DRB1-DPB1	11
		A-B-DRB1-DQB1	237
		B-C-DRB1-DQB1	16
		DRB1-DPA1-DPB1-DQA1	14
		DRB1-DPB1-DQA1-DQB1	86
3	209	A-B-C	532
		A-B-DRB1	2,461
		B-C-DRB1	36
		B- DRB1-DQB1	22
		DRB1-DPA1-DPB1	17
		DRB1-DPB1-DQB1	145
		DRB1-DQA1-DQB1	1,476
		DRB1-DRB3-DQB1	9
		DRB1-DRB4-DQB1	6
		DRB1-DRB5-DQB1	6
2	128	A-B	936
		A-C	295
		B-C	700
		B-DRB1	388
		DPA1-DPB1	168
		DRB1-DPB1	116
		DQA1-DQB1	217
		DRB1-DQA1	23
		DRB1-DQB1	652
		DRB1-DRB3	8
Total	403		8,916

4.3.2 Occurrence of high-resolution HLA alleles

To investigate the occurrence of HLA alleles submitted to the AFND, a meta-analysis was carried out by examining the number of confirmations of HLA alleles typed at high-resolution (variants at protein level). Figure 4.1 shows the number of alleles in all three classical loci for Class I (HLA-A, -B, -C) divided by geographical regions. As expected, the majority of confirmations for Class I loci were reported in the HLA-B locus, which is the most polymorphic locus within the HLA region with 2,068 alleles described in release 3.3.0 in the IMGT/HLA database. Asia was the geographical region with more confirmations of high-resolution alleles for Class I (459 alleles), mainly due to the fact that more individuals were typed in this region compared to the rest. Due to the heterogeneity of data compiled at different years it was not possible to compare ratios between geographic regions. For instance, different laboratories may have used a different allele list as reference. However, Figure 4.1 serves as an indication of the occurrence of alleles at a particular geographical region.

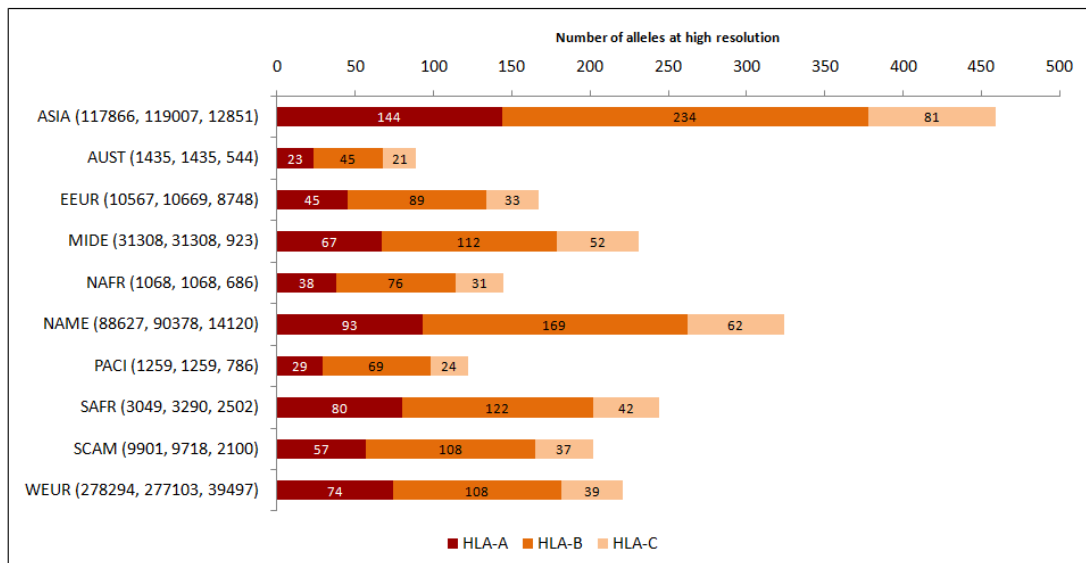


Figure 4.1: Occurrence of high-resolution alleles of HLA Class I loci by region.

In this figure, horizontal bars indicate the number of alleles reported at high-resolution organised by geographical region. Numbers in brackets represent the total number of individuals typed for each locus. ASIA=Asia, AUST=Australia, EEUR=Eastern Europe, MIDE=Middle East, NAFR=North Africa, NAME=North America, PACI=Pacific, SAFR=Sub-Saharan Africa, SCAM=South and Central America, WEUR=Western Europe.

A similar analysis was performed in classical loci of Class II (HLA-DRB1: 873 alleles, -DPA1: 28 alleles, -DPB1: 145 alleles, -DQA1: 35 alleles and -DQB1: 144 alleles

described in the IMGT/HLA as of release 3.3.0) shown in Figure 4.2. As expected, the highly polymorphic HLA-DRB1 locus contained the top number of confirmations of alleles at high-resolution with 775 alleles. Again, the majority of the high-resolution allele confirmations were observed in Asia (294 alleles in total).

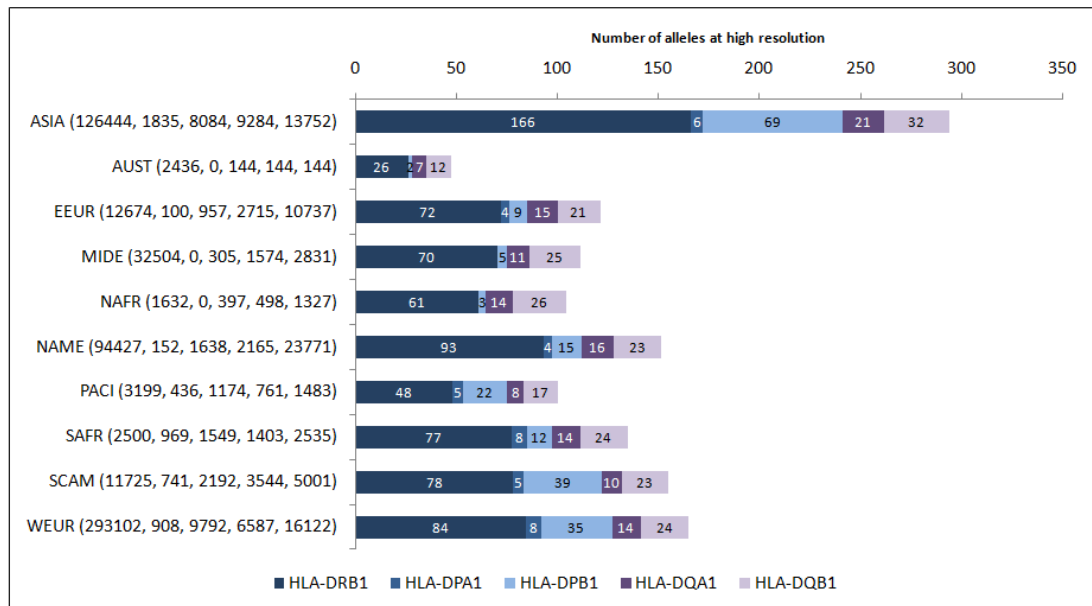


Figure 4.2: Occurrence of high-resolution alleles of HLA Class II loci by region.

In this figure, horizontal bars indicate the number of alleles reported at high-resolution organised by geographical region. Numbers in brackets represent the total number of individuals typed for each locus. ASIA=Asia, AUST=Australia, EEUR=Eastern Europe, MIDE=Middle East, NAFR=North Africa, NAME=North America, PACI=Pacific, SAFR=Sub-Saharan Africa, SCAM=South and Central America, WEUR=Western Europe.

4.3.3 Online applications for the analysis of HLA alleles

Based on the need to investigate the frequency distribution of HLA frequencies among different groups (e.g. most frequent alleles and/or overall frequencies by geographical region, country, ethnic group, etc.), two main software applications were implemented in the AFND website: (i) the Allele Frequency Calculator and (ii) the Allele Frequency Search.

4.3.3.1 Estimation of Global Allele Frequencies

As described in the literature, HLA frequencies vary significantly among geographical regions and ethnic groups. In order to provide the research community with a flexible application to estimate the overall allele frequencies of HLA data available, the *Allele Frequency Calculator* (AFC) tool was implemented in the AFND website (Figure 4.3). This searching mechanism allows individuals to examine the overall frequencies depending on criteria inputted by users. Figure 4.3 shows an example of the overall frequencies of fifteen HLA-A alleles sorted by the number of reports in descending order.

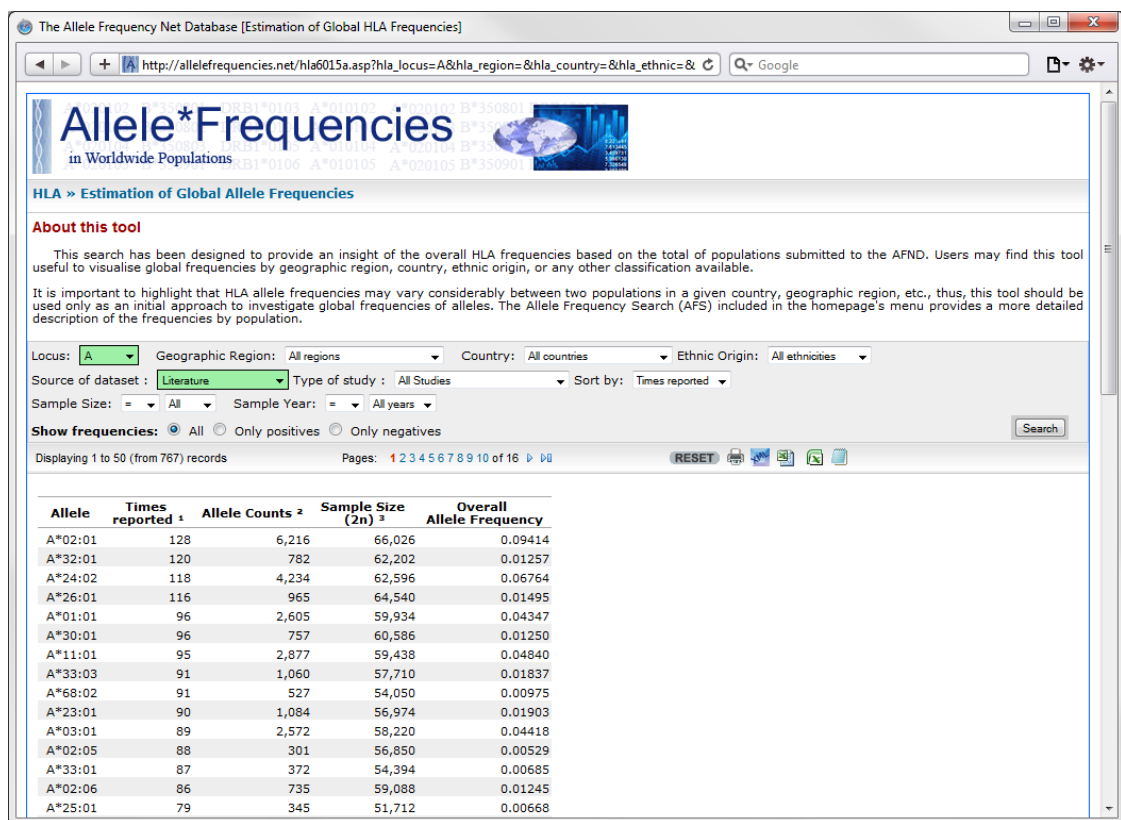


Figure 4.3: Example of the Allele Frequency Calculator (AFC).

The AFC is available for alleles typed at high-resolution only. In the example of Figure 4.3, HLA-A*02:01 was the allele with most reports for HLA-A locus (reported in 128 populations). The difference in the number of reports of HLA-A*02:01 (128 times) and HLA-A*32:01 (120 times) can be explained by the fact that HLA-A*32:01 contained considerable lower frequencies (Figure 4.4), thus, several authors may have not found this allele in their studies. As shown in Figure 4.3, several filters can be applied to the search including the source of data (literature, unpublished data), type of study

(Anthropology Study, Bone Marrow Registry, Solid Organ Unrelated Donors, Controls for Disease Study, Blood Donors, Other), sample size and year of the sample. The output list comprises the allele name, number of times reported in the AFND, number of allele copies, sample size (2n) and an approximation of the overall frequency. The search can be customised to examine the frequencies at different levels: Overall frequency → Geographical region → Country → Ethnic origin.

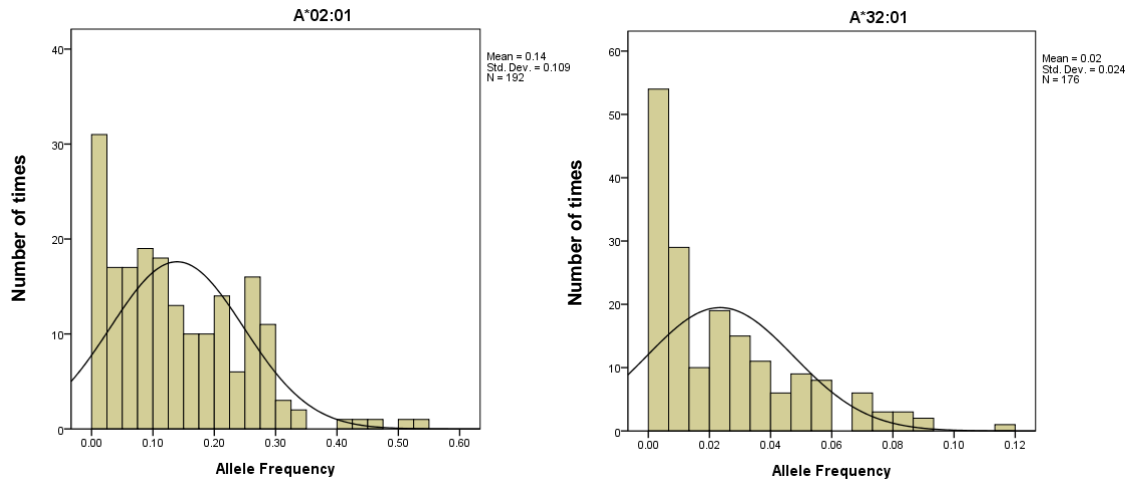


Figure 4.4: Frequency distribution of the HLA-A*02:01 and HLA-A*32:01 alleles.

4.3.3.2 The Allele Frequency Search

The AFC application, shown in Figure 4.3, was designed for the estimation of overall frequencies in a given ethnicity or geographical region. However, HLA alleles differ in populations within these groups. To analyse the frequency distribution of alleles in the different populations, the AFS (Reviewed in Section 3.4.1), was implemented as the core of the searching mechanisms in the HLA section.

4.3.4 Software for analysis of HLA haplotype distribution

Haplotype frequency data collected in the AFND was significantly lower (~10 times) than allele frequency information as shown in Section 4.2.2. The HLA-A-B-DRB1 and HLA-DRB1-DQA1-DQB1 datasets were the haplotypes that contained the highest number of frequency records with 2,461 and 1,476 respectively (Table 4.6). From these two datasets, HLA-DRB1-DQA1-DQB1 haplotypes, which contained more data at

high-resolution, were further examined using the HFS described in Section 3.4.2. Figure 4.5 shows one of the haplotypes with more reports in the AFND, DRB1*14:02-DQA1*05:01-DQB1*03:01, which was reported in 13 different populations. In this example, the highest frequencies were found in Amerindian populations [Athabaskan in Canada = 34.7% (Monsalve, Edin & Devine 1998), Xavantes in Brazil = 25.7% (Cerna et al. 1993) and Wichis in Argentina = 22.4% (Cerna et al. 1993)]. The high percentages shown in several haplotypes, for instance in Athabaskan, are the result of selection of individuals with minimal admixture with other populations. Using the map generation tool available in the HFS, the frequency distribution of this haplotype across the world is illustrated in Figure 4.6. Although little haplotype data was submitted for this haplotype, a pattern in the occurrence of haplotypes can be observed which may be useful to identify possible patterns of human migration. To extend the usefulness of the HFS, hyperlinks were implemented in the map generation tool to examine the occurrence of the specific alleles of the haplotype. As shown in Figure 4.7A-C similar patterns observed in the HFS were confirmed with data from alleles, demonstrating that the interaction of the two searching mechanisms may assist users when low haplotype frequency data is available.

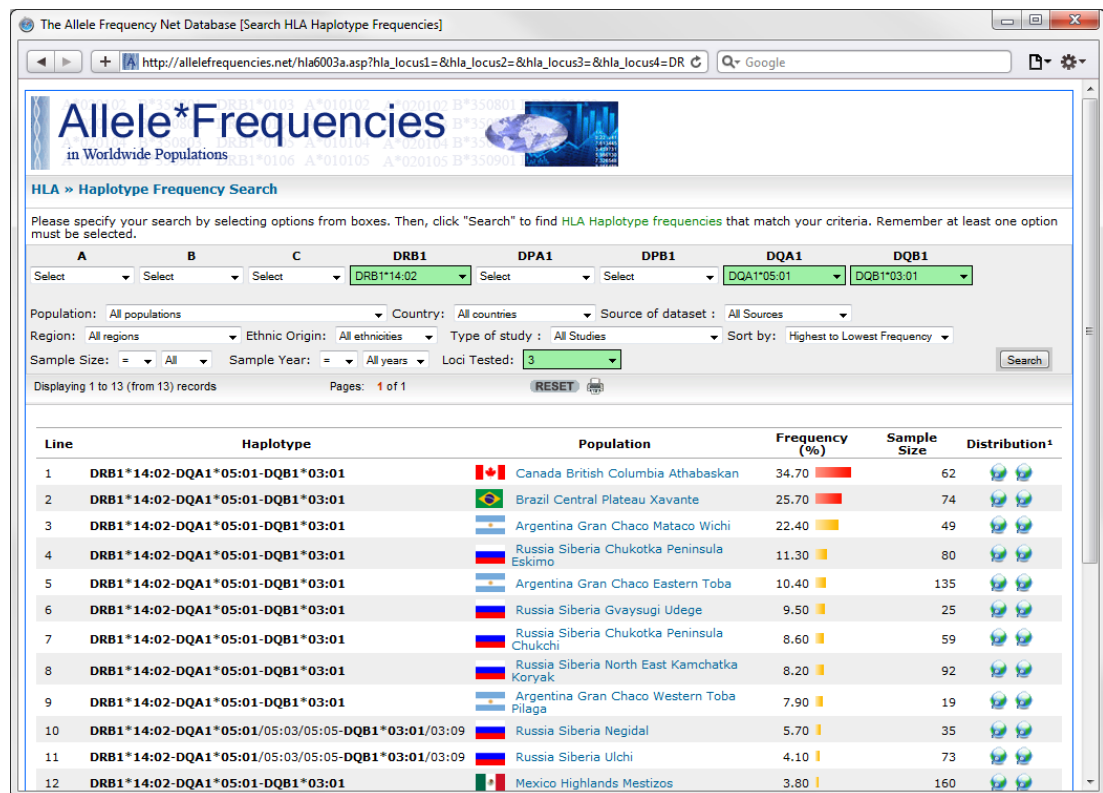


Figure 4.5: Example of the HLA haplotype frequency search.

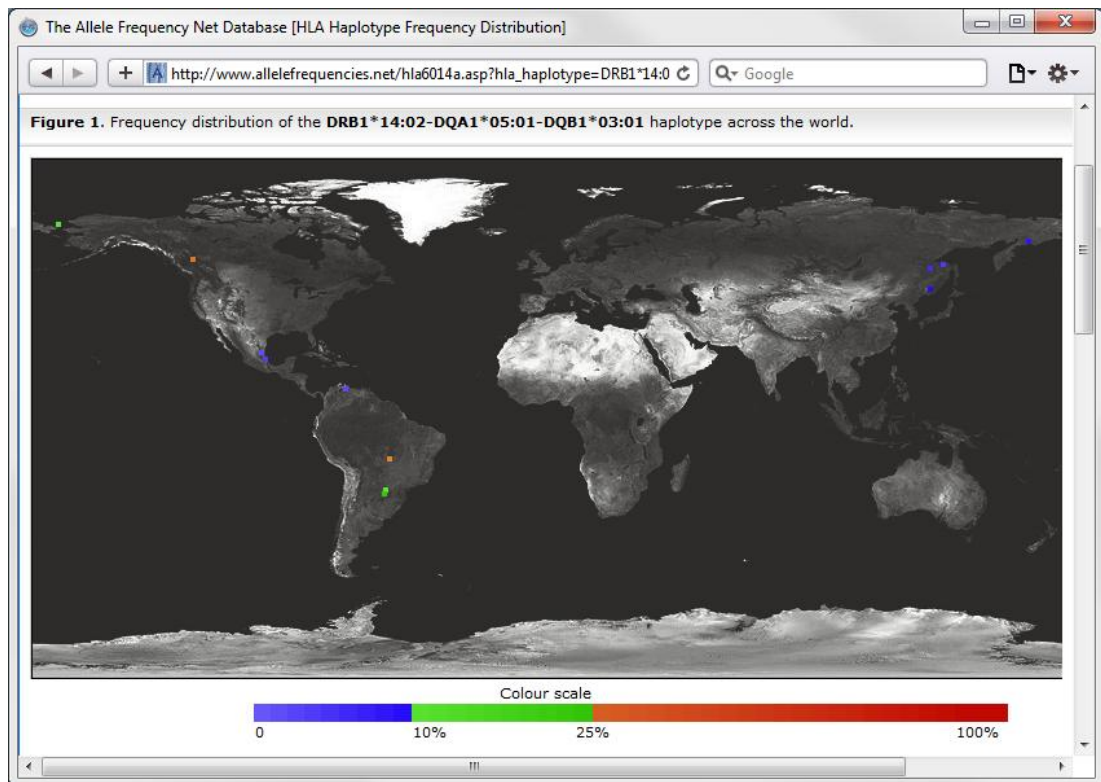


Figure 4.6: Distribution of the HLA-DRB1*14:02-DQA1*05:01-DQB1*03:01 haplotype.

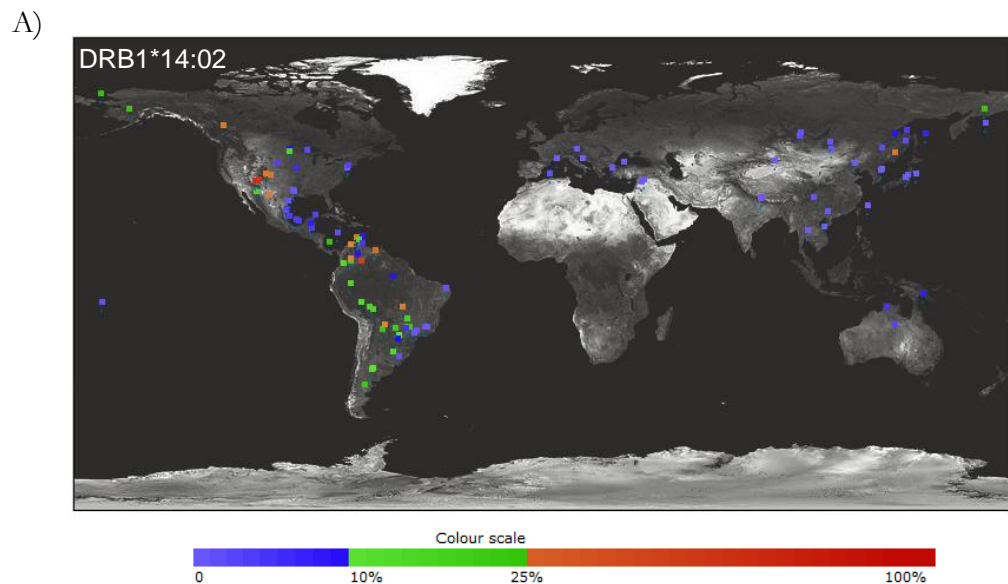


Figure 4.7: Distribution of the (A) DRB1*14:02, (B) DQA1*05:01 and (C) DQB1*03:01 alleles.

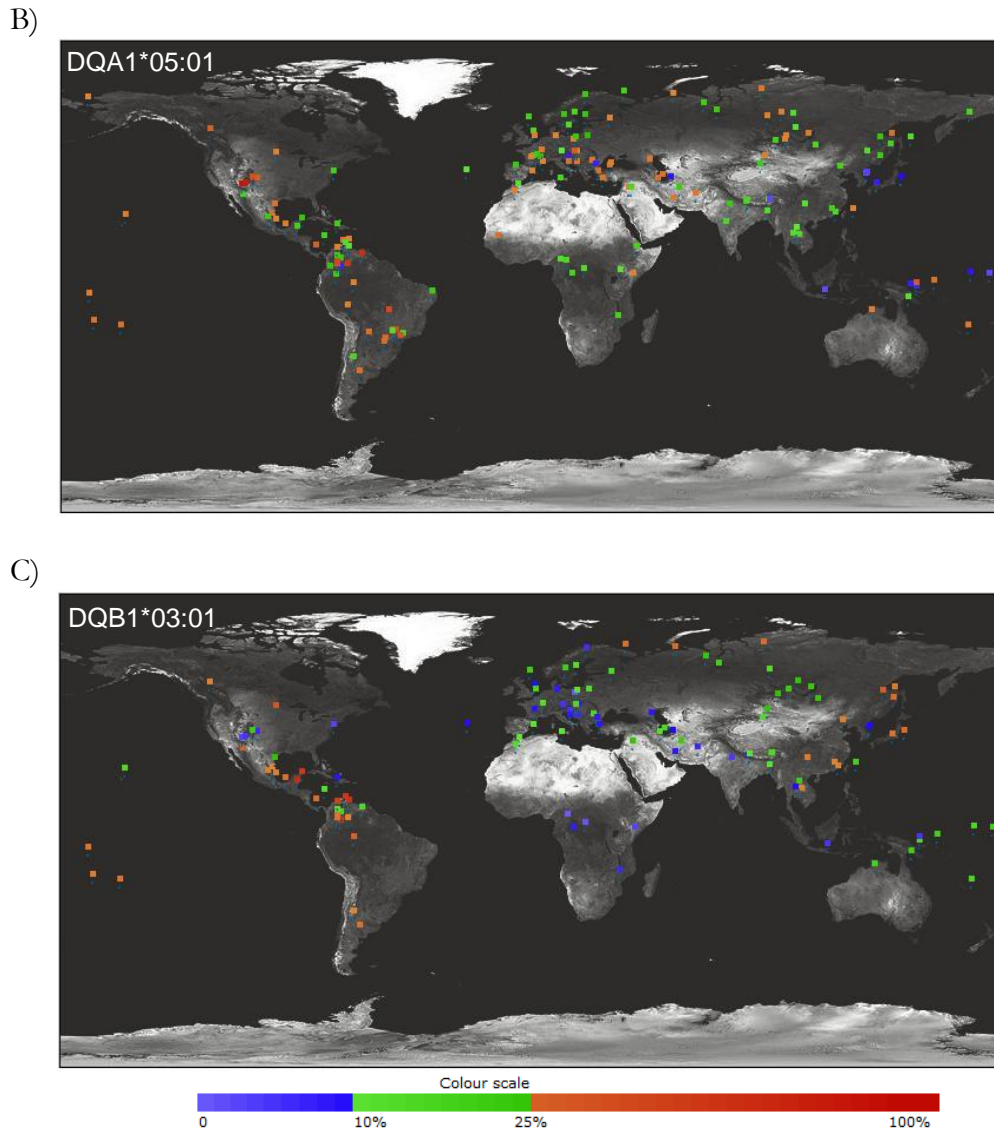


Figure 4.7: Distribution of the (A) DRB1*14:02, (B) DQA1*05:01 and (C) DQB1*03:01 alleles (Continued).

4.4 Discussion

There is a tremendous amount of data available in the literature on the frequencies of HLA alleles in populations from different geographical regions. In this study, more than 800 populations were compiled in the AFND comprising the largest collection of HLA data from different ethnic groups across the world. Having allele and haplotype frequencies of HLA genes in a readily available electronic format has been shown to be a valuable resource for meta-analysis. Greater knowledge of the distribution of HLA and other immunogenetic gene alleles in human populations may aid matching of

donors and recipients in stem cell transplantation and facilitate tracing of the history and origins of human populations. To the best of knowledge, the dbMHC and ALFRED databases are the only publicly available resources with HLA data. The amount of data available in these two databases is considerably lower than in the AFND and these databases have not had the addition of new populations in several years (See discussion in Chapter 3). The major difficulty in the compilation and analysis of HLA information on a large scale was the highly diverse type of data reported in the literature by different authors. Mechanisms of data collection implemented in the AFND website allowed the validation of data content, particularly in the capture of HLA allele names using the official nomenclature from the IMGT/HLA database. The series of nomenclature updates from IMGT/HLA are synchronously reflected into the AFND making the website an up-to-date resource for the analysis of HLA alleles.

Different meta-analyses have been performed principally on data collected for anthropological components in IHWSs (Mack et al. 2007; Nunes et al. 2010; Sanchez-Mazas 2007). The widest meta-analysis covering HLA data from different sources was carried out by Solberg and colleagues in which 497 populations with HLA frequency data were analysed for comparisons of balancing selection and heterogeneity including data from the literature, IHWSs and 70 populations from AFND (Solberg et al. 2008). The analysis of HLA mechanisms of selection are beyond the scope of this investigation. The analysis of HLA data was oriented to the development of computational approaches and mechanisms for the examination of frequencies in different populations.

The analysis carried out in this chapter was focussed on the examination of the high-resolution alleles to provide an insight into the occurrence of alleles within the entire collection and within a specific geographical region. The division of geographical regions described in this chapter may lead to controversy. The classification used in the AFND differs from the groups provided by Meyer and Single who performed analysis using data from the dbMHC database (Meyer et al. 2006). For example, in AFND, Asia encompasses South West, South East and North Asia used by Meyer and Single. This may lead to an overestimation of the confirmation of high-resolution alleles in certain geographical regions. However, in other cases, AFND covered a more specific area, for example Europe, which was divided into Eastern and Western compared to the division

provided by Meyer and Single. The advantage of software online tools and stored procedures implemented in the AFND facilitates the expansion and recalculations of the geographical divisions by the assignment of each country to the new geographical region subdivision. Interestingly, despite the high polymorphism observed in alleles of the HLA genes, less than 50% the alleles were confirmed at high-resolution, leading to the possibility that many of these alleles could be dubious. To analyse the rarity of the HLA alleles in depth, Chapter 5 presents an extensive analysis by compiling data from different sources.

The three main searching mechanisms (AFC, AFS and HFS) referenced in this chapter are expected to increase the usefulness of the AFND in the examination of allele and haplotype frequencies. The AFC may help individuals in the rapid consultation of those alleles more frequently observed in a given geographical region, country and/or ethnicity. However, calculation of global frequencies should be considered only as an approximation due to some alleles may have not been considered in old publications leading to an under/over estimation of real frequencies. As frequencies may vary significantly among populations from the same geographical region, the AFS may serve as the best approach in the understanding of the occurrence of a given allele in a particular geographical region and/or ethnic group. Additionally, the AFS can run semantic algorithms to include information on low-resolution when a high-resolution alleles is searched. This software implementation was of great importance since ~20% of the data submitted to the AFS relates to low-resolution typing (Table 4.3 and Table 4.4). The generation of automatic overlaid maps has also been widely used in the visualisation of the occurrence of a specific allele or haplotype across the world. Hyperlinks among the HFS and AFS have also been important in several cases when the availability of data is limited, particularly in haplotype data. Haplotypes can be more useful in the understanding of relationships among populations than analysing separate alleles. However, one of the major problems found in the compilation of haplotype frequency data from the literature is that many investigators do not look at haplotypes frequencies due to the number of subjects to obtain haplotypes is small. In other cases, authors tend to report only those haplotype frequencies which are highest in their population. Although the structure of the AFND can support the addition of raw data (individuals' genotypes), at present no information has been submitted to the database.

The use of raw data will increase considerably the number of analysis (e.g. Hardy-Weinberg, tests of neutrality, etc.) in the website.

The most consulted section in AFND has been the HLA frequency database. This section has facilitated the analysis in numerous projects in a wide range of contexts. First, population genetics researchers have used the AFND for studying local or global distributions of particular alleles. As one example, Solberg and colleagues used AFND for performing a population meta-analysis to search for evidence of balancing selection and heterogeneity in the HLA genes (Solberg et al. 2008). Also, Zhang and co-workers described a meta-analysis to investigate the relationships between HLA-DRB1 and multiple sclerosis in Caucasians using several datasets listed in the website (Zhang et al. 2011). The database has also provided assistance in the analysis of evolution of allele families as referred by Martinez-Laso and collaborators who analysed the HLA-B*15 allele (Martinez-Laso et al. 2011). Second, immunology researchers working to understand the basic mechanisms of self and non-self recognition, e.g. Ziegler's group who reported current theories for T cell selection and compared the amino acid sequence of different HLA alleles using frequency data available in AFND (Ziegler et al. 2009). Third, researchers working in infection biology or immunology used AFND to determine how common particular alleles are in a specific population, e.g. Wahl et al. studied influenza peptides presented by the HLA-B*07:02 allele and used AFND to demonstrate its frequency in North American populations (Wahl et al. 2009). Finally, researchers working in pharmacogenetics who have used frequency data to investigate distributions of alleles which are known to react differently to drug treatments, i.e. co-amoxiclav in liver injury (Donaldson et al. 2010), or examining effects such as cutaneous adverse drug reactions (Aihara 2011). Additionally, AFND has become a daily tool for many histocompatibility laboratories to consult which alleles appear in other populations and search for locations of possible bone marrow donors for their patients.

4.5 Conclusions

The results presented in this chapter provided an insight into the frequencies of HLA alleles and haplotypes across worldwide populations. Data compiled in the AFND corresponds to the largest collection of HLA frequencies in different worldwide

populations comprising more than 800 population samples in more than four million individuals. Results described in this chapter included the examination of data by geographical region and the use of software applications for the estimation of global allele frequencies. One of the major outcomes of the analysis was that despite the high polymorphism that has been shown in genes of the HLA region, less than a half of the alleles at high resolution were reported in the data compiled. Thus, the analysis of the occurrence of HLA alleles will be discussed in detail in the following chapter.

Chapter 5

A bioinformatics approach to ascertain the rarity of HLA alleles

5.1 Introduction

As mentioned in Section 1.4.1, the increase in the number of HLA alleles, particularly since the implementation of molecular techniques has been colossal. Nearly 6,000 HLA alleles from classical genes (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 and -DPB1) had been reported as of January 2011 at release 3.3.0 in the IMGT/HLA database.

The sequences of new HLA alleles are submitted to the IMGT/HLA database which assigns the official name for the novel allele according to a set of guidelines that are maintained by the Nomenclature Committee for Factors of the HLA System (Marsh et al. 2010). After the initial submission, other individual laboratories may report to the IMGT/HLA database if the same sequence was found in a different individual. In this case, the second submission is reported by the Nomenclature Committee as a confirmatory sequence.

Despite the massive number of alleles, many of these sequences have only been reported in one individual, which corresponds to the original sequence submitted to the IMGT/HLA database. Also, some alleles may have been found in more than one individual but reported by the same laboratory, leading to the possibility that the original submission and confirmatory sequences could have been erroneous. Thus, the confirmation of HLA alleles by different laboratories across the world was expected to provide more evidence on the occurrence of the HLA alleles described in the IMGT/HLA database.

The aim of this chapter is to describe the computational approach which was used to ascertain the rarity of HLA alleles based on the number of confirmations reported by different sources (databases and individual laboratories). This chapter includes a complete description of the datasets that were used for the confirmation of HLA alleles and presents a summary of the findings related to the occurrence of the HLA alleles. The present chapter includes an update of this work which was published in 2009 in *Tissue Antigens* (Middleton et al. 2009).

5.2 Materials and methods

5.2.1 Datasets

The information gathered for the analysis of the rarity of HLA alleles was compiled from three databases and submissions from individual laboratories.

- **IMGT/HLA.** This database was used for compiling information on the initial sequence of the allele and whether the allele was reported in another individual by the same laboratory or by a different group. The information was provided in a tab-separated text file including the name of the allele, the number of cells in which the allele was seen, the number of groups that confirmed the allele, and whether the allele was *confirmed* or *unconfirmed* based on the number of cell confirmations, i.e. confirmed (cells > 1), unconfirmed (cells = 1). The file is deposited in a server machine after every new release from the IMGT/HLA database. The analyses described in this chapter are based on release 3.3.0.
- **National Marrow Donor Program (NMDP).** A spreadsheet file containing a list of alleles which were described as rare alleles by the NMDP was included in the compilation of data. The information described in the file consisted of the name of the allele and the number of times that the allele was seen in the United States classified in six ethnic groups [African Americans (AFA), Asian and Pacific Islanders (API), Caucasians (CAU), Hispanics (HIS), Native Americans (NAM) and Others (OTH)]. This file is generated biannually and publicly available at the following link:

(http://bioinformatics.nmdp.org/HLA/Rare_Allele_Lists/Biannual_Rare_Allele_Lists.aspx)

- **AFND.** Information on the number of times that the alleles have been reported in the Allele Frequency Net Database was also included in the collection of data.

- **Individual laboratories.**

The information from individual laboratories consisted of four groups which were interested in participating in the confirmation of alleles.

- 19 laboratories who submitted data from January 2008 to August 2008 as part of a project presented at the 15th IHWS in Buzios, Brazil in September 2008.
- 10 laboratories from the National Health Service (NHS) in the United Kingdom and Ireland who submitted data from January 2010 to May 2011.
- 14 laboratories participating in the European Network HLA-NET who submitted data from January 2010 to May 2011.
- 17 laboratories who sent data from September 2009 to May 2011 and who were classified as participants of the 16th IHWS which will be hosted in Liverpool, UK in May 2012.

5.2.2 Collection of confirmatory data from individual laboratories

Generation of the rare alleles list

An initial list was generated using data available from the IMGT/HLA, NMDP and AFND databases. The number of confirmations of these three sources was combined to identify those alleles which were found on less than four occasions in unrelated individuals. The alleles that matched that criterion were temporarily classified as *rare alleles* and the output list was exported into a spreadsheet file. The list consisted of 1700 alleles for classical loci (HLA-A, -B, -C, -DRB1, -DPA1, -DPB1, -DQA1 and -DQB1). Afterwards, the file was sent to each laboratory who had expressed interest in participating in the confirmation of the rare alleles. Each laboratory was given the instruction to review the list and report the alleles that may have been typed in its

installations. The list was initially sent to laboratories participating in the 15th IHWS. Then, to extend the analysis of rare alleles in specific geographical locations, the list was sent to participants from the HLA-NET group in Europe and to NHS tissue typing laboratories located in the United Kingdom and Ireland.

Information required for rare allele confirmations

If an allele was reported by a laboratory, a questionnaire was sent asking for minimum information on the allele such as the ethnicity of the individual in whom the allele was found, name of the population from where the individual came, reason for which the individual was typed (e.g. disease association study, renal patient, bone marrow related donor, bone marrow unrelated donor, bone marrow patient, kidney donor or other), and the method used to detect the allele (e.g. SBT, SSP, SSO, direct sequencing or any other method). Additionally, details of the laboratory submitting the confirmation of the rare allele were required for each submission.

Furthermore, the haplotype or genotype of the individual was required to complete the submission. If the haplotype was known by pedigree analysis, only the haplotype containing the rare allele was entered to the AFND. If the haplotype was deduced by the maximum likelihood method or any other estimation method (e.g. resampling) then both haplotypes were required. However, if haplotypes were not known, the genotype of the individual was requested.

Online submissions

At the beginning of the project, each participating laboratory sent the information using a preformatted spreadsheet file. These confirmations were subsequently entered to the AFND via an online form available on the website (Figure 5.1). Official names and ethnicities of the individuals were validated through this form.

On some occasions, laboratories decided to send a file containing raw data (genotypes of individuals) due to the large amount of data available in their centres. For these large datasets, the *Rare Allele Detector* (RAD) application was used to identify those alleles

which were considered as rare. The functionalities of the RAD module are described in this chapter in Section 5.3.6.

The screenshot shows a web browser window with the title "The Allele Frequency Net Database [Rare HLA Allele Submission Form]". The address bar shows the URL "http://allelefrequencies.net/hla2003a.asp?all_name=C*08:12". The page content includes an "Instructions" section, followed by a form with the following fields:

- * Allele:** A dropdown menu with "C*08:12" selected.
- As assigned by WHO HLA nomenclature committee:** A text box containing "HAN-1027".
- * ID of individual possessing allele:** A text box containing "HAN-1027" with "(e.g. HAN-1002)" as a hint.
- Give designated name of Cell / DNA used:** A text box.
- * Ethnicity:** A dropdown menu with "Oriental" selected.
- * Population:** A text box containing "USA Asians" with "(if not known type Unknown)" as a hint.
- * Individual type:** A dropdown menu with "Bone marrow unrelated donor" selected.
- Give registry donor:** A text box containing "NMDP".
- Haplotypes:** A dropdown menu with "Unknown" selected.
- * Haplotype confirmation:** A dropdown menu.
- HLA Phenotype:** A text box containing "A*0301/0201; B*1402/5701; C*0812/0602; DRB1*0102/1401; DQB1*0501/0503; Please enter full phenotype. e.g. A*01:01/01:06; B*08/13; C*07/04; DRB1*03/07; ... e.g. A*01:01".
- Method(s) used to detect allele:** A row of checkboxes for "SBT", "SSP", "SSO", "Sequencing", and "Other".
- Submitter:** A text box at the bottom of the form.

Figure 5.1: Screenshot of the online submission form for rare allele confirmations.

5.3 Results

5.3.1 Summary of submissions

A total of 1180 allele confirmations were submitted to the AFND by participating laboratories from 24 different countries (Table 5.1). Nearly half of the confirmations (45.5%) corresponded to submissions from participants of the 15th IHWS. The rest of the submissions were 23.8% from UK laboratories, 22.1% from HLA-NET and 8.3% from participants of the 16th IHWS. In some occasions, several rare alleles were confirmed more than once from the same laboratory. For example, 61 submissions were sent by a laboratory in Curitiba, Brazil for the 15th IHWS, however, only 28 different rare alleles were found from those submissions. From the entire datasets, 356 alleles were reported only once in the accumulated data as shown in Table 5.1.

Table 5.1: Rare alleles reported to AFND by individual laboratories

Dataset	City	Country	Total Sent	Different alleles in each lab	Alleles reported only once in the accumulated data
15 th IHWS	Buenos Aires	Argentina	3	3	2
	West Australia	Australia	1	1	0
	Curitiba	Brazil	61	28	9
	Rio de Janeiro	Brazil	13	6	2
	Sao Paulo	Brazil	4	2	0
	Ottawa	Canada	13	12	8
	Helsinki	Finland	25	11	1
	Paris	France	12	8	5
	L'Aquila	Italy	25	21	15
	Milan	Italy	5	3	1
	Pavia	Italy	29	15	4
	Barcelona	Spain	22	19	12
	Geneva	Switzerland	14	14	9
	Hualien	Taiwan	19	19	16
	Newcastle	United Kingdom	8	8	5
	Aurora	United States	100	17	2
	Houston	United States	101	31	7
	Oklahoma	United States	43	24	13
Seattle	United States	40	24	12	
16 th IHWS	Adelaide	Australia	1	1	0
	Perth	Australia	10	9	6
	Curitiba	Brazil	22	20	11
	Montreal	Canada	4	4	3
	Pilsen	Czech Republic	6	3	1
	Paris	France	3	2	0
	Hyderabad	India	8	5	2
	Cagliari	Italy	3	2	1
	Milan	Italy	5	5	2
	Roma	Italy	3	3	2
	Uppsala	Sweden	26	26	11
	Taipei	Taiwan	6	6	4
	Aurora	United States	1	1	1
	Portland	United States	1	1	0
HLA-NET	Vienna	Austria	26	26	16
	Sofia	Bulgaria	6	6	3
	Zagreb	Croatia	3	3	1
	Helsinki	Finland	3	3	1
	Besançon	France	14	13	9
	Lyon	France	50	45	30
	Athens	Greece	7	3	0
	Milan	Italy	38	30	5

Table 5.1: Rare alleles reported to AFND by individual laboratories (Continued)

Dataset	City	Country	Total Sent	Different alleles in each lab	Alleles reported only once in the accumulated data
HLA-NET	Leiden	Netherlands	76	66	35
	Oslo	Norway	5	5	2
	Lisbon	Portugal	27	27	18
	Ljubljana	Slovenia	4	3	1
	Geneva	Switzerland	1	1	0
	Glasgow	United Kingdom	1	1	1
UK Labs	Dublin	Ireland	25	22	11
	Birmingham	United Kingdom	7	4	0
	Cambridge	United Kingdom	2	2	0
	Cardiff	United Kingdom	122	55	14
	Glasgow	United Kingdom	5	5	1
	Leeds	United Kingdom	13	10	3
	London ^a	United Kingdom	19	19	3
	London ^b	United Kingdom	47	44	28
	Manchester	United Kingdom	34	34	2
	Sheffield	United Kingdom	8	8	5
Total			1180		356

^a Barts and the London NHS Trust, ^b Anthony Nolan Research Institute.

Table 5.2 shows a summary of the number of times an allele was reported by individual laboratories. A total of 584 different alleles were reported among the 1700 rare alleles that were initially asked for. As shown in Table 5.2, the allele C*07:18 was confirmed 44 times indicating a low representation of some of the alleles in the IMGT/HLA, AFND and NMDP databases. Interestingly, 90.4% of the alleles that were considered as ‘rare’ in the initial spreadsheet were confirmed only 1-3 times considering all submissions.

Table 5.2: Number of allele confirmations by times seen

Number of alleles	Times seen	Number of alleles	Times seen	
1	(C*07:18)	44	4	8
1	(B*35:16)	24	6	7
1	(DQB1*03:19)	22	7	6
1	(B*07:06)	18	9	5
1	(C*18:02)	14	18	4
1	(C*15:09)	13	47	3
1	(B*35:14:01)	12	125	2
3		11	356	1
2		10		
Total	584			

5.3.2 Classification of HLA alleles

An analysis was performed according to whether the alleles had been sequenced only in one cell (unconfirmed) or sequenced and reported in more than one cell (confirmed) in the IMGT/HLA database. Each of these two categories was compared against confirmations from the other three datasets (AFND, NMDP and other laboratories) to examine differences. Results are shown in Figure 5.2. This figure is quite revealing in several ways. First, only 35.7% of the submissions to the IMGT/HLA had been confirmed in more than one cell. Second, 47.1% of the unconfirmed alleles (only one cell) described in the IMGT/HLA database have also never been found in any other source (AFND, NMDP and other laboratories) as of May 2011. Surprisingly, 15.6% alleles which were ‘confirmed’ in the IMGT/HLA database have never been found in any of the three datasets. From Figure 5.2 one can also conclude that, whereas the majority of alleles only sequenced once had never been reported in any other source of data (47.1%), those alleles which had been sequenced and reported in a second individual in the IMGT/HLA database were likely to be found on more than three occasions in the other datasets (11.9%). The high percentage (15.6%) of the confirmed alleles that were not found in any of the three sources may be the result of new sequences included in the IMGT/HLA database.

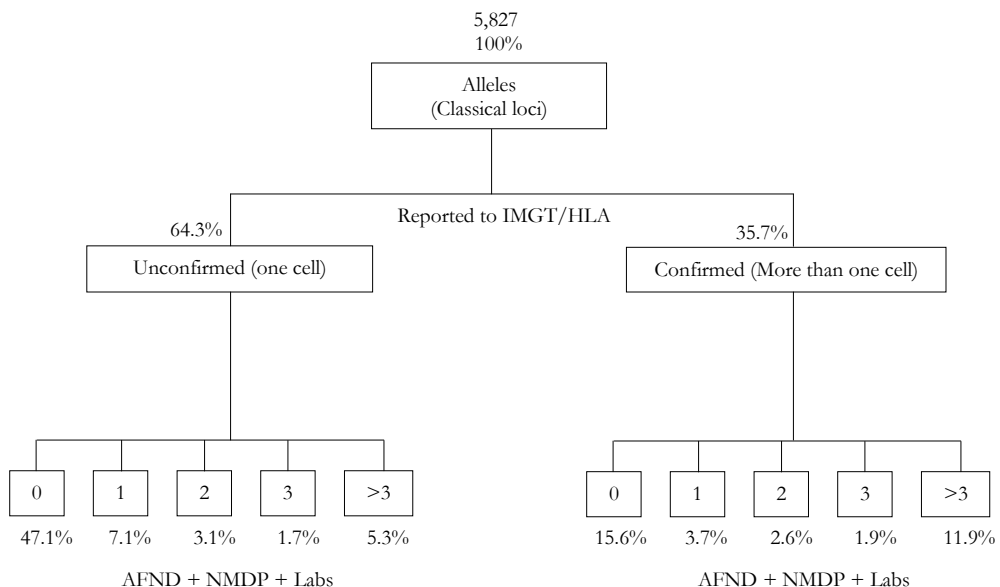


Figure 5.2: Summary of findings of HLA alleles.

During the course of discussions in the 15th IHWS between participating laboratories, it was decided that alleles only found once (initial sequencing reported to IMGT/HLA) and not on any other source would be called ‘*very rare*’, those found on 1-3 occasions (excluding the initial sequence), would be called ‘*rare*’, and those found on more than three occasions (excluding the initial sequence) called non-rare or ‘*frequent*’. Following this classification, 47.1% of all alleles were never reported after the first sequence report, thus very rare (Table 5.3). The percentage was similar for all HLA Class I classical genes (HLA-A, -B, -C) with a range from 48.1-50.8%. For HLA Class II classical genes (HLA-DRB1, -DQA1, -DQB1, -DPA1 and -DPB1) there was a wide variation with a range from 17.1-52.8% possibly due to a lower number of alleles described in these genes. Notably, only 23.6% of the alleles of classical genes have been reported as frequent (more than three times excluding the initial sequence).

Table 5.3: Summary of results of rare alleles

HLA locus	Number of alleles	Very rare (%)	Rare (%)	Frequent (%)
All	5,827	2,745 (47.1)	1,704 (29.2)	1,378 (23.6)
A	1,518	763 (50.3)	434 (28.6)	321 (21.1)
B	2,068	994 (48.1)	578 (27.9)	496 (24.0)
C	1,016	516 (50.8)	319 (31.4)	181 (17.8)
DRB1	873	335 (38.4)	276 (31.6)	262 (30.0)
DQA1	35	6 (17.1)	13 (37.1)	16 (45.7)
DQB1	144	76 (52.8)	36 (25.0)	32 (22.2)
DPA1	28	12 (42.9)	6 (21.4)	10 (35.7)
DPB1	145	43 (29.7)	42 (29.0)	60 (41.4)

Very rare: 0 times; Rare: 1, 2, 3 times excluding the initial confirmation; Frequent: more than 3 times excluding the initial confirmation

Table 5.4 shows the percentage of the confirmations by the different sources for those alleles which were reported 1-3 times after their initial sequence report to the IMGT/HLA database. The majority of the confirmations (61.6-69.9%) for alleles of the HLA-A, -B, -C, and -DRB1 loci were reported to the IMGT/HLA database as a second confirmatory sequence. For alleles of the HLA-DQA1, -DPA1 and -DPB1 loci the highest percentages of confirmations were found in the AFND and in individual laboratories for HLA-DQB1.

Table 5.4: Origin of rare alleles (1, 2 and 3 times)

HLA locus	Total	IMGT/HLA (%)	NMDP ^a (%)	AFND (%)	Labs (%)
All	1704	65.4	32.2	15.1	15.6
A	434	62.9	33.9	8.5	13.4
B	578	61.6	35.8	11.8	13.8
C	319	69.9	25.1	10.0	12.9
DRB1	276	48.2	41.3	18.1	18.5
DQA1	13	30.8		69.2	30.8
DQB1	36	33.3		38.9	44.4
DPA1	6	33.3		66.7	16.7
DPB1	42	19.0		69.0	35.7

^a NMDP only sent data for HLA-A, -B, -C, -DRB1

Considering the fact that several of the HLA alleles share identical sequences over exons 2 and 3 for Class I and exon 2 for Class II (See Figure 1.4 in Chapter 1), very rare alleles were analysed according to whether HLA alleles encode the same protein (P) or had identical nucleotide sequences (G) for the peptide binding domains. Table 5.5 shows that from the 2745 alleles which were identified as very rare, 1752 alleles did not contain identical nucleotide or protein sequences. Therefore, although individuals may attribute that some alleles may have been confirmed in another identical sequence, the number of very rare alleles was still high considering G and P groups.

Table 5.5: Very rare alleles by identical sequences

HLA locus	Very Rare	Identical sequences (G)	Identical sequences (P)	G or P	Not G and not P
All	2,745	255	977	993	1,752
A	763	80	266	274	489
B	994	81	321	327	667
C	516	54	184	186	330
DRB1	335	12	139	139	196
DQA1	6	3	4	4	2
DQB1	76	22	43	43	33
DPA1	12	0	7	7	5
DPB1	43	3	13	13	30

G=Identical nucleotide sequences over exons 2 and 3 for Class I and exon 2 for Class II; P=Identical protein sequences over exons 2 and 3 for Class I and exon 2 for Class II.

Finally, an analysis was performed to identify the year in which these very alleles were described in the IMGT/HLA. Figure 5.3 indicates that 1343 out of the 2745 very rare alleles were reported in 2010. However, more than a half of the very rare alleles reported

in previous years have never been found in any of the sources (NMDP, AFND, individual laboratories). For instance, the HLA-A*02:31 allele was submitted to the IMGT/HLA database in February 1999 after being found in an individual of African background (Ellis et al. 2000). However, since then, the allele has not been reported in any of the data sources.

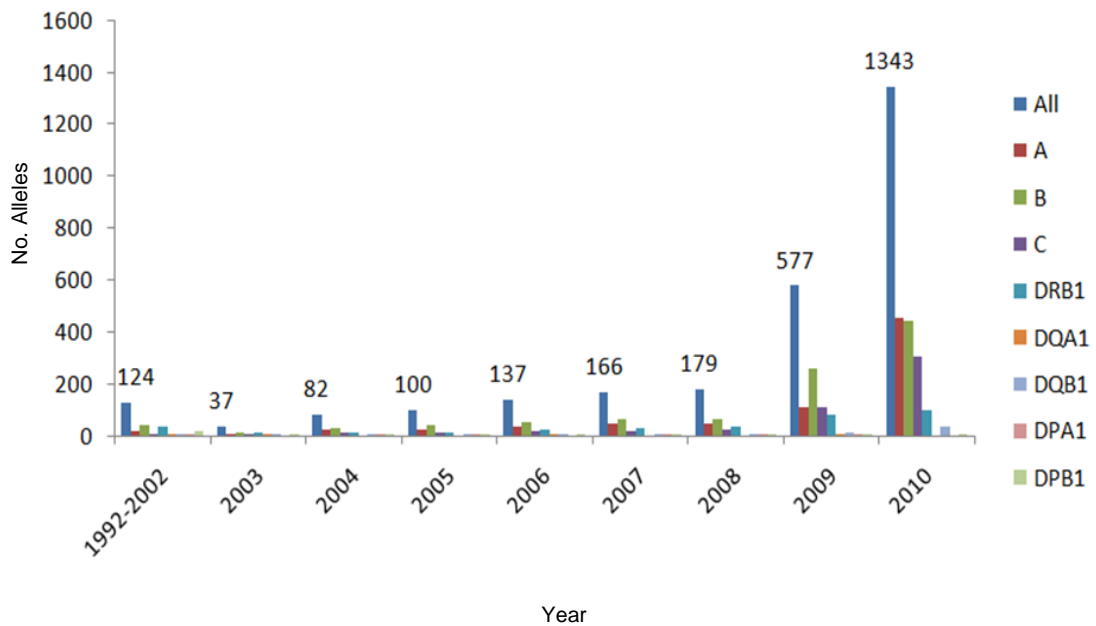


Figure 5.3: Very rare alleles by year of submission.

5.3.3 Investigation of the rarity of HLA alleles in UK laboratories

To investigate the rarity of HLA alleles in a specific country, data submitted by UK laboratories was analysed. A total of 290 submissions were compiled from UK laboratories (Table 5.6). Nearly half of the submissions corresponded to confirmations of alleles performed in the Welsh Transplantation and Immunogenetics Laboratory in Cardiff. In contrast, Belfast, Edinburgh, Harefield, Leicester, Liverpool and Plymouth laboratories did not report any rare allele due to only low resolution typing was available. Of the 290 submissions, 117 alleles were reported on only one occasion in the accumulated data.

Table 5.6: Rare allele submissions by UK laboratories

City	Number of submissions	Alleles reported only once in the accumulated data
Belfast	0	0
Birmingham	7	0
Cambridge	2	0
Cardiff	122	21
Dublin	25	11
Edinburgh	0	0
Glasgow	5	2
Harefield	0	0
Leeds	13	4
Leicester	0	0
Liverpool	0	0
London ^a	47	35
London ^b	19	8
Manchester	34	22
Newcastle ^c	8	6
Plymouth	0	0
Sheffield	8	8
Total	290	117

^a Anthony Nolan Research Institute, ^b Barts and The London;

^c Newcastle, which belonged to data from the 15th IHWS, was also considered in the analysis.

Table 5.7 shows how often, in this short study, alleles which were previously considered as very rare or rare, were found. Not surprisingly, alleles that had been considered as rare prior to this study (i.e. found on 1-3 times excluding the initial sequence) were the main alleles found on several occasions.

Table 5.7: Summary of alleles by number of times seen

Number of alleles	Previous Status	Times reported
1	1 R	11
1	1 R	10
2	2 R	7
5	5 R	6
2	2 R	5
2	2 R	4
12	3 VR, 9 R	3
27	4 VR, 23 R	2
117	30 VR, 87 R	1
Total 169		

The contribution of data from UK laboratories in the confirmation of rare and very rare alleles is shown in Table 5.8. From the 169 different alleles reported by UK laboratories, the status of 34 very rare alleles was changed to rare, three alleles were changed from very rare to frequent, 114 rare alleles were changed from rare to frequent and 18 alleles remained rare. From this table one can conclude that by using data from different laboratories within the same country the number of rare and very rare alleles can be reduced significantly.

Table 5.8: Number of changes in allele status after submissions by UK labs

Locus	Alleles	Previous status		Changes				Current status	
		VR	R	VR -> R	VR -> F	R -> F	Remained R	VR	R
Total	5,827	2,782	1,784	34	3	114	18	2,745	1,704
A	1,518	770	451	7	0	24	4	763	434
B	2,068	1,005	609	9	2	40	5	994	578
C	1,016	520	341	4	0	26	0	516	319
DRB1	873	341	285	6	0	15	4	335	276
DQA1	35	7	12	1	0	0	0	6	13
DQB1	144	76	42	0	0	6	2	76	36
DPA1	28	12	6	0	0	0	1	12	6
DPB1	145	51	38	7	1	3	2	43	42

VR=Very rare allele, R=Rare allele, F=Frequent allele

5.3.4 The rare allele search (RAS)

Despite the previous suggested criteria to classify the rarity of HLA, a tool called *Rare Allele Search* (RAS) was developed and incorporated on the AFND website to allow investigators to decide whether an HLA allele should be considered as rare according to their own criteria (Figure 5.4). For instance, users can specify the allele or locus of interest and filter results by the number of times that the allele has been reported from the existing sources (e.g. IMGT/HLA Cells < 2, AFND < 2, NMDP < 2 and Labs < 2). Also, these four data sources can be combined and filtered as shown in Figure 5.4. In the example shown, 1456 alleles from the HLA-B gene were found to be reported on less than four occasions combining the four datasets. Additionally, the search includes an option to filter results by the year in which the initial confirmation was submitted to the IMGT/HLA database. Furthermore, the output list can be also reduced by grouping alleles which are identical over exons 2 and 3 for Class I and exon 2 for Class II.

IMGT/HLA	Allele Frequencies Website		IMGT/HLA (See details)		NMDP (See details)						Other Labs (See details)					Total	ASHI							
Allele	Pops with allele in website	See lower resolution	See sequences identical over exons 2 + 3	Allele initially sequenced in population	Year	Sequence confirmed	Cells	Groups	Length	Total	AFA	API	CAU	HIS	NAM	OTH	Total	BLA	CAU	MES	HIS	OTH	Total	ASHI
391 B*15:219	0				2010	Unconfirmed	1	1	Partial	0							-						1	
392 B*15:221	0				2010	Unconfirmed	1	1	Partial	0							-						1	
393 B*18:01:02	0	B*18:01		Caucasoid - Unknown	2001	Unconfirmed	1	1	Partial	-							1				1		2	CWD
394 B*18:01:03	0	B*18:01	B*18:01:01, B*18:17N, B*18:53	Hispanic - Unknown	2006	Unconfirmed	1	1	Partial	-							-						1	
395 B*18:01:04	0	B*18:01			2009	Unconfirmed	1	1	Partial	-							-						1	
396 B*18:01:05	0	B*18:01			2009	Unconfirmed	1	1	Partial	-							-						1	
397 B*18:01:07	0	B*18:01			2009	Unconfirmed	1	1	Partial	-							-						1	
398 B*18:01:08	0	B*18:01			2009	Unconfirmed	1	1	Partial	-							-						1	
399 B*18:01:09	0	B*18:01			2010	Unconfirmed	1	1	Partial	-							-						1	
400 B*18:01:10	0	B*18:01			2010	Confirmed	2	1	Partial	-							-						2	
401 B*18:01:11	0	B*18:01			2010	Unconfirmed	1	1	Partial	-							-						1	
402 B*18:07:01	0	B*18:07			1998	Confirmed	2	2	Partial	-							1		1				3	
403 B*18:07:02	0	B*18:07			2009	Confirmed	3	1	Partial	-							-						3	
404 B*18:12	0			Caucasoid - Unknown	2000	Confirmed	2	1	Partial	0							-						2	
405 B*18:13	0			Caucasoid - Unknown	2000	Confirmed	2	1	Partial	0							-						2	
406 B*18:14	1			Caucasoid - North America	2001	Confirmed	2	2	Partial	0							-						3	
407 B*18:17N	0		B*18:01:01, B*18:01:03, B*18:53	Caucasoid - Aosta, Italy, Europe	2001	Confirmed	2	1	Full	0							-						2	

Figure 5.4: Example of the Rare Allele Search (RAS).

In this figure, the example shows the list of HLA-B alleles found on less than 4 occasions in the different data sources. Hyphens (-) in ‘NMDP’ indicate that the allele was not reported due to the level of resolution whereas hyphens in ‘Other Labs’ were used for those alleles that have not been confirmed.

The resulting list consisted of several columns describing the information provided by the different data sources. For example, data from IMGT/HLA included the official allele name, the ethnicity of the individual(s) in which the allele was found, year of the first sequence report, whether the sequence was confirmed or unconfirmed, the number of cells and groups and whether the allele was partially or fully sequenced. Data from AFND consisted of the number of confirmations submitted to the website, whether the allele was seen in a lower resolution in the database (e.g. B*18:01:02 may have been reported as B*18:01) and whether the allele belonged to a set of alleles that shared identical sequences (An example is shown in Figure 5.5). Data from NMDP and from individual laboratories consisted of the number of confirmations grouped by some of the ethnic origins provided by these two sources. Finally, an additional column was included to summarise the confirmations of the four datasets (IMGT/HLA, NMDP, AFND and other laboratories). Additionally, by clicking in the ‘Total Other lab’ column, users can access the data content submitted by laboratories (Figure 5.6).

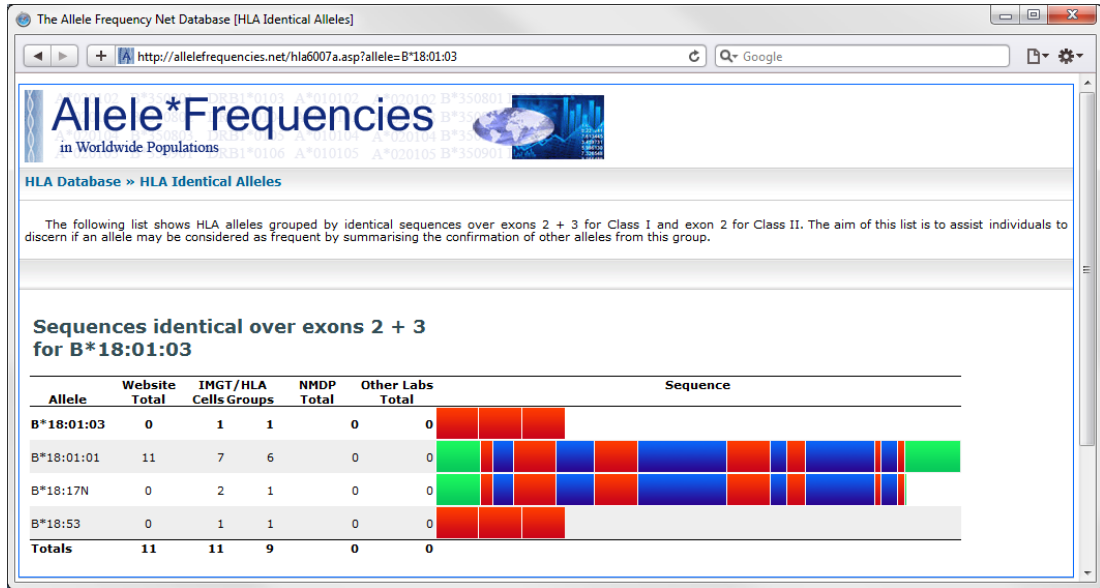


Figure 5.5: Example of identical HLA alleles over exons 2 and 3.

The figure shows an example of HLA-B*18:01:03 which was reported as very rare. However, identical sequences of this allele over exons 2 and 3 contained several confirmations (e.g. B*18:01:01 reported eleven times in the AFND).

ID: 176

Rare HLA Alleles Submission Form

Information

* **Allele Name:** **B*18:01:02**
As assigned by WHO HLA nomenclature committee

* **ID of individual possessing allele:** **03/703**
Give designated name of Cell / DNA used

Ethnicity: Hispanic
Population: Venezuela
Individual type: Bone marrow patient

Haplotype confirmation: **Confirmed by family studies**

Haplotypes

Locus	Haplotype 1
HLA-A	0201
HLA-B	180102
HLA-C	0701
HLA-DRA1	
HLA-DRB1	1104
HLA-DQA1	0505
HLA-DQB1	0301
HLA-DPA1	
HLA-DPB1	0201

Other family members with same allele:
Individuals typed to the same resolution as the individual with the rare allele: 5686
Method(s) used to detect allele: SSP

Figure 5.6: Information provided by individual laboratories on rare alleles.

5.3.5 AFND and confirmations in the US

A parallel project to identify the rarity of HLA alleles was initiated at the same time in the USA by the American Society for Histocompatibility and Immunogenetics (ASHI) (Cano et al. 2007). In that study, the acronym ‘C’ was used to represent *common alleles* that contained a frequency greater than 0.001 in the USA population and the acronym ‘WD’ used for alleles that are *well documented* in the literature. The published list consisted of 693 alleles for HLA-A, -B, -C, -DRB1, -DQB1, -DRB3/4/5, -DQA1 and -DPB1 which were described as well-documented. The list was subsequently updated to include more alleles found in the literature and the addition of the HLA-DPA1 locus. Based on this list, an analysis was performed to determine whether the alleles reported as common and well documented (CWD) in USA populations were also reported as frequent according to data provided by IMGT/HLA, NMDP, AFND and individual laboratories. Table 5.9 shows a comparison between the two groups. 340 out of the 693 well documented alleles were reported as common in the United States. Of these, only 276 were described in release 3.3.0 of the IMGT/HLA database as the rest were subsequently split at a higher resolution. As expected, nearly all alleles considered as CWD which were in the 3.3.0 release were confirmed as frequent in the AFND after the compilation of the datasets.

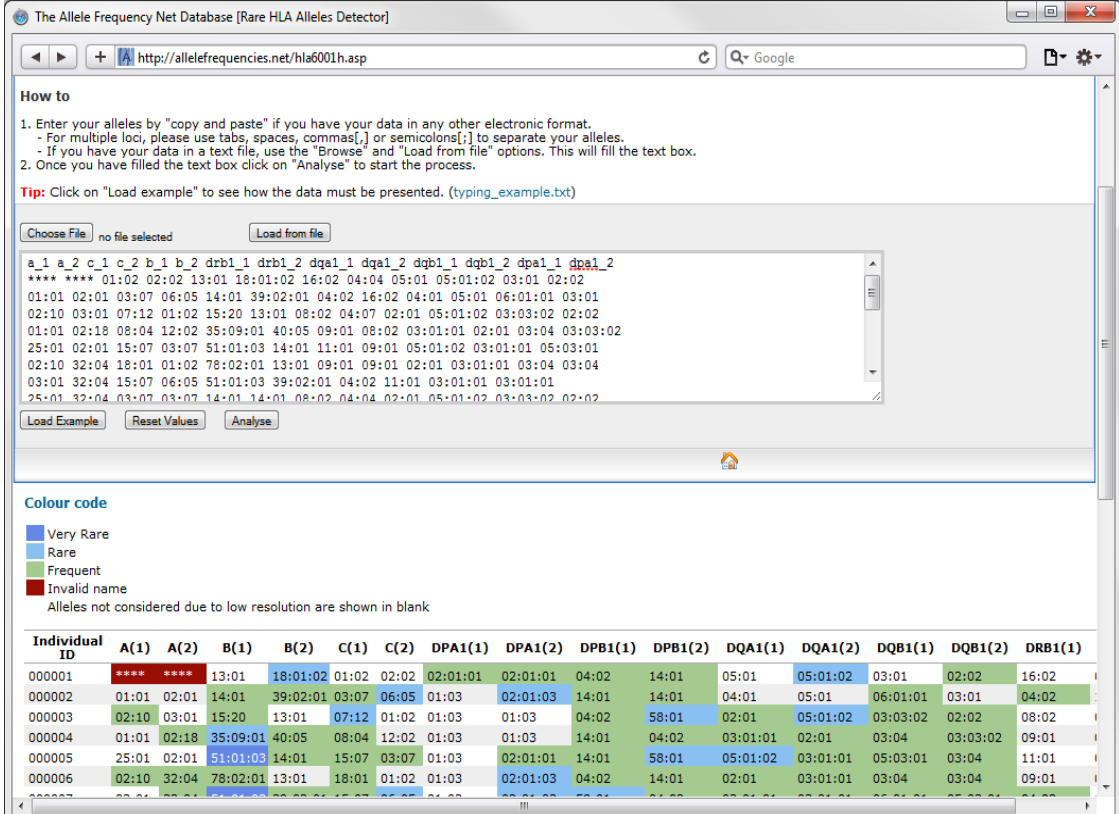
Table 5.9: Common and well-documented alleles and data in AFND

Locus	WD ^a	CWD ^b	CWD current nomenclature ^c	Frequent in AFND ^d (%)
All	732	340	276	267
A	137	53	44	43 (97.7)
B	252	118	83	79 (95.2)
C	90	44	36	36 (100)
DPA1	6	6	6	6 (100)
DPB1	53	25	24	23 (95.8)
DQA1	13	13	13	13 (100)
DQB1	28	19	16	15 (93.8)
DRB1	153	62	54	52 (96.3)

^a WD=Well documented alleles; ^b CWD=Common and well documented alleles; ^c Common and well documented alleles listed in the release 3.3.0 in the IMGT/HLA database; ^d Frequent=Alleles found more than three times in the Allele Frequency Net Database considering the four data sources (IMGT/HLA, AFND, NMDP and other laboratories).

5.3.6 Rare allele detector (RAD)

As shown in the previous section (5.3.4), the RAS was only optimal when looking for a specific allele or a set of alleles that matched given criteria. To facilitate the analysis of large datasets (i.e. 1000 individuals typed) the *Rare Allele Detector* tool was developed and incorporated to the AFND website (Figure 5.7). This tool simplified the detection of alleles which were considered to be very rare, rare or frequent from submissions from individual laboratories which contained raw data (genotype of the individuals). In this module users can upload their own raw data file in a tab-separated text file or by inputting the allele(s) of interest directly into a text box as shown in Figure 5.7. The output list includes the detection of those alleles which were invalid according to the official nomenclature of the IMGT/HLA, whether the alleles were not considered in the analysis due to being typed at a lower resolution, and whether the alleles belonged to a very rare, rare or frequent category.



How to

- Enter your alleles by "copy and paste" if you have your data in any other electronic format.
 - For multiple loci, please use tabs, spaces, commas[,] or semicolons[,] to separate your alleles.
 - If you have your data in a text file, use the "Browse" and "Load from file" options. This will fill the text box.
- Once you have filled the text box click on "Analyse" to start the process.

Tip: Click on "Load example" to see how the data must be presented. ([typing_example.txt](#))

Choose File: no file selected Load from file

```
a_1 a_2 c_1 c_2 b_1 b_2 drb1_1 drb1_2 dqa1_1 dqa1_2 dqb1_1 dqb1_2 dpa1_1 dpa1_2
**** ** 01:02 02:02 13:01 18:01:02 16:02 04:04 05:01 05:01:02 03:01 02:02
01:01 02:01 03:07 06:05 14:01 39:02:01 04:02 16:02 04:01 05:01 06:01:01 03:01
02:10 03:01 07:12 01:02 15:20 13:01 08:02 04:07 02:01 05:01:02 03:03:02 02:02
01:01 02:18 08:04 12:02 35:09:01 40:05 09:01 08:02 03:01:01 02:01 03:04 03:03:02
25:01 02:01 15:07 03:07 51:01:03 14:01 11:01 09:01 05:01:02 03:01:01 05:03:01
02:10 32:04 18:01 01:02 78:02:01 13:01 09:01 09:01 02:01 03:01:01 03:04 03:04
03:01 32:04 15:07 06:05 51:01:03 39:02:01 04:02 11:01 03:01:01 03:01:01
75:01 32:04 03:07 03:07 14:01 14:01 08:02 04:04 02:01 05:01:02 03:03:02 02:02
```

Load Example Reset Values Analyse

Colour code

- Very Rare
- Rare
- Frequent
- Invalid name

Alleles not considered due to low resolution are shown in blank

Individual ID	A(1)	A(2)	B(1)	B(2)	C(1)	C(2)	DPA1(1)	DPA1(2)	DPB1(1)	DPB1(2)	DQA1(1)	DQA1(2)	DQB1(1)	DQB1(2)	DRB1(1)
000001	****	****	13:01	18:01:02	01:02	02:02	02:01:01	02:01:01	04:02	14:01	05:01	05:01:02	03:01	02:02	16:02
000002	01:01	02:01	14:01	39:02:01	03:07	06:05	01:03	02:01:03	14:01	14:01	04:01	05:01	06:01:01	03:01	04:02
000003	02:10	03:01	15:20	13:01	07:12	01:02	01:03	01:03	04:02	58:01	02:01	05:01:02	03:03:02	02:02	08:02
000004	01:01	02:18	35:09:01	40:05	08:04	12:02	01:03	01:03	14:01	04:02	03:01:01	02:01	03:04	03:03:02	09:01
000005	25:01	02:01	51:01:03	14:01	15:07	03:07	01:03	02:01:01	14:01	58:01	05:01:02	03:01:01	05:03:01	03:04	11:01
000006	02:10	32:04	78:02:01	13:01	18:01	01:02	01:03	02:01:03	04:02	14:01	02:01	03:01:01	03:04	03:04	09:01

Figure 5.7: Screenshot of the Rare Allele Detector (RAD) module.

5.4 Discussion

The number of HLA alleles that have been sequenced and reported to the IMGT/HLA database in recent years has been immense. This has led to a problem in differentiating these alleles in Histocompatibility Laboratories. Some researchers have expressed scepticism whether all these alleles have been correctly determined. The results described in this chapter show that, despite a worldwide analysis which included four different datasets (NMDP, AFND, IMGT/HLA and individual laboratories) containing confirmations of HLA alleles, many alleles have not been confirmed after the initial sequence reported to the IMGT/HLA database. It is interesting to highlight that 47.1% of all alleles have only been reported once. This percentage shows an increase from previous results published in *Tissue Antigens* in 2009 (Middleton et al. 2009) which was 40.6% at that time. The percentage was similar for the HLA Class I classical genes (HLA-A, -B and -C). However, in Class II classical genes (HLA-DRB1, -DQA1, -DQB1, -DPA1 and -DPB1) the percentages were more diverse. A possible reason was that the number of alleles in the Class II genes was considerable lower than in those of Class I and that very less diversity is observed for HLA-DR alleles.

One of the outcomes of the 15th IHWS was the allocation of HLA alleles into three groups: *very rare* (only sequenced once), *rare* (found on one, two or three occasions in addition to the initial sequence) and *frequent* (found on more than three times excluding the initial sequence). However, rather than presenting a specific status for each allele, an online searching mechanism was included in the AFND website to allow individuals to decide the criteria for defining a rare allele. This online tool also provides an up-to-date resource for consultation of the alleles as it is automatically maintained by inputs from IMGT/HLA, NMDP, laboratories and submissions of populations in the AFND.

At the beginning of this project, the majority of the confirmations were reported to NMDP, however, in recent years there was a significantly increase in the number of confirmatory sequences to the IMGT/HLA database which has implemented more requirements in the confirmation of alleles. However, the confirmations in the IMGT/HLA database are limited to sequence-based typing.

Many researches have argued that very rare and rare alleles may correspond to limitation of number of individuals typed. It is worth mentioning that although the number of confirmations from laboratories is yet lower than confirmations from NMDP, the number of submissions exceeded one thousand and it is expected that this number will increase considerably over the next years. Thus, the incorporation of data from large databases such as bone marrow donor registries and the analysis of the alleles at different levels, including a specific continent such as Europe (part of the HLA-NET project) or a specific country (United Kingdom) are expected to provide more evidence in the confirmation of the rarity of HLA alleles. In the future, when enough data is available, it will be possible to analyse the rarity of alleles by continents, geographic regions and eventually by countries.

Another important factor to consider is that some individual laboratories may have reported the same data to more than one of the sources considered in this analysis. For example, confirmation of a sequence may have been submitted to IMGT/HLA and the type submitted to NMDP. In these cases, it was not possible to validate duplications. However, for laboratory confirmations the genotype of the individual was required to alleviate the problem at least in this source of data. An ongoing process includes the validation of duplications in NMDP and IMGT/HLA which will be part of the rare project presented in the 16th IHWS in 2012.

One of the issues that emerged during the presentation of the data in the 15th IHWS was that many of the alleles may have been reported in another allele sharing an identical sequence over exons 2 and 3 for Class I and exon 2 for Class II. However, Table 5.5 indicated that 1752 alleles were still considered to be very rare taking into account this consideration.

A parallel project about the rarity of HLA alleles was performed by Cano and colleagues to identify *common* HLA alleles based on the frequencies of the alleles observed in individuals from the United States (Cano et al. 2007). The analysis also included the report of alleles that were *well-documented* in the literature. One of the drawbacks of this study was that frequency analysis was restricted to individuals from the US. Thus, the data presented in this chapter corresponds to the most extensive analysis of rare alleles performed in worldwide populations. Additionally, the rare allele project has also helped

as a reference in the report of new alleles (Chu et al. 2010) and in the investigation of allele lineages (Martinez-Laso et al. 2011).

Rare alleles may be the result of errors in the detection of alleles using a particular PCR method, mutations, genetic recombination, or the lack of enough data from a specific geographic location or population. Thus, the investigation of rare alleles requires an extensive analysis of which location these alleles may be found. One of the major problems from data reported in the IMGT/HLA database is that does not record the location/ethnicity in a standardised format. In the majority of the cases the ethnicity of the individual is stored using only 9 major categories which are difficult to interpret and in very few occasions the country is included in the report. Thus, future analysis will include the investigation of the rarity of alleles by ethnicity or geographic location which will require the standardisation of the name of the ethnic groups and geographic locations utilised in the NMDP, IMGT/HLA and AFND databases.

One of the latest advances in the determination of HLA alleles has been the use of new techniques such as next-generation sequencing (Erlich et al. 2011). However, these technologies may be limited to a few laboratories due to the high-costs and large numbers for cost efficiency. The number of different populations tested may be effected by this as it is likely that second generation typing would be applied to registry typing. A significant effort would be needed to determine all possible alleles occurring in worldwide populations, if it ever occurs. Thus, it is expected that the AFND will serve as a primary source for the deposition of confirmatory sequences in the next future assisting in the investigation of HLA variants.

5.5 Conclusions

One of the most important contributions of this research was the analysis of the rarity of HLA alleles by examining large amount of data from different sources including reports in the AFND, data from the National Marrow Donor Program in the US and confirmations by individual laboratories. The structure of the AFND and the availability of the data in electronic format facilitated the compilation of such extensive amount of information. This study, which was presented as a project in the 15th IHWS in 2008, has

encouraged the active participation of individual laboratories and International organisations in the confirmation of rare alleles. A summary of results was described in this chapter including a suggested classification of alleles based on the number of confirmations by the different sources. Additionally, the chapter described a computational approach to ascertain the rarity of these alleles by including an up-to-date searching mechanism implemented in the AFND. The results presented in the current analysis indicated that nearly half of the alleles described in the IMGT/HLA database have never been found in any of the four different sources. This important finding encourages the continuous collaboration between the different databases to exchange data to increase the accuracy of information.

Chapter 6

KIR genes and genotype frequencies

6.1 Introduction

Chapters 4 and 5 presented an extensive analysis on the examination of HLA genes and their corresponding alleles in worldwide human populations. The analysis encompassed an overview of the frequency distribution of the HLA alleles among populations and described a computational approach as a method to analyse the rarity of the existing HLA alleles.

This chapter describes a set of analyses that were performed in the Killer-cell Immunoglobulin-like Receptors (KIR) genes which are expressed in NK cells and a subset of T cells. As explained in Section 1.5, the products of some of the KIR genes are well known for their interactions with products of the HLA molecules and have been a topic of investigation by researchers. The genomic region comprising these genes has also been shown to present an elevated polymorphism with more than 600 alleles described in the IPD-KIR database as of release 2.4.0, April 2011. However, the observed polymorphism in the KIR genes is lower than that of the HLA system.

The genes from the KIR family present a set of peculiarities which differ from the HLA system. One of the conspicuous differences is that an individual may possess the presence or absence of some of the KIR genes, termed the *KIR genotype*, which defines the gene content profile of the individual. Additionally, some of the KIR genes are considered to be habitually present and called framework genes, although a few individuals reported in some populations do not have these genes. As mentioned in Section 1.5.1, an individual's KIR genotype is formed by the combination of an A and B haplotype leading to the importance of investigating the structure of the KIR complex based on its gene content. Similarly to HLA, KIR genes also present wide variation among individuals from different populations bringing the attention of investigators to

examine the occurrence of genes and alleles among these groups. Therefore, in an endeavour to disseminate the known polymorphisms in the KIR gene cluster on a large scale, nearly two hundred KIR population samples listed in the AFND were analysed.

This chapter presents an outline of the frequencies of the seventeen KIR genes which have been reported in the AFND. Furthermore, the chapter includes an overview of the KIR Genotype Frequency Database implemented in the AFND which contains the largest compilation of the KIR genotypes and their frequencies in worldwide populations. The results presented in this chapter provide an insight into the occurrence of KIR genes, their corresponding alleles and genotypes based on the analysis of the information using the AFND and tools included in the website. Finally, several tables and figures included in the chapter present an update of those results published as part of a review article in the *Immunology* Journal (Middleton & Gonzalez 2010).

6.2 Materials and methods

6.2.1 Population datasets

A total of 194 population studies covering 23,204 unrelated healthy individuals with data on KIR frequencies collected by the AFND were included in this analysis. The compilation of datasets consisted of 161 population samples reported in fourteen peer-review Journals from January 2001 to December 2010 and 33 unpublished datasets which were submitted directly to the AFND. The criteria applied for the selection of populations are described in Section 2.2.3. 75 out of the 194 population samples corresponded to individuals that were typed at allele level (3, 5 and/or 7 digits). 108 out of the 194 population samples contained genotype frequency information on 12,291 individuals. In order to compare the frequency distribution of genotypes among different populations, samples were classified in ten geographical regions: Asia (ASIA), Australia (AUST), Eastern Europe (EEUR), Middle East (MIDE), North Africa (NAFR), North America (NAME), Pacific (PACI), Sub-Saharan Africa (SAFR), South and Central America (SCAM) and Western Europe (WEUR).

6.2.2 Frequency data and terminology

Frequency datasets used for the analysis of KIR genes were available in two formats: (i) gene/allele frequencies and (ii) genotype frequencies.

Gene/allele frequencies

A total of 4,922 records containing data on frequencies at gene or allele level were included in the analysis. Frequency information for each KIR gene was given as *gene frequency* (percentage of individuals carrying the gene) or *allele frequency* (proportion of allele copies in the gene). Allele frequencies submitted to the AFND were given in four-decimal format whereas gene frequencies were available in percentages with two decimals of precision.

Genotype frequencies

Genotype data comprised 2,600 genotype frequency records (percentage of individuals carrying the presence or absence of KIR genes) which were included in the analysis. KIR genotypes consisted of the presence or absence of the following genes: *KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL4*, *KIR2DL5*, *KIR2DP1*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5*, *KIR3DL1*, *KIR3DL2*, *KIR3DL3*, *KIR3DP1* and *KIR3DS1*.

6.2.3 Construction of the KIR genotype database

The KIR genotype database was implemented in the AFND based on the concept of a genotype comparison performed by Rajalingam and colleagues which consisted of a preliminary list of 47 distinct genotypes from twelve different populations (Rajalingam et al. 2002). In 2008, the list was subsequently extended to include more data from which 179 distinct genotypes in 42 populations were reported and supplied as an Excel spreadsheet file (Rajalingam et al. 2008). In order to provide the scientific community with an up-to-date list and allow automated submission of new genotype findings, an online genotype capture module was developed and included in the AFND website (Figure 6.1). Additionally, all populations listed in the AFND which had been inputted

for gene/allele frequency data were verified for genotype data. Then, genotype data reported in literature were inputted via the online module. The first entries corresponded to the 179 genotypes in the 42 populations referenced in Rajalingam’s publication (Rajalingam et al. 2008). KIR genotypes were identified in the AFND by the assignment of a consecutive identifier (ID) number. The ID numbers reported by Rajalingam and colleagues were used as a referencing index where possible. All genotypes included in the original list published by Rajalingam and colleagues were validated for duplications. During this step, eight genotypes were detected to present duplications and thus were excluded from the initial list. The IDs of the duplicated genotypes were subsequently occupied by novel genotypes submitted to the AFND. For each new genotype, a consecutive number was assigned to the new entry. Internally, genotypes were stored as binary data to indicate the presence (1) or absence of the gene (0). However, in some populations the typing of certain genes was not reported by authors. In these cases, a letter “n” was used to denote that the genotype was ‘not typed’ at that particular locus.

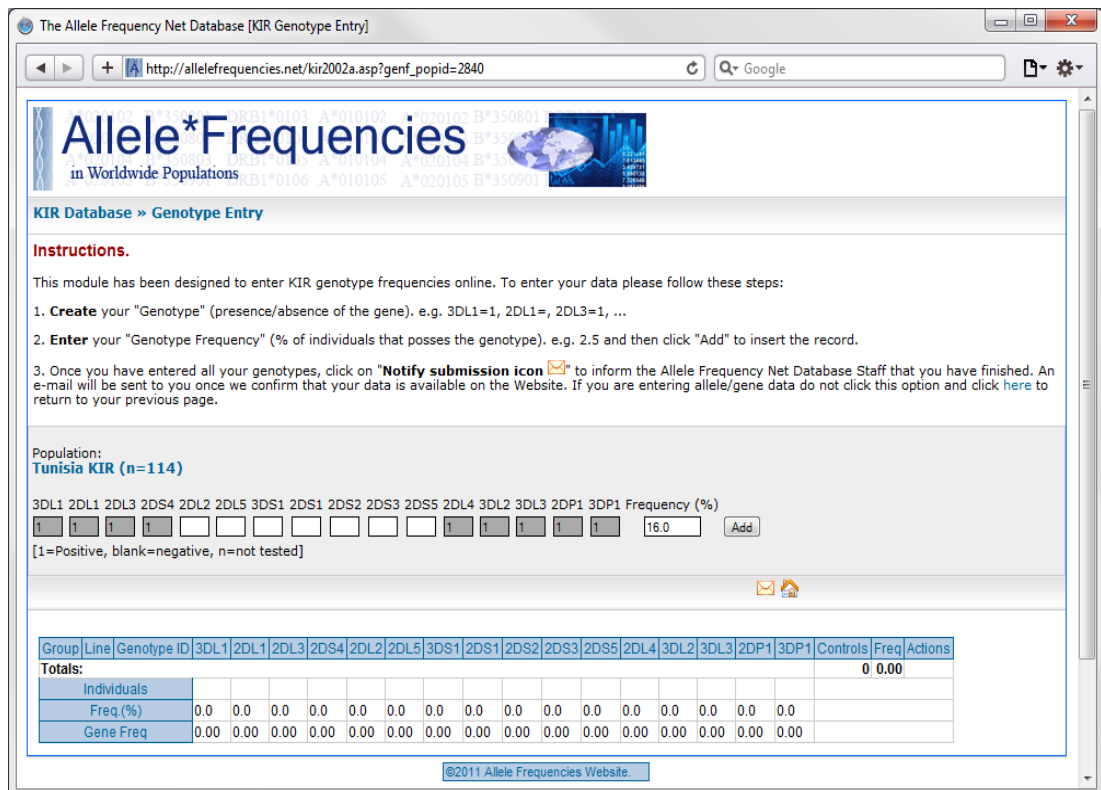


Figure 6.1: Screenshot of the KIR genotype submission form.

6.2.4 Data analysis

Demographic and frequency datasets were administered using the MySQL database system which constitutes the back-end of the AFND. To investigate the frequency distribution of KIR genes, statistical analyses were carried out using the Predictive Analytics Software (PASW) Version 18.

6.2.5 Data visualisation

To illustrate the geographical location and occurrence of the KIR genes across the world, overlaid maps were automatically generated using the frequency distribution map tool described in Section 3.4.1. The automated maps were based on the coordinates available for each of the populations with KIR data.

6.3 Results

6.3.1 Summary of data available in the AFND for KIR

To explore the availability and type of data submitted to the AFND, all 194 KIR population samples were classified according to the level of resolution (gene level or at alleles with 3, 5, 7 digits) employed in the allele determination. Table 6.1 shows the organisation of KIR data according to the number of populations and individuals typed by level of resolution. For instance, 166 population samples contained frequency data for the *KIR2DL1* gene of which only nine populations had alleles determined at 3-digits resolution. *KIR3DS1* was the locus with the most data available at gene level with 171 population samples. Conversely, *KIR2DL5A* and *KIR2DL5B* were the genes containing the least data with only fourteen and fifteen population samples respectively. The low number of submissions for *KIR2DL5A* and *KIR2DL5B* can be explained by the fact that these genes are relatively new (in terms of allele detection) and submissions were reported in the past as *KIR2DL5*. Also, pseudogenes *KIR2DP1* and *KIR3DP1* were reported in less than a half of the populations due to many publications excluding the

analysis of these genes. *KIR3DL1* and *KIR3DS1* were the genes with more information at allele level, the main reason was due to one publication by Norman and colleagues being performed to investigate selection on the *KIR3DL1* and *KIR3DS1* genes using the data from twenty-eight populations (Norman et al. 2007). Interestingly, only five populations have included data on alleles typed at seven-digit resolution. Although 198 different alleles with seven digits were described in the IPD-KIR database as of release 2.4.0 (April 2011), only nineteen of these alleles have been reported in the AFND (data not shown in table).

Table 6.1: KIR populations in the AFND by locus and level of resolution

KIR locus	Gene level		3-digits		5-digits		7-digits	
	Pops	Sample Size	Pops	Sample Size	Pops	Sample Size	Pops	Sample Size
<i>KIR2DL1</i>	166	17,767	9	810	4	461	1	100
<i>KIR2DL2</i>	168	18,897	15	1,342	-	-	-	-
<i>KIR2DL3</i>	168	18,897	14	1,345	-	-	-	-
<i>KIR2DL4</i>	131	15,750	20	2,080	20	2,080	3	202
<i>KIR2DL5</i>	123	14,464	-	-	-	-	-	-
<i>KIR2DL5A</i>	14	1,613	6	687	1	100	2	175
<i>KIR2DL5B</i>	15	1,713	6	687	1	100	2	175
<i>KIR2DP1</i>	95	10,666	-	-	-	-	-	-
<i>KIR2DS1</i>	168	17,997	8	738	1	100	-	-
<i>KIR2DS2</i>	168	17,997	2	175	2	175	-	-
<i>KIR2DS3</i>	140	16,441	5	592	3	329	-	-
<i>KIR2DS4</i>	136	16,132	17	1,989	12	1,321	-	-
<i>KIR2DS5</i>	136	16,200	9	839	1	100	-	-
<i>KIR3DL1</i>	167	18,852	54	7,408	54	7,408	-	-
<i>KIR3DL2</i>	134	15,864	15	1,826	2	286	-	-
<i>KIR3DL3</i>	120	14,658	2	169	2	169	-	-
<i>KIR3DP1</i>	88	9,702	5	668	-	-	-	-
<i>KIR3DS1</i>	171	20,990	48	5,066	41	4,310	-	-

Data available in the AFND as of May 2011

The analysis of data availability was extended to examine whether alleles described in release 2.4.0 of the IPD-KIR database (Robinson et al. 2005) were reported in the submissions to the AFND. Not surprisingly, only one third of the alleles described in release 2.4.0 were found in populations from AFND. *KIR2DL5B* and *KIR3DS1* were the genes with more allele confirmations containing more than eighty percent of the alleles confirmed each (Table 6.2). On the contrary, *KIR3DP1* and *KIR2DP1* pseudogenes contained the lowest number of confirmations. In the case of *KIR2DP1*

none of the alleles was confirmed in the AFND. The lack of allele information in several genes can be explained by the relatively new definition of many of the KIR alleles. For instance, the first release (1.0.0) generated in 2003 in the IPD-KIR database consisted of only 87 alleles (Marsh et al. 2003). However, in release 2.4.0, more than one hundred novel alleles were reported compared to the previous release 2.3.0 (August 2010).

Table 6.2: Confirmations of KIR alleles in the AFND

KIR locus	Number of alleles in IPD-KIR ^a	Alleles reported to AFND ^b (%)
<i>KIR2DL1</i>	43	28 (65.1)
<i>KIR2DL2</i>	27	5 (18.5)
<i>KIR2DL3</i>	32	7 (21.9)
<i>KIR2DL4</i>	47	30 (63.8)
<i>KIR2DL5A</i>	15	7 (46.7)
<i>KIR2DL5B</i>	25	22 (88.0)
<i>KIR2DP1</i>	22	0 (0.0)
<i>KIR2DS1</i>	15	7 (46.7)
<i>KIR2DS2</i>	22	7 (31.8)
<i>KIR2DS3</i>	13	10 (76.9)
<i>KIR2DS4</i>	30	14 (46.7)
<i>KIR2DS5</i>	16	11 (68.8)
<i>KIR3DL1</i>	70	47 (67.1)
<i>KIR3DL2</i>	84	17 (20.2)
<i>KIR3DL3</i>	101	51 (50.5)
<i>KIR3DP1</i>	23	2 (8.7)
<i>KIR3DS1</i>	16	13 (81.3)

^a Number of alleles reported as of release 2.4.0 April 2011 in the IPD-KIR database (Robinson et al. 2005);

^b Number of alleles confirmed in the AFND as of May 2011.

6.3.2 Frequency distribution of KIR genes

There is a wide variation in the frequencies of the KIR genes between worldwide populations. Based on this premise, *gene frequencies* (percentage of individuals carrying the gene) were compiled to examine the overall frequency distribution of the seventeen KIR genes. The analysis was divided into two sections: (i) estimation of the overall frequencies by gene and (ii) investigation of the occurrence of a particular gene by geographical region.

6.3.2.1 Overall frequency of KIR genes

The first analysis consisted of the examination of frequencies of KIR genes (Figure 6.2). The analysis showed that framework genes (*KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1*) were present in nearly all individuals with few exceptions.

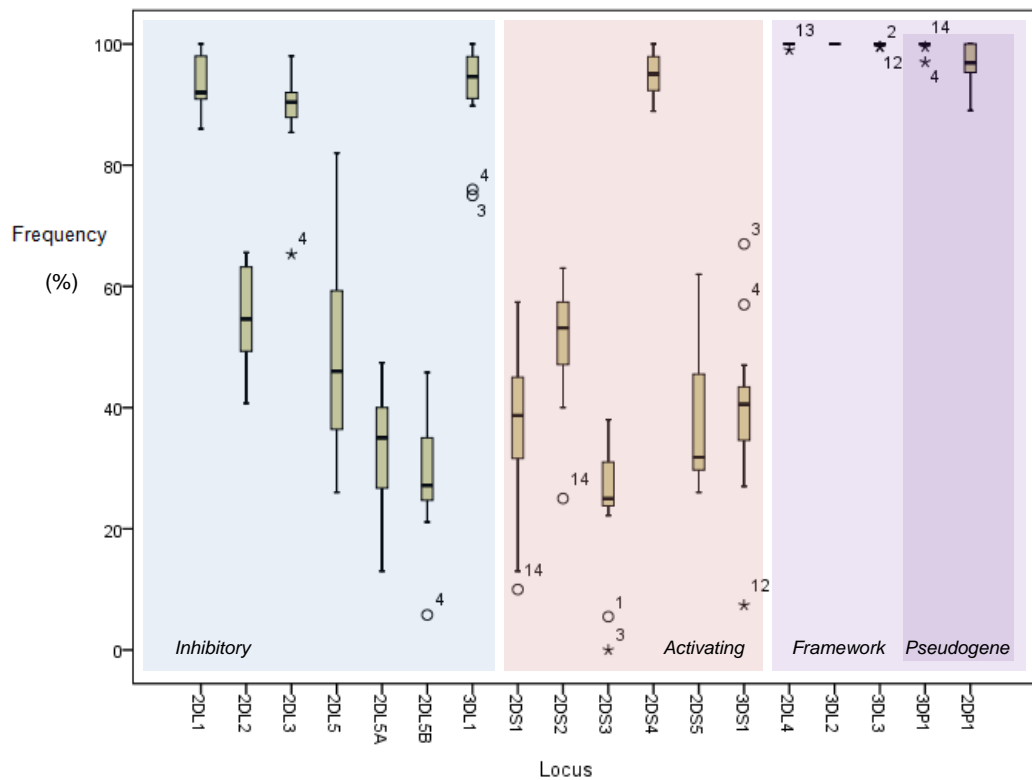


Figure 6.2: Overall gene frequencies in KIR populations on AFND.

Figure showing the overall gene frequencies organised by gene type: inhibitory (blue), activating (light pink), framework gene (violet) or pseudogene (purple). Outliers (values outside range) are denoted by a circle (o) and the number of occurrences. Extreme values (values exceeding more than three times the inter-quartile range) are displayed as asterisks (*) along with the number of cases. For instance, in *KIR2DL3*, the majority of frequencies were between 83-99% with four cases of frequencies around 65%.

As shown in Figure 6.2, inhibitory KIR genes *KIR2DL1*, *KIR2DL3* and *KIR3DL1* were found in the majority of the individuals whereas *KIR2DL2* and *KIR2DL5* (including *KIR2DL5A* and *KIR2DL5B*) contained the lowest frequencies among the inhibitory genes. Activating KIR genes (*KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5* and *KIR3DS1*) showed great variation in the presence/absence in the populations studied. All activating genes presented overall frequencies below 60%, except *KIR2DS4* which had frequencies over 90%. The high frequencies observed in *KIR2DL1*, *KIR2DL3*, *KIR3DL1* and *KIR2DS4* can be explained by the fact that these genes composed the A

haplotype group which is present in all populations (See Section 6.3.3). Also of interest, the box-and-whisker diagram, shown in Figure 6.2, depicts certain genes such as *KIR2DS1*, *KIR2DS2* and *KIR3DS1* with many occurrences of outliers and/or extreme values (values outside the expected range) leading to the investigation of the possible origin of these frequencies.

6.3.2.2 Frequency distribution of KIR genes by geographical region

As frequencies for each specific gene may vary considerably between individuals from different populations, the analysis was further extended to estimate frequencies by grouping populations belonging to a similar geographical region. Simple statistical analyses based on the generation of histograms and box-and-whisker plots were used to examine the frequencies of each KIR gene.

Inhibitory genes

KIR2DL1 presented very high frequencies (90-100%) in more than eighty percent of the populations (Figure 6.3A). However, three populations submitted to the AFND contained low frequencies: one population from Taiwan (65.6%) (Yen et al. 2006), one population from Yucpa people in Venezuela (70.5%) (Gendzekhadze et al. 2006) and one population from Aboriginal people in Australia (72%) (Toneva et al. 2001). It can be observed from Figure 6.3A that two Australian populations had the lowest overall frequencies by geographical region. This also was illustrated after the generation of a map showing the occurrence of the *KIR2DL1* gene (Figure 6.4A). *KIR2DL2* gene showed great variation whereby the majority of frequencies (91 cases) were reported from 40-60% (Figure 6.3B). Low frequencies were principally observed in three Oriental populations – two populations from Japan with 8.5% and 11.4% (Single et al. 2007a; Yawata et al. 2006) and one population with Atayal people from the Taroko Region in Taiwan (Single et al. 2007a), which surprisingly, presented complete absence of this gene. On the other hand, Nasioi people from Papua New Guinea were reported to contain the highest frequency with 95.5% (Single et al. 2007a). A high variation in frequencies was observed in Asian populations as shown in (Figure 6.3B). This variability can also be graphically illustrated in Figure 6.4B in which low frequencies are

mainly observed in East Asia, contrary to the high frequencies reported in West Asia. ***KIR2DL3*** was found to contain high frequencies (85-95%) in more than half of the populations reported in the database with this gene (Figure 6.3C). The two lowest frequencies were observed in two very geographically-distant populations: one population from Romania (56.5%) (Constantinescu et al. 2006) and the other in Nasioi individuals in Papua New Guinea (59.1%) (Single et al. 2007a). ***KIR2DL5*** was found to present a wide range of frequencies, the higher incidences being in twenty-nine populations with frequencies from 55-60% (Figure 6.3D). The lowest frequencies were observed in a population from South Asian individuals (Indian) living in England (26%) (Cook, Moss & Briggs 2003) and a population from Uzbekistan (28.4%) (Unpublished data). The highest frequencies were found in populations from Kanikar, people from the Tamil Nadu State in India (86%) (Rajalingam et al. 2008), West Papua Province (formerly known as Irian Jaya) in the New Guinea Island (86%) (Velickovic et al. 2009) and the Amazon Region in Brazil (85%) (Ewerton et al. 2007). Despite the low data available for ***KIR2DL5A*** and ***KIR2DL5B***, these loci were also included in the analysis to investigate possible patterns of these two similar structured genes (Figure 6.3E and Figure 6.3F). Most frequencies for ***KIR2DL5A*** were observed to vary from 30-40%, with the lowest frequencies found in two Sub-Saharan African populations – Comoros (13.0%) (Frassati et al. 2006) and Bubi people from Bioko Island in Equatorial Guinea (3.5%) (Unpublished data). The highest frequencies observed in this gene were found in one population from Uruguay (47.9%) (Unpublished data), and in one population from the Terceira Island in Azores (46.2%) (Fialho et al. 2009). ***KIR2DL5B*** was found to present frequencies from 20-30% in eight of the fifteen population samples containing data for this gene. The lowest frequency (5.8%) for the ***KIR2DL5B*** gene was found in Han individuals from the Zhejiang Province in China (Jiang et al. 2005) whereas the peak frequency (45.8%) was observed in Bubi people from Equatorial Guinea. Finally, ***KIR3DL1*** was observed to present high frequencies (90-100%) in 132 of the 167 populations with data on this gene (Figure 6.3G). However, low frequencies were found in one Australian and two Pacific Islands populations – Papua New Guinea Irian Jaya (46.0%) (Velickovic et al. 2009), Australia South Aborigine (55.0%) (Toneva et al. 2001) and Papua New Guinea Nasioi (59.1%) (Single et al. 2007a).

Activating genes

Activating genes were observed to present a much wider range of frequencies. ***KIR2DS1*** was found in 58 populations in a range from 35-45% (Figure 6.3H). As illustrated in Figure 6.4F, low frequencies were predominantly seen in African populations. In contrast, the highest frequencies were mainly reported in indigenous populations in Australia and Pacific Islands, in which Nasiois from Papua New Guinea presented the highest incidence with 90.9% (Single et al. 2007a). ***KIR2DS2*** was found to vary from 40-60% in nearly half of the populations typed for this locus (Figure 6.3I). Several populations in East Asia presented the lowest frequencies as shown in Figure 6.4G. For instance, a population from Japan reported the presence of this gene in 8.5% of the individuals (Single et al. 2007a). More outstanding was the report in the same study from a population of Taiwan in which the full absence of this gene was reported (Single et al. 2007a). The highest frequencies in the *KIR2DS2* gene were shown in one Australian and one Pacific Island populations – Australia South Aborigine (84.0%) (Toneva et al. 2001) and Papua New Guinea Nasioi (90.9%) (Single et al. 2007a). ***KIR2DS3*** appeared to have much lower frequencies than the rest of the activating genes – with frequencies from seventy populations between 20-35% (Figure 6.3J). Interestingly, it was observed that four Amerindian populations presented the complete absence of this gene – two Venezuelan populations (Yucpa and Bari) (Gendzekhadze et al. 2006), one population from Argentina (Wichis from the Salta Region) (Unpublished data) and one population corresponding to Tarahumara people located in the North of Mexico (Gutierrez-Rodriguez et al. 2006). The highest frequency (81%) was observed in Aboriginal individuals living in the South of Australia (Toneva et al. 2001). ***KIR2DS4*** contained the highest frequencies reported in activating genes in which ninety-six populations reported frequencies over 90% (Figure 6.3K). Surprisingly, a very low frequency (31%) was observed in individuals from the Guanacaste Province in Costa Rica (Carrington et al. 2005). ***KIR2DS5*** was mainly found in frequencies from 20-40% (Figure 6.3L). As shown in Figure 6.4J, the lower frequencies were found in South East Asia and in some Austronesian Islands, e.g. Borneo Kalimantan (10%) (Velickovic et al. 2009). Higher frequencies were mainly presented in India and in the majority of the Amerindian populations, such as in individuals from the Amazon Region in Brazil which reported a frequency of 90% (Ewernton et al. 2007). In the case of ***KIR3DS1***,

frequencies were generally found to range from 35-45% as reported by seventy-eight populations (Figure 6.3M). Lower frequencies were observed in all populations from Sub-Saharan Africa as depicted in Figure 6.4K. Conversely, highest frequencies corresponded to one population in Australia (78%) (Toneva et al. 2001) and one population from Karitiana people from the Rondonia State in Brazil (80%) (Single et al. 2007a). The extraordinary similarities of gene occurrence with the *KIR2DS1* gene are graphically represented in the overlaid maps (Figure 6.4F and Figure 6.4K).

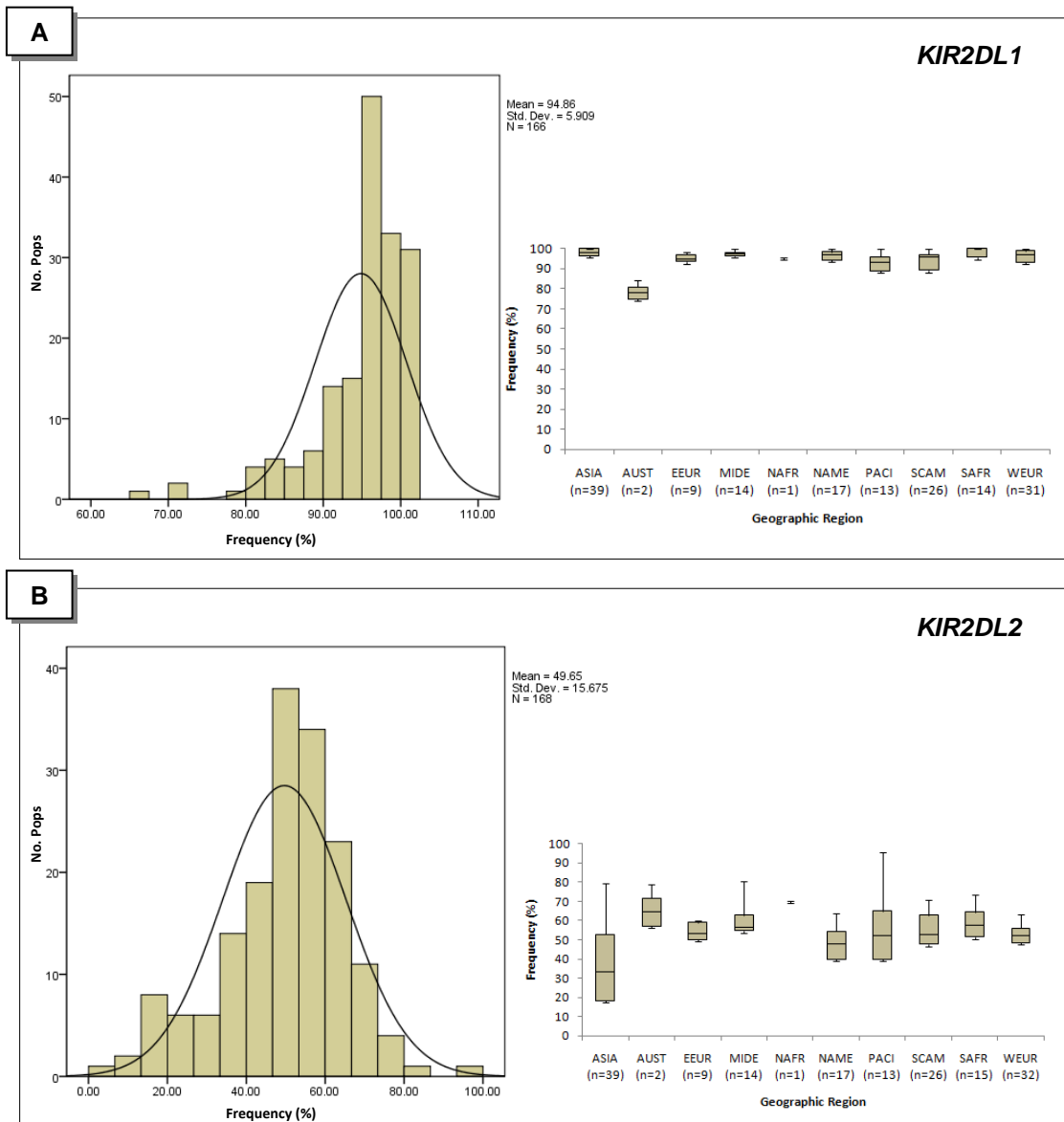


Figure 6.3: Frequency distribution of the KIR genes by geographical region.

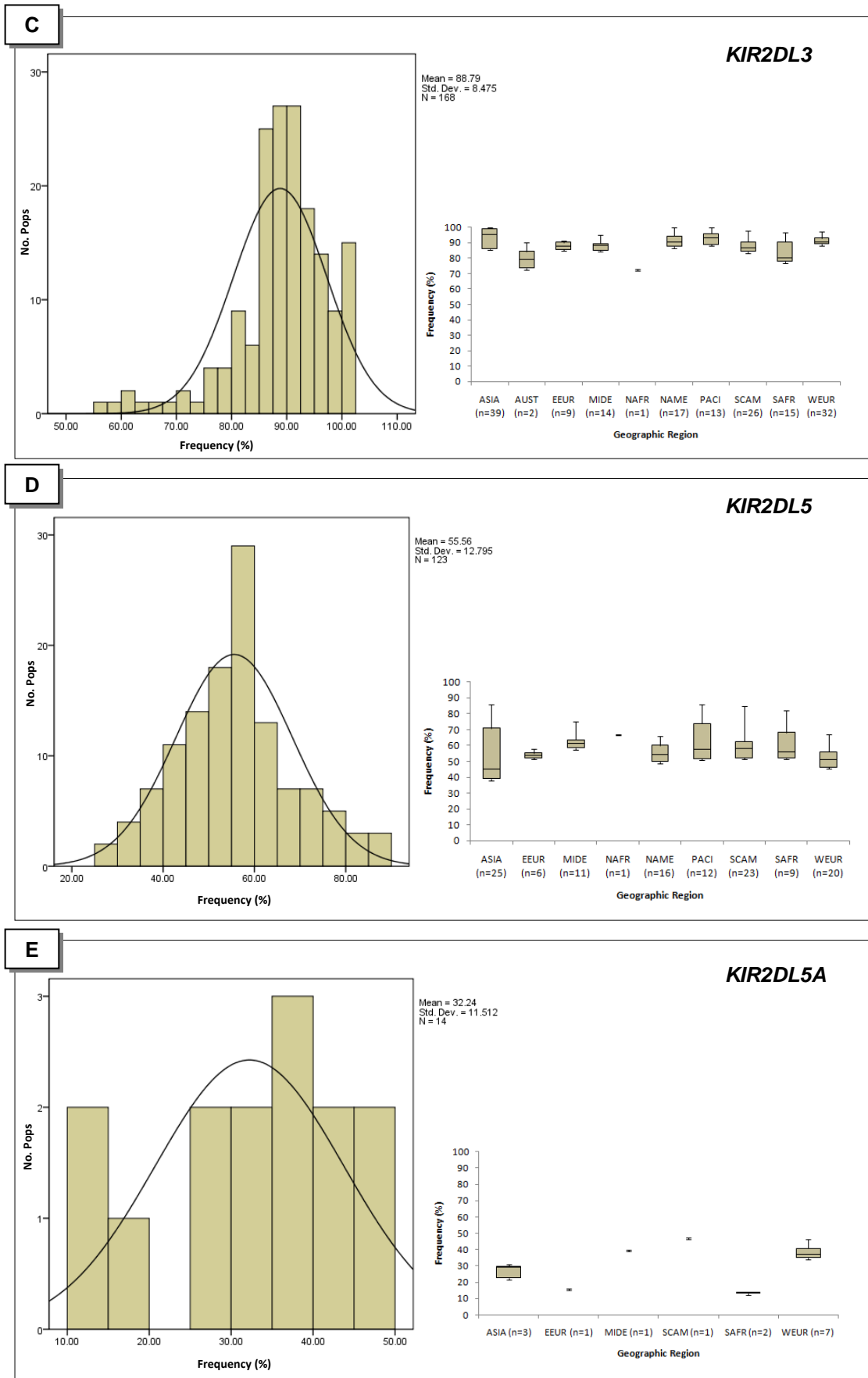


Figure 6.3: Frequency distribution of the KIR genes by geographical region (*Continued*).

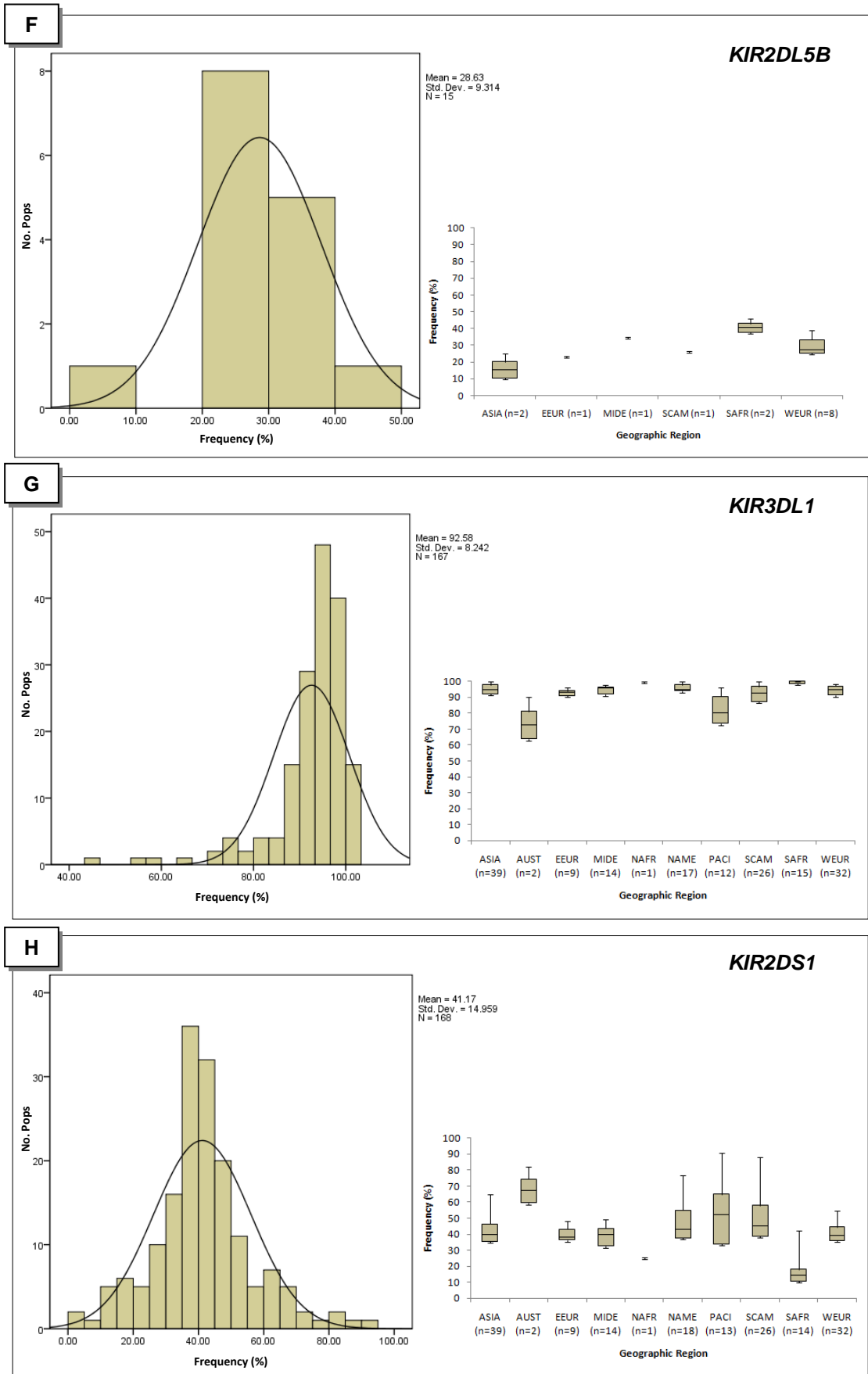


Figure 6.3: Frequency distribution of the KIR genes by geographical region (*Continued*).

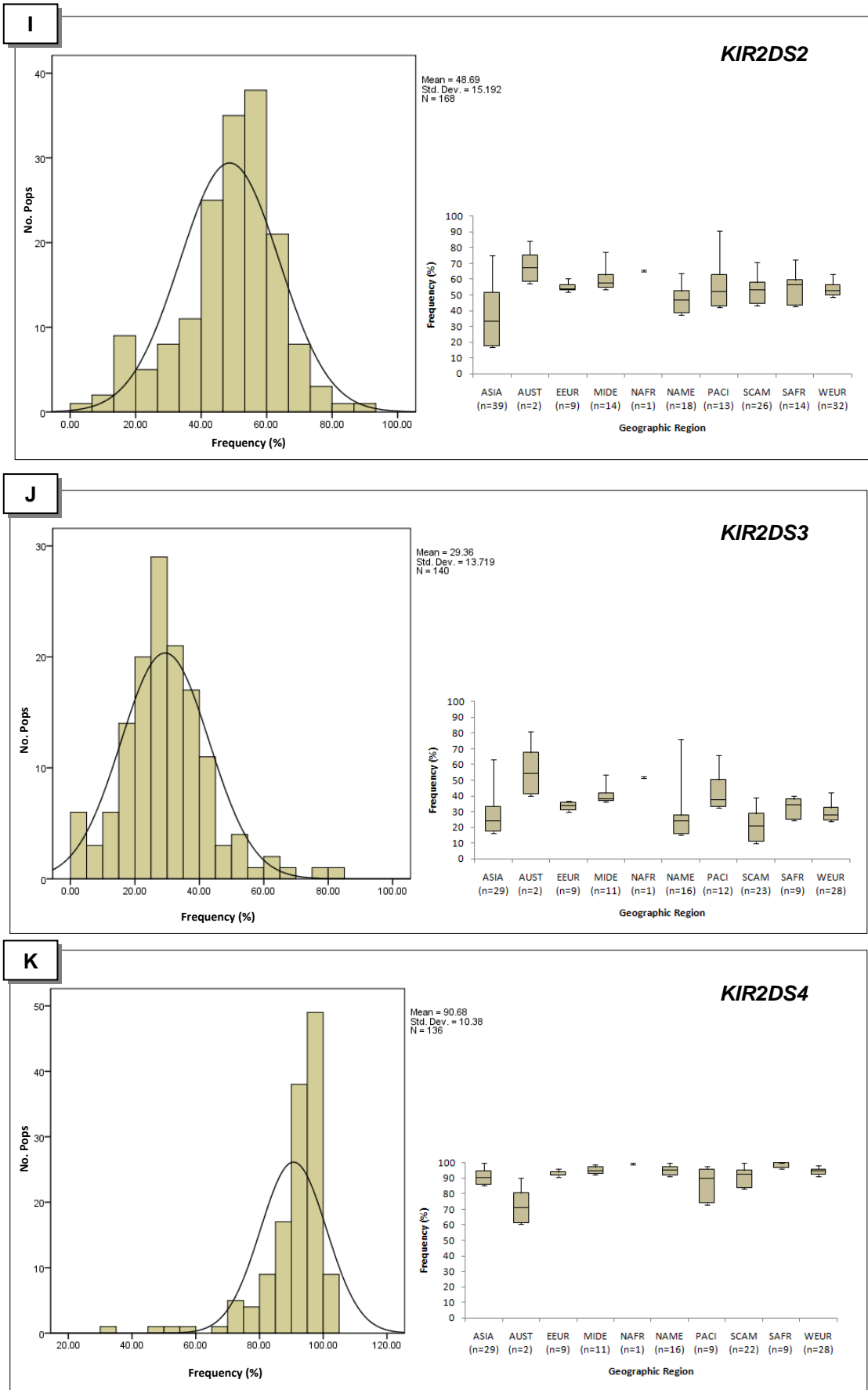


Figure 6.3: Frequency distribution of the KIR genes by geographical region (*Continued*).

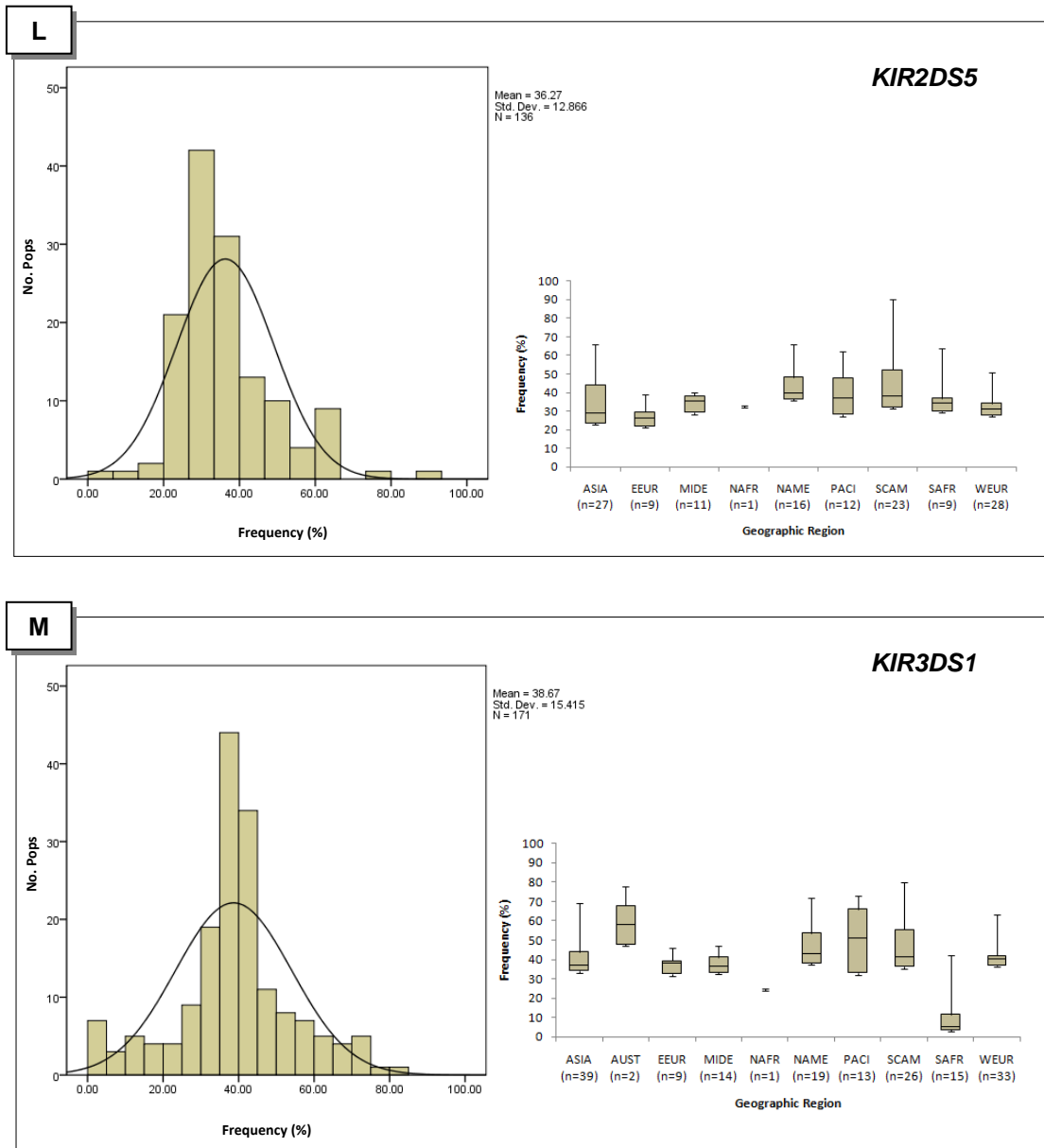


Figure 6.3: Frequency distribution of the KIR genes by geographical region (Continued).

This figure shows the counts of frequencies in the *KIR2DL1* gene (left) and the variation by geographical region (right). ASIA=Asia; AUST=Australia; EEUR=Eastern Europe; MIDE=Middle East; NAFR=North Africa; NAME=North America; PACI=Pacific; SAFR=Sub-Saharan Africa; SCAM=South and Central America; WEUR=Western Europe. In the case of *KIR2DL5A* and *KIR2DL5B*, gaps may be explained by limited number of populations typed for these two genes.

Framework genes

KIR2DL4, *KIR3DL2*, *KIR3DL3* and *KIR3DP1* framework genes have been recognised by their presence in nearly all individuals. However, several studies submitted to the AFND reported the absence of one of these genes in some individuals. Table 6.3 describes a summary of the individuals that were negative for these specific framework

genes. From this table can be observed that the *KIR2DL4* gene was found to be absent in ten individuals in seven populations, *KIR3DL2* in fourteen individuals in six populations and *KIR3DL3* in six individuals in five populations. Finally, the pseudogene *KIR3DP1*, which is also recognised as a framework gene presented the highest number of cases with gene absence with twenty-nine individuals in twelve populations.

Table 6.3: Populations with the absence of a KIR framework gene

Locus	Population	Sample Size	Gene Frequency (%)	Individuals with gene absence	Reference
2DL4	Poland KIR	690	99.7	2	(Majorczyk et al. 2007)
2DL4	Trinidad South Asians KIR	108	99	1	(Norman et al. 2002)
2DL4	Pakistan Karachi KIR	78	99	1	(Norman et al. 2002)
2DL4	Brazil Rio de Janeiro Mixed KIR	166	99.4	1	(NR)
2DL4	Turkey South KIR	200	99	2	(+)
2DL4	Solomon Islands KIR	40	95	2	(Taniguchi & Kawabata 2009)
2DL4	Equatorial Guinea Bioko Island Bubi KIR	95	99	1	(NR)
3DL2	Japan Tokyo KIR	239	99.2	2	(Miyashita et al. 2006)
3DL2	Brazil Amazon KIR	40	98	1	(Ewerton et al. 2007)
3DL2	Japan Kyoto KIR	240	99.6	1	(Mogami et al. 2007)
3DL2	Turkey South KIR	200	99.5	1	(+)
3DL2	Azores Terceira Island KIR	117	95.7	5	(Fialho et al. 2009)
3DL2	Taiwan KIR	96	95.8	4	(Yen et al. 2006)
3DL3	Tokelau KIR	47	98	1	(Velickovic, Velickovic & Dunkley 2006)
3DL3	Morocco Chaouya KIR	67	97	2	(NR)
3DL3	Brazil Rio de Janeiro Mixed KIR	166	99.4	1	(NR)
3DL3	Turkey South KIR	200	99.5	1	(+)
3DL3	Venezuela Mestizo KIR	205	99.5	1	(Conesa et al. 2010)
3DP1	France West KIR	108	97	3	(Denis et al. 2005)
3DP1	Spain Basque KIR	71	99	1	(Santin et al. 2006)
3DP1	Cook Islands KIR	48	94	3	(Velickovic, Velickovic & Dunkley 2006)
3DP1	Samoa KIR	50	98	1	(Velickovic, Velickovic & Dunkley 2006)
3DP1	Tokelau KIR	47	96	2	(Velickovic, Velickovic & Dunkley 2006)
3DP1	Uzbekistan KIR	67	98.5	1	(NR)
3DP1	Canada Iranian KIR	68	95.6	3	(NR)
3DP1	Turkey South KIR	200	99	2	(+)
3DP1	Solomon Islands KIR	40	95	2	(Taniguchi & Kawabata 2009)
3DP1	Macedonia KIR	214	99.5	1	(Djulejic et al. 2010)
3DP1	Equatorial Guinea Bioko Island Bubi KIR	95	99	1	(NR)
3DP1	Azores Terceira Island KIR	117	92.3	9	(Fialho et al. 2009)

NR=Populations submitted to the AFND with no associated bibliographic reference. +DOI: 10.1007/s11033-011-0945-5

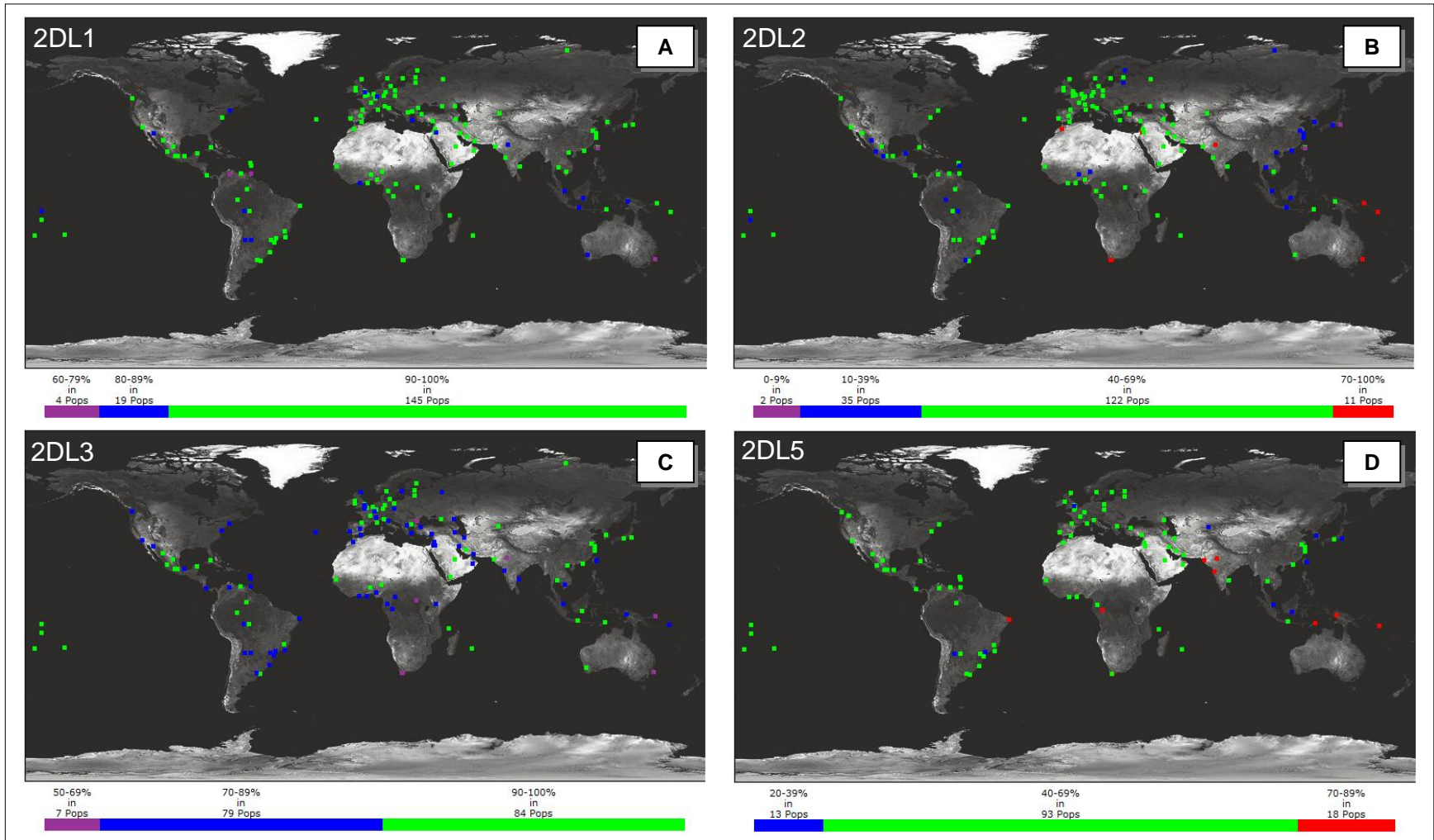


Figure 6.4: Occurrence of KIR genes in worldwide populations.

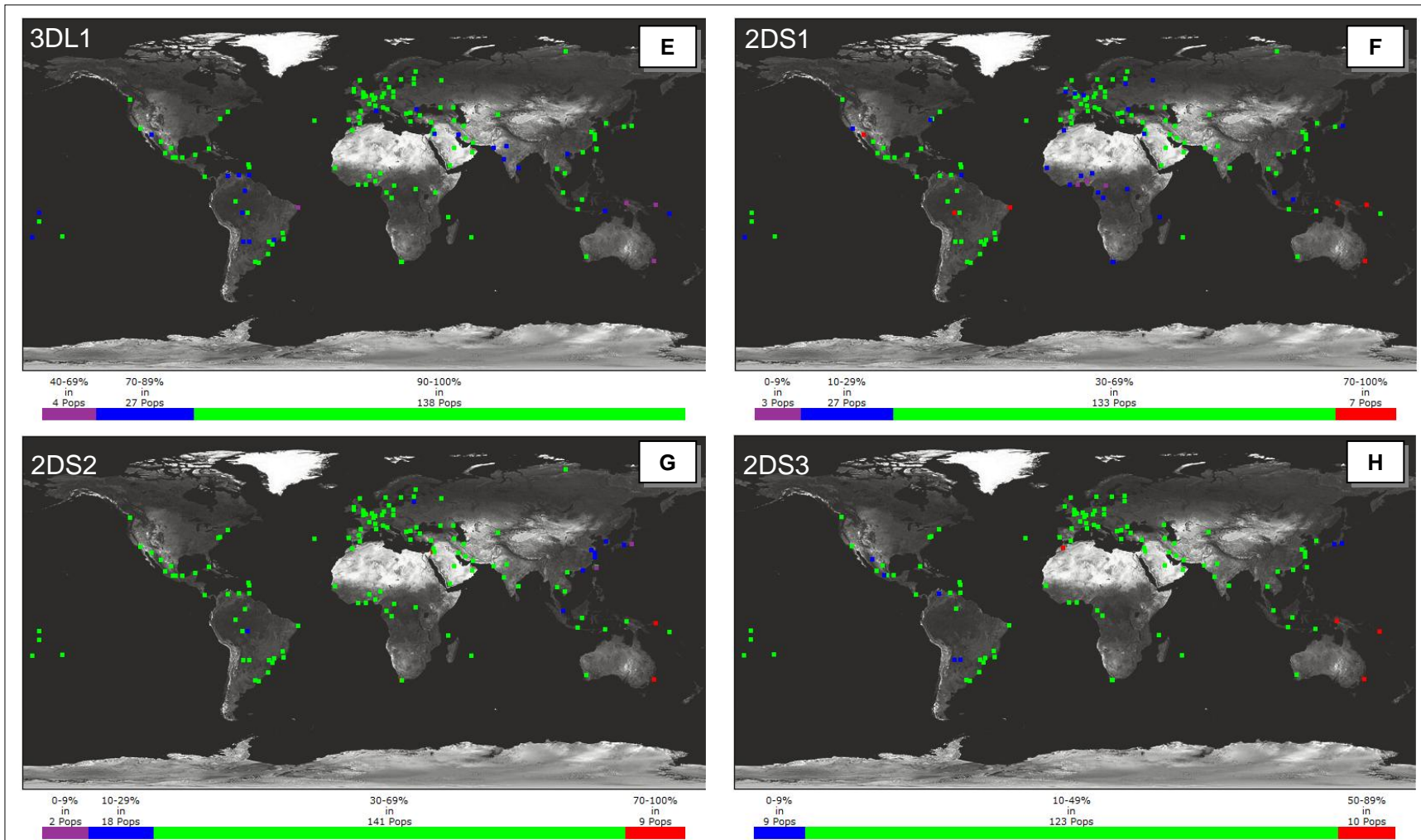


Figure 6.4: Occurrence of KIR genes in worldwide populations (*Continued*).

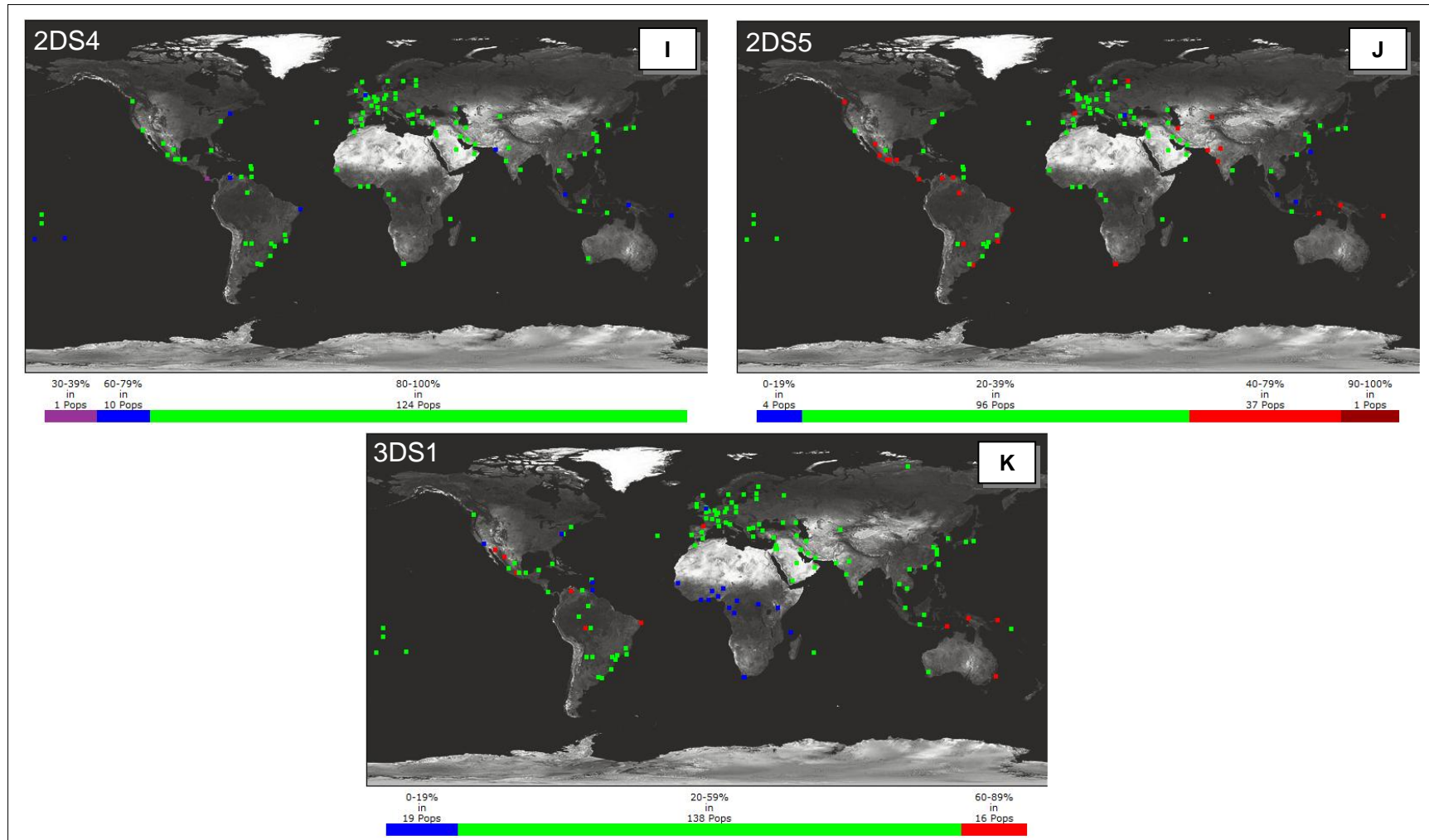


Figure 6.4: Occurrence of KIR genes in worldwide populations (*Continued*).

6.3.3 The KIR genotype database

A key concept in the investigation of the KIR gene content is the analysis of the genotypes (presence / absence of the gene) which are formed by the combination of A and/or B haplotypes. KIR haplotypes can be identified by familial studies. However, when the ability to perform such assays is not accessible, KIR genotypes can serve as a useful resource to understand the composition of an individual's profile. To investigate the KIR gene content organisation, 108 populations including 12,291 individuals with KIR genotype information were submitted to the AFND. From this compilation, a total of 396 different genotypes were identified. Figure 6.5 displays ten of the most common genotypes in the AFND representing 74.1% of the individuals tested. Based on the nomenclature adopted to identify the combination of haplotypes (AA and Bx), it was shown that one genotype AA (Genotype 1) was present in all populations in ~30% of individuals. The other nine common genotypes identified as Bx represented ~44% of the individuals.

Haplotype Genotype																	Pops	Indiv		
Group	ID	3DL1	2DL1	2DL3	2DS4	2DL2	2DL5	3DS1	2DS1	2DS2	2DS3	2DS5	2DL4	3DL2	3DL3	2DP1			3DP1	
AA	1	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	108	3,686
Bx	4	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	99	1,209
Bx	2	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	97	1,265
Bx	5	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	90	837
Bx	3	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	88	622
Bx	6	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	86	482
Bx	8	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	73	249
Bx	7	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	72	347
Bx	9	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	63	170
Bx	71	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	62	245

Figure 6.5: Most common KIR genotypes in worldwide populations

KIR genotypes are displayed by the presence (grey) or absence (blank) of each gene.

Despite the high percentage of individuals possessing one of these common combinations, 206 genotypes were identified to be unique occurring in only one individual in one population (Figure 6.6). Some of these genotypes could be the result of inaccurate typing. On the contrary, fifty-six genotypes representing ninety percent of the individuals typed were found in more than ten populations. Occasionally, AA genotypes had one of the genes normally present on an A haplotype missing (Figure 6.7). To aid the assessment of novel genotypes, the Genotype Frequency Search described in

Section 3.4.3 is used as a practical approach by displaying the closest genotypes (those differing in one gene) when an input profile is not listed after a search (See Figure 3.12).

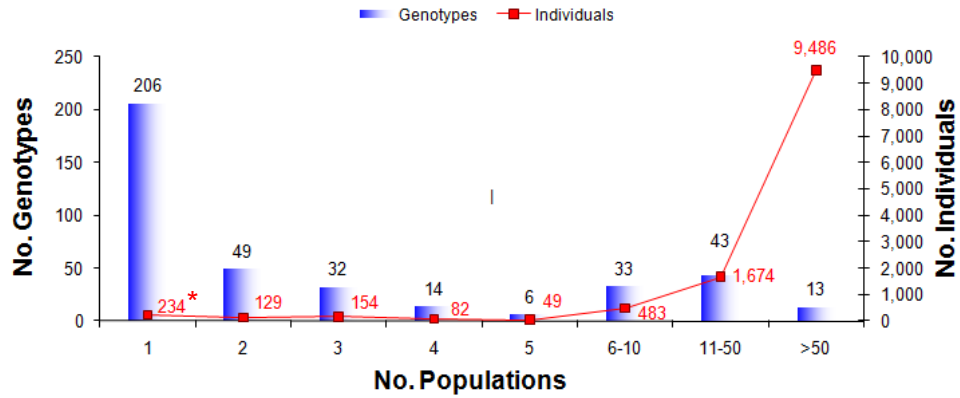


Figure 6.6: Occurrence of distinct genotypes in populations and individuals.

Number of genotypes (bars) and individuals (lines) in different populations are shown in this figure. * 185 genotypes occurred in one individual, 15 genotypes in two, 5 in three and 1 in four individuals.

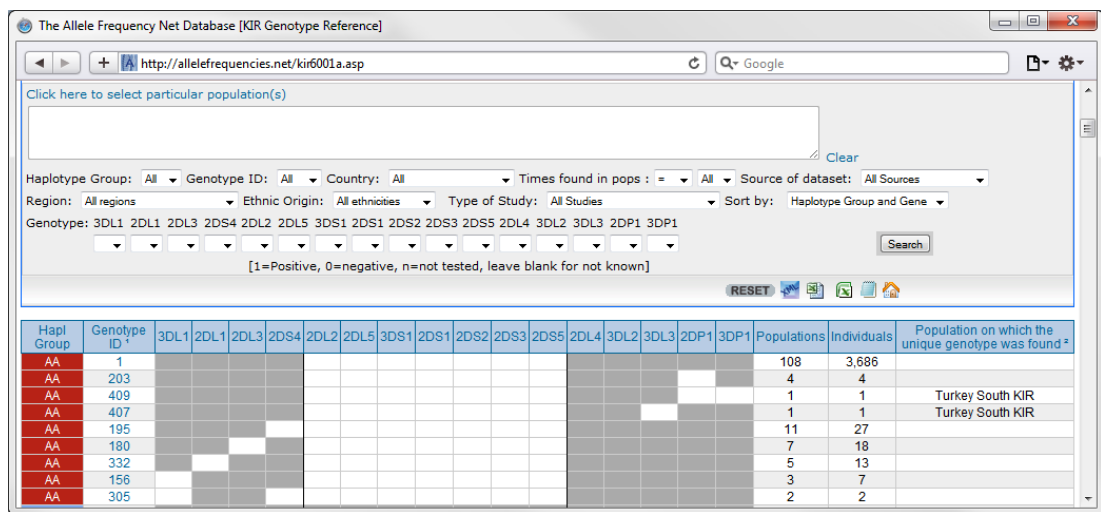


Figure 6.7: Example of KIR AA genotypes with one gene missing.

The analysis of genotype data was extended to examine the occurrence of these genotypes by geographical region (Table 6.4). Approximately 40% of the existing genotypes were reported in Asian individuals. Interestingly, the number of genotypes found in Eastern and Western Europe differed slightly (149 and 141 respectively) considering the number of populations reported in both regions (7 and 19 respectively). The number of unique genotypes was not necessarily proportional to the number of individuals tested. For example, twenty-three distinct genotypes were specific to

Western Europe populations and twenty-two in populations from the Pacific despite nearly seven fold more individuals from Western Europe being typed.

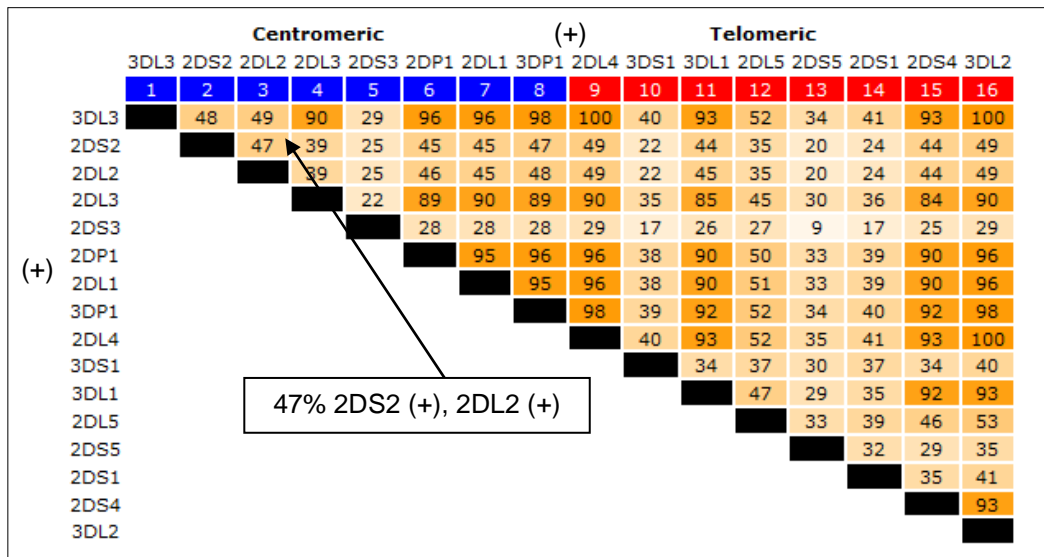
Table 6.4: Distribution of KIR genotypes by geographical region

Geographical region	Populations	Total genotypes	Unique genotypes	Individuals
Asia	24	160	34	2,326
Australia	1	6	0	42
Eastern Europe	7	149	37	1,267
Middle East	11	122	29	1,390
North Africa	1	22	2	67
North America	13	111	17	1,536
Pacific	9	92	22	427
South and Central America	16	136	31	1,868
Sub-Saharan Africa	7	60	11	461
Western Europe	19	141	23	2,907
Total	108		206	12,291

Another analysis performed in the examination of the gene content was the investigation of the presence/absence of two KIR genes within a given genotype by pairwise comparisons. Figure 6.8 shows a summary of the analysis performed in the sixteen KIR genes (*KIR2DL5A* and *KIR2DL5B* data combined), according to the reports from genotype data. This figure presents an outline of the linkage disequilibrium found in the KIR gene cluster. For instance, in the 108 populations with KIR genotype frequency data, 47% of the times that *KIR2DS2* was present, *KIR2DL2* was also present (Figure 6.8A) and 50% of the times that *KIR2DS2* was negative *KIR2DL2* was also negative (Figure 6.8B) illustrating the high LD between these two genes [See reports of linkage disequilibrium in (Middleton 2005)]. As expected, the highest percentages were found in the framework genes *KIR2DL4*, *KIR3DL2* and *KIR3DL3*, except *KIR3DP1* which was not determined in many of the populations. The analysis can also be extended to determine how many individuals had one gene and not the other. Thus, this tool can be useful for those individuals wishing to understand the LD between certain genes. Results displayed as percentages in this figure are available in the KIR breakdowns section in the AFND website at:

<http://allelefreqencies.net/kir6004a.asp>.

A)



B)

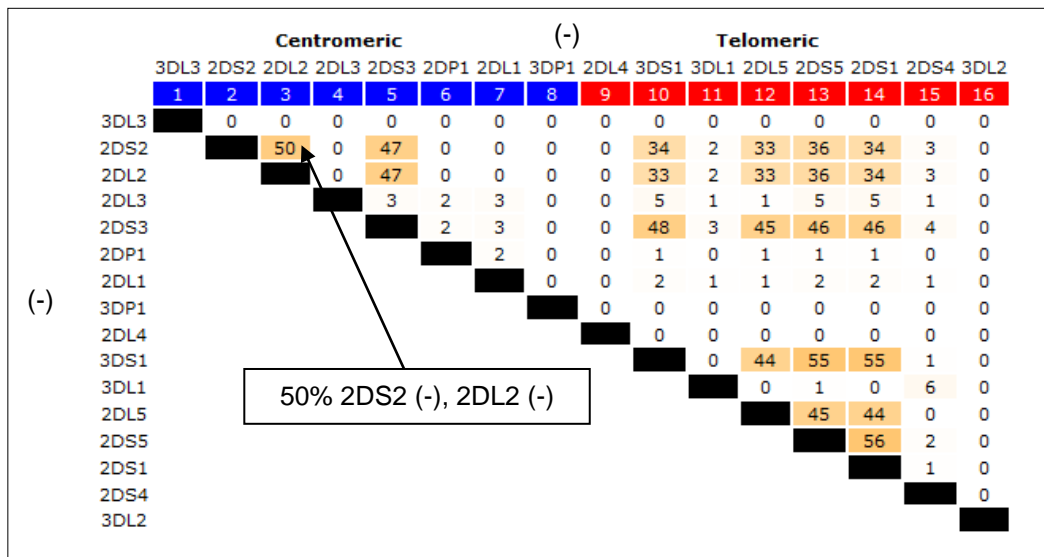


Figure 6.8: Linkage disequilibrium in the KIR genes.

6.4 Discussion

The design of a database to contain frequency data on the frequencies of KIR genes, complemented by searching mechanisms, has been shown to provide a convenient approach in the investigation of KIR genes among worldwide human populations. The KIR populations available in the AFND encompass the largest collection of frequency

data of this genomic region. These data has also assisted in the confirmation of KIR alleles described in the IPD-KIR database which stores the DNA sequence variants. Although it would be premature to use AFND to determine the rarity of KIR alleles, this resource can be considered as an initial approach to ascertain the presence of the existing alleles in the same fashion as in HLA alleles.

In the last decade, researchers have performed several analyses to investigate the occurrence of genes in different populations. The most extensive analysis involving comparisons of different populations was a study carried out by Single and colleagues in which thirty populations across the world were genotyped to investigate diversity and coevolution between KIR genes and HLA ligands (Single et al. 2007a). Those populations were also submitted to the AFND and included in the analysis described in this chapter. The enrichment of the database with data from 194 populations has increased significantly the amount of information to perform wider analysis of KIR genes. The results presented in this section provided more evidence on the characteristics of certain genes such as the absence of framework genes (*KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1*) observed in several individuals in various populations across the world (see Table 6.3). This information was used as a reference in a study by Nowak and colleagues in which a comparison of individuals lacking the *KIR2DL4* gene was used to demonstrate that the absence of this gene is not essential for human reproduction (Nowak et al. 2011).

One of the major outcomes from the results presented in this chapter was the Genotype Frequency Database, which represents an initial reference for those investigators interested in the analysis of the KIR gene content expressed in different individuals. Presented as a section in a review in *Immunology* (Middleton & Gonzalez 2010), the genotype list has been frequently adopted by researchers as a referencing index in the reporting of existing and novel genotypes (Djulejic et al. 2010; Shi et al. 2011; Zhu et al. 2010; Zhu et al. 2011). It is expected that the use of genotype IDs specified in the AFND will be considered as official by international accreditation bodies in the future. Another interesting finding in the KIR Genotype Database was that from the 396 genotypes reported, 206 were found to be unique to one population. These genotypes may be the product of an erroneous tissue typing, however it was not possible in this study to assess the accuracy of the typing of these genotypes. However, the GFS

described in Section 3.4.3 may help in the identification of false positive and false negative of KIR genes.

Linkage disequilibrium in the KIR region has also been under investigation by scientists in recent years (Gourraud et al. 2010). Pairwise comparisons of KIR genes using the online tool described in Section 6.3.3 can be used as a complementary resource of the study performed by Gourraud and colleagues in which the use of familial data was employed to investigate the LD in individuals from Northern Ireland (Gourraud et al. 2010). The LD analysis tool can also be expanded to include additional criteria such as geographical region, country, etc., increasing the usefulness of this automatic search. Also of interest was the analysis of the occurrence of KIR genes in the different geographical regions with the use of overlaid maps. These maps may be used as an illustrative representation of relationships between KIR genes such as the strong LD observed in *KIR2DL2* and *KIR2DS2* (Figure 6.4B and 6.4G).

The KIR module has been used in a wide range of projects being cited in more than twenty peer-reviewed publications according to recent figures from the Web of Knowledge as of August 2011. Some of the studies included the investigation of the structure of ligand recognition by NK cells receptors (Joyce & Sun 2011), understanding the stratification and cell function in European populations (Guinan et al. 2010), among many others.

Ongoing developments include the incorporation of virtual haplotypes into the KIR Genotype Database section. Based on data generated from a previous study in which KIR haplotypes were estimated using allele resolution from Northern Ireland families (Middleton, Meenagh & Gourraud 2007), this tool will allow users to enter their own genotype data and obtain the corresponding predicted haplotypes. Because gene content is suspected to have different effects in infectious and autoimmune diseases, the module will also provide analysis of these genotypes into frequencies of AA and Bx haplotype groups.

Future work in the investigation of KIR genes includes the continuation of a project of the 15th IHWS which collected KIR genotypes and the corresponding HLA ligand from individuals in twenty three worldwide populations (Hollenbach et al. 2010). Although

KIR and HLA genes are on different chromosomes it appears that associations between the presence/absence of KIR genes and the corresponding presence/absence of the HLA ligand (Parham 2005b). Thus, the collection of these data and development of the corresponding modules will allow investigators to examine the linkage between KIR and HLA genes, through the analysis of raw data (KIR individual's genotypes) from worldwide populations.

6.5 Conclusions

This chapter described a set of analyses which were performed for the investigation of the frequencies of KIR genes among individuals from different populations. The results presented in this chapter provide an insight into the extensive variability presented in this genomic region. The analyses carried out in this investigation included the dissemination of data available in the AFND in different formats such as allele, gene and genotype frequency. The analysis of frequency data was strongly supported by the use of maps generated by tools available in the AFND website. Results described in the chapter provided a wide and up-to-date outline of the relationships found among the presence of KIR genes in different geographical regions. At present, the KIR section in the AFND encompasses the largest source of data related to this polymorphism.

Chapter 7

Other immune genes: MIC and Cytokine gene polymorphisms

7.1 Introduction

Chapters 4 and 6 presented an outline on the frequencies of two of the most polymorphic regions in the human genome, HLA and KIR. The information expounded in these two chapters described a compendium of the frequencies observed in genes and corresponding alleles among worldwide human populations.

This chapter focuses on the investigation of two additional set of immune genes which were included in the AFND: (i) the major histocompatibility complex Class I chain-related genes (MIC) and (ii) several cytokine gene polymorphisms. The aim of this chapter is to present a similar synopsis on the frequencies of these genes among populations and describe a set of online modules that were implemented in the AFND website for the investigation of these genes.

As mentioned in Section 1.6.1, MIC genes encode a set of proteins expressed at the surface of tissue cells and fibroblasts. The MIC genomic region comprises seven loci in which MICA and MICB are the only genes which encode for proteins. MICA and MICB genes also contain a considerable polymorphism with 73 and 31 alleles respectively as of release 3.3.0 in the IMGT/HLA database. Prior to the official nomenclature of MICA, alleles were classified into eight variants (A4, A5, A5.1, A6, A7, A9, A9.1 and A10) based on microsatellites differences at exon 5. Thus, the examination of the different types of MIC frequency data (microsatellites and official alleles) published in the literature has been a primary need for researchers. With the availability of sequence data provided in the IMGT/HLA database, the mapping of MICA microsatellites and MICA official alleles can be performed by examining the specific

codons within the sequence at exon 5. For instance, MICA A4 presents the same four Alanine repeats between codons 291-304 as the official MICA*001, *007:01, *007:03, *012:01, *018:01, *018:02, *029, *043, *045, *051 and *061 alleles. Therefore, a MIC module was implemented in the AFND to assist scientists in the analysis of microsatellites and official allele nomenclatures. Additionally, the analysis of several diseases has been expanded to investigate the relationships between MICA and HLA-B due to the proximity of these two loci in the chromosome. To provide scientists with a computational tool for the examination of MICA and HLA-B association frequencies, this chapter also includes an online searching mechanism to consult data reported in more than fifty worldwide human populations.

To extend the polymorphism covered in the AFND, cytokine genes, which are protein molecules secreted by different cells in the human body, were also included in the database. As reviewed in Section 1.6.2, cytokines are believed to participate in the regulation of innate and adaptive immune response and are known to present high polymorphism in specific regions such as SNPs or microsatellites leading to the importance of investigating the variability of these genes in different individuals. Following a similar scheme for the investigation of gene polymorphisms as described in previous chapters, a module in the AFND was incorporated to examine the frequencies of cytokine gene polymorphism across worldwide populations. As such, this chapter includes a summary of the cytokine gene polymorphisms reported in more than one hundred populations and a set of online tools that were implemented in the AFND for the analysis of the polymorphisms.

7.2 Materials and methods

7.2.1 Population datasets

The criteria applied for the selection of population datasets are described in the list of protocols reviewed in Section 2.2.3. Both datasets (MIC and Cytokine gene polymorphisms) were captured via the online submission form described in Section 2.2.2.

MIC population samples

Fifty-eight population samples covering 7,649 healthy unrelated individuals containing data on MIC frequencies were included in the analysis. The collection of datasets consisted of fifty-two populations reported in fifteen peer-review Journals from January 1997 to December 2010, five populations available in the Proceedings of the 13th IHWS and one unpublished population directly submitted to the AFND. In order to compare the frequency distribution of MIC genes among different demographic groups, samples were classified in nine geographical regions: Asia (ASIA), Eastern Europe (EEUR), Middle East (MIDE), North Africa (NAFR), North America (NAME), Pacific (PACI), Sub-Saharan Africa (SAFR), South and Central America (SCAM) and Western Europe (WEUR). None of the populations submitted to the AFND contained data on individuals from Australia.

MICA and HLA-B association datasets

In order to examine the relationships between MICA and HLA-B, 20 of the 58 populations containing data on MICA and HLA-B frequencies were submitted to the AFND. The compilation of MICA and HLA-B data consisted of 257 frequency records reported in 2,426 individuals.

Cytokine population samples

One hundred and thirteen population samples in 18,246 healthy unrelated individuals containing data on frequencies of cytokine polymorphisms were submitted to the AFND and included in the analysis. The assemblage of datasets corresponded to seventy-five publications from a review of fifteen peer-reviewed Journals from January 2001 to December 2010, eighteen population samples from Proceedings from the 13th IHWS and twenty unpublished dataset. All populations containing cytokine frequency data were classified into ten geographical regions for further analysis: Asia (ASIA), Australia (AUST), Eastern Europe (EEUR), Middle East (MIDE), North Africa (NAFR), North America (NAME), Pacific (PACI), Sub-Saharan Africa (SAFR), South and Central America (SCAM) and Western Europe (WEUR).

7.2.2 MIC and cytokine gene frequency data

Frequency datasets used in the analysis of MIC genes were available at microsatellite (e.g. A4, A5, A5.1, etc.) and/or allele level (e.g. MICA*001, MICA*002:01, MICB*001, etc). Frequency information for each MIC gene was given as *genotype frequency* (percentage of individuals carrying the allele/microsatellite variant) or *allele frequency* (proportion of allele/microsatellite variant copies in the gene). 709 of the 718 frequency records submitted to the AFND contained data on frequencies of MICA gene and only one population included the report of nine MICB alleles.

The nomenclature used in the construction of the cytokine module consisted of three main components: (i) the gene name followed by a slash symbol (/), (ii) the nucleotide position of the SNP or microsatellite variant and (iii) the genotype generated by nucleotide combinations. For instance, the codes ‘AIF-1/-932CC’, ‘AIF-1/-932CT’ and ‘AIF-1/-932TT’ were used to represent the genotype polymorphisms of the Allograft inflammatory factor 1 gene at position -932. Using this notation, a total of 27 cytokine genes comprising 51 genotype polymorphisms were submitted to the AFND. Names of cytokine genes were stored in the AFND based on the reports from literature, however, all cytokine genes were internally mapped to the current notation described in the Entrez Gene database (Table 7.1) (Maglott et al. 2011).

Table 7.1: Cytokine names and references used in the AFND

Cytokine name	AFND ID	Gene Entrez Name / ID
Allograft inflammatory factor 1	AIF-1	AIF1 / 199
Basic fibroblast growth factor 2	bFGF	FGF2 / 2247
Epidermal growth factor	EGF	EGF / 1950
Colony stimulating factor 2 (granulocyte-macrophage)	GM-CSF	CSF2 / 1437
Interferon gamma	IFN γ	IFNG / 3458
Insulin-like growth factor 1	IGF-1	IGF1 / 3479
Interleukin 1 alpha	IL-1alpha	IL1A / 3552
Interleukin 1 beta	IL-1beta	IL1B / 3553
Interleukin 1 receptor type I	IL1R psti 1970	IL1R1 / 3554
Interleukin 1 receptor type I	IL1RA mspa 111100	IL1R1 / 3554
Interleukin 2	IL-2	IL2 / 3558

Table 7.1: Cytokine names and references used in the AFND

Cytokine name	AFND ID	Gene Entrez Name / ID
Interleukin 4	IL-4	IL4 / 3565
Interleukin 4 receptor	IL-4R	IL4R / 3566
Interleukin 6	IL-6	IL6 / 3569
Interleukin 10	IL-10	IL10 / 3586
Interleukin 12A p35	IL-12	IL12A / 3592
Interleukin 12B p40	IL-12p40	IL12B / 3593
Interleukin 13	IL-13	IL13 / 3596
Interleukin 15	IL-15	IL15 / 3600
Interleukin 18	IL-18	IL18 / 3606
Nerve growth factor	NGF	NGF / 4803
Platelet-derived growth factor beta polypeptide	PDGFB	PDGFB / 5155
Regulated upon Activation Normal T-cell Expressed and Secreted	RANTES	CCL5 / 6352
Transforming growth factor beta 1	TGFbeta1	TGFB1 / 7040
Tumour necrosis factor	TNFalpha	TNF / 7124
Lymphotoxin alpha TNF superfamily member 1	TNFbeta	LTA / 4049
Vascular endothelial growth factor A	VEGF	VEGFA / 7422

Frequency data for cytokine gene polymorphisms was entered as *genotype frequencies* (number of individuals carrying the specific cytokine polymorphism genotype). The compilation of genotype frequency data consisted of 3,603 frequency records.

7.3 Results

7.3.1 Major histocompatibility Class I chain-related genes

Three main subjects were investigated in the analysis of MIC frequency data: (i) the availability of MICA and MICB according to the level of resolution performed in the allele determination, (ii) the frequencies of MIC genes in different populations at allele and microsatellite level and (iii) the examination of the occurrence of MICA and HLA-B associations.

7.3.1.1 Availability of MIC data in the AFND

After the submission of MIC frequency information in the AFND, all 58 population datasets were classified according to the level of resolution of allele typing. Table 7.2 illustrates the availability of MICA and MICB data at microsatellite and allele level (3, 5 or 7 digits). Not surprisingly, the vast majority of reports comprised data on MICA alleles. 25 of the 58 populations were found to contain frequency data at microsatellite level whereas 36 described frequencies using the official nomenclature. Three populations reported data at both levels (microsatellites and official alleles). Very limited data on MICB was found in the literature. The unique population reported in the AFND with MICB data consisted of a population from Thailand and the corresponding frequencies are described in this chapter in Section 7.3.1.2.

Table 7.2: Availability of MIC data by level of resolution

Locus	Microsatellites		3 digits		5 digits		7 digits	
	Pops	Indiv	Pops	Indiv	Pops	Indiv	Pops	Indiv
MICA	25	3,319	36	4,733	21	2,503	N/A	N/A
MICB	-	-	1	100	1	100	-	-

N/A=Not available as MICA locus does not contain alleles at 7 digits.

To extend the analysis of data available, all official alleles described in the IMGT/HLA database were examined according to the number of confirmations reported in the AFND. Approximately 74% (54/73) of the MICA alleles described in the IMGT/HLA database (as of release 3.3.0) was reported in the AFND whereas only six alleles of 54 were found for MICB as of May 2011.

7.3.1.2 Polymorphism and diversity of MIC genes among populations

Overall frequencies of MICA microsatellites

All twenty-five populations with MICA microsatellite data were analysed to estimate their overall frequencies. As shown in Figure 7.1, MICA A6 was observed to contain the highest incidence in that 13 of the 25 populations typed for MICA A6 were found to vary from 25-50%. Conversely, MICA A4 was the microsatellite with lowest frequencies (5.6-15% in 20 of the 25 populations). The mutational variant MICA A9.1 was reported

in only one individual who was from Korea (Sohn et al. 2010), similarly to MICA A10 which was also found in only one individual from the Hunan Province in China (Tian et al. 2006). The A7 microsatellite, which was originally found in a Spanish individual from a cohort of celiac disease subjects (Rueda et al. 2002), was not found in information submitted to the AFND. Interestingly, MICA*050, the allele counterpart of A7 microsatellite, was also reported in one individual from Murcia in Spain from a sample of 154 healthy unrelated individuals (Lucas et al. 2008).

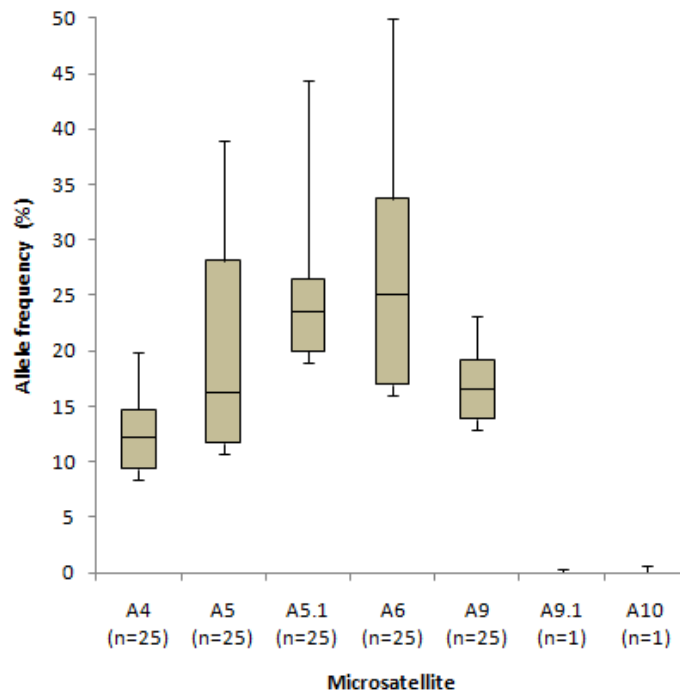


Figure 7.1: Frequency distribution of MICA microsatellites.

MICA microsatellites by geographical region

To examine the occurrence of MICA microsatellites by geographical location, all populations were grouped in only four regions due to limited data available (Figure 7.2). In this figure a consistency can be observed in frequencies of MICA A4 and A9 in all four geographical regions. However, the uniformity in frequencies of these two microsatellites contrasts with the significant variation shown in MICA A5, A5.1 and A6.

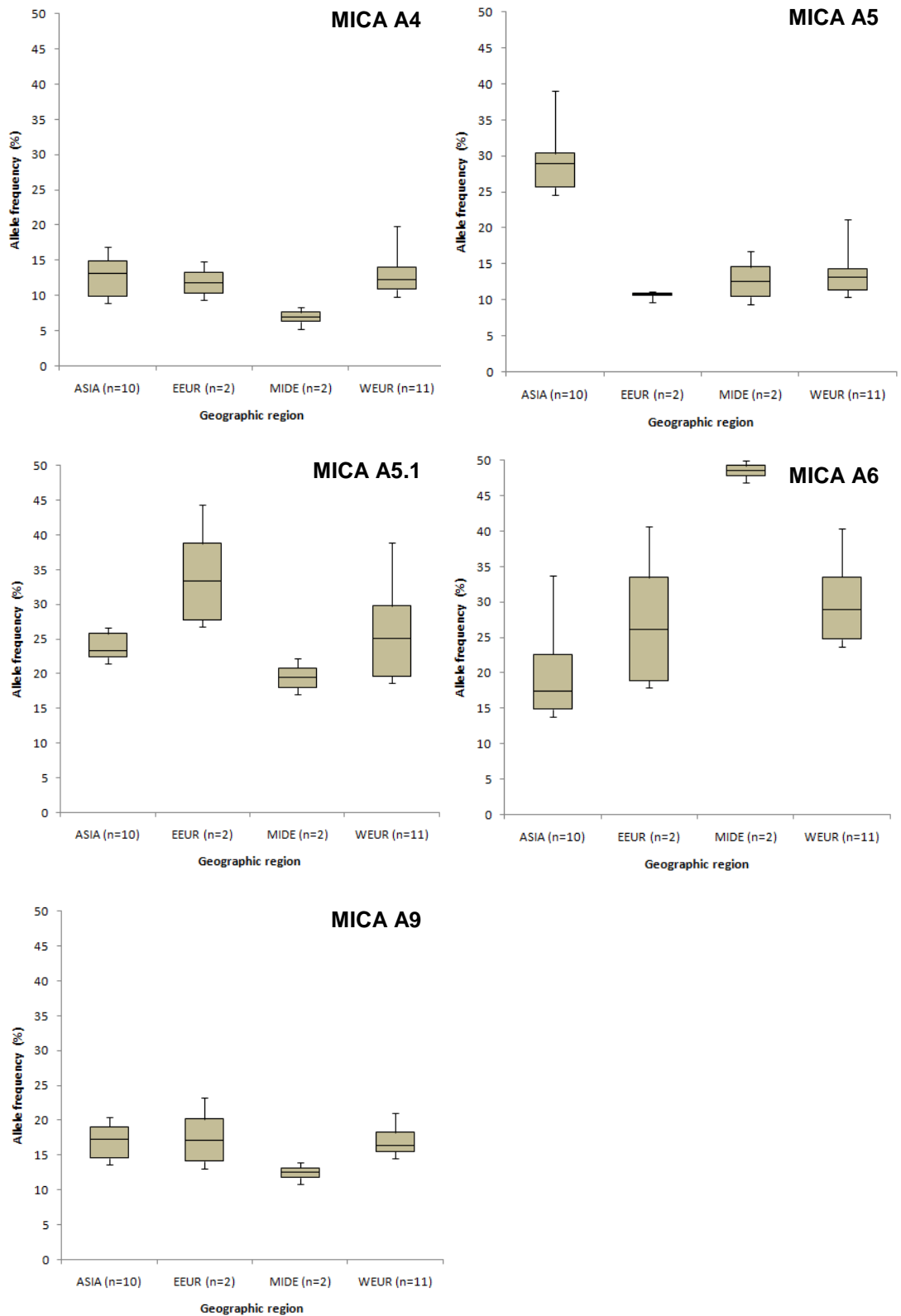


Figure 7.2: Frequency distribution of MICA microsatellites by geographical region.

This figure shows the overall allele frequencies by geographical region. Number of populations considered in the analysis is indicated with an 'n'. ASIA=Asia; Geographical regions: EEUR=Eastern Europe; MIDE=Middle East; WEUR=Western Europe.

For instance, in Figure 7.2 it can be observed that Asian populations presented very high frequencies for MICA A5 compared to those reported in Eastern and Western Europe and Middle East. Likewise, MICA A6 was found to be much more frequent (~50%) in the two populations from Middle East than in the rest of geographical regions. These high frequencies may be associated to the fact that in these two Middle East populations the number of individuals sampled was very small (n=18).

Overall frequencies of MICA alleles

A third analysis was carried out to examine the occurrence of MICA alleles reported in the AFND using the official nomenclature. Results were organised according to the number of reports in the AFND which are shown in Figure 7.3. MICA*004 and MICA*010 were the more prevalent alleles with 36 and 30 reports respectively. However, in terms of frequencies, the highest frequencies corresponded to the MICA*002:01 and MICA*008:01 alleles whereas the lowest frequencies of the top 10 seen alleles were found in MICA*016 and MICA*011 alleles.

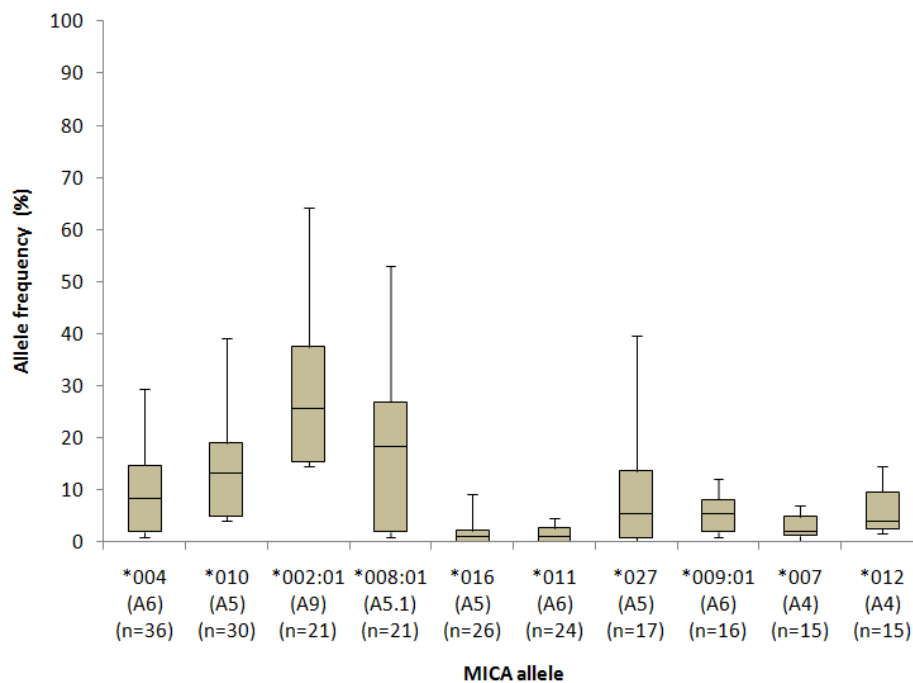


Figure 7.3: Frequency distribution of top 10 MICA alleles.

The first two lines in the x-axis legend correspond to the MICA official allele name and their corresponding microsatellite equivalence. Numbers of populations considered in the analysis are indicated with an 'n'.

As expected, the top ten official MICA alleles reported in the AFND corresponded to the most frequent MICA microsatellites: A4 (MICA*007, *012), A5 (MICA*010, *016, *027), A5.1 (MICA*008:01), A6 (MICA*004, *009:01, *011) and A9 (MICA*002:01). Using this approach data in the past containing microsatellite variants can be linked to the most likely official allele by examining the corresponding allelic equivalence at exon 5 and the frequency of the allele in a given population.

Frequencies of MICB alleles

Data on MICB frequencies was narrowed to only one population from Thailand containing nine frequencies: MICB*002 (59%), MICB*003 (1%), MICB*004 (23%), MICB*005:01 (0%), MICB*005:02 (62%), MICB*005:03 (6%), MICB*008 (17%), MICB*013 (4%), MICB*014 (7%) (Jumnainsong et al. 2008).

Software to aid the examination of MICA and MICB alleles

In order to allow users the examination of MICA/MICB alleles and MICA microsatellite frequencies, an online search was implemented in the AFND. The searching tool was based on the structure of the allele frequency search described in Chapter 3 in Section 3.4.1. The tool can be accessed using the following link:

<http://www.allelefrequencies.net/mic6001a.asp>

7.3.1.3 Occurrence of MICA and HLA-B associations among populations

To aid the examination of MICA and HLA-B relationships, an online module was implemented in the AFND website under the MIC section (Figure 7.4). Following a similar scheme of multi-filters as in the HLA haplotype frequency search described in Section 3.4.2, this searching mechanism allows users to input a particular MICA or HLA-B allele and find the strongest relationships from 257 records available in the AFND. For instance, Figure 7.4 shows 10 of the 37 records with associations between MICA*004 and any HLA-B allele. From this figure it can be observed that the highest frequency match corresponded to the allele HLA-B*42:01 which was observed in seven of the 39 individuals typed in a sample of African Americans in the United States (Tian

et al. 2003). As shown in Figure 7.4, the frequency of MICA and HLA-B associations may vary considerably depending in the ethnicity/geographical location of the individuals. For example, all three populations from Spain displayed in Figure 7.4 were found to present high associations (7.3-12.1%) with the HLA-B*44 allele, similar to two Korean (8.3-11.5%) and one Moroccan (10.6%) populations. In Efik individuals from Nigeria, the highest “MICA*004-HLA-B*” frequency association (10.9%) was observed with the HLA-B*49:01 allele. Thus, by using the HLA-MIC association frequency tool researchers can examine the most likely occurrences within MICA and HLA-B loci.

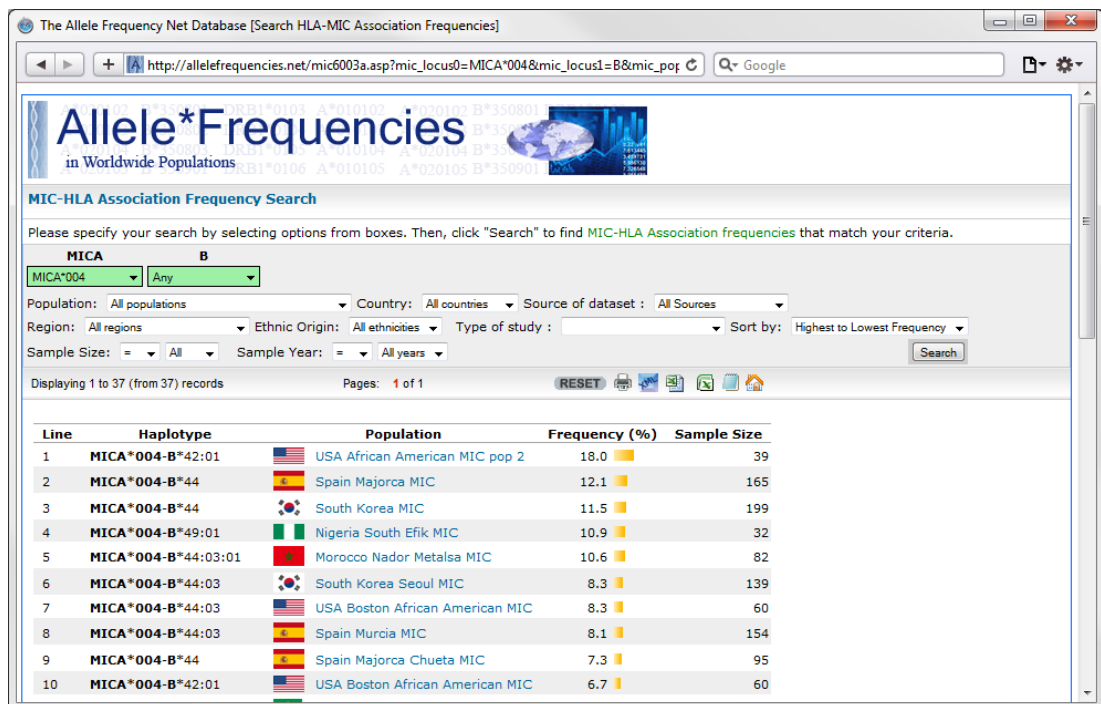


Figure 7.4: MICA and HLA-B association frequency search.

7.3.2 Cytokine gene polymorphisms

The analysis of cytokine gene polymorphisms was divided into two main categories: (i) dissemination of the availability of frequency data by cytokine gene polymorphism and geographical region and (ii) the implementation of an online searching mechanism for the investigation of polymorphisms among worldwide populations.

7.3.2.1 Availability of Cytokine data in the AFND

The initial analysis for the investigation of cytokine gene polymorphisms consisted of the classification of all 113 cytokine population frequency datasets grouping the populations by geographical region and gene. As shown in Table 7.3, frequency data was available for all geographical regions with at least one population reporting polymorphism frequencies. From data compilation, 51 cytokine gene polymorphisms were found in the submissions. The majority of the submissions corresponded to the report of Interleukin 6 (IL-6), Transforming Growth Factor beta 1 (TGFbeta1) and Tumour Necrosis Factor (TNFalpha).

Table 7.3: Availability of cytokine gene polymorphisms by geographical region

Gene / Polymorphism	Geographical region							
	ASIA Pops/ Indiv	EEUR Pops/ Indiv	MIDE Pops/ Indiv	NAFR Pops/ Indiv	NAME Pops/ Indiv	SAFR Pops/ Indiv	SCAM Pops/ Indiv	WEUR Pops/ Indiv
AIF-1/ -932 (C T)					2/395			
bFGF/ 223 (C T)								1/500
EGF/ 61 (A G)								2/760
GM-CSF/ -677 (A C) -1916 (C T)								1/500
IFNgamma/ 874 (A T) utr5644 (A T)	9/506	14/1387	7/583		19/2698		7/776	14/3103
IGF-1/ -383 (C T) -1089 (C T)								1/500
IL-10/ -592 (A C) -819 (C T) -1082 (A G)	13/1039	15/1487	8/741		22/4163	1/86	8/878	23/6102
IL-12/ -1188 (A C)	9/506	9/846	6/535				1/99	8/1288
IL-12p40/ -1287 (C T)								1/220
IL-13/ -1055 (C T) -1459 (A C)								2/720
IL-15/ 3utr (A T) 5utr (C T)								1/500
IL-18/ -137 (C G) -607 (A C) -656 (G T) 1248 (A G)	1/143				2/395			1/220
IL-1alpha/ -889 (C T)	9/506	10/946	6/535			2/233	1/99	11/1640
IL-1beta/ -31 (C T) -511 (C T) 3962 (C T)	10/690	9/846	6/535		3/1574		1/99	11/1678
IL1R/ PstI 1970 (C T)	9/506	9/846	6/535				1/99	5/831

Table 7.3: Availability of cytokine gene polymorphisms by geographical region (Continued)

Gene / Polymorphism	Geographical region							
	ASIA	EEUR	MIDE	NAFR	NAME	SAFR	SCAM	WEUR
	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv	Pops/ Indiv
IL1RA/ MspAII 11100 (C T)	9/506	9/846	6/535				1/99	6/938
IL-2/ -330 (G T) 166 (G T)	10/589	11/1151	6/575		4/608	1/86	1/99	12/2170
IL-4/ -33 (C T) -34 (C T) -589 (C T) -590 (C T) -1098 (G T)	11/1241	11/1151	5/495		4/1747	2/233	1/99	12/2384
IL-4Ralpha/ 223 (A G) 1092 (A G)	9/506	9/846	6/535		3/322	2/233	1/99	7/1158
IL-6/ -174 (C G) 565 (A G)	11/1158	16/1692	7/623		23/4312	3/319	8/878	20/3355
NGF/ -198 (C T)								1/500
PDGF B/ 1135 (A C)								1/500
RANTES/ -109 (C T)								1/500
TGFbeta1/ -509 (C T) -800 (A G) codon10 (C T) codon25 (C T)	11/706	11/1064	6/543		18/2591	2/233	8/878	11/1946
TNFalpha/ -238 (A G) -308 (A G) -857 (C T) -863 (A C) -1031 (C T)	15/2087	17/1853	8/741	1/157	26/4630	3/319	8/878	25/4308
TNFbeta/ 252 (A G)	1/83	3/331	1/80	1/157	3/255	1/86		4/705
VEGF/ -7 (C T) -1001 (C G) -1154 (A G) -1455 (C T) -2578 (A C)					3/364			3/2860

Asia (ASIA), Australia (AUST), Eastern Europe (EEUR), Middle East (MIDE), North Africa (NAFR), North America (NAME), Pacific (PACI), Sub-Saharan Africa (SAFR), South and Central America (SCAM) and Western Europe (WEUR).

7.3.2.2 Diversity of cytokine gene polymorphisms among populations

To assess the variation of the different cytokine gene polymorphisms among populations across the world, an online searching mechanism was implemented in the AFND website. Figure 7.5 shows an example of the frequency search displaying twelve of the 309 records containing data for the Tumour necrosis factor (TNFalpha) gene with a polymorphism being reported at the -308 nucleotide position. In the example, the polymorphism consists of genotype combinations of Adenine (A) and Guanine (G).

From the figure it can be observed that the first four populations ordered alphabetically, presented very low frequencies ($\leq 2\%$) for the genotype AA contrary to the other two genotypes AG (9.3-26.2%) and GG (73.8-90.7%). The output results include the cytokine gene polymorphism, the name of the population, the genotype frequency (in percentages) and sample size of the population. Additionally, a reference to the Entrez-Gen database was incorporated in the output list to provide users with the ability to examine the gene sequence. Moreover, a hyperlink was added to the results to generate automatic overlaid maps based on frequency data. The frequency map tool, which was also implemented in each of the polymorphic regions available in the AFND, represents comparisons of frequencies of cytokine gene polymorphism across the world. Figure 7.6 shows an example of the variation present in the genotype AG at -308 nucleotide position of the TNFalpha gene. The figure illustrates that the highest frequencies for this genotype were mainly found in populations of Caucasian origin whereas the lowest corresponded principally to populations in North and South America.

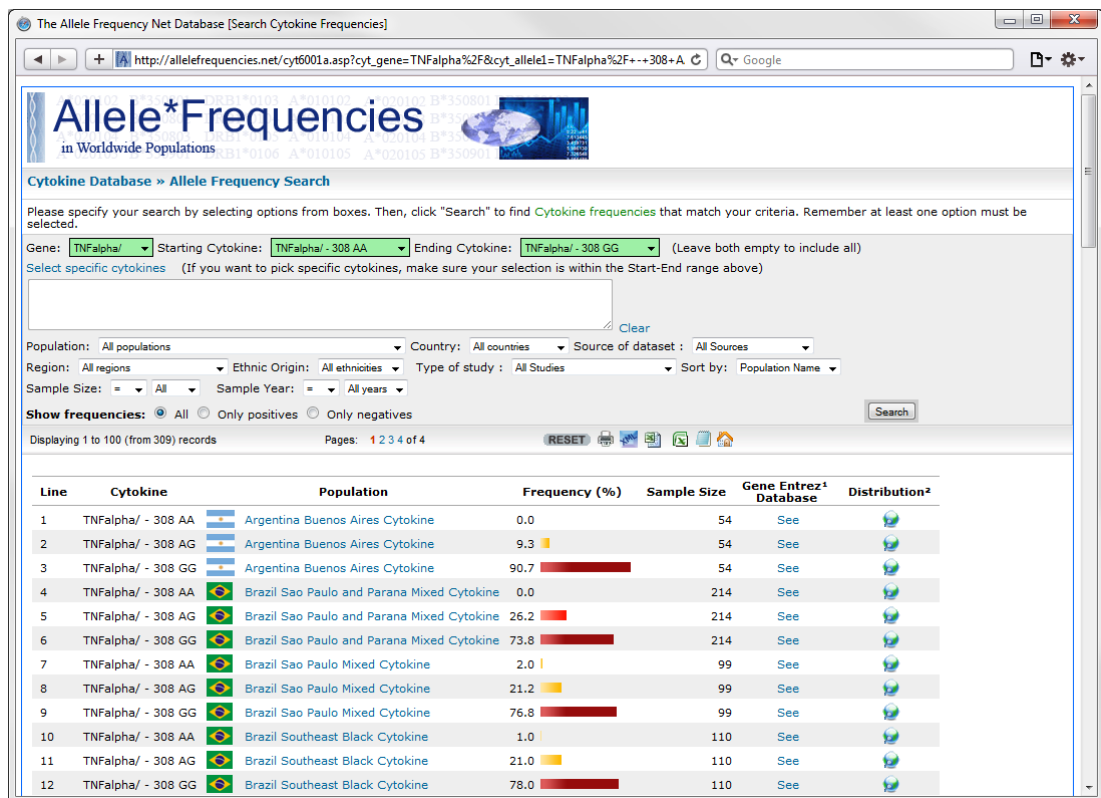


Figure 7.5: Example of the frequency search for cytokine gene polymorphisms.

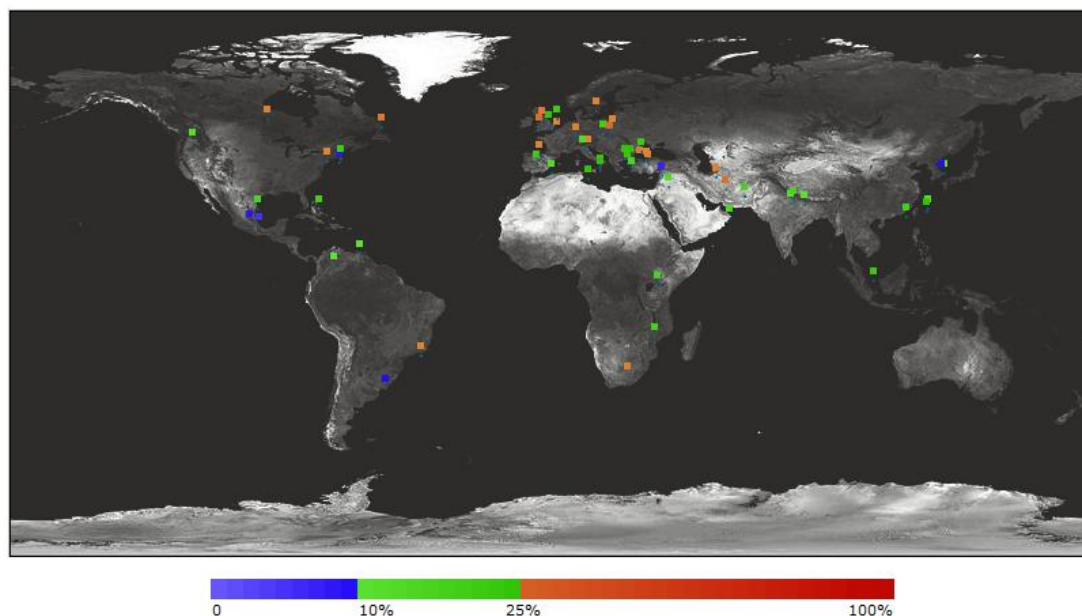


Figure 7.6: Frequency distribution of the 'TNFalpha/-308 AG' polymorphism.

7.4 Discussion

The use of computational approaches implemented in the AFND has been shown to provide a useful resource for the analysis of frequency data. This chapter described the analysis of the polymorphisms observed in two additional set of immune genes, MIC and cytokine genes. The analysis consisted in the examination of frequencies reported in the literature in different formats (allele, microsatellite and genotype). Although the polymorphism present in these two set of genes is not as extensive as HLA and KIR, the implementation of the searching mechanisms in a similar fashion allows the evaluation of a generic software framework to investigate other immune genes of interest.

In the case of MIC genes, the design of a new module to contain MIC frequency data provides researchers with a valuable electronic resource for the investigation of frequencies at allele and/or microsatellite level. A review presented by Collins and colleagues in 2004 showed the comparisons of frequencies of MICA microsatellites and official alleles in eleven population samples from six different ethnic groups reported in the literature at that time (Collins 2004). Thus, the results described in this chapter correspond to a wider and up-to-date analysis of MIC frequencies comprising data from fifty-eight populations available in the AFND.

One of the drawbacks of the old MICA microsatellite nomenclature was that exons 2-4 (encoding the extracellular domains) were not considered in the analysis of populations, limiting the examination of frequencies to allelic variability at exon 5 only. However, with the emergence of the official nomenclature for MIC alleles, further studies have reported the analysis of populations using alleles described in the IMGT/HLA database. Therefore, with the availability of sequence data, investigators are able to compare frequency data of microsatellites to the corresponding equivalences of official alleles by examining the polyalanine repeats in the protein sequence at exon 5. The searching methods used in the online tools were developed to cover both formats (microsatellite and official alleles). Additionally, data and searching options available in the AFND provides a convenient approach to investigate the presence of very rare variants. For instance, in the review by Stephens, an enquiry was raised concerning the high frequencies of the MICA*008 (A5.1), which corresponds to an 'aberrant' allele (Stephens 2001). This topic was also followed in the review by Collins who compared different population including one South American population which, interestingly, was found to present a very low frequency for MICA*008 (Collins 2004). Low frequencies in South American populations were subsequently confirmed by Oliveira (Oliveira et al. 2008). These findings can be clearly illustrated using the online tool for generating overlaid maps (http://www.allele-frequencies.net/mic6004a.asp?mic_allele=MICA*008:01). Also of interest, is that the MICA microsatellite variant A8 which was reported to present eight GTC (Alanine) repeats (Gambelunghe et al. 2006), was not found in any of the reports submitted to the AFND including the official allele counterpart (MICA*055). Originally, the A8 variant was found in one individual from a cohort of 1100 Italian subjects leading to the possibility of a wrong sequence or that more studies are needed to clarify the existence of this variation.

The number of MIC alleles has increased in the last years but not comparable to the numbers presented in HLA and KIR. The first release in the IMGT/HLA database describing MIC sequences in January 2001 comprised 51 MICA alleles. The list was updated to 73 alleles as of release 3.3.0, January 2011 (Robinson et al. 2001). Limited data had previously been reported for MICB, possibly due to the lack of an official nomenclature. MICB official alleles were initially incorporated in April 2005 containing

18 of the 31 alleles reported in release 3.3.0. However, it is expected that in the coming years more data will be available for MICB.

Due to the proximity of MICA and HLA-B within the chromosome, a number of studies have focussed on the investigation of associations of these two loci to several diseases (Stephens 2001). However, few reports with frequencies of other ethnicities have been reported in the literature. Therefore, data available in the AFND may be valuable for further analysis. Additionally, sizes of several samples in the association studies described in previous publications were small leading to the over/underestimation of 'real' values, thus, the increase of the amount of data in associations may result in better insights into the relationships between the MICA and HLA-B loci.

In the case of cytokines, the study of the gene polymorphism has been an important topic in the analysis of the mechanisms involved in the immune response. The lack of a central source to collect frequency data led to the incorporation of a module in the AFND. An online bibliographic reference published in 1999 by Bidwell provided an extraordinary resource comprising more than 1000 cytokine reviews for cytokine associations in human diseases (Bidwell et al. 1999), however the resource was discontinued in recent years. Additionally, the study of cytokine gene polymorphism frequencies was not included in that repository.

This chapter included the description of cytokine gene polymorphism frequency data from a compilation of more than one hundred populations. Different studies have been carried out to investigate the polymorphism present in populations from different continents (Delaney et al. 2004; Larcombe et al. 2005; Middleton et al. 2002). The creation of an electronic cytokine gene polymorphism warehouse may help in the analysis of variations in these genes. The Cytokine Polymorphism Frequency tool described in the chapter may provide a useful resource in the assistance of population comparisons at large scale.

In cytokine gene polymorphisms the number of citations to the website is considerably lower than in HLA and KIR (6 citations as of August 2011). Currently, this section has been principally used for referencing a given population. However, it is expected that

the addition of novel tools for the analysis of more than one hundred cytokine populations may increase the usability of this section.

Although at present there is not a standardised nomenclature for representing cytokine gene polymorphisms, a suggested nomenclature was presented in this chapter based on of the name of cytokine genes described in the Gene-Entrez database (Maglott et al. 2011). Future work will include the dissemination of a standardised notation and the incorporation of more polymorphism in cytokine genes.

7.5 Conclusions

This chapter described an outline on the frequencies of two additional set of immune genes, MIC and cytokine gene polymorphisms. Two modules were implemented in the AFND to store the information of each polymorphism and provide the scientific community with an online resource for the analysis of the occurrence of the polymorphisms in worldwide human populations. The chapter included the dissemination of the information available for MIC and cytokine genes comprising data available at various formats: microsatellite, allele and genotype level. In this chapter, an analysis was presented to examine the different MICA microsatellites and most frequent official MICA alleles from a compilation of 58 population samples. The results included in the chapter can be used as a useful resource in the ascertaining of a given allele in a particular geographical region or ethnicity. Cytokine gene polymorphism frequency data was also analysed and summarised in the chapter by classifying data available in each polymorphism. In addition, a set of online tools for the analysis and visualisation of frequency data was also described in this section.

Chapter 8

Discussion, general conclusions and future work

The aim of this chapter is to summarise the content of this thesis, extend the discussion of the analysis carried out in each chapter and describe ongoing projects and future work related to this research. Finally, general conclusions of this work are outlined at the end of this chapter.

8.1 Summary of thesis

This thesis was focussed on the development of a database to provide the scientific community with an electronic resource for the investigation of the frequencies of several immune genes such as HLA, KIR, MIC and cytokine genes across worldwide populations. The work carried out in this research can be summarised in three main components: first, the compilation of immune gene frequencies and the design of the AFND for the storage of frequency data at allele, haplotype and genotype level, second, the development of customised searching mechanisms and other software applications for the examination of data and, third, the description of several analyses that were performed to outline the extensive polymorphism found in these immune genes.

Chapter 1 provided an insight into several genes involved in the immune response and introduced the reader to the field of immunogenetics and bioinformatics as a means to investigate the genetic basis and function of these genes. The immune genes analysed in this research included the most polymorphic region in the human genome: the HLA system. Additionally, other extremely polymorphic regions such as KIR, MIC and several cytokine gene polymorphisms were also included in this study. To analyse such extraordinary polymorphisms in human populations, Chapter 2 described a database

schema developed to contain frequencies available in different formats: allele, haplotype and genotype level. In addition, Chapter 2 included a full description of the criteria and methods applied for the input and validation of datasets. Chapter 3 described a set of bioinformatics online tools that were implemented in the AFND website for the examination of data. Based on a multi-filter scheme, these tools allow flexibility in the retrieval of information performed by users. The generic model implemented in the searching mechanisms allowed the replication of the same technique for the analysis of additional genes. In Chapter 4, the analysis of the frequencies of HLA alleles and haplotypes was carried out with the use of software applications implemented in the AFND website to estimate the overall frequencies by geographical region, country or ethnic group. Despite the polymorphism present in genes of the HLA system comprising more than 6,000 alleles as of January 2011, many of these alleles were not reported in the compilation of HLA datasets. Thus, the rarity of the HLA alleles was examined in Chapter 5 by collecting data from other electronic sources such as the IMGT/HLA database, the NMDP database and from individual laboratories. Rivalling the polymorphism of HLA, the variability present in KIR genes was examined in Chapter 6. The analysis carried out in KIR genes included the examination of gene frequencies and the investigation of the organisation of KIR genes by inspecting the presence or absence of the gene in an individual's profile (KIR genotype). The latter led to the most extensive compilation of KIR genotype data in the world. Finally, in Chapter 7 other polymorphic regions such as MIC and several cytokine polymorphisms that were submitted to the AFND were also examined. The incorporation of these two additional set of genes allowed the enrichment of immune gene data and confirmed the design of the database schema as a generic model for the storage of other immune genes in the future.

8.2 Discussion

This section includes an extension of the discussion described in each chapter and includes a critical appraisal of alternative methodologies, techniques and other resources available for the investigation of similar immune genes.

8.2.1 The Allele Frequency Net Database

The vast amount of information available on the occurrence of several immune genes and their corresponding alleles and its significance in different fields of research were the motivations for the development of the AFND. The AFND comprises the most extensive electronic archive for the investigation of immune genes across worldwide human populations. The compilation of data in electronic format has provided the research community with a valuable instrument to perform meta-analysis which would not be feasible without this resource. The applications described in this thesis have increased the versatility in the examination of data. It is expected that the range of functionalities will assist expert and non-specialist users in the investigation of data. Several case studies were described in the 'Discussion' section in each chapter. One of the most recent analysis using data from AFND was a study published in *Science* by Abi-Rached and colleagues, led by Peter Parham FRS, who compared the HLA-B*73 allele between archaic and modern humans demonstrating the admixture across these species (Abi-Rached et al. 2011).

8.2.2 Alternative approaches in the design of immune gene databases

The first objective of this research was to utilise a software engineering approach in the development of a freely accessible web-based interface to assist researchers in the analysis of immune genes in an interactive manner. To facilitate the consultation of such large amount of data, multiple filter options were implemented in all searching mechanisms in the AFND. The addition of customised searching options allows users to optimise their queries and obtain the closest matching results. As reviewed in Chapter 1, several technologies are available for the design of web-based repositories. In this work, all software applications were implemented using CSS and XHTML standards to guarantee the correct visualisation in any operating system used in the client machine. In terms of the selection of the programming language, it was decided that ASP would cover the requirements for data querying and it would speed the development of the different modules. Presently, ASP has covered all the needs for the development of the current searching tools including the analysis of data at amino acid level implemented in the AFAAT tool described in Section 3.5. However, more advanced queries/analyses

may require the use of other programming languages for optimal performance of the analysis such as Java or C++. This is particularly important when the DBMS cannot cover the initial query request and a secondary process needs to be executed on the server/client machine before the display of results. Nevertheless, one of the advantages of the web-based approach employed in this research is that there is no restriction on the implementation of secondary interfaces. For instance, additional interfaces can be implemented in alternative platforms (e.g. operating system and/or programming language) and data can be accessed via a particular connection protocol. In the case of the AFND, MySQL, which has been used as the back-end, includes supporting connections for the most of the popular programming languages. Also, the use of a web service that can process the information on an external server can be used as an alternative approach. In this scenario, users of the database would send a job (i.e. data for a given analysis) to the secondary server and processed data would be sent back to the server and subsequently displayed in the client machine.

Data annotation

In recent years, one implementation that has been introduced in the design of biological repositories is the inclusion of websites called *wikis*. These websites facilitate the review process by providing online tools for edition and management of the information (Brohee, Barriot & Moreau 2010). In the case of the AFND, the use of wikis may help in the validation of current datasets listed in the website by allowing investigators to perform annotations on a specific population. As discussed in Chapter 2, at present, users rely on data available in the AFND which has normally followed a peer-reviewed process. However, on some occasions, different queries concerning data accuracy are sent to curators of the database by e-mail. Thus, it is important to allow users the incorporation of queries, annotations or the inclusion of supplementary material which may help in the evaluation of the current datasets and serve others in the interpretation of their data and/or analyses performed.

8.2.3 Use of relational models and DBMS

The relational model approach used in the implementation of the AFND has allowed the management and query of data in an easy and interactive way. The simplicity and robustness of this methodology has improved significantly the management of datasets. The use of MySQL as the relational DBMS has various advantages, first, the flexibility of this software to interact with other design tools such as MySQL Workbench and PowerDesigner in the generation of the database schemas and the definition of rules and constraints of the data, second, the excellent performance in the implementation of simple and complex queries and, third, the distribution of this software as a cross-platform application released under the GNU General Public License.

8.2.4 Archive of raw data

One of the characteristics that distinguish AFND from other immune gene frequency databases (dbMHC and ALFRED) is the ability to allow users the submission of their own data. To do this, an online submission form for data input was created. However, less than ten percent of data available in the AFND has been entered by users only, possibly due to the fact that users do not directly receive any credit for their submission, as they would for a published paper. Therefore, future plans in the AFND include the design of an online journal for the compilation of raw data which will serve to two purposes: first, it should increase the number of individuals submitting their own data and, second, the incorporation of raw data will increase the types of analysis that can be performed in the AFND. Due to the lack of raw data, at present, it is not possible to ascertain how individuals analysed their data, which protocols they used in the publication of their results, the level of resolution employed in the calculation of frequencies, etc. Thus, the addition of raw data is expected to provide better understanding of how frequency data has been compiled.

8.2.5 Standards in immune gene frequencies

The use of controlled vocabulary and definition of standards for data exchange plays an important role in the development of biological repositories. One example of the application of data standards has been in the field of proteomics performed by the Human Proteome Organisation via the Proteomics Standards Initiative which is in charge of the definition of standards to facilitate the exchange, comparison and analysis of proteomics data (Orchard, Hermjakob & Apweiler 2003). In the case of HLA, KIR, MIC and cytokine gene polymorphisms the definition of standards is in the early stages. Several attempts to produce standard formats have been under development in recent years. For instance, the NMDP bioinformatics group has designed a XML-based prototype called HML (Histoimmunogenetics Markup Language) to report HLA testing results. However, the HML project has not had recent modifications and the version 0.3, which is the most recent version available, has not been distributed or recognised in the field of immunogenetics. Recently, two international groups, the HLA European Network (HLA-NET) and the Immunogenomics Data-Analysis Working Group (IDAWG), have initiated the definition of standards and minimal data recommended in the reporting of HLA data. Due to the AFND comprising the largest repository of HLA data, this research work has been in continuous interaction with these initiatives via the formulation of examples of the heterogeneity of data reported by investigators.

8.3 Future work

There are a number of ongoing and future projects that will be implemented in the AFND. The addition of these modules is expected to increase the usefulness of the website and may help investigators better understand the function of the immune genes.

8.3.1 Quality control

Quality control is perhaps the most important concern in the collection of immune gene frequency data. To date, there has been a lack of standardisation in recording HLA, KIR, MIC and cytokine gene polymorphisms results. For instance, in HLA, data is often

produced from individual laboratories using different methodologies which do not necessarily detect the same alleles and/or the same resolution, and individuals make different non-documented assumptions to record typing data. Thus, analysis of data may be prone to errors in the absence of a standardised approach for documenting data. Future implementations include the collection of raw data and a software pipeline to verify allele frequency calculations and request a standard set of metadata to improve consistency for analyses. Datasets will be validated automatically to ensure that the information submitted comply with the minimal standard guidelines. To define the required minimal metadata, the AFND project will actively participate with international organisations, such as the HLA-NET the IDAWG groups for the standardisation and validation of frequency datasets, controlled vocabulary (e.g. ethnic origins, geographical regions, etc.), database schemas and data exchange.

8.3.2 Software toolkit for data sharing and complementary analysis

To extend the set of tools provided in the AFND, future developments include the incorporation of statistical routines to allow users to apply advanced analyses for all populations available. This will be done by the integration of AFND output formats into external programs such as PyPop (Lancaster et al. 2003), GENEPOP (Rousset 2008) and Arlequin (Excoffier, Laval & Schneider 2005) for test of Hardy-Weinberg and neutrality, and PHYLIP (Retief 2000) for the display of dendrograms and phylogenetic trees. In the context of data sharing, future developments will include data portability in XML format. This will be done by the implementation of a standardised XML layer and definition of controlled vocabularies. Additionally, a minimum reporting requirements document will be also generated to ensure that all studies deposited in AFND are described consistently.

8.3.3 KIR and HLA ligands

Ongoing developments include the implementation of the display of KIR gene frequencies correlated with their HLA ligand frequencies as a continuation of the KIR anthropology component of the 15th IHWS (Hollenbach et al. 2010) and a section for the collection of KIR and disease studies. It is believed that the addition of these

software applications on the website will be invaluable to researchers interested in the investigation of KIR genes.

8.4 General conclusions

The Allele Frequency Net Database has provided an important resource for the histocompatibility and immunogenetics community. The design of the database to store immune gene frequencies in different populations has moved the analysis of immune genes from publication-driven into digital resources. The development of novel mechanisms of querying and the incorporation of new polymorphisms have enriched the examination of the genes available. Acting as a primary source, the AFND contains the most extensive archive of immune gene frequencies. As of August 2011, 168 peer-reviewed publications have cited the website according to the records in the Web of Knowledge (<http://apps.webofknowledge.com/>). From 4 August 2010 to 4 August 2011, the website was accessed by 15,784 distinct users from 2,758 cities in 136 countries, demonstrating the large international user base. To date, the users are spread across domains in clinical, biomedical and biological research and future developments should further increase the international impact, especially for fundamental immunology research, human genetics and groups working to understand host responses to infectious disease.

The nature of databases involves continuous maintenance of datasets and implementations of new modules for the coverage of new methods and techniques which are constantly published by researchers. Perhaps, the problem with databases such as AFND is that they are usually the work of one individual or one group and the success of the databases on many occasions depends on the interest of Research Councils to fund the costs. However, it is expected that the design and structure presented on this work may serve a wide community interested in the analyses of these genes for more years. Full details of searches and datasets can be freely explored by accessing the www.allelefrequencies.net website.

Bibliography

- 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. & Parham, P. (2011) 'The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans', *Science*.
- Achard, F., Vaysseix, G. & Barillot, E. (2001) 'XML, bioinformatics and data integration', *Bioinformatics*, vol. 17, no. 2, pp. 115-125.
- Aihara, M. (2011) 'Pharmacogenetics of cutaneous adverse drug reactions', *Journal of Dermatology*, vol. 38, no. 3, pp. 246-254.
- Alper, C.A. & Larsen, C.E. (2004) 'Immunogenetics: Human immunogenetics', *Current Opinion in Immunology*, vol. 16, no. 5, pp. 623-625.
- Apweiler, R. & Consortium, U. (2010) 'The Universal Protein Resource (UniProt) in 2010', *Nucleic Acids Research*, vol. 38, pp. D142-D148.
- Arnaiz-Villena, A., Siles, N., Moscoso, J., Zamora, J., Serrano-Vela, J.I., Gomez-Casado, E., Castro, M.J. & Martinez-Laso, J. (2005) 'Origin of Aymaras from Bolivia and their relationship with other Amerindians according to HLA genes', *Tissue Antigens*, vol. 65, no. 4, pp. 379-390.
- Bahram, S. (2000) 'MIC genes: from genetics to biology', *Adv Immunol*, vol. 76, pp. 1-60.
- Balding, D.J. (2006) 'A tutorial on statistical methods for population association studies', *Nat Rev Genet*, vol. 7, no. 10, pp. 781-791.
- Bartlett, J.M. & Stirling, D. (2003) 'A short history of the polymerase chain reaction', *Methods Mol Biol*, vol. 226, pp. 3-6.
- Beaver, J.E., Bourne, P.E. & Ponomarenko, J.V. (2007) 'EpitopeViewer: a Java application for the visualization and analysis of immune epitopes in the Immune Epitope Database and Analysis Resource (IEDB)', *Immunome Res*, vol. 3, p. 3.
- Begovich, A.B., McClure, G.R., Suraj, V.C., Helmuth, R.C., Fildes, N., Bugawan, T.L., Erlich, H.A. & Klitz, W. (1992) 'Polymorphism, recombination, and linkage disequilibrium within the HLA class II region', *J Immunol*, vol. 148, no. 1, pp. 249-258.
- 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. & Klitz, W. (2001) 'Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations', *Tissue Antigens*, vol. 57, no. 5, pp. 424-439.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2011) 'GenBank', *Nucleic Acids Research*, vol. 39, pp. D32-D37.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2008) 'GenBank', *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D25-30.
- Bidwell, J., Keen, L., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J., Pociot, F., Hardt, C. & D'Alfonso, S. (1999) 'Cytokine gene polymorphism in human disease: on-line databases', *Genes Immun*, vol. 1, no. 1, pp. 3-19.

- Bidwell, J., Keen, L., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J., Pociot, F., Hardt, C. & D'Alfonso, S. (2001) 'Cytokine gene polymorphism in human disease: on-line databases, supplement 1', *Genes Immun*, vol. 2, no. 2, pp. 61-70.
- Birney, E. & Clamp, M. (2004) 'Biological database design and implementation', *Brief Bioinform*, vol. 5, no. 1, pp. 31-38.
- Blackwell, J.M., Jamieson, S.E. & Burgner, D. (2009) 'HLA and infectious diseases', *Clin Microbiol Rev*, vol. 22, no. 2, pp. 370-385.
- Bluestone, J.A., Herold, K. & Eisenbarth, G. (2010) 'Genetics, pathogenesis and clinical interventions in type 1 diabetes', *Nature*, vol. 464, no. 7293, pp. 1293-1300.
- Bontrop, R.E., Kasahara, M. & Watkins, D.I. (1999) 'Immunogenetics: changes and challenges', *Immunogenetics*, vol. 50, no. 5, pp. 243-243.
- Bornberg-Bauer, E. & Paton, N.W. (2002) 'Conceptual data modelling for bioinformatics', *Brief Bioinform*, vol. 3, no. 2, pp. 166-180.
- Brohee, S., Barriot, R. & Moreau, Y. (2010) 'Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases', *Bioinformatics*, vol. 26, no. 17, pp. 2210-2211.
- Buus, S., Lauemoller, S.L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A. & Brunak, S. (2003) 'Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach', *Tissue Antigens*, vol. 62, no. 5, pp. 378-384.
- Cano, P., Klitz, W., Mack, S.J., Maiers, M., Marsh, S.G.E., Noreen, H., Reed, E.F., Senitzer, D., Setterholm, M., Smith, A. & Fernandez-Vina, M. (2007) 'Common and Well-Documented HLA Alleles: Report of the Ad-Hoc Committee of the American Society for Histocompatibility and Immunogenetics', *Human Immunology*, vol. 68, no. 5, pp. 392-417.
- Carrington, M., Wang, S., Martin, M.P., Gao, X.J., Schiffman, M., Cheng, J., Herrero, R., Rodriguez, A.C., Kurman, R., Mortel, R., Schwartz, P., Glass, A. & Hildesheim, A. (2005) 'Hierarchy of resistance to cervical neoplasia mediated by combinations of killer immunoglobulin-like receptor and human leukocyte antigen loci', *Journal of Experimental Medicine*, vol. 201, no. 7, pp. 1069-1075.
- Cerna, M., Falco, M., Friedman, H., Raimondi, E., Maccagno, A., Fernandezvina, M. & Stastny, P. (1993) 'Differences in Hla Class-II Alleles of Isolated South-American Indian Populations from Brazil and Argentina', *Human Immunology*, vol. 37, no. 4, pp. 213-220.
- Chen, P.L., Fann, C.S.J., Chu, C.C., Chang, C.C., Chang, S.W., Hsieh, H.Y., Lin, M., Yang, W.S. & Chang, T.C. (2011) 'Comprehensive Genotyping in Two Homogeneous Graves' Disease Samples Reveals Major and Novel HLA Association Alleles', *Plos One*, vol. 6, no. 1.
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. (2003) 'Multiple sequence alignment with the Clustal series of programs', *Nucleic Acids Res*, vol. 31, no. 13, pp. 3497-3500.
- Cheung, K.H., Osier, M.V., Kidd, J.R., Pakstis, A.J., Miller, P.L. & Kidd, K.K. (2000) 'ALFRED: an allele frequency database for diverse populations and DNA polymorphisms', *Nucleic Acids Res*, vol. 28, no. 1, pp. 361-363.
- Chinen, J. & Buckley, R.H. (2010) 'Transplantation immunology: solid organ and bone marrow', *J Allergy Clin Immunol*, vol. 125, no. 2 Suppl 2, pp. S324-335.
- Chu, C.C., Lee, K.C., Lee, H.L., Lai, S.K. & Lin, M. (2010) 'Identification of the novel allele HLA-B*13:01:03 by sequence-based typing method in a Taiwanese individual', *Tissue Antigens*, vol. 76, no. 6, pp. 496-497.

- Collins, R.W. (2004) 'Human MHC class I chain related (MIC) genes: their biological function and relevance to disease and transplantation', *Eur J Immunogenet*, vol. 31, no. 3, pp. 105-114.
- Conesa, A., Fernandez-Mestre, M., Padron, D., Toro, F., Silva, N., Tassinari, P., Blanca, I., Martin, M.P., Carrington, M. & Layrisse, Z. (2010) 'Distribution of killer cell immunoglobulin-like receptor genes in the mestizo population from Venezuela', *Tissue Antigens*, vol. 75, no. 6, pp. 724-729.
- Constantinescu, I., Nedelcu, F.D., Toader, M.A., Vasile, D., Zaharia, M., Harza, M. & Sinescu, I. (2006) 'Evaluation of KIR genotypes and cytokine gene polymorphism in Romanian kidney transplant recipients: impact on acute and chronic allograft rejection', *Tissue Antigens*, vol. 67, no. 6, pp. 505-505.
- Cook, M.A., Moss, P.A.H. & Briggs, D.C. (2003) 'The distribution of 13 killer-cell immunoglobulin-like receptor loci in UK blood donors from three ethnic groups', *European Journal of Immunogenetics*, vol. 30, no. 3, pp. 213-221.
- Crawford, D.C. & Nickerson, D.A. (2005) 'Definition and clinical importance of haplotypes', *Annu Rev Med*, vol. 56, pp. 303-320.
- Delaney, N.L., Esquenazi, V., Lucas, D.P., Zachary, A.A. & Leffell, M.S. (2004) 'TNF-alpha, TGF-beta, IL-10, IL-6, and INF-gamma alleles among African Americans and Cuban Americans. Report of the ASHI Minority Workshops: Part IV', *Human Immunology*, vol. 65, no. 12, pp. 1413-1419.
- Delves, P.J. & Roitt, I.M. (2000a) 'The immune system. First of two parts', *N Engl J Med*, vol. 343, no. 1, pp. 37-49.
- Delves, P.J. & Roitt, I.M. (2000b) 'The immune system. Second of two parts', *N Engl J Med*, vol. 343, no. 2, pp. 108-117.
- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) 'Maximum likelihood estimation from incomplete data via the em algorithm', *J Royal Stat Soc*, vol. 39, pp. 1-38.
- Denis, L., Sivula, J., Gourraud, P.A., Kerdudou, N., Chout, R., Ricard, C., Moisan, J.P., Gagne, K., Partanen, J. & Bignon, J.D. (2005) 'Genetic diversity of KIR natural killer cell markers in populations from france, guadeloupe, finland, senegal and reunion', *Tissue Antigens*, vol. 66, no. 4, pp. 267-276.
- Dinarello, C.A. (2007) 'Historical insights into cytokines', *European Journal of Immunology*, vol. 37, pp. S34-S45.
- Djulejic, E., Petlichkovski, A., Trajkov, D., Hristomanova, S., Middleton, D. & Spiroski, M. (2010) 'Distribution of killer cell immunoglobulinlike receptors in the Macedonian population', *Human Immunology*, vol. 71, no. 3, pp. 281-288.
- Donaldson, P.T., Daly, A.K., Henderson, J., Graham, J., Pirmohamed, M., Bernal, W., Day, C.P. & Aithal, G.P. (2010) 'Human leucocyte antigen class II genotype in susceptibility and resistance to co-amoxiclav-induced liver injury', *Journal of Hepatology*, vol. 53, no. 6, pp. 1049-1053.
- Dudley, J.T. & Butte, A.J. (2009) 'A quick guide for developing effective bioinformatics programming skills', *PLoS Comput Biol*, vol. 5, no. 12, p. e1000589.
- Duquesnoy, R.J. (2002) 'HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. I. Description of the algorithm', *Hum Immunol*, vol. 63, no. 5, pp. 339-352.
- Duquesnoy, R.J. & Askar, M. (2007) 'HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. V. Eplet matching for HLA-DR, HLA-DQ, and HLA-DP', *Hum Immunol*, vol. 68, no. 1, pp. 12-25.
- Duquesnoy, R.J., Howe, J. & Takemoto, S. (2003) 'HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. IV. An alternative strategy to increase the number of compatible donors for highly sensitized patients', *Transplantation*, vol. 75, no. 6, pp. 889-897.

- Duquesnoy, R.J. & Marrari, M. (2002) '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', *Hum Immunol*, vol. 63, no. 5, pp. 353-363.
- Duquesnoy, R.J., Takemoto, S., de Lange, P., Doxiadis, II, Schreuder, G.M., Persijn, G.G. & Claas, F.H. (2003) 'HLAmatchmaker: a molecularly based algorithm for histocompatibility determination. III. Effect of matching at the HLA-A,B amino acid triplet level on kidney transplant survival', *Transplantation*, vol. 75, no. 6, pp. 884-889.
- Eberhard, H.P., Feldmann, U., Bochtler, W., Baier, D., Rutt, C., Schmidt, A.H. & Muller, C.R. (2010) 'Estimating unbiased haplotype frequencies from stem cell donor samples typed at heterogeneous resolutions: a practical study based on over 1 million German donors', *Tissue Antigens*, vol. 76, no. 5, pp. 352-361.
- Ellis, J.M., Henson, V., Slack, R., Ng, J., Hartzman, R.J. & Katovich Hurley, C. (2000) 'Frequencies of HLA-A2 alleles in five U.S. population groups. Predominance Of A*02011 and identification of HLA-A*0231', *Hum Immunol*, vol. 61, no. 3, pp. 334-340.
- Erlich, H.A., Opelz, G. & Hansen, J. (2001) 'HLA DNA typing and transplantation', *Immunity*, vol. 14, no. 4, pp. 347-356.
- Erlich, R.L., Jia, X., Anderson, S., Banks, E., Gao, X., Carrington, M., Gupta, N., DePristo, M.A., Henn, M.R., Lennon, N.J. & de Bakker, P.I. (2011) 'Next-generation sequencing for HLA typing of class I loci', *BMC Genomics*, vol. 12, p. 42.
- Evseeva, I., Nicodemus, K.K., Bonilla, C., Tonks, S. & Bodmer, W.F. (2010) 'Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe', *Eur J Hum Genet*, vol. 18, no. 8, pp. 924-932.
- Ewerton, P.D., Leite, M.D., Magalhaes, M., Sena, L. & dos Santos, E.J.M. (2007) 'Amazonian Amerindians exhibit high variability of KIR profiles', *Immunogenetics*, vol. 59, no. 8, pp. 625-630.
- Excoffier, L., Laval, G. & Schneider, S. (2005) 'Arlequin (version 3.0): an integrated software package for population genetics data analysis', *Evol Bioinform Online*, vol. 1, pp. 47-50.
- Excoffier, L. & Slatkin, M. (1995) 'Maximum-Likelihood-Estimation of Molecular Haplotype Frequencies in a Diploid Population', *Molecular Biology and Evolution*, vol. 12, no. 5, pp. 921-927.
- Fialho, R.N., Martins, L., Pinheiro, J.P., Bettencourt, B.F., Couto, A.R., Santos, M.R., Peixoto, M.J., Garrett, F., Leal, J., Tomas, A.M. & Bruges-Armas, J. (2009) 'Role of human leukocyte antigen, killer-cell immunoglobulin-like receptors, and cytokine gene polymorphisms in leptospirosis', *Human Immunology*, vol. 70, no. 11, pp. 915-920.
- Fourment, M. & Gillings, M.R. (2008) 'A comparison of common programming languages used in bioinformatics', *BMC Bioinformatics*, vol. 9, pp. -.
- Frassati, C., Touinssi, M., Picard, C., Segura, M., Galicher, V., Papa, K., Gagne, K., Vivier, E., Degioanni, A., Botsch, G., Mercier, P., Vely, F., de Micco, P., Reviron, D. & Chiaroni, J. (2006) 'Distribution of killer-cell immunoglobulin-like receptor (KIR) in Comoros and Southeast France', *Tissue Antigens*, vol. 67, no. 5, pp. 356-367.
- Galperin, M.Y. & Cochrane, G.R. (2011) 'The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection', *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D1-6.

- Gambelunghe, G., Brozzetti, A.L., Ghaderi, M., Tortoioli, C. & Falorni, A. (2006) 'MICA A8: a new allele within MHC class I chain-related a transmembrane region with eight GCT repeats', *Human Immunology*, vol. 67, no. 12, pp. 1005-1007.
- Gendzekhadze, K., Norman, P.J., Abi-Rached, L., Layrisse, Z. & Parham, P. (2006) 'High KIR diversity in Amerindians is maintained using few gene-content haplotypes', *Immunogenetics*, vol. 58, no. 5-6, pp. 474-480.
- Good, P.I. (2005) Estimating Population Parameters, in P.I. Good (ed.), *Resampling Methods. A practical guide to data analysis*, 3rd Edition edn, Birkhäuser, Boston, USA, pp. 1-4.
- Gourraud, P.A., Meenagh, A., Cambon-Thomsen, A. & Middleton, D. (2010) 'Linkage disequilibrium organization of the human KIR superlocus: implications for KIR data analyses', *Immunogenetics*, vol. 62, no. 11-12, pp. 729-740.
- Guinan, K.J., Cunningham, R.T., Meenagh, A., Dring, M.M., Middleton, D. & Gardiner, C.M. (2010) 'Receptor systems controlling natural killer cell function are genetically stratified in Europe', *Genes Immun*, vol. 11, no. 1, pp. 67-78.
- Gutierrez-Rodriguez, M.E., Sandoval-Ramirez, L., Diaz-Flores, M., Marsh, S.G.E., Valladares-Salgado, A., Madrigal, J.A., Mejia-Arangure, J.M., Garcia, C.A., Huerta-Zepeda, A., Ibarra-Cortes, B., Ortega-Camarillo, C. & Cruz, M. (2006) 'KIR gene in ethnic and mestizo populations from Mexico', *Human Immunology*, vol. 67, no. 1-2, pp. 85-93.
- Hattersley, A.T. & McCarthy, M.I. (2005) 'What makes a good genetic association study?', *The Lancet*, vol. 366, no. 9493, pp. 1315-1323.
- Haukim, N., Bidwell, J.L., Smith, A.J., Keen, L.J., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J., Pociot, F., Hardt, C. & D'Alfonso, S. (2002) 'Cytokine gene polymorphism in human disease: on-line databases, supplement 2', *Genes Immun*, vol. 3, no. 6, pp. 313-330.
- Hellerstein, J.M., Stonebraker, M. & Hamilton, J. (2007) 'Architecture of a Database System', *Found. Trends databases*, vol. 1, no. 2, pp. 141-259.
- Helmberg, W., Dunivin, R. & Feolo, M. (2004) 'The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences', *Nucleic Acids Res*, vol. 32, no. Web Server issue, pp. W173-175.
- Helmberg, W. & Feolo, M. (2007) Anthropology/human genetic diversity population reports, in J.A. Hansen (ed.), *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference*, vol. 1, IHWG Press, Seattle, WA, pp. 502-509.
- Helmberg, W.M.C., Feolo, M.L., Hoffman, D.J. & Mack, S. (2004) 'HLA typing data of the IHWG anthropology project at DBMHC', *Genes and Immunity*, vol. 5, pp. S55-S55.
- Hill, A.V., Allsopp, C.E., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J. & Greenwood, B.M. (1991) 'Common west African HLA antigens are associated with protection from severe malaria', *Nature*, vol. 352, no. 6336, pp. 595-600.
- Hogeweg, P. (1978) 'Simulating Growth of Cellular Forms', *Simulation*, vol. 31, no. 3, pp. 90-96.
- Hollegaard, M.V. & Bidwell, J.L. (2006) 'Cytokine gene polymorphism in human disease: on-line databases, Supplement 3', *Genes Immun*, vol. 7, no. 4, pp. 269-276.
- Hollenbach, J.A., Meenagh, A., Sleator, C., Alaez, C., Bengoche, M., Canossi, A., Contreras, G., Creary, L., Evseeva, I., Gorodezky, C., Hardie, R.A., Karlsen, T.H., Lie, B., Luo, M., Martinetti, M., Navarette, C., de Oliveira, D.C., Ozzella, G., Pasi, A., Pavlova, E., Pinto, S., Porto, L.C., Santos, P., Slavcev, A., Srinak,

- D., Tavoularis, S., Tonks, S., Trachtenberg, E., Vejbaesya, S. & Middleton, D. (2010) 'Report from the killer immunoglobulin-like receptor (KIR) anthropology component of the 15th International Histocompatibility Workshop: worldwide variation in the KIR loci and further evidence for the co-evolution of KIR and HLA', *Tissue Antigens*, vol. 76, no. 1, pp. 9-17.
- Hollenbach, J.A., Thomson, G., Cao, K., Fernandez-Vina, M., Erlich, H.A., Bugawan, T.L., Winkler, C., Winter, M. & Klitz, W. (2001) 'HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives', *Hum Immunol*, vol. 62, no. 4, pp. 378-390.
- Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., Buus, S. & Nielsen, M. (2009) 'NetMHCpan, a method for MHC class I binding prediction beyond humans', *Immunogenetics*, vol. 61, no. 1, pp. 1-13.
- Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R., Almeida, J., Forbes, S., Gilbert, J., Halls, K., Harrow, J., Hart, E., Howe, K., Jackson, D., Palmer, S., Roberts, A., Sims, S., Stewart, C., Traherne, J., Trevanion, S., Wilming, L., Rogers, J., de Jong, P., Elliott, J., Sawcer, S., Todd, J., Trowsdale, J. & Beck, S. (2008) 'Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project', *Immunogenetics*, vol. 60, no. 1, pp. 1-18.
- Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Jr., Wright, M.W., Wain, H.M., Trowsdale, J., Ziegler, A. & Beck, S. (2004) 'Gene map of the extended human MHC', *Nat Rev Genet*, vol. 5, no. 12, pp. 889-899.
- Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. & Rhee, S.Y. (2008) 'Big data: The future of biocuration', *Nature*, vol. 455, no. 7209, pp. 47-50.
- Hsu, K.C., Chida, S., Geraghty, D.E. & Dupont, B. (2002a) 'The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism', *Immunol Rev*, vol. 190, pp. 40-52.
- Hsu, K.C., Liu, X.R., Selvakumar, A., Mickelson, E., O'Reilly, R.J. & Dupont, B. (2002b) 'Killer Ig-like receptor haplotype analysis by gene content: evidence for genomic diversity with a minimum of six basic framework haplotypes, each with multiple subsets', *J Immunol*, vol. 169, no. 9, pp. 5118-5129.
- Hviid, T.V.F. (2006) 'HLA-G in human reproduction: aspects of genetics, function and pregnancy complications', *Human Reproduction Update*, vol. 12, no. 3, pp. 209-232.
- Irwin, M.R. (1976) 'The beginnings of immunogenetics', *Immunogenetics*, vol. 3, no. 1, pp. 1-13.
- Jiang, K., Zhu, F.M., Lv, Q.F. & Yan, L.X. (2005) 'Distribution of killer cell immunoglobulin-like receptor genes in the Chinese Han population', *Tissue Antigens*, vol. 65, no. 6, pp. 556-563.
- Joyce, M.G. & Sun, P.D. (2011) 'The Structural Basis of Ligand Recognition by Natural Killer Cell Receptors', *Journal of Biomedicine and Biotechnology*.
- Jumnainsong, A., Jearanaikoon, P., Khahmahpahte, S., Wongsena, W., Romphruk, A.V., Chumworathayi, B., Vaeteewoottacharn, K., Ponglikitmongkol, M., Romphruk, A. & Leelayuwat, C. (2008) 'Associations of MICB with cervical cancer in north-eastern Thais: identification of major histocompatibility complex class I chain-related gene B motifs influencing natural killer cell activation', *Clinical and Experimental Immunology*, vol. 153, no. 2, pp. 205-213.
- Kanterakis, S., Magira, E., Rosenman, K.D., Rossman, M., Talsania, K. & Monos, D.S. (2008) 'SKDM human leukocyte antigen (HLA) tool: A comprehensive HLA and disease associations analysis software', *Hum Immunol*, vol. 69, no. 8, pp. 522-525.

- Kinnersley, B. (2011) *The Language List* [Online], Available from: <http://people.ku.edu/~nkinners/LangList/Extras/langlist.htm> (Accessed: 01/05/2011).
- Klein, J. & Sato, A. (2000a) 'The HLA system. First of two parts', *N Engl J Med*, vol. 343, no. 10, pp. 702-709.
- Klein, J. & Sato, A. (2000b) 'The HLA system. Second of two parts', *N Engl J Med*, vol. 343, no. 11, pp. 782-786.
- Kollman, C., Maiers, M., Gragert, L., Muller, C., Setterholm, M., Oudshoorn, M. & Hurley, C.K. (2007) 'Estimation of HLA-A, -B, -DRB1 haplotype frequencies using mixed resolution data from a national registry with selective retyping of volunteers', *Human Immunology*, vol. 68, no. 12, pp. 950-958.
- Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pilar, M., Pastor, G., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W.M. & Apweiler, R. (2007) 'EMBL Nucleotide Sequence Database in 2006', *Nucleic Acids Research*, vol. 35, pp. D16-D20.
- Kumar, S., Nei, M., Dudley, J. & Tamura, K. (2008) 'MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences', *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 299-306.
- Kumar, V. & McNerney, M.E. (2005) 'A new self: MHC-class-I-independent natural-killer-cell self-tolerance', *Nat Rev Immunol*, vol. 5, no. 5, pp. 363-374.
- Lancaster, A., Nelson, M.P., Meyer, D., Thomson, G. & Single, R.M. (2003) 'PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data', *Pac Symp Biocomput*, pp. 514-525.
- Lancaster, A.K., Single, R.M., Solberg, O.D., Nelson, M.P. & Thomson, G. (2007) 'PyPop update--a software pipeline for large-scale multilocus population genomics', *Tissue Antigens*, vol. 69 Suppl 1, pp. 192-197.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaanty, K.D., Miner, T.L., Delehaanty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E.,

- Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. & Chen, Y.J. (2001) 'Initial sequencing and analysis of the human genome', *Nature*, vol. 409, no. 6822, pp. 860-921.
- Larcombe, L., Rempel, J.D., Dembinski, I., Tinckam, K., Rigatto, C. & Nickerson, P. (2005) 'Differential cytokine genotype frequencies among Canadian Aboriginal and Caucasian populations', *Genes Immun*, vol. 6, no. 2, pp. 140-144.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, vol. 23, no. 21, pp. 2947-2948.
- Lefranc, M.P. (2004) 'IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics', *Mol Immunol*, vol. 40, no. 10, pp. 647-660.
- Listgarten, J., Brumme, Z., Kadie, C., Xiaojiang, G., Walker, B., Carrington, M., Goulder, P. & Heckerman, D. (2008) 'Statistical resolution of ambiguous HLA typing data', *PLoS Comput Biol*, vol. 4, no. 2, p. e1000016.
- Little, A.M. (2007) 'An overview of HLA typing for hematopoietic stem cell transplantation', *Methods Mol Med*, vol. 134, pp. 35-49.
- Little, A.M. & Parham, P. (1999) 'Polymorphism and evolution of HLA class I and II genes and molecules', *Rev Immunogenet*, vol. 1, no. 1, pp. 105-123.
- Lucas, D., Campillo, J.A., Lopez-Hernandez, R., Martinez-Garcia, P., Lopez-Sanchez, M., Botella, C., Salgado, G., Minguela, A., Alvarez-Lopez, M.R. & Muro, M. (2008) 'Allelic diversity of MICA gene and MICA/HLA-B haplotypic variation in a population of the Murcia region in southeastern Spain', *Human Immunology*, vol. 69, no. 10, pp. 655-660.
- Mack, S.J., Erlich, H.A., Feolo, M., Fernandez-Vina, M., Gourraud, P.A., Helmberg, W., Kanga, U., Kupatawintu, P., Lancaster, A., Maiers, M., Maldonado-Torres, H., Marsh, S.G.E., Meyer, D., Middleton, D., Mueller, C.R., Nathalang, O., Park, M.H., Single, R.M., Tait, B., Thomson, G., Varney, M. & Hollenbach, J. (2009) 'Idawg - the Immunogenomic Data-Analysis Working Group.', *Human Immunology*, vol. 70, pp. S86-S86.

- Mack, S.J., Sanchez-Mazas, A., Single, R.M., Meyer, D., Hill, J., Dron, H.A., Jani, A.J., Thomson, G. & Erlich, H.A. (2007) 'Population samples and genotyping technology', *Tissue Antigens*, vol. 69, pp. 188-191.
- Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. (2011) 'Entrez Gene: gene-centered information at NCBI', *Nucleic Acids Research*, vol. 39, pp. D52-D57.
- Maier, C., Long, J., Hemminger, B. & Giddings, M. (2009) 'Ultra-Structure database design methodology for managing systems biology data and analyses', *Bmc Bioinformatics*, vol. 10, no. 1, p. 254.
- Majorczyk, E., Pawlik, A., Luszczek, W., Nowak, I., Wisniewski, A., Jasek, M. & Kusnierczyk, P. (2007) 'Associations of killer cell immunoglobulin-like receptor genes with complications of rheumatoid arthritis', *Genes and Immunity*, vol. 8, no. 8, pp. 678-683.
- Marsh, S.G., Parham, P., Dupont, B., Geraghty, D.E., Trowsdale, J., Middleton, D., Vilches, C., Carrington, M., Witt, C., Guethlein, L.A., Shilling, H., Garcia, C.A., Hsu, K.C. & Wain, H. (2003) 'Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002', *Hum Immunol*, vol. 64, no. 6, pp. 648-654.
- Marsh, S.G.E., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Dupont, B., Erlich, H.A., Fernández-Viña, M., Geraghty, D.E., Holdsworth, R., Hurley, C.K., Lau, M., Lee, K.W., Mach, B., Maiers, M., Mayr, W.R., Müller, C.R., Parham, P., Petersdorf, E.W., Sasazuki, T., Strominger, J.L., Svejgaard, A., Terasaki, P.I., Tiercy, J.M. & Trowsdale, J. (2010) 'Nomenclature for factors of the HLA system, 2010', *Tissue Antigens*, vol. 75, no. 4, pp. 291-455.
- Martinez-Laso, J., Herraiz, M.A., Vidart, J.A., Penaloza, J., Barbolla, M.L., Jurado, M.L. & Cervera, I. (2011) 'Polymorphism of the HLA-B*15 group of alleles is generated following 5 lineages of evolution', *Human Immunology*, vol. 72, no. 5, pp. 412-421.
- McCluskey, J., Kanaan, C. & Diviney, M. (2003) 'Nomenclature and serology of HLA class I and class II alleles', *Curr Protoc Immunol*, vol. Appendix 1, p. Appendix 1S.
- McQueen, K.L., Dorigi, K.M., Guethlein, L.A., Wong, R., Sanjanwala, B. & Parham, P. (2007) 'Donor-recipient combinations of group A and B KIR haplotypes and HLA class I ligand affect the outcome of HLA-matched, sibling donor hematopoietic cell transplantation', *Hum Immunol*, vol. 68, no. 5, pp. 309-323.
- Meyer, D., Single, R.M., Mack, S.J., Erlich, H.A. & Thomson, G. (2006) 'Signatures of demographic history and natural selection in the human major histocompatibility complex loci', *Genetics*, vol. 173, no. 4, pp. 2121-2142.
- Middelton, D. (2005) 'HLA Typing from Serology to Sequencing Era', *Iran J Allergy Asthma Immunol*, vol. 4, no. 2, pp. 53-66.
- Middleton, D. (2005) 'KIR allele and gene polymorphism group (KAG)', *Mol Immunol*, vol. 42, no. 4, pp. 455-457.
- Middleton, D., Curran, M. & Maxwell, L. (2002) 'Natural killer cells and their receptors', *Transpl Immunol*, vol. 10, no. 2-3, pp. 147-164.
- Middleton, D. & Gonzalez, F. (2010) 'The extensive polymorphism of KIR genes', *Immunology*, vol. 129, no. 1, pp. 8-19.
- Middleton, D., Gonzalez, F., Fernandez-Vina, M., Tiercy, J.M., Marsh, S.G.E., Aubrey, M., Bicalho, M.G., Canossi, A., Carter, V., Cate, S., Guerini, F.R., Loiseau, P., Martinetti, M., Moraes, M.E., Morales, V., Perasaari, J., Setterholm, M., Sprague, M., Tavoularis, S., Torres, M., Vidal, S., Witt, C., Wohlwend, G. & Yang, K.L. (2009) 'A bioinformatics approach to ascertaining the rarity of HLA alleles', *Tissue Antigens*, vol. 74, no. 6, pp. 480-485.

- Middleton, D., Meenagh, A. & Gourraud, P.A. (2007) 'KIR haplotype content at the allele level in 77 Northern Irish families', *Immunogenetics*, vol. 59, no. 2, pp. 145-158.
- Middleton, D., Meenagh, A., Moscoso, J. & Arnaiz-Villena, A. (2008) 'Killer immunoglobulin receptor gene and allele frequencies in Caucasoid, Oriental and Black populations from different continents', *Tissue Antigens*, vol. 71, no. 2, pp. 105-113.
- Middleton, D., Meenagh, A., Williams, F., Ross, O.A., Patterson, C., Gorodezky, C., Hammond, M. & Leheny, W.A. (2002) 'Frequency of cytokine polymorphisms in populations from Western Europe, Africa, Asia, the Middle East and South America', *Human Immunology*, vol. 63, no. 11, pp. 1055-1061.
- Middleton, D., Menchaca, L., Rood, H. & Komerofsky, R. (2003) 'New allele frequency database: www.allelefrequencies.net', *Tissue Antigens*, vol. 61, no. 5, pp. 403-407.
- Miyashita, R., Tsuchiya, N., Yabe, T., Kobayashi, S., Hashimoto, H., Ozaki, S. & Tokunaga, K. (2006) 'Association of killer cell immunoglobulin-like receptor genotypes with microscopic polyangiitis', *Arthritis and Rheumatism*, vol. 54, no. 3, pp. 992-997.
- Mogami, S., Hasegawa, G., Nakayama, I., Asano, M., Hosoda, H., Kadono, M., Fukui, M., Kitagawa, Y., Nakano, K., Ohta, M., Obayashi, H., Yoshikawa, T. & Nakamura, N. (2007) 'Killer cell immunoglobulin-like receptor genotypes in Japanese patients with type 1 diabetes', *Tissue Antigens*, vol. 70, no. 6, pp. 506-510.
- Monsalve, M.V., Edin, G. & Devine, D.V. (1998) 'Analysis of HLA class I and class II in Na-Dene and Amerindian populations from British Columbia, Canada', *Human Immunology*, vol. 59, no. 1, pp. 48-55.
- Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O. & Buus, S. (2007) 'NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence', *PLoS One*, vol. 2, no. 8, p. e796.
- Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S. & Lund, O. (2008) 'Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan', *PLoS Comput Biol*, vol. 4, no. 7, p. e1000107.
- Norman, P.J., Abi-Rached, L., Gendzekhadze, K., Korbelt, D., Gleimer, M., Rowley, D., Bruno, D., Carrington, C.V., Chandanayingyong, D., Chang, Y.H., Crespi, C., Saruhan-Direskeneli, G., Fraser, P.A., Hameed, K., Kamkamidze, G., Koram, K.A., Layrisse, Z., Matamoros, N., Mila, J., Park, M.H., Pitchappan, R.M., Ramdath, D.D., Shiau, M.Y., Stephens, H.A., Struik, S., Verity, D.H., Vaughan, R.W., Tyan, D., Davis, R.W., Riley, E.M., Ronaghi, M. & Parham, P. (2007) 'Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans', *Nat Genet*, vol. 39, no. 9, pp. 1092-1099.
- Norman, P.J., Carrington, C.V.F., Byng, M., Maxwell, L.D., Curran, M.D., Stephens, H.A.F., Chandanayingyong, D., Verity, D.H., Hameed, K., Ramdath, D.D. & Vaughan, R.W. (2002) 'Natural killer cell immunoglobulin-like receptor (KIR) locus profiles in African and South Asian populations', *Genes and Immunity*, vol. 3, no. 2, pp. 86-95.
- Nowak, I., Majorczyk, E., Ploski, R., Senitzer, D., Sun, J.Y. & Kusnierczyk, P. (2011) 'Lack of KIR2DL4 gene in a fertile Caucasian woman', *Tissue Antigens*, vol. 78, no. 2, pp. 115-119.
- Nunes, J.M. (2007) 'Tools for analysing ambiguous HLA data', *Tissue Antigens*, vol. 69 Suppl 1, pp. 203-205.

- Nunes, J.M., Riccio, M.E., Buhler, S., Di, D., Currat, M., Ries, F., Almada, A.J., Benhamamouch, S., Benitez, O., Canossi, A., Fadhlou-Zid, K., Fischer, G., Kervaire, B., Loiseau, P., de Oliveira, D.C., Papasteriades, C., Piancatelli, D., Rahal, M., Richard, L., Romero, M., Rousseau, J., Spiroski, M., Sulcebe, G., Middleton, D., Tiercy, J.M. & Sanchez-Mazas, A. (2010) 'Analysis of the HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics Workshop by using the Gene[rate] computer tools accommodating ambiguous data (AHPD project report)', *Tissue Antigens*, vol. 76, no. 1, pp. 18-30.
- Oliveira, L.A., Ribas, F., Bicalho, M.G., Tsuneto, L.T. & Petzl-Erler, M.L. (2008) 'High frequencies of alleles MICA*020 and MICA*027 in Amerindians and evidence of positive selection on exon 3', *Genes and Immunity*, vol. 9, no. 8, pp. 697-705.
- Orchard, S., Hermjakob, H. & Apweiler, R. (2003) 'The proteomics standards initiative', *Proteomics*, vol. 3, no. 7, pp. 1374-1376.
- Osier, M.V., Cheung, K.H., Kidd, J.R., Pakstis, A.J., Miller, P.L. & Kidd, K.K. (2001) 'ALFRED: an allele frequency database for diverse populations and DNA polymorphisms--an update', *Nucleic Acids Res*, vol. 29, no. 1, pp. 317-319.
- Ouzounis, C.A. & Valencia, A. (2003) 'Early bioinformatics: the birth of a discipline--a personal view', *Bioinformatics*, vol. 19, no. 17, pp. 2176-2190.
- Parham, P. (2005a) 'Influence of KIR diversity on human immunity', *Adv Exp Med Biol*, vol. 560, pp. 47-50.
- Parham, P. (2005b) 'MHC class I molecules and KIRs in human history, health and survival', *Nat Rev Immunol*, vol. 5, no. 3, pp. 201-214.
- Park, I. & Terasaki, P. (2000) 'Origins of the first HLA specificities', *Hum Immunol*, vol. 61, no. 3, pp. 185-189.
- Petersdorf, E.W. (2008) 'Optimal HLA matching in hematopoietic cell transplantation', *Curr Opin Immunol*, vol. 20, no. 5, pp. 588-593.
- Prugnolle, F., Manica, A., Charpentier, M., Guegan, J.F., Guernier, V. & Balloux, F. (2005) 'Pathogen-driven selection and worldwide HLA class I diversity', *Curr Biol*, vol. 15, no. 11, pp. 1022-1027.
- Rajalingam, R., Du, Z., Meenagh, A., Luo, L., Kavitha, V.J., Pavithra-Arulvani, R., Vidhyalakshmi, A., Sharma, S.K., Balazs, I., Reed, E.F., Pitchappan, R.M. & Middleton, D. (2008) 'Distinct diversity of KIR genes in three southern Indian populations: comparison with world populations revealed a link between KIR gene content and pre-historic human migrations', *Immunogenetics*, vol. 60, no. 5, pp. 207-217.
- Rajalingam, R., Krausa, P., Shilling, H.G., Stein, J.B., Balamurugan, A., McGinnis, M.D., Cheng, N.W., Mehra, N.K. & Parham, P. (2002) 'Distinctive KIR and HLA diversity in a panel of north Indian Hindus', *Immunogenetics*, vol. 53, no. 12, pp. 1009-1019.
- Rajeevan, H., Osier, M.V., Cheung, K.H., Deng, H., Druskin, L., Heinzen, R., Kidd, J.R., Stein, S., Pakstis, A.J., Tosches, N.P., Yeh, C.C., Miller, P.L. & Kidd, K.K. (2003) 'ALFRED: the ALlele FREquency Database. Update', *Nucleic Acids Res*, vol. 31, no. 1, pp. 270-271.
- Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.A. & Stevanovic, S. (1999) 'SYFPEITHI: database for MHC ligands and peptide motifs', *Immunogenetics*, vol. 50, no. 3-4, pp. 213-219.
- Retief, J.D. (2000) 'Phylogenetic analysis using PHYLIP', *Methods Mol Biol*, vol. 132, pp. 243-258.
- Rob, P. & Coronel, C. (2008) *Database systems: design, implementation, and management*, 8th edn, vol. 1, Cengage Learning, Inc, Boston, MA, USA.

- Robinson, J., Mistry, K., McWilliam, H., Lopez, R. & Marsh, S.G. (2010) 'IPD--the Immuno Polymorphism Database', *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D863-869.
- Robinson, J., Perez-Rodriguez, M., Waller, M.J., Cuillerier, B., Bahram, S., Yao, Z., Albert, E.D., Madrigal, J.A. & Marsh, S.G. (2001) 'MICA sequences 2000', *Immunogenetics*, vol. 53, no. 2, pp. 150-169.
- Robinson, J., Waller, M.J., Fail, S.C., McWilliam, H., Lopez, R., Parham, P. & Marsh, S.G. (2009) 'The IMGT/HLA database', *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D1013-1017.
- Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P. & Marsh, S.G. (2003) 'IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex', *Nucleic Acids Res*, vol. 31, no. 1, pp. 311-314.
- Robinson, J., Waller, M.J., Stoehr, P. & Marsh, S.G. (2005) 'IPD--the Immuno Polymorphism Database', *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D523-526.
- Rose, P.W., Beran, B., Bi, C.X., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M. & Bourne, P.E. (2011) 'The RCSB Protein Data Bank: redesigned web site and web services', *Nucleic Acids Research*, vol. 39, pp. D392-D401.
- Rousset, F. (2008) 'genepop'007: a complete re-implementation of the genepop software for Windows and Linux', *Molecular Ecology Resources*, vol. 8, no. 1, pp. 103-106.
- Rueda, B., Pascual, M., Lopez-Nevot, M.A., Gonzalez, E. & Martin, J. (2002) 'A new allele within the transmembrane region of the human MICA gene with seven GCT repeats', *Tissue Antigens*, vol. 60, no. 6, pp. 526-528.
- Sanchez-Mazas, A. (2007) 'An apportionment of human HLA diversity', *Tissue Antigens*, vol. 69 Suppl 1, pp. 198-202.
- Sanchez-Mazas, A., Djoulah, S., Busson, M., Le Monnier de Gouville, I., Poirier, J.C., Dehay, C., Charron, D., Excoffier, L., Schneider, S., Langaney, A., Dausset, J. & Hors, J. (2000) 'A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families', *Eur J Hum Genet*, vol. 8, no. 1, pp. 33-41.
- Sanchez-Mazas, A., Nunes, J.M., Vidan-Jeras, B., Fischer, G., Little, A.M., Bekmane, U., Buhler, S., Buus, S., Dormoy, A., Dubois, V., Eberhard, H.P., Eglite, E., Gonzalez-Galarza, F., Grubic, Z., Ivanova, M., Lie, B., Ligeiro, D., Lokki, M.L., Torres, H.M., Marsh, S.G.E., da Silva, B.M., Martorell, J., Mendonca, D., Middleton, D., Muller, C.R., Papasteriades, C., Poli, F., Sulcebe, G., Tonks, S., Nevessignsky, M.T., van Walraven, A.M. & Tiercy, J.M. (2010) 'HLA-NET, an EU-funded network for research on HLA diversity', *Tissue Antigens*, vol. 75, no. 5, pp. 505-506.
- Santin, I., de Nanclares, G.P., Calvo, B., Gaafar, A., Castano, L., Bilbao, J.R. & Grp, G.N. (2006) 'Killer cell immunoglobulin-like receptor (KIR) genes in the Basque population: Association study of KIR gene contents with type 1 diabetes mellitus', *Human Immunology*, vol. 67, no. 1-2, pp. 118-124.
- Schuler, M.M., Nastke, M.D. & Stevanovic, S. (2007) 'SYFPEITHI: database for searching and T-cell epitope prediction', *Methods Mol Biol*, vol. 409, pp. 75-93.
- Shah, A.A., Barthel, D., Lukasiak, P., Blazewicz, J. & Krasnogor, N. (2008) 'Web and grid technologies in bioinformatics, computational and systems biology: A review', *Current Bioinformatics*, vol. 3, no. 1, pp. 10-31.

- Sheldon, S. & Poulton, K. (2006) 'HLA typing and its influence on organ transplantation', *Methods Mol Biol*, vol. 333, pp. 157-174.
- Shi, L., Shi, L., Tao, Y., Lin, K., Liu, S., Yu, L., Yang, Z., Yi, W., Huang, X., Sun, H., Chu, J. & Yao, Y. (2011) 'Distribution of killer cell immunoglobulin-like receptor genes and combinations with HLA-C ligands in an isolated Han population in southwest China', *Tissue Antigens*, vol. 78, no. 1, pp. 60-64.
- Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J.K. (2009) 'The HLA genomic loci map: expression, interaction, diversity and disease', *J Hum Genet*, vol. 54, no. 1, pp. 15-39.
- Shilling, H.G., Guethlein, L.A., Cheng, N.W., Gardiner, C.M., Rodriguez, R., Tyan, D. & Parham, P. (2002) 'Allelic polymorphism synergizes with variable gene content to individualize human KIR genotype', *J Immunol*, vol. 168, no. 5, pp. 2307-2315.
- Shlomchik, W.D. (2007) 'Graft-versus-host disease', *Nat Rev Immunol*, vol. 7, no. 5, pp. 340-352.
- Simpson, E., Scott, D., James, E., Lombardi, G., Cwynarski, K., Dazzi, F., Millrain, M. & Dyson, R. (2002) 'Minor H antigens: genes and peptides', *Transplant Immunology*, vol. 10, no. 2-3, pp. 115-123.
- Single, R.M., Martin, M.P., Gao, X., Meyer, D., Yeager, M., Kidd, J.R., Kidd, K.K. & Carrington, M. (2007a) 'Global diversity and evidence for coevolution of KIR and HLA', *Nat Genet*, vol. 39, no. 9, pp. 1114-1119.
- Single, R.M., Meyer, D., Mack, S.J., Lancaster, A., Erlich, H.A. & Thomson, G. (2007b) '14th International HLA and Immunogenetics Workshop: Report of progress in methodology, data collection, and analyses', *Tissue Antigens*, vol. 69, pp. 185-187.
- Slatkin, M. (2008) 'Linkage disequilibrium--understanding the evolutionary past and mapping the medical future', *Nat Rev Genet*, vol. 9, no. 6, pp. 477-485.
- Sohn, Y.H., Cha, C.H., Oh, H.B., Kim, M.H., Choi, S.E. & Kwon, O.J. (2010) 'MICA polymorphisms and haplotypes with HLA-B and HLA-DRB1 in Koreans', *Tissue Antigens*, vol. 75, no. 1, pp. 48-55.
- Solberg, O.D., Mack, S.J., Lancaster, A.K., Single, R.M., Tsai, Y., Sanchez-Mazas, A. & Thomson, G. (2008) 'Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies', *Hum Immunol*, vol. 69, no. 7, pp. 443-464.
- Stein, L.D. (2003) 'Integrating biological databases', *Nat Rev Genet*, vol. 4, no. 5, pp. 337-345.
- Stephens, H.A.F. (2001) 'MICA and MICB genes: can the enigma of their polymorphism be resolved?', *Trends in Immunology*, vol. 22, no. 7, pp. 378-385.
- Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007) 'MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0', *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596-1599.
- Taniguchi, M. & Kawabata, M. (2009) 'KIR3DL1/S1 genotypes and KIR2DS4 allelic variants in the AB KIR genotypes are associated with Plasmodium-positive individuals in malaria infection', *Immunogenetics*, vol. 61, no. 11-12, pp. 717-730.
- Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H. & Gojobori, T. (2002) 'DNA Data Bank of Japan (DDBJ) for genome scale research in life science', *Nucleic Acids Research*, vol. 30, no. 1, pp. 27-30.
- Terasaki, P.I. (2007) 'A brief history of HLA', *Immunol Res*, vol. 38, no. 1-3, pp. 139-148.
- Terasaki, P.I. & Cai, J. (2008) 'Human leukocyte antigen antibodies and chronic rejection: from association to causation', *Transplantation*, vol. 86, no. 3, pp. 377-383.

- The MHC sequencing consortium (1999) 'Complete sequence and gene map of a human major histocompatibility complex', *Nature*, vol. 401, no. 6756, pp. 921-923.
- Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2002) 'Multiple sequence alignment using ClustalW and ClustalX', *Curr Protoc Bioinformatics*, vol. Chapter 2, p. Unit 2.3.
- Thorsby, E. (2009) 'A short history of HLA', *Tissue Antigens*, vol. 74, no. 2, pp. 101-116.
- Thorsby, E. & Lie, B.A. (2005) 'HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms', *Transpl Immunol*, vol. 14, no. 3-4, pp. 175-182.
- Tian, W., Boggs, D.A., Uko, G., Essiet, A., Inyama, M., Banjoko, B., Adewold, T., Ding, W.Z., Mohseni, M., Fritz, R., Chen, D.F., Palmer, L.J. & Fraser, P.A. (2003) 'MICA, HLA-B haplotypic variation in five population groups of sub-Saharan African ancestry', *Genes and Immunity*, vol. 4, no. 7, pp. 500-505.
- Tian, W., Li, L.X., Wang, F., Luo, Q.Z., Yan, M.Y., Yu, P., Guo, S.S. & Cao, Y. (2006) 'MICA-STR, HLA-B haplotypic diversity and linkage disequilibrium in the Hunan Han population of southern China', *International Journal of Immunogenetics*, vol. 33, no. 4, pp. 241-245.
- Toneva, M., Lepage, V., Lafay, G., Dulphy, N., Busson, M., Lester, S., Vu-Trien, A., Michaylova, A., Naumova, E., McCluskey, J. & Charron, D. (2001) 'Genomic diversity of natural killer cell receptor genes in three populations', *Tissue Antigens*, vol. 57, no. 4, pp. 358-362.
- Traherne, J.A. (2008) 'Human MHC architecture and evolution: implications for disease association studies', *Int J Immunogenet*, vol. 35, no. 3, pp. 179-192.
- Trowsdale, J. (2001) 'Genetic and functional relationships between MHC and NK receptor genes', *Immunity*, vol. 15, no. 3, pp. 363-374.
- Uhrberg, M., Parham, P. & Wernet, P. (2002) 'Definition of gene content for nine common group B haplotypes of the Caucasoid population: KIR haplotypes contain between seven and eleven KIR genes', *Immunogenetics*, vol. 54, no. 4, pp. 221-229.
- Uhrberg, M., Valiante, N.M., Shum, B.P., Shilling, H.G., Lienert-Weidenbach, K., Corliss, B., Tyan, D., Lanier, L.L. & Parham, P. (1997) 'Human diversity in killer cell inhibitory receptor genes', *Immunity*, vol. 7, no. 6, pp. 753-763.
- Varmus, H. (2003) 'Genomic empowerment: the importance of public databases', *Nat Genet*, vol. 35 Suppl 1, p. 3.
- Velickovic, M., Velickovic, Z. & Dunckley, H. (2006) 'Diversity of killer cell immunoglobulin-like receptor genes in Pacific Islands populations', *Immunogenetics*, vol. 58, no. 7, pp. 523-532.
- Velickovic, M., Velickovic, Z., Panigoro, R. & Dunckley, H. (2009) 'Diversity of killer cell immunoglobulin-like receptor genes in Indonesian populations of Java, Kalimantan, Timor and Irian Jaya', *Tissue Antigens*, vol. 73, no. 1, pp. 9-16.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W.,

- Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001) 'The Sequence of the Human Genome', *Science*, vol. 291, no. 5507, pp. 1304-1351.
- Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. & Peters, B. (2010) 'The immune epitope database 2.0', *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D854-862.
- Wahl, A., Schafer, F., Bardet, W., Buchli, R., Air, G.M. & Hildebrand, W.H. (2009) 'HLA class I molecules consistently present internal influenza epitopes', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 2, pp. 540-545.
- Wende, H., Colonna, M., Ziegler, A. & Volz, A. (1999) 'Organization of the leukocyte receptor cluster (LRC) on human chromosome 19q13.4', *Mamm Genome*, vol. 10, no. 2, pp. 154-160.
- Williams, F., Meenagh, A., Darke, C., Acosta, A., Daar, A.S., Gorodezky, C., Hammond, M., Nascimento, E. & Middleton, D. (2001) 'Analysis of the distribution of HLA-B alleles in populations from five continents', *Hum Immunol*, vol. 62, no. 6, pp. 645-650.
- Williams, T.M. (2001) 'Human leukocyte antigen gene polymorphism and the histocompatibility laboratory', *J Mol Diagn*, vol. 3, no. 3, pp. 98-104.
- Yawata, M., Yawata, N., Draghi, M., Little, A.M., Partheniou, F. & Parham, P. (2006) 'Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection

- and modulation of effector function', *Journal of Experimental Medicine*, vol. 203, no. 3, pp. 633-645.
- Yen, J.H., Lin, C.H., Tsai, W.C., Wu, C.C., Ou, T.T., Hu, C.J. & Liu, H.W. (2006) 'Killer cell immunoglobulin-like receptor gene's repertoire in rheumatoid arthritis', *Scandinavian Journal of Rheumatology*, vol. 35, no. 2, pp. 124-127.
- Yuliwulandari, R., Kashiwase, K., Nakajima, H., Uddin, J., Susmiarsih, T.P., Sofro, A.S. & Tokunaga, K. (2009) 'Polymorphisms of HLA genes in Western Javanese (Indonesia): close affinities to Southeast Asian populations', *Tissue Antigens*, vol. 73, no. 1, pp. 46-53.
- Zhang, Q.O., Lin, C.Y., Dong, Q., Wang, J. & Wang, W. (2011) 'Relationship between HLA-DRB1 polymorphism and susceptibility or resistance to multiple sclerosis in Caucasians: A meta-analysis of non-family-based studies', *Autoimmunity Reviews*, vol. 10, no. 8, pp. 474-481.
- Zhu, B.F., Wang, H.D., Shen, C.M., Deng, Y.J., Yang, G.A., Wu, Q.J., Xu, P., Qin, H.X., Fan, S.L., Huang, P., Deng, L.B., Lucas, R. & Wang, Z.Y. (2010) 'Killer cell immunoglobulin-like receptor gene diversity in the Tibetan ethnic minority group of China', *Human Immunology*, vol. 71, no. 11, pp. 1116-1123.
- Zhu, B.F., Wang, H.D., Shen, C.M., Fan, A.Y., Yang, G., Qin, H.X., Jin, T.B., Xie, T., Deng, L., Lucas, R. & Lian, Z.M. (2011) 'Diversity of Killer Cell Immunoglobulin-like Receptor Genes in the Bai Ethnic Minority of Yunnan, China', *Scandinavian Journal of Immunology*, vol. 73, no. 4, pp. 284-292.
- Ziegler, A., Muller, C.A., Bockmann, R.A. & Uchanska-Ziegler, B. (2009) 'Low-affinity peptides and T-cell selection', *Trends in Immunology*, vol. 30, no. 2, pp. 53-60.
- Zou, Y.Z. & Stastny, P. (2009) 'The role of major histocompatibility complex class I chain-related gene A antibodies in organ transplantation', *Current Opinion in Organ Transplantation*, vol. 14, no. 4, pp. 414-418.

Index

A

Active Server Pages, 30
adaptive immune system, 3
Allele frequencies, 52
Allele Frequency Calculator, 100
Allele Frequency Search, 71
alleles, 6
Amino Acid Frequency Analysis Tool, 81
antigens, 6
Asynchronous JavaScript and XML language, 30
attributes, 25

B

bioinformatics, 1
Biological databases, 23

C

classes, 26
classical HLA genes, 6
common alleles, 124
Cytokines, 22

D

data curation, 32
Database Management System, 25
dynamic websites, 29

E

Entity-Relationship Diagram, 25
Extensible Markup Language, 30

F

File Transfer Protocol, 29

frequent alleles, 117

G

genotype, 11
Genotype frequencies, 53
Genotype Frequency Search, 77

H

Haplotype Frequency Search, 75
Haplotypes frequencies, 52
heterozygous, 12
HLA allele family, 8
homozygous, 12
Human Leukocyte Antigen, 4
Hypertext Transfer Protocol, 29

I

immune system, 3
immunogenetics, 1
innate immune system, 3
International Histocompatibility Workshops, 1

J

JavaScript, 30

K

keys, 25
Killer-cell Immunoglobulin-like Receptors, 16
KIR framework genes, 17
KIR genotypes, 16

L

Leukocyte Receptor Complex, 16
linkage disequilibrium, 13

loci, 6

M

Major Histocompatibility Complex, 3

Major histocompatibility complex class I chain related,
21

map overlay display, 73

methods, 26

MHC Class I, 4

MHC Class II, 4

N

non-classical HLA genes, 6

O

Object oriented databases, 26

objects, 26

P

peptides, 4

public databases, 2

R

Rare Allele Detector, 125

Rare Allele Search, 121

rare alleles, 117

Relational databases, 25

relational model, 25

S

Simple Object Access Protocol, 29

Static websites, 29

Structured Query Language, 27

synonymous DNA substitution, 8

T

tables, 25

tuples, 25

U

Uniform Resource Locator, 29

V

very rare alleles, 117

views, 24

W

Web applications, 29

web browsers, 29

well documented alleles, 124

X

XML Schemas, 31

Appendix A

Schemas of the Allele Frequency Net Database

The following set of diagrams represents the logical and physical model schemas that were generated for the design of the AFND. Additionally, the SQL schema which can be used for the generation of the database is also included in this section.

Logical Data Model

Demographic data

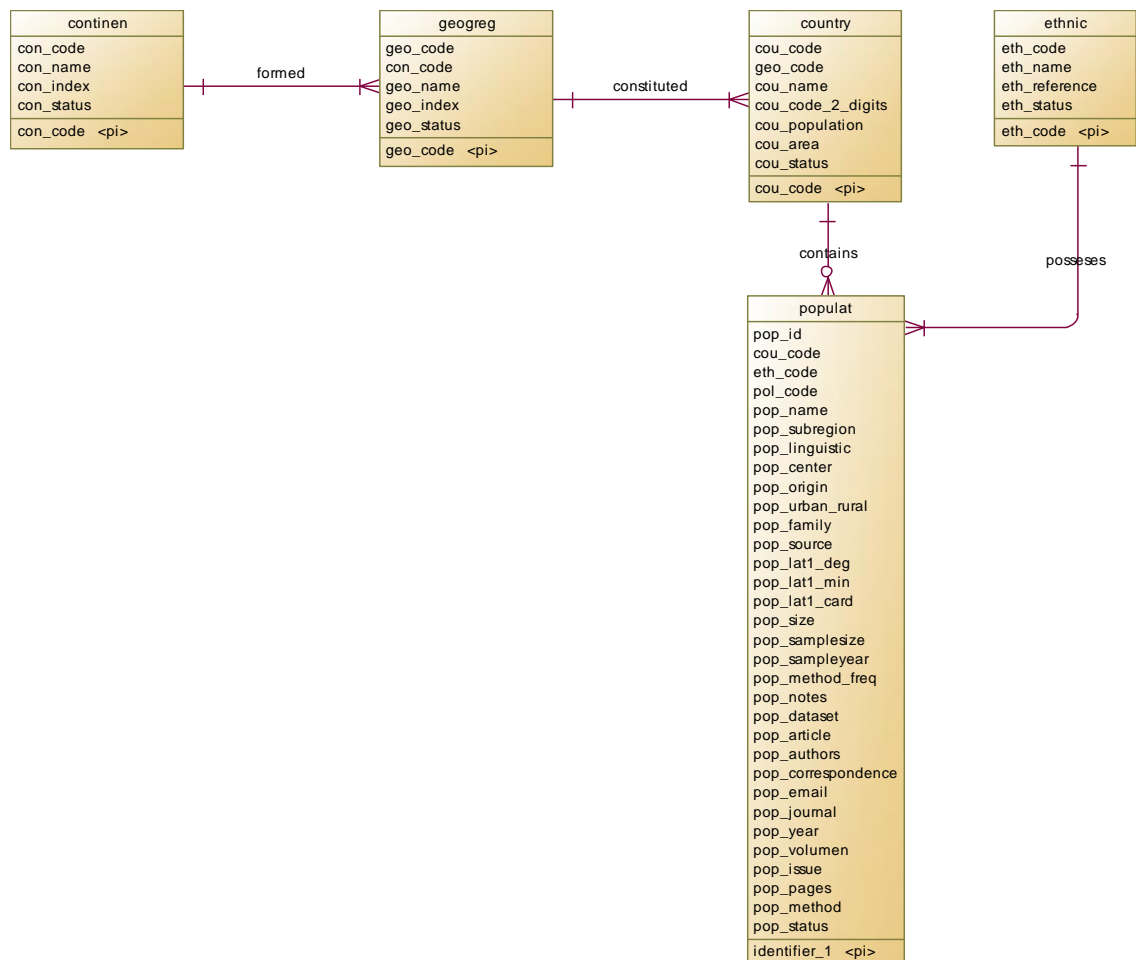


Figure A.1: Logical data model of the AFND: Demographic data.

Frequency data

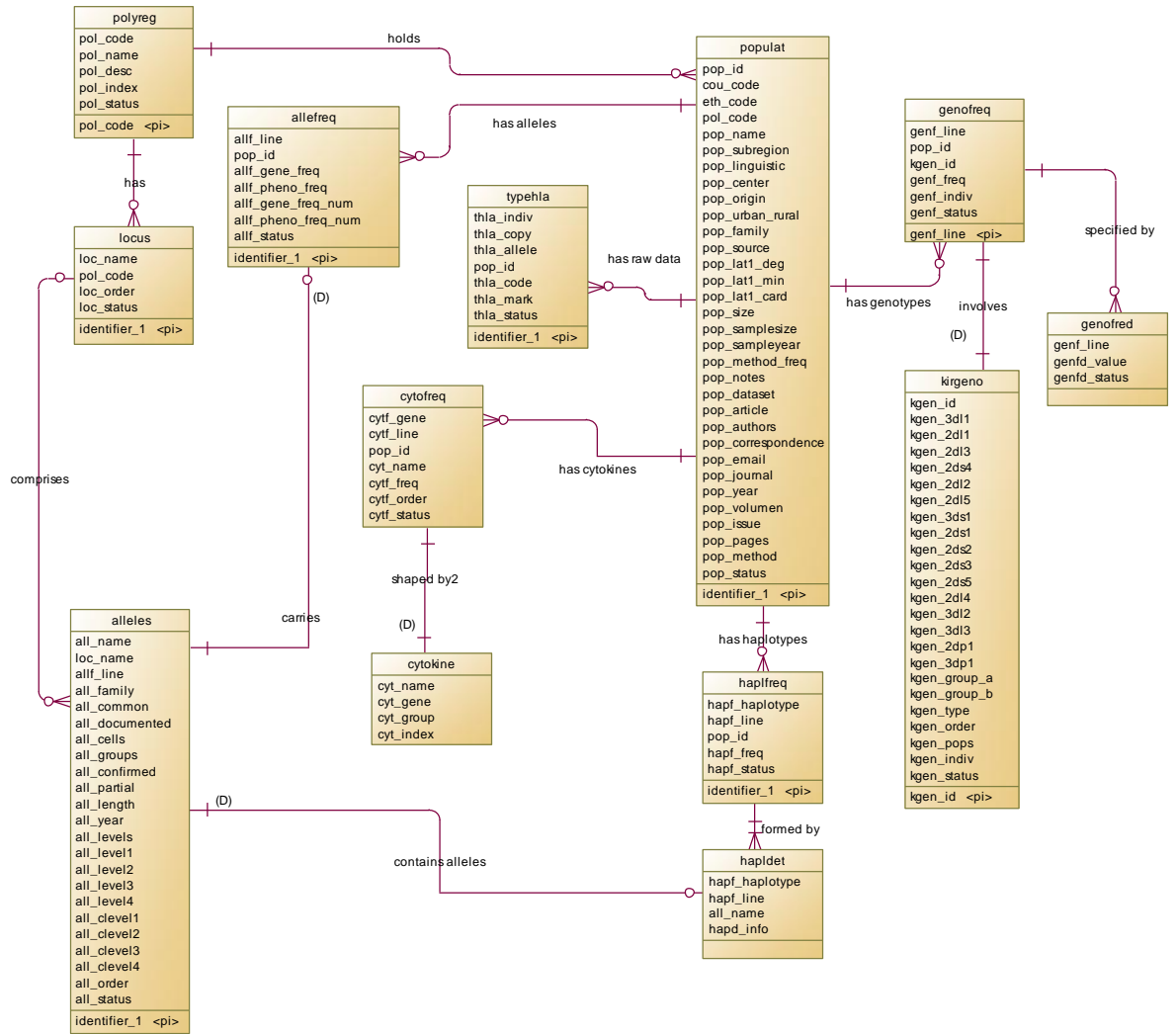


Figure A.2: Logical data model of the AFND: Frequency data.

DNA sequence data

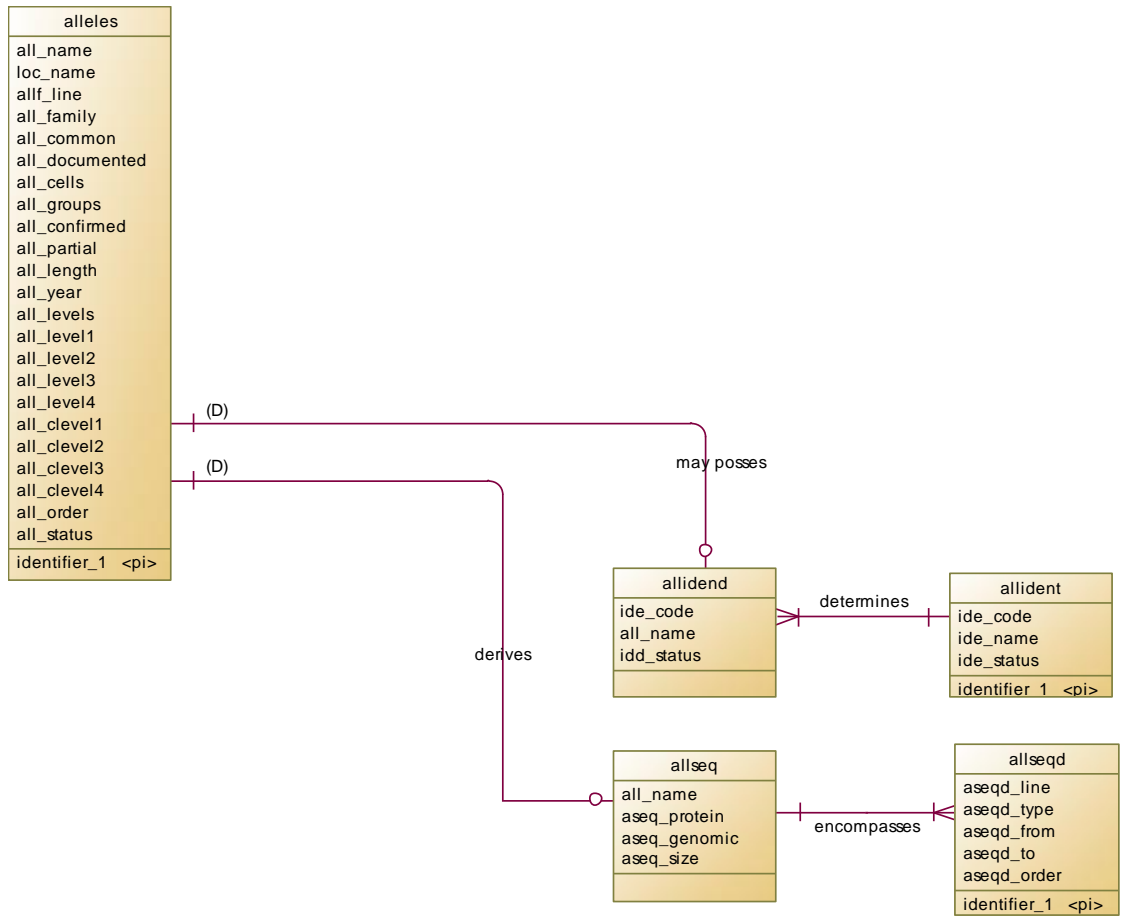


Figure A.3: Logical data model of the AFND: DNA sequence data.

Rare allele data

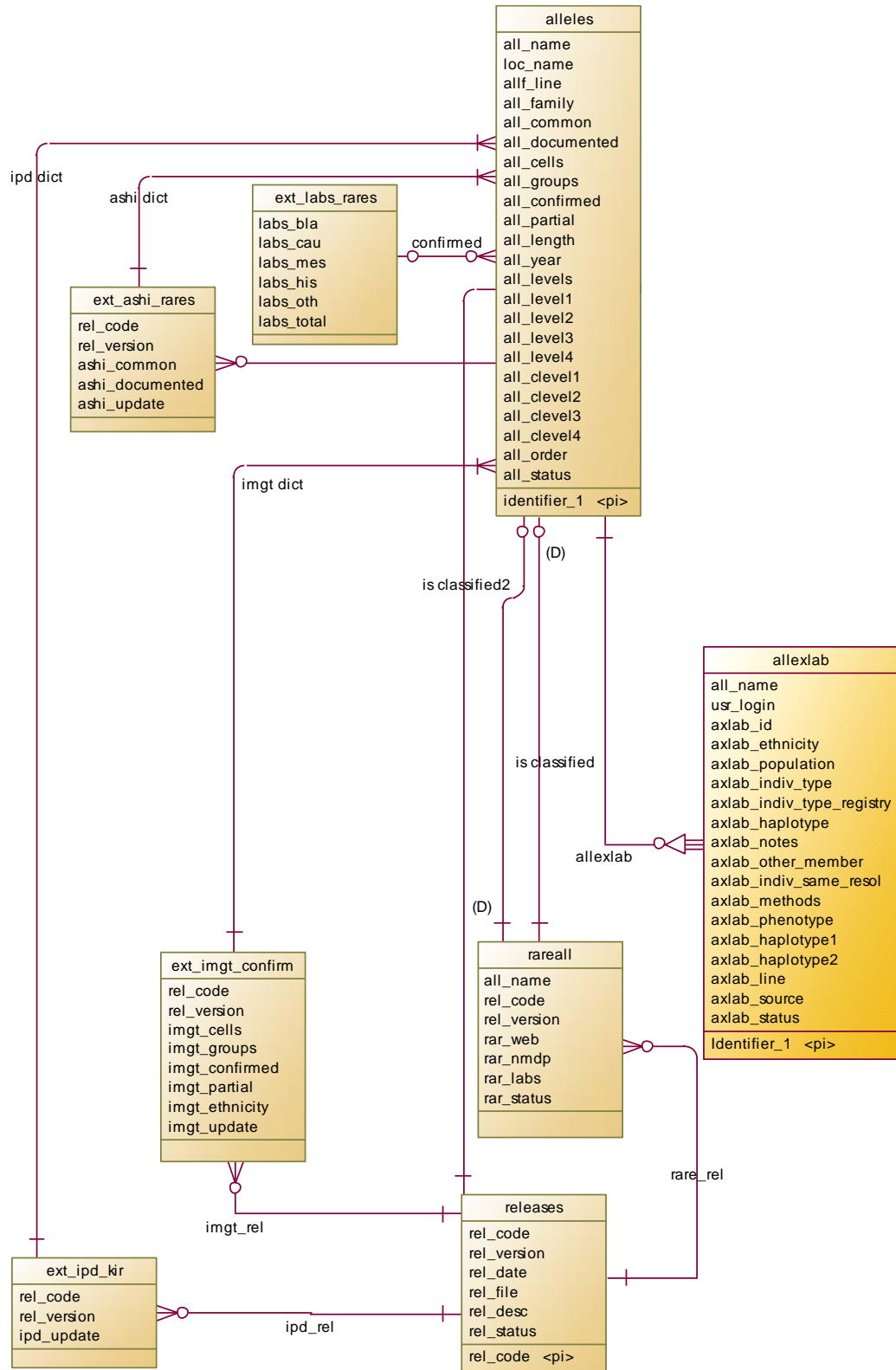


Figure A.4: Logical data model of the AFND: Rare allele data

Users and Security Access

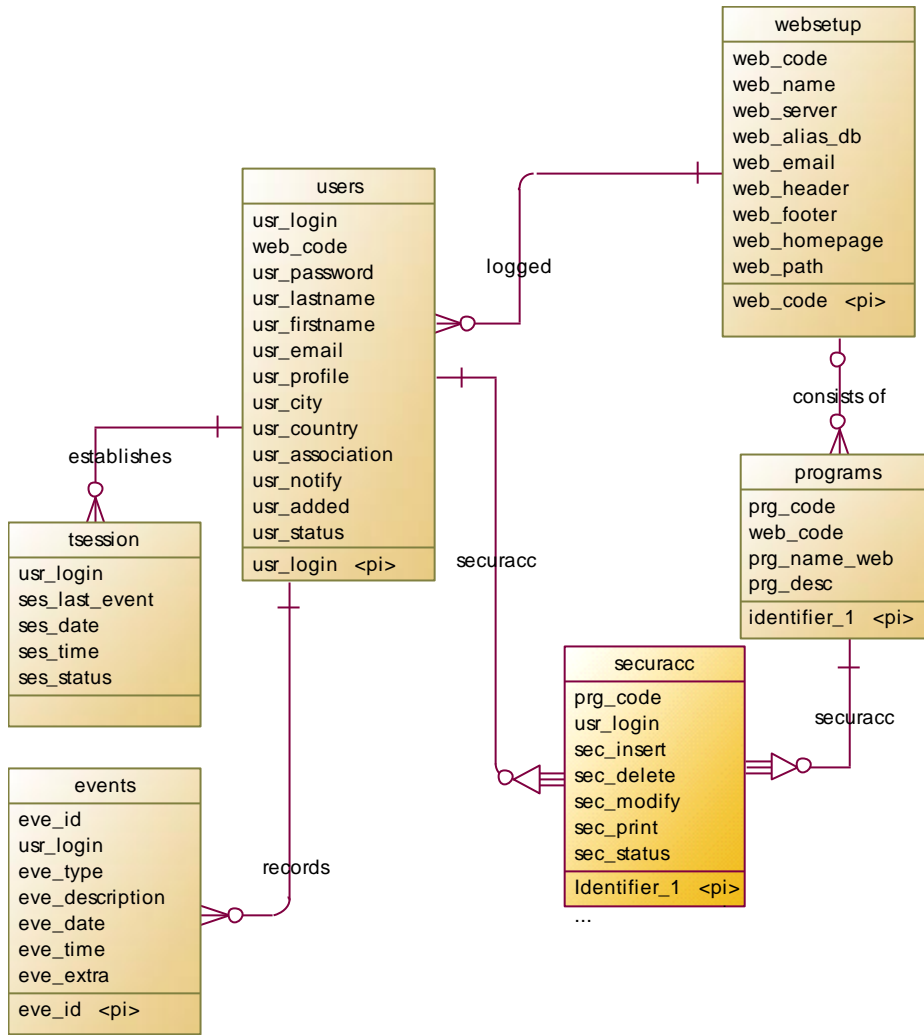


Figure A.5: Logical data model of the AFND: Users and security access

Physical Data Model

Demographic data

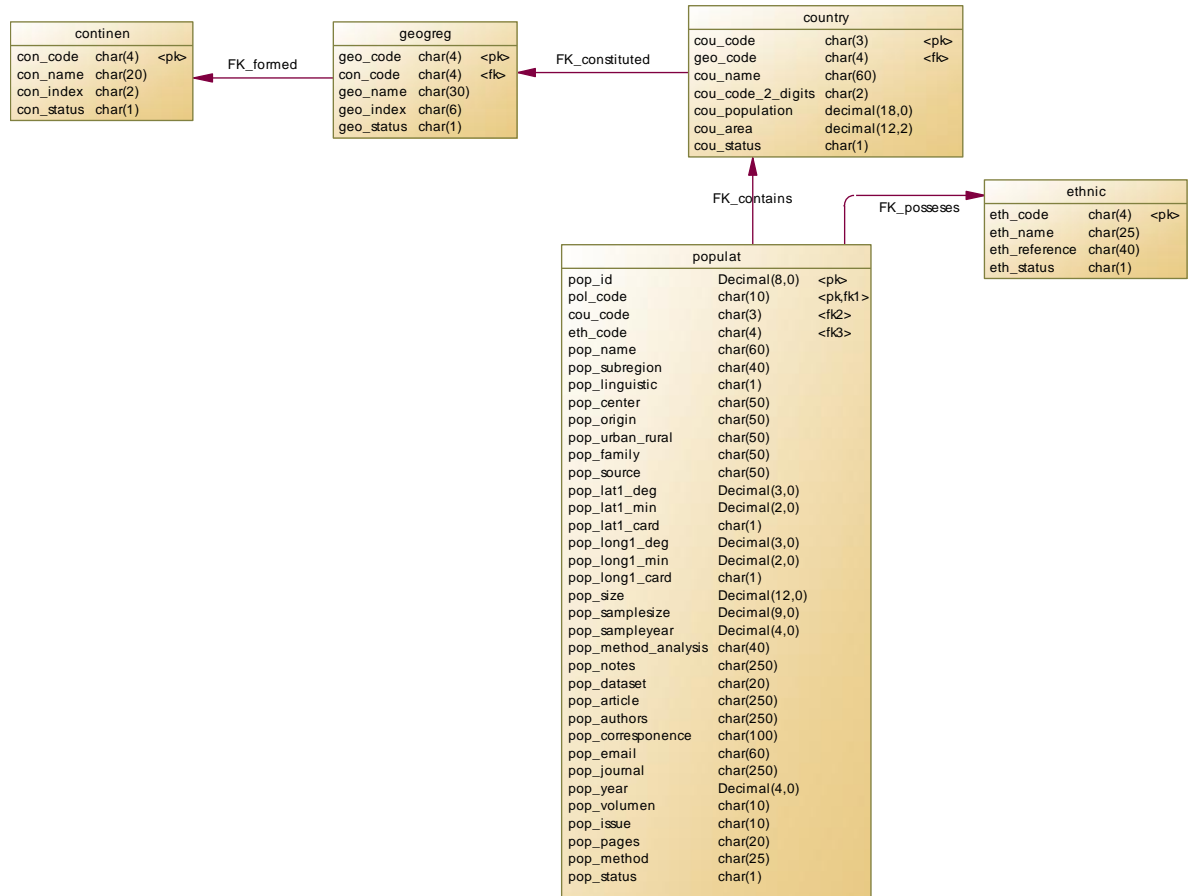


Figure A.6: Physical data model of the AFND: Demographic data

Frequency Data

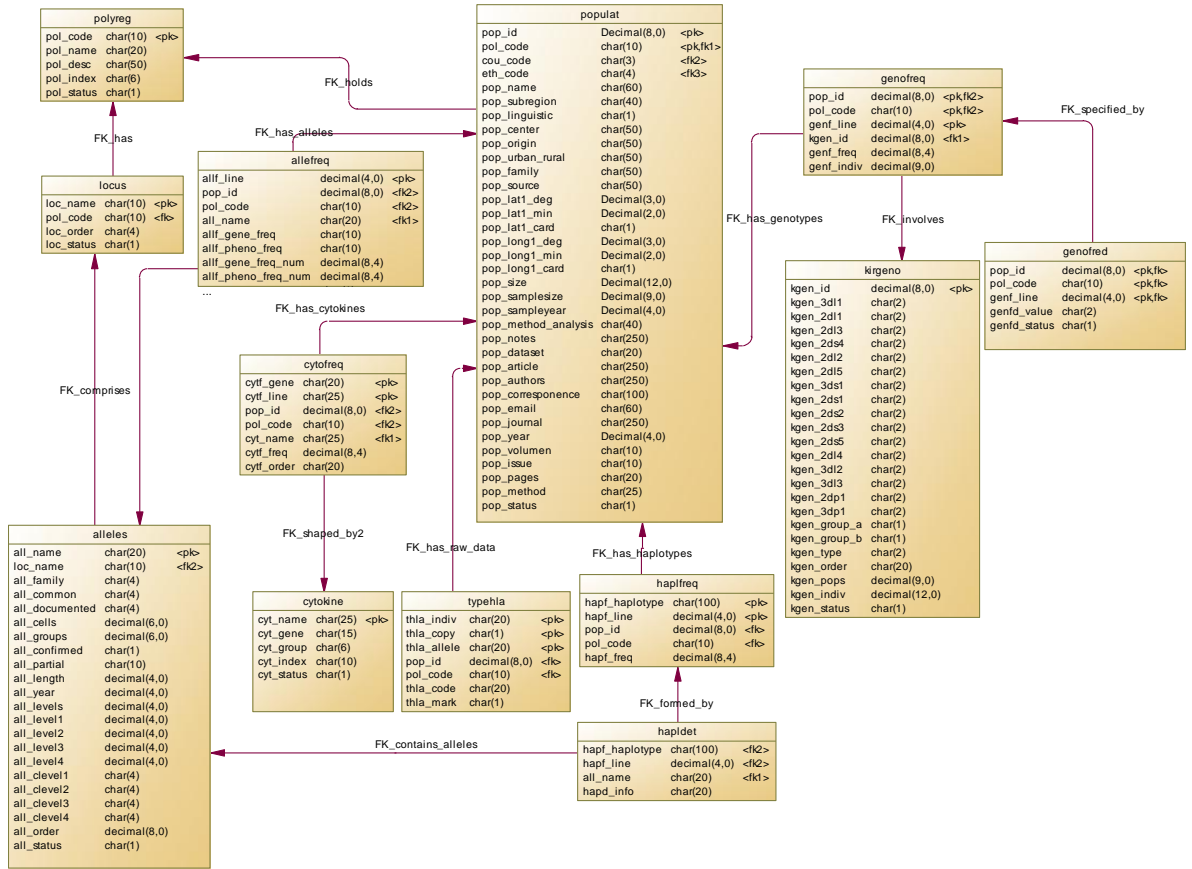


Figure A.7: Physical data model of the AFND: Frequency data

DNA Sequence data

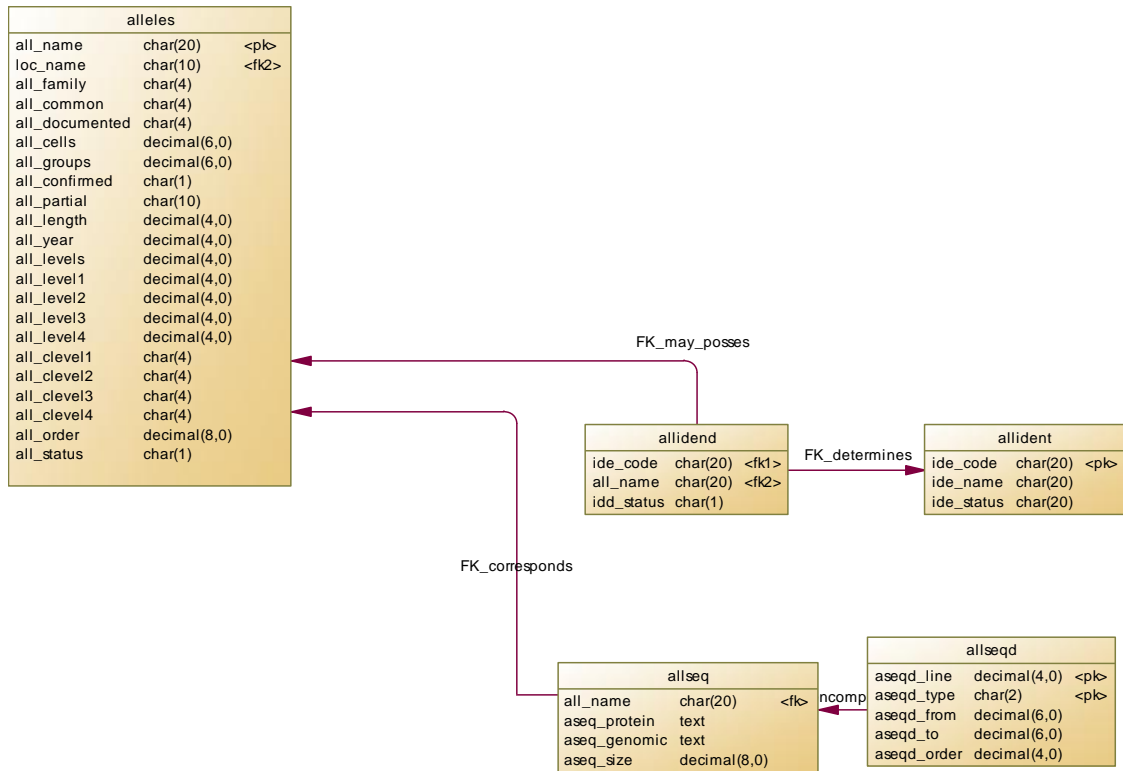


Figure A.8: Physical data model of the AFND: DNA sequence data

Users and Security Access

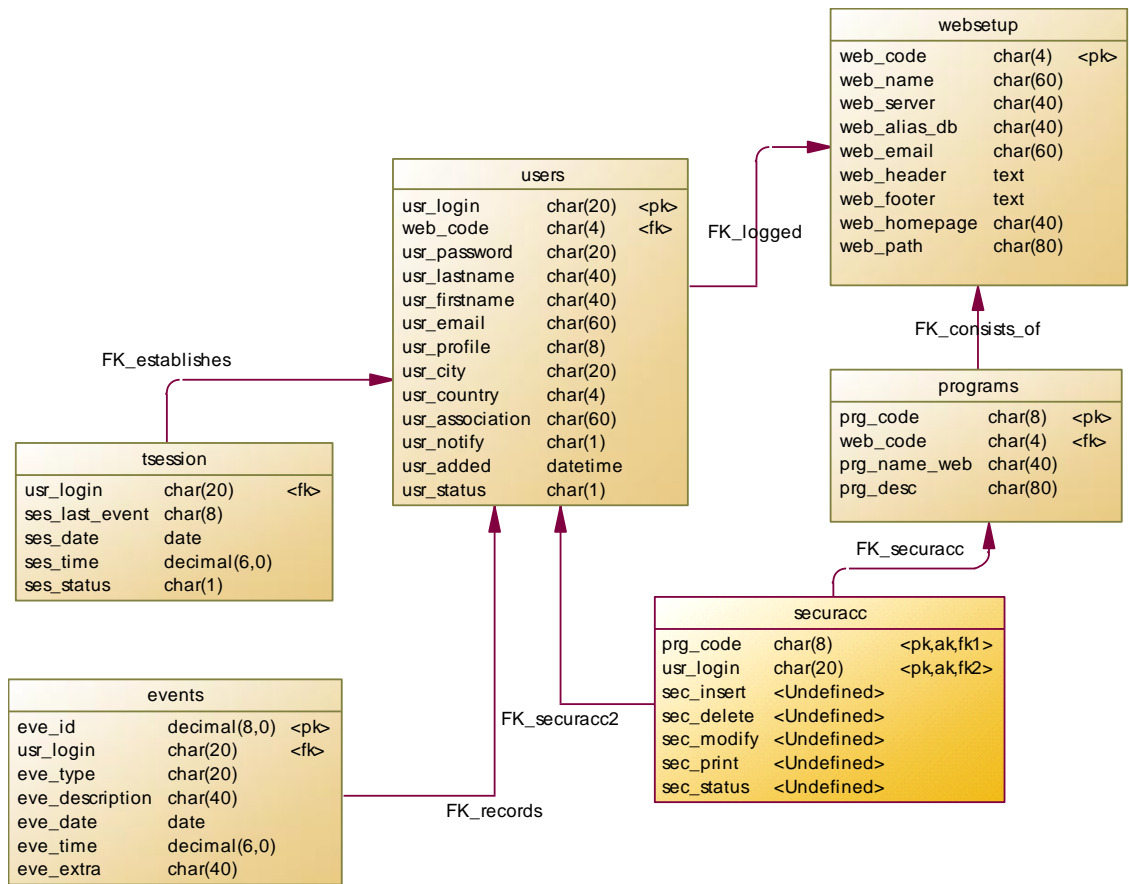


Figure A.10: Physical data model of the AFND: Users and security access

SQL schema of the AFND

```

CREATE TABLE alldesig(
  alld_allele nvarchar(20) Not NULL,
  alld_type nvarchar(20) NULL,
  alld_code nvarchar(20) NULL,
  alld_polyreg nvarchar(255) Not NULL,
  CONSTRAINT PK_alldesig PRIMARY KEY CLUSTERED
  (
    alld_allele Asc,
    alld_polyreg Asc
  )
)

CREATE TABLE allelfreq(
  allf_popid decimal(8, 0) Not NULL,
  allf_polyreg nvarchar(10) Not NULL,
  allf_locus nvarchar(20) Not NULL,
  allf_allele nvarchar(20) Not NULL,
  allf_line decimal(4, 0) NULL,
  allf_gene_freq nvarchar(10) NULL,
  allf_pheno_freq nvarchar(10) NULL,
  allf_added datetime NULL,
  allf_updated datetime NULL,
  allf_user nvarchar(20) NULL,
  allf_status nvarchar(1) NULL,
  allf_indiv decimal(9, 0) NULL,
  allf_chrom decimal(9, 0) NULL,
  allf_sample_size decimal(9, 0) NULL,
  allf_notes nvarchar(250) NULL,
  CONSTRAINT PK_allelfreq PRIMARY KEY CLUSTERED
  (
    allf_popid Asc,
    allf_polyreg Asc,
    allf_locus Asc,
    allf_allele Asc
  )
)

CREATE TABLE alleles(
  all_name nvarchar(20) Not NULL,
  all_name2 nvarchar(20) NULL,
  all_region nvarchar(10) NULL,
  all_locus nvarchar(20) NULL,
  all_family nvarchar(4) NULL,
  all_ethnic_origin nvarchar(Max) NULL,
  all_class nvarchar(15) NULL,
  all_common nvarchar(4) NULL,
  all_documented nvarchar(4) NULL,
  all_cells decimal(6, 0) NULL,
  all_groups decimal(6, 0) NULL,
  all_confirmed nvarchar(12) NULL,
  all_start decimal(6, 0) NULL,
  all_end decimal(6, 0) NULL,
  all_partial nvarchar(10) NULL,
  all_updated datetime NULL,
  all_pops decimal(9, 0) NULL,
  all_total_nmdp decimal(9, 0) NULL,
  all_total_labs decimal(9, 0) NULL,
  all_total_identical_cell decimal(9, 0)
  NULL,
  all_total_identical_group decimal(9, 0)
  NULL,
  all_total_identical_pops decimal(9, 0)
  NULL,
  all_total_identical_nmdp decimal(9, 0)
  NULL,
  all_total_identical_labs decimal(9, 0)
  NULL,
  all_length decimal(4, 0) NULL,
  all_status nvarchar(1) NULL,
  all_year decimal(4, 0) NULL,
  all_levels decimal(4, 0) NULL,
  all_level1 decimal(4, 0) NULL,
  all_level2 decimal(4, 0) NULL,
  all_level3 decimal(4, 0) NULL,
  all_level4 decimal(4, 0) NULL,
  all_clevel1 nvarchar(4) NULL,
  all_clevel2 nvarchar(4) NULL,
  all_clevel3 nvarchar(4) NULL,
  all_clevel4 nvarchar(4) NULL,
  all_order decimal(8, 0) NULL,
  CONSTRAINT PK_alleles PRIMARY KEY CLUSTERED
  (
    all_name Asc
  )
)

CREATE TABLE allexlab(
  axlab_allele nvarchar(20) Not NULL,
  axlab_indiv_id nvarchar(20) Not NULL,
  axlab_user nvarchar(20) Not NULL,
  axlab_ethnicity nvarchar(50) NULL,
  axlab_population nvarchar(50) NULL,
  axlab_indiv_type nvarchar(30) NULL,
  axlab_indiv_type_registry nvarchar(40)
  NULL,
  axlab_indiv_type_other nvarchar(40) NULL,
  axlab_haplotype nvarchar(40) NULL,
  axlab_notes nvarchar(40) NULL,
  axlab_other_member nvarchar(40) NULL,
  axlab_indiv_same_resol decimal(9, 0) NULL,
  axlab_methods nvarchar(60) NULL,
  axlab_sbt nvarchar(1) NULL,
  axlab_ssp nvarchar(1) NULL,
  axlab_sso nvarchar(1) NULL,
  axlab_seq nvarchar(255) NULL,
  axlab_oth nvarchar(1) NULL,
  axlab_type nvarchar(200) NULL,
  axlab_phenotype nvarchar(200) NULL,
  axlab_haplotype1 nvarchar(200) NULL,
  axlab_haplotype2 nvarchar(10) NULL,
  axlab_hap1_hla_a nvarchar(30) NULL,
  axlab_hap2_hla_a nvarchar(30) NULL,
  axlab_hap1_hla_b nvarchar(30) NULL,
  axlab_hap2_hla_b nvarchar(30) NULL,
  axlab_hap1_hla_c nvarchar(30) NULL,
  axlab_hap2_hla_c nvarchar(30) NULL,
  axlab_hap1_hla_dra1 nvarchar(30) NULL,
  axlab_hap2_hla_dra1 nvarchar(30) NULL,
  axlab_hap1_hla_drb1 nvarchar(30) NULL,
  axlab_hap2_hla_drb1 nvarchar(30) NULL,
  axlab_hap1_hla_dqa1 nvarchar(30) NULL,
  axlab_hap2_hla_dqa1 nvarchar(30) NULL,
  axlab_hap1_hla_dqb1 nvarchar(30) NULL,
  axlab_hap2_hla_dqb1 nvarchar(30) NULL,
  axlab_hap1_hla_dp1a1 nvarchar(30) NULL,
  axlab_hap2_hla_dp1a1 nvarchar(30) NULL,
  axlab_hap1_hla_dp1b1 nvarchar(30) NULL,
  axlab_hap2_hla_dp1b1 nvarchar(30) NULL,
  axlab_date_added datetime NULL,
  axlab_line decimal(9, 0) NULL,
  axlab_status nvarchar(1) NULL,
  axlab_authors nvarchar(200) NULL,
  axlab_add1 nvarchar(80) NULL,
  axlab_add2 nvarchar(80) NULL,
  axlab_city nvarchar(20) NULL,
  axlab_county nvarchar(20) NULL,
  axlab_country nvarchar(4) NULL,
  axlab_postcode nvarchar(10) NULL,
  axlab_fax nvarchar(25) NULL,
  axlab_email nvarchar(60) NULL,
  axlab_ihw nvarchar(1) NULL,
  CONSTRAINT PK_allexlab PRIMARY KEY CLUSTERED
  (
    axlab_allele Asc,
    axlab_indiv_id Asc,
    axlab_user Asc
  )
)

CREATE TABLE allidend(
  idd_code nvarchar(20) Not NULL,
  idd_allele nvarchar(20) Not NULL,
  idd_status nvarchar(1) NULL,
  CONSTRAINT PK_allidend PRIMARY KEY CLUSTERED
  (
    idd_code Asc,
    idd_allele Asc
  )
)

CREATE TABLE allidendr(
  idr_allele nvarchar(20) Not NULL,
  idr_identical nvarchar(20) Not NULL,
  CONSTRAINT PK_allidendr PRIMARY KEY CLUSTERED
  (
    idr_allele Asc,
    idr_identical Asc
  )
)

CREATE TABLE allident(
  ide_code nvarchar(20) Not NULL,
  ide_name nvarchar(20) Not NULL,
  ide_status nvarchar(1) NULL,
  CONSTRAINT PK_allident PRIMARY KEY CLUSTERED
  (
    ide_code Asc,
    ide_name Asc
  )
)

CREATE TABLE allseq(
  aseq_allele nvarchar(20) Not NULL,
  aseq_protein nvarchar(Max) NULL,
  aseq_genomic nvarchar(Max) NULL,
  aseq_size decimal(8, 0) NULL,

```



```

CONSTRAINT PK_allseq PRIMARY KEY CLUSTERED
(
    aseq_allele Asc
)
)
CREATE TABLE allseqd(
    aseqd_allele nvarchar(20) Not NULL,
    aseqd_line decimal(4, 0) Not NULL,
    aseqd_type nvarchar(2) Not NULL,
    aseqd_from decimal(6, 0) NULL,
    aseqd_to decimal(6, 0) NULL,
    aseqd_order decimal(4, 0) NULL,
    CONSTRAINT PK_allseqd PRIMARY KEY CLUSTERED
    (
        aseqd_allele Asc,
        aseqd_line Asc,
        aseqd_type Asc
    )
)
CREATE TABLE allsynon(
    asyn_parent nvarchar(20) Not NULL,
    asyn_child nvarchar(20) Not NULL,
    asyn_parent_size decimal(4, 0) NULL,
    asyn_child_size decimal(4, 0) NULL,
    CONSTRAINT PK_allsynon PRIMARY KEY CLUSTERED
    (
        asyn_parent Asc,
        asyn_child Asc
    )
)
CREATE TABLE cellline(
    cell_id decimal(8, 0) Not NULL,
    cell_dna_no nvarchar(20) NULL,
    cell_ihw_no nvarchar(20) NULL,
    cell_ipd_code nvarchar(20) NULL,
    cell_consang nvarchar(1) NULL,
    cell_2d11_1 nvarchar(20) NULL,
    cell_2d11_2 nvarchar(20) NULL,
    cell_2dp1_1 nvarchar(20) NULL,
    cell_2d13_1 nvarchar(20) NULL,
    cell_2d13_2 nvarchar(20) NULL,
    cell_2ds4_1 nvarchar(20) NULL,
    cell_2ds4_2 nvarchar(20) NULL,
    cell_3d11_1 nvarchar(20) NULL,
    cell_3d11_2 nvarchar(20) NULL,
    cell_3d11_1_s nvarchar(20) NULL,
    cell_3d11_2_s nvarchar(20) NULL,
    cell_3ds1_1 nvarchar(20) NULL,
    cell_3ds1_1_s nvarchar(20) NULL,
    cell_2ds1_1 nvarchar(20) NULL,
    cell_2ds2_1 nvarchar(20) NULL,
    cell_2ds3_1 nvarchar(20) NULL,
    cell_2ds3_2 nvarchar(20) NULL,
    cell_2ds5_1 nvarchar(20) NULL,
    cell_2d12_1 nvarchar(20) NULL,
    cell_2d12_2 nvarchar(20) NULL,
    cell_2d15a_1 nvarchar(20) NULL,
    cell_2d15a_2 nvarchar(20) NULL,
    cell_2d15a_s nvarchar(20) NULL,
    cell_2d15b_1 nvarchar(20) NULL,
    cell_2d15b_2 nvarchar(20) NULL,
    cell_2d15b_s nvarchar(20) NULL,
    cell_2d14_1 nvarchar(20) NULL,
    cell_2d14_2 nvarchar(20) NULL,
    cell_3d12_1 nvarchar(20) NULL,
    cell_3d12_2 nvarchar(20) NULL,
    cell_3d13_1 nvarchar(20) NULL,
    cell_3dpl_1 nvarchar(20) NULL,
    cell_status nvarchar(1) NULL,
    CONSTRAINT PK_cellline PRIMARY KEY CLUSTERED
    (
        cell_id Asc
    )
)
CREATE TABLE codons(
    cod_id decimal(4, 0) Not NULL,
    cod_aminoacid nvarchar(20) NULL,
    cod_letter nvarchar(1) NULL,
    cod_3letter nvarchar(3) NULL,
    cod_rna nvarchar(3) NULL,
    cod_dna nvarchar(3) NULL,
    cod_polarity nvarchar(5) NULL,
    CONSTRAINT PK_codons PRIMARY KEY CLUSTERED
    (
        cod_id Asc
    )
)
CREATE TABLE continen(
    con_code nvarchar(4) Not NULL,
    con_name nvarchar(20) NULL,
    con_index nvarchar(2) NULL,
    con_status nvarchar(1) NULL,
    CONSTRAINT PK_continen PRIMARY KEY CLUSTERED
    (
        con_code Asc
    )
)
CREATE TABLE country(
    cou_code nvarchar(3) Not NULL,
    geo_code nvarchar(4) Not NULL,
    cou_name nvarchar(60) NULL,
    cou_code2 nvarchar(2) NULL,
    cou_number decimal(6, 0) NULL,
    cou_name2 nvarchar(60) NULL,
    cou_tld nvarchar(6) NULL,
    cou_timezone nvarchar(60) NULL,
    cou_phone_code nvarchar(5) NULL,
    cou_population decimal(18, 0) NULL,
    cou_pop_date datetime NULL,
    cou_pop_source nvarchar(60) NULL,
    cou_area decimal(12, 2) NULL,
    cou_area_date datetime NULL,
    cou_area_source nvarchar(60) NULL,
    cou_capital nvarchar(60) NULL,
    cou_latitude nvarchar(40) NULL,
    cou_longitude nvarchar(40) NULL,
    cou_status nvarchar(1) NULL,
    CONSTRAINT PK_country PRIMARY KEY CLUSTERED
    (
        cou_code Asc
    )
)
CREATE TABLE couxgeo(
    cogr_geo_region_code nvarchar(4) Not NULL,
    cogr_country_code nvarchar(3) Not NULL,
    cogr_geo_region_name nvarchar(25) NULL,
    cogr_country_name nvarchar(60) NULL,
    cogr_status nvarchar(1) NULL,
    CONSTRAINT PK_couxgeo PRIMARY KEY CLUSTERED
    (
        cogr_geo_region_code Asc,
        cogr_country_code Asc
    )
)
CREATE TABLE cytofreq(
    cytf_popid decimal(8, 0) Not NULL,
    cytf_gene nvarchar(20) Not NULL,
    cytf_cytokine nvarchar(25) Not NULL,
    cytf_line decimal(4, 0) NULL,
    cytf_freq decimal(8, 4) NULL,
    cytf_added datetime NULL,
    cytf_updated datetime NULL,
    cytf_user nvarchar(20) NULL,
    cytf_order decimal(12, 4) NULL,
    cytf_status nvarchar(1) NULL,
    cytf_indiv decimal(9, 0) NULL,
    cytf_sample_size decimal(9, 0) NULL,
    CONSTRAINT PK_cytofreq PRIMARY KEY CLUSTERED
    (
        cytf_popid Asc,
        cytf_gene Asc,
        cytf_cytokine Asc
    )
)
CREATE TABLE cytokine(
    cyt_name nvarchar(25) Not NULL,
    cyt_gene nvarchar(15) NULL,
    cyt_group nvarchar(6) NULL,
    cyt_updated datetime NULL,
    cyt_index nvarchar(10) NULL,
    cyt_status nvarchar(1) NULL,
    CONSTRAINT PK_cytokine PRIMARY KEY CLUSTERED
    (
        cyt_name Asc
    )
)
CREATE TABLE ethnalle(
    etha_allele nvarchar(20) Not NULL,
    etha_ethnic_origin nvarchar(200) NULL,
    etha_first nvarchar(4) NULL,
    CONSTRAINT PK_ethnalle PRIMARY KEY CLUSTERED
    (
        etha_allele Asc
    )
)
CREATE TABLE ethnic(
    eth_code nvarchar(4) Not NULL,
    eth_name nvarchar(25) NULL,
    eth_status nvarchar(1) NULL,
    eth_reference nvarchar(40) NULL,
    CONSTRAINT PK_ethnic PRIMARY KEY CLUSTERED
    (
        eth_code Asc
    )
)
CREATE TABLE ethxcou(
    ethc_country_code nvarchar(3) Not NULL,
    ethc_ethnic_code nvarchar(4) Not NULL,
    ethc_country_name nvarchar(60) NULL,
    ethc_ethnic_name nvarchar(25) NULL,
    ethc_status nvarchar(1) NULL,
    CONSTRAINT PK_ethxcou PRIMARY KEY CLUSTERED

```

```

(
    ethc_country_code Asc,
    ethc_ethnic_code Asc
)
)
)
CREATE TABLE ethxgeo(
    ethg_geo_region_code nvarchar(4) Not NULL,
    ethg_ethnic_code nvarchar(4) Not NULL,
    ethg_geo_region_name nvarchar(25) NULL,
    ethg_ethnic_name nvarchar(25) NULL,
    ethg_status nvarchar(1) NULL,
    CONSTRAINT PK_ethxgeo PRIMARY KEY CLUSTERED
(
    ethg_geo_region_code Asc,
    ethg_ethnic_code Asc
)
)
)
CREATE TABLE event(
    eve_id decimal(8, 0) Not NULL,
    eve_event nvarchar(20) NULL,
    eve_description nvarchar(40) NULL,
    eve_user nvarchar(20) NULL,
    eve_program nvarchar(8) NULL,
    eve_date datetime NULL,
    eve_time decimal(6, 0) NULL,
    eve_extra nvarchar(40) NULL,
    CONSTRAINT PK_event PRIMARY KEY CLUSTERED
(
    eve_id Asc
)
)
)
CREATE TABLE ext_ashi_rares(
    ashi_allele nvarchar(20) Not NULL,
    ashi_release nvarchar(10) Not NULL,
    ashi_common nvarchar(2) NULL,
    ashi_documented nvarchar(2) NULL,
    ashi_update datetime NULL,
    CONSTRAINT PK_ext_ashi_rares PRIMARY KEY CLUSTERED
(
    ashi_allele Asc,
    ashi_release Asc
)
)
)
CREATE TABLE ext_imgt_confirm(
    imgt_allele nvarchar(20) Not NULL,
    imgt_release nvarchar(10) Not NULL,
    imgt_cells decimal(6, 0) NULL,
    imgt_groups decimal(6, 0) NULL,
    imgt_confirmed nvarchar(15) NULL,
    imgt_start decimal(6, 0) NULL,
    imgt_end decimal(6, 0) NULL,
    imgt_partial nvarchar(10) NULL,
    imgt_ethnicity nvarchar(200) NULL,
    imgt_update datetime NULL,
    CONSTRAINT PK_ext_imgt_confirm PRIMARY KEY CLUSTERED
(
    imgt_allele Asc,
    imgt_release Asc
)
)
)
CREATE TABLE ext_ipd_kir(
    ipd_allele nvarchar(20) Not NULL,
    ipd_release nvarchar(10) Not NULL,
    ipd_update datetime NULL,
    CONSTRAINT PK_ext_ipd_kir PRIMARY KEY CLUSTERED
(
    ipd_allele Asc,
    ipd_release Asc
)
)
)
CREATE TABLE ext_labs_rares(
    labs_allele nvarchar(20) Not NULL,
    labs_bla decimal(9, 0) NULL,
    labs_cau decimal(9, 0) NULL,
    labs_mes decimal(9, 0) NULL,
    labs_his decimal(9, 0) NULL,
    labs_oth decimal(9, 0) NULL,
    labs_total decimal(9, 0) NULL,
    CONSTRAINT PK_ext_labs_rares PRIMARY KEY CLUSTERED
(
    labs_allele Asc
)
)
)
CREATE TABLE ext_nmdp_rares(
    nmdp_allele nvarchar(20) Not NULL,
    nmdp_release nvarchar(10) Not NULL,
    nmdp_afa decimal(6, 0) NULL,
    nmdp_api decimal(6, 0) NULL,
    nmdp_cau decimal(6, 0) NULL,
    nmdp_his decimal(6, 0) NULL,
    nmdp_nam decimal(6, 0) NULL,
    nmdp_oth decimal(6, 0) NULL,
    nmdp_total decimal(9, 0) NULL,
    nmdp_rare nvarchar(1) NULL,
    nmdp_zero nvarchar(1) NULL,
    nmdp_update datetime NULL,
    CONSTRAINT PK_ext_nmdp_rares PRIMARY KEY CLUSTERED
(
    nmdp_allele Asc,
    nmdp_release Asc
)
)
)
CREATE TABLE extratbl(
    ext_table nvarchar(4) Not NULL,
    ext_code nvarchar(4) Not NULL,
    ext_value nvarchar(60) NULL,
    ext_value2 nvarchar(60) NULL,
    ext_status nvarchar(1) NULL,
    CONSTRAINT PK_extratbl PRIMARY KEY CLUSTERED
(
    ext_table Asc,
    ext_code Asc
)
)
)
CREATE TABLE genofred(
    genfd_popid decimal(8, 0) Not NULL,
    genfd_polyreg nvarchar(4) Not NULL,
    genfd_genotype decimal(8, 0) Not NULL,
    genfd_locus nvarchar(20) Not NULL,
    genfd_line_h decimal(4, 0) NULL,
    genfd_line decimal(4, 0) NULL,
    genfd_value nvarchar(2) NULL,
    genfd_status nvarchar(1) NULL,
    CONSTRAINT PK_genofred PRIMARY KEY CLUSTERED
(
    genfd_popid Asc,
    genfd_polyreg Asc,
    genfd_genotype Asc,
    genfd_locus Asc
)
)
)
CREATE TABLE genofreq(
    genf_popid decimal(8, 0) Not NULL,
    genf_polyreg nvarchar(10) Not NULL,
    genf_genotype decimal(8, 0) Not NULL,
    genf_line decimal(4, 0) NULL,
    genf_phenotype nvarchar(10) NULL,
    genf_genotype_b2 nvarchar(20) NULL,
    genf_genotype_b3 nvarchar(50) NULL,
    genf_freq decimal(8, 4) NULL,
    genf_added datetime NULL,
    genf_updated datetime NULL,
    genf_status nvarchar(1) NULL,
    genf_individuals decimal(9, 0) NULL,
    genf_sample_size decimal(9, 0) NULL,
    CONSTRAINT PK_genofreq PRIMARY KEY CLUSTERED
(
    genf_popid Asc,
    genf_polyreg Asc,
    genf_genotype Asc
)
)
)
CREATE TABLE geogreg(
    geo_code nvarchar(4) Not NULL,
    con_code nvarchar(4) NULL,
    geo_name nvarchar(25) NULL,
    geo_index nvarchar(6) NULL,
    geo_status nvarchar(1) NULL,
    CONSTRAINT PK_geogreg PRIMARY KEY CLUSTERED
(
    geo_code Asc
)
)
)
CREATE TABLE gracolor(
    gcol_gene nvarchar(20) Not NULL,
    gcol_range nvarchar(10) Not NULL,
    gcol_color nvarchar(10) NULL,
    gcol_color_value nvarchar(10) NULL,
    gcol_bar nvarchar(10) NULL,
    gcol_minv decimal(4, 0) NULL,
    gcol_maxv decimal(4, 0) NULL,
    CONSTRAINT PK_gracolor PRIMARY KEY CLUSTERED
(
    gcol_gene Asc,
    gcol_range Asc
)
)
)
CREATE TABLE gradient(
    gra_value decimal(4, 0) Not NULL,
    gra_min decimal(4, 0) NULL,
    gra_max decimal(4, 0) NULL,
    CONSTRAINT PK_gradient PRIMARY KEY CLUSTERED
(
    gra_value Asc
)
)
)

```

```

CREATE TABLE hallele(
  hla_allele nvarchar(20) Not NULL,
  hla_locus nvarchar(10) NULL,
  hla_level1 decimal(4, 0) NULL,
  hla_level2 decimal(4, 0) NULL,
  hla_level3 decimal(4, 0) NULL,
  hla_level4 decimal(4, 0) NULL,
  hla_order decimal(8, 0) NULL,
  CONSTRAINT PK_hallele PRIMARY KEY CLUSTERED
(
  hla_allele Asc
)
)

CREATE TABLE hallele2(
  hla2_allele nvarchar(20) Not NULL,
  hla2_allele2 nvarchar(20) NULL,
  hla2_locus nvarchar(10) NULL,
  hla2_length decimal(3, 0) NULL,
  hla2_source nvarchar(4) NULL,
  hla2_resol nvarchar(1) NULL,
  hla2_level1 decimal(4, 0) NULL,
  hla2_level2 decimal(4, 0) NULL,
  hla2_level3 decimal(4, 0) NULL,
  hla2_level4 decimal(4, 0) NULL,
  hla2_order decimal(8, 0) NULL,
  CONSTRAINT PK_hallele2 PRIMARY KEY CLUSTERED
(
  hla2_allele Asc
)
)

CREATE TABLE hapldet(
  hapd_popid decimal(8, 0) Not NULL,
  hapd_line decimal(4, 0) Not NULL,
  hapd_allele nvarchar(20) Not NULL,
  hapd_locus nvarchar(10) Not NULL,
  hapd_info nvarchar(20) NULL,
  CONSTRAINT PK_hapldet PRIMARY KEY CLUSTERED
(
  hapd_popid Asc,
  hapd_line Asc,
  hapd_allele Asc,
  hapd_locus Asc
)
)

CREATE TABLE haplfreq(
  hapf_popid decimal(8, 0) Not NULL,
  hapf_region nvarchar(10) Not NULL,
  hapf_line decimal(4, 0) Not NULL,
  hapf_haplotype nvarchar(100) Not NULL,
  hapf_haplotype2 nvarchar(100) Not NULL,
  hapf_freq decimal(8, 4) NULL,
  hapf_added datetime NULL,
  hapf_updated datetime NULL,
  hapf_user nvarchar(20) NULL,
  hapf_status nvarchar(1) NULL,
  hapf_loci decimal(4, 0) NULL,
  CONSTRAINT PK_haplfreq PRIMARY KEY CLUSTERED
(
  hapf_popid Asc,
  hapf_region Asc,
  hapf_line Asc,
  hapf_haplotype Asc,
  hapf_haplotype2 Asc
)
)

CREATE TABLE hladict(
  hlad_allele nvarchar(20) Not NULL,
  hlad_expert nvarchar(20) NULL,
  hlad_who nvarchar(20) NULL,
  hlad_status nvarchar(1) NULL,
  CONSTRAINT PK_hladict PRIMARY KEY CLUSTERED
(
  hlad_allele Asc
)
)

CREATE TABLE hlaprot(
  hlap_allele nvarchar(20) Not NULL,
  hlap_line decimal(8, 0) NULL,
  hlap_sequence nvarchar(Max) NULL,
  hlap_reference nvarchar(Max) NULL,
  CONSTRAINT PK_hlaprot PRIMARY KEY CLUSTERED
(
  hlap_allele Asc
)
)

CREATE TABLE kallele(
  kir_allele nvarchar(20) Not NULL,
  CONSTRAINT PK_kallele PRIMARY KEY CLUSTERED
(
  kir_allele Asc
)
)

CREATE TABLE kallele2(
  kir2_allele nvarchar(20) Not NULL,
  kir2_locus nvarchar(20) NULL,
  kir2_length decimal(3, 0) NULL,
  kir2_source nvarchar(4) NULL,
  kir2_resol nvarchar(1) NULL,
  CONSTRAINT PK_kallele2 PRIMARY KEY CLUSTERED
(
  kir2_allele Asc
)
)

CREATE TABLE kirgeno(
  kgen_id decimal(8, 0) Not NULL,
  kgen_3dl1 nvarchar(2) NULL,
  kgen_2dl1 nvarchar(2) NULL,
  kgen_2dl3 nvarchar(2) NULL,
  kgen_2ds4 nvarchar(2) NULL,
  kgen_2dl2 nvarchar(2) NULL,
  kgen_2dl5 nvarchar(2) NULL,
  kgen_3ds1 nvarchar(2) NULL,
  kgen_2ds1 nvarchar(2) NULL,
  kgen_2ds2 nvarchar(2) NULL,
  kgen_2ds3 nvarchar(2) NULL,
  kgen_2ds5 nvarchar(2) NULL,
  kgen_2dl4 nvarchar(2) NULL,
  kgen_3dl2 nvarchar(2) NULL,
  kgen_3dl3 nvarchar(2) NULL,
  kgen_2dp1 nvarchar(2) NULL,
  kgen_3dp1 nvarchar(2) NULL,
  kgen_group_a nvarchar(1) NULL,
  kgen_group_b nvarchar(1) NULL,
  kgen_type nvarchar(2) NULL,
  kgen_order nvarchar(20) NULL,
  kgen_status nvarchar(1) NULL,
  kgen_phenotype nvarchar(10) NULL,
  kgen_pops decimal(9, 0) NULL,
  kgen_individuals decimal(12, 0) NULL,
  kgen_a_genes decimal(4, 0) NULL,
  kgen_b_genes decimal(4, 0) NULL,
  kgen_pseudo_genes decimal(4, 0) NULL,
  kgen_total_genes decimal(4, 0) NULL,
  CONSTRAINT PK_kirgeno PRIMARY KEY CLUSTERED
(
  kgen_id Asc
)
)

CREATE TABLE locus(
  loc_region nvarchar(10) Not NULL,
  loc_name nvarchar(20) Not NULL,
  loc_chrom_name nvarchar(20) NULL,
  loc_chrom_number decimal(2, 0) NULL,
  loc_chrom_position nvarchar(1) NULL,
  loc_chrom_band decimal(3, 0) NULL,
  loc_chrom_subband decimal(3, 0) NULL,
  loc_chrom_ending nvarchar(4) NULL,
  loc_hla_class nvarchar(10) NULL,
  loc_hla_type nvarchar(20) NULL,
  loc_kir_domains decimal(2, 0) NULL,
  loc_kir_cyto_tail nvarchar(10) NULL,
  loc_kir_gene_number nvarchar(2) NULL,
  loc_kir_type nvarchar(10) NULL,
  loc_kir_code nvarchar(4) NULL,
  loc_order nvarchar(4) NULL,
  loc_status nvarchar(1) NULL,
  CONSTRAINT PK_locus PRIMARY KEY CLUSTERED
(
  loc_region Asc,
  loc_name Asc
)
)

CREATE TABLE mallele(
  mic_allele nvarchar(20) Not NULL,
  mic_locus nvarchar(10) NULL,
  mic_level1 decimal(4, 0) NULL,
  mic_level2 decimal(4, 0) NULL,
  mic_level3 decimal(4, 0) NULL,
  mic_level4 decimal(4, 0) NULL,
  mic_order decimal(8, 0) NULL,
  CONSTRAINT PK_mallele PRIMARY KEY CLUSTERED
(
  mic_allele Asc
)
)

CREATE TABLE mallele2(
  mic2_allele nvarchar(20) Not NULL,
  mic2_allele2 nvarchar(20) NULL,
  mic2_locus nvarchar(10) NULL,
  mic2_length decimal(4, 0) NULL,
  mic2_source nvarchar(4) NULL,
  mic2_resol nvarchar(1) NULL,
  mic2_level1 decimal(4, 0) NULL,
  mic2_level2 decimal(4, 0) NULL,
  mic2_level3 decimal(4, 0) NULL,
  mic2_level4 decimal(4, 0) NULL,
  mic2_order decimal(8, 0) NULL,
  CONSTRAINT PK_mallele2 PRIMARY KEY CLUSTERED
(
  mic2_allele Asc
)
)

```

```

)
CREATE TABLE metadata(
    met_type nvarchar(4) Not NULL,
    met_table nvarchar(30) Not NULL,
    met_number decimal(6, 0) Not NULL,
    met_name nvarchar(40) NULL,
    met_namedisp nvarchar(20) NULL,
    met_desc nvarchar(40) NULL,
    met_datatype nvarchar(20) NULL,
    met_length nvarchar(8) NULL,
    met_mask nvarchar(20) NULL,
    met_help nvarchar(40) NULL,
    met_common nvarchar(1) NULL,
    met_upper nvarchar(1) NULL,
    met_order decimal(6, 0) NULL,
    met_status nvarchar(1) NULL,
    CONSTRAINT PK_metadata PRIMARY KEY CLUSTERED
    (
        met_type Asc,
        met_table Asc,
        met_number Asc
    )
)

CREATE TABLE polyreg(
    pol_code nvarchar(10) Not NULL,
    pol_name nvarchar(20) NULL,
    pol_desc nvarchar(50) NULL,
    pol_index nvarchar(6) NULL,
    pol_status nvarchar(1) NULL,
    CONSTRAINT PK_polyreg PRIMARY KEY CLUSTERED
    (
        pol_code Asc
    )
)

CREATE TABLE poplocus(
    popl_poly nvarchar(20) Not NULL,
    popl_popid decimal(8, 0) Not NULL,
    popl_locus nvarchar(20) Not NULL,
    CONSTRAINT PK_poplocus PRIMARY KEY CLUSTERED
    (
        popl_poly Asc,
        popl_popid Asc,
        popl_locus Asc
    )
)

CREATE TABLE popshla(
    phla_popid decimal(8, 0) Not NULL,
    phla_population nvarchar(50) NULL,
    phla_a decimal(6, 0) NULL,
    phla_b decimal(6, 0) NULL,
    phla_cw decimal(6, 0) NULL,
    phla_drb1 decimal(6, 0) NULL,
    phla_dqb1 decimal(6, 0) NULL,
    phla_dqal decimal(6, 0) NULL,
    phla_dpai decimal(6, 0) NULL,
    phla_dpb1 decimal(6, 0) NULL,
    phla_others decimal(6, 0) NULL,
    phla_total decimal(6, 0) NULL,
    phla_hap decimal(6, 0) NULL,
    phla_a_p decimal(6, 0) NULL,
    phla_b_p decimal(6, 0) NULL,
    phla_cw_p decimal(6, 0) NULL,
    phla_drb1_p decimal(6, 0) NULL,
    phla_dqb1_p decimal(6, 0) NULL,
    phla_dqal_p decimal(6, 0) NULL,
    phla_dpai_p decimal(6, 0) NULL,
    phla_dpb1_p decimal(6, 0) NULL,
    phla_others_p decimal(6, 0) NULL,
    phla_total_p decimal(6, 0) NULL,
    CONSTRAINT PK_popshla PRIMARY KEY CLUSTERED
    (
        phla_popid Asc
    )
)

CREATE TABLE popskir(
    pkir_popid decimal(8, 0) Not NULL,
    pkir_population nvarchar(50) NULL,
    pkir_2dl1 decimal(6, 0) NULL,
    pkir_2dl2 decimal(6, 0) NULL,
    pkir_2dl3 decimal(6, 0) NULL,
    pkir_2dl4 decimal(6, 0) NULL,
    pkir_2dl5 decimal(6, 0) NULL,
    pkir_2dl5a decimal(6, 0) NULL,
    pkir_2dl5b decimal(6, 0) NULL,
    pkir_2ds1 decimal(6, 0) NULL,
    pkir_2ds2 decimal(6, 0) NULL,
    pkir_2ds3 decimal(6, 0) NULL,
    pkir_2ds4 decimal(6, 0) NULL,
    pkir_2ds5 decimal(6, 0) NULL,
    pkir_3ds1 decimal(6, 0) NULL,
    pkir_3dl1 decimal(6, 0) NULL,
    pkir_3dl2 decimal(6, 0) NULL,
    pkir_3dl3 decimal(6, 0) NULL,
    pkir_2dp1 decimal(6, 0) NULL,
    pkir_3dp1 decimal(6, 0) NULL,
    pkir_total decimal(6, 0) NULL,
    pkir_gen decimal(6, 0) NULL,
    CONSTRAINT PK_popskir PRIMARY KEY CLUSTERED
    (
        pkir_popid Asc
    )
)

CREATE TABLE populat(
    pop_id decimal(8, 0) Not NULL,
    pop_polyreg nvarchar(20) Not NULL,
    pop_name nvarchar(60) NULL,
    pop_name2 nvarchar(60) NULL,
    pop_country nvarchar(50) NULL,
    pop_geog_region nvarchar(25) NULL,
    pop_subregion nvarchar(50) NULL,
    pop_ethnic_origin nvarchar(50) NULL,
    pop_linguistic nvarchar(50) NULL,
    pop_center nvarchar(50) NULL,
    pop_origin nvarchar(50) NULL,
    pop_urban_rural nvarchar(50) NULL,
    pop_family nvarchar(50) NULL,
    pop_source nvarchar(50) NULL,
    pop_isolated nvarchar(1) NULL,
    pop_admixture nvarchar(1) NULL,
    pop_lat1_deg decimal(3, 0) NULL,
    pop_lat1_min decimal(2, 0) NULL,
    pop_lat1_card nvarchar(1) NULL,
    pop_long1_deg decimal(3, 0) NULL,
    pop_long1_min decimal(2, 0) NULL,
    pop_long1_card nvarchar(1) NULL,
    pop_coord_conf nvarchar(1) NULL,
    pop_size decimal(10, 0) NULL,
    pop_samplesize decimal(9, 0) NULL,
    pop_sampleyear nvarchar(6) NULL,
    pop_method_analysis nvarchar(20) NULL,
    pop_notes nvarchar(250) NULL,
    pop_verified nvarchar(1) NULL,
    pop_status nvarchar(1) NULL,
    pop_submitted_date datetime NULL,
    pop_submitted_by nvarchar(20) NULL,
    pop_reviewed_date datetime NULL,
    pop_reviewed_by nvarchar(20) NULL,
    pop_update datetime NULL,
    pop_alleles decimal(6, 0) NULL,
    pop_phenotypes decimal(6, 0) NULL,
    pop_genotypes decimal(6, 0) NULL,
    pop_haplotypes decimal(6, 0) NULL,
    pop_loci_test_hap decimal(6, 0) NULL,
    pop_main_author nvarchar(250) NULL,
    pop_reference1 nvarchar(250) NULL,
    pop_reference2 nvarchar(50) NULL,
    pop_reference3 nvarchar(250) NULL,
    pop_dataset nvarchar(20) NULL,
    pop_article nvarchar(250) NULL,
    pop_authors nvarchar(250) NULL,
    pop_correspondence nvarchar(100) NULL,
    pop_email nvarchar(50) NULL,
    pop_journal nvarchar(250) NULL,
    pop_year decimal(4, 0) NULL,
    pop_volumen nvarchar(10) NULL,
    pop_issue nvarchar(10) NULL,
    pop_pages nvarchar(20) NULL,
    pop_doi nvarchar(50) NULL,
    pop_pubmed nvarchar(20) NULL,
    pop_file_name nvarchar(100) NULL,
    pop_method nvarchar(25) NULL,
    pop_ssp nvarchar(1) NULL,
    pop_ssop nvarchar(1) NULL,
    pop_sscp nvarchar(1) NULL,
    pop_sbt nvarchar(1) NULL,
    pop_seg nvarchar(1) NULL,
    pop_rsca nvarchar(1) NULL,
    pop_rflp nvarchar(1) NULL,
    pop_oth nvarchar(1) NULL,
    pop_country_lab nvarchar(50) NULL,
    pop_dna_origin nvarchar(20) NULL,
    pop_kit nvarchar(50) NULL,
    pop_certify nvarchar(6) NULL,
    pop_reference4 nvarchar(250) NULL,
    pop_has_alleles nvarchar(1) NULL,
    pop_has_haplotypes nvarchar(1) NULL,
    pop_has_cytokines nvarchar(1) NULL,
    CONSTRAINT PK_populat PRIMARY KEY CLUSTERED
    (
        pop_id Asc,
        pop_polyreg Asc
    )
)

CREATE TABLE programs(
    prg_code nvarchar(8) Not NULL,
    prg_desc nvarchar(80) NULL,
    prg_desc2 nvarchar(Max) NULL,
    prg_desc3 nvarchar(80) NULL,
    prg_desc4 nvarchar(40) NULL,
    prg_desc5 nvarchar(40) NULL,
    prg_type nvarchar(4) NULL,
    prg_module nvarchar(20) NULL,
    prg_status nvarchar(1) NULL,
    prg_author nvarchar(8) NULL,
    prg_created datetime NULL,
    prg_last_update datetime NULL,
)

```

```

prg_user_update nvarchar(8) NULL,
prg_version nvarchar(10) NULL,
prg_display nvarchar(1) NULL,
prg_web nvarchar(1) NULL,
CONSTRAINT PK_programs PRIMARY KEY CLUSTERED
(
    prg_code Asc
)
)
CREATE TABLE rareall(
    rar_allele nvarchar(20) Not NULL,
    rar_release nvarchar(10) Not NULL,
    rar_web decimal(8, 0) NULL,
    rar_nmdp decimal(8, 0) NULL,
    rar_labs decimal(8, 0) NULL,
    rar_status nvarchar(1) NULL,
CONSTRAINT PK_rareall PRIMARY KEY CLUSTERED
(
    rar_allele Asc,
    rar_release Asc
)
)
CREATE TABLE releases(
    rel_code nvarchar(8) Not NULL,
    rel_name nvarchar(20) Not NULL,
    rel_date datetime NULL,
    rel_ftp nvarchar(100) NULL,
    rel_file nvarchar(100) NULL,
    rel_http nvarchar(100) NULL,
    rel_desc nvarchar(40) NULL,
    rel_total_rows nvarchar(20) NULL,
    rel_status nvarchar(1) NULL,
CONSTRAINT PK_releases PRIMARY KEY CLUSTERED
(
    rel_code Asc,
    rel_name Asc
)
)
CREATE TABLE resolut(
    res_parent nvarchar(20) Not NULL,
    res_child nvarchar(20) Not NULL,
    res_parent_size decimal(2, 0) NULL,
    res_child_size decimal(2, 0) NULL,
CONSTRAINT PK_resolut PRIMARY KEY CLUSTERED
(
    res_parent Asc,
    res_child Asc
)
)
CREATE TABLE securacc(
    sec_user nvarchar(20) Not NULL,
    sec_program nvarchar(8) Not NULL,
    sec_password nvarchar(20) NULL,
    sec_master_ins nvarchar(1) NULL,
    sec_master_del nvarchar(1) NULL,
    sec_master_mod nvarchar(1) NULL,
    sec_detail_ins nvarchar(1) NULL,
    sec_detail_del nvarchar(1) NULL,
    sec_detail_mod nvarchar(1) NULL,
    sec_print nvarchar(1) NULL,
    sec_apply nvarchar(1) NULL,
    sec_undo nvarchar(1) NULL,
    sec_open nvarchar(1) NULL,
    sec_cancel nvarchar(1) NULL,
    sec_lock nvarchar(1) NULL,
    sec_unlock nvarchar(1) NULL,
    sec_export nvarchar(1) NULL,
    sec_page1 nvarchar(1) NULL,
    sec_page2 nvarchar(1) NULL,
    sec_page3 nvarchar(1) NULL,
    sec_page4 nvarchar(1) NULL,
    sec_page5 nvarchar(1) NULL,
    sec_page6 nvarchar(1) NULL,
    sec_page7 nvarchar(1) NULL,
    sec_page8 nvarchar(1) NULL,
    sec_page9 nvarchar(1) NULL,
    sec_page10 nvarchar(1) NULL,
CONSTRAINT PK_securacc PRIMARY KEY CLUSTERED
(
    sec_user Asc,
    sec_program Asc
)
)
CREATE TABLE statshla(
    shla_locus nvarchar(20) Not NULL,
    shla_imgt decimal(9, 0) NULL,
    shla_website decimal(9, 0) NULL,
    shla_entries decimal(9, 0) NULL,
CONSTRAINT PK_statshla PRIMARY KEY CLUSTERED
(
    shla_locus Asc
)
)
CREATE TABLE statskir(
    skir_locus nvarchar(20) Not NULL,
    skir_ipd decimal(9, 0) NULL,
    skir_website decimal(9, 0) NULL,
    skir_entries decimal(9, 0) NULL,
CONSTRAINT PK_statskir PRIMARY KEY CLUSTERED
(
    skir_locus Asc
)
)
CREATE TABLE statsmic(
    smic_locus nvarchar(10) Not NULL,
    smic_imgt decimal(9, 0) NULL,
    smic_website decimal(9, 0) NULL,
    smic_entries decimal(9, 0) NULL,
CONSTRAINT PK_statsmic PRIMARY KEY CLUSTERED
(
    smic_locus Asc
)
)
CREATE TABLE statspop(
    spop_polyreg nvarchar(10) Not NULL,
    spop_pops decimal(9, 0) NULL,
    spop_alleles decimal(9, 0) NULL,
    spop_haplotypes decimal(9, 0) NULL,
    spop_genotypes decimal(9, 0) NULL,
    spop_family decimal(9, 0) NULL,
    spop_both decimal(9, 0) NULL,
CONSTRAINT PK_statspop PRIMARY KEY CLUSTERED
(
    spop_polyreg Asc
)
)
CREATE TABLE tblallele(
    tall_popid decimal(8, 0) Not NULL,
    tall_polyreg nvarchar(10) Not NULL,
    tall_locus nvarchar(10) Not NULL,
    tall_allele nvarchar(20) Not NULL,
    tall_line decimal(10, 0) NULL,
    tall_gene_freq nvarchar(10) NULL,
    tall_status nvarchar(1) NULL,
    tall_user nvarchar(15) NULL,
    tall_pheno_freq nvarchar(10) NULL,
    tall_added datetime NULL,
    tall_updated datetime NULL,
    tall_gene_has_data nvarchar(1) NULL,
    tall_gene_freq_num decimal(8, 4) NULL,
    tall_freq_has_data nvarchar(1) NULL,
    tall_pheno_freq_num decimal(8, 4) NULL,
    tall_has_data nvarchar(1) NULL,
    tall_indiv decimal(8, 0) NULL,
    tall_chrom decimal(8, 0) NULL,
    tall_sample_size decimal(8, 0) NULL,
    tall_notes nvarchar(250) NULL,
CONSTRAINT PK_tblallele PRIMARY KEY CLUSTERED
(
    tall_popid Asc,
    tall_polyreg Asc,
    tall_locus Asc,
    tall_allele Asc
)
)
CREATE TABLE tblcytokines(
    tcyt_popid decimal(8, 0) Not NULL,
    tcyt_gene nvarchar(20) Not NULL,
    tcyt_cytokine nvarchar(25) Not NULL,
    tcyt_genotype nvarchar(4) Not NULL,
    tcyt_line decimal(10, 0) NULL,
    tcyt_order decimal(8, 4) NULL,
    tcyt_freq decimal(8, 4) NULL,
    tcyt_added datetime NULL,
    tcyt_update datetime NULL,
    tcyt_status nvarchar(1) NULL,
    tcyt_user nvarchar(20) NULL,
    tcyt_indiv decimal(9, 0) NULL,
    tcyt_sample_size decimal(9, 0) NULL,
CONSTRAINT PK_tblcytokines PRIMARY KEY CLUSTERED
(
    tcyt_popid Asc,
    tcyt_gene Asc,
    tcyt_cytokine Asc,
    tcyt_genotype Asc
)
)
CREATE TABLE tblhaplotypes(
    thap_popid decimal(8, 0) Not NULL,
    thap_region nvarchar(20) Not NULL,
    thap_haplotype nvarchar(90) Not NULL,
    thap_line decimal(10, 0) NULL,
    thap_haplotype2 nvarchar(90) NULL,
    thap_frequency decimal(8, 4) NULL,
    thap_status nvarchar(1) NULL,
    thap_loci decimal(4, 0) NULL,
CONSTRAINT PK_tblhaplotypes PRIMARY KEY CLUSTERED
(
    thap_popid Asc,
    thap_region Asc,
    thap_haplotype Asc
)
)

```


Appendix B

Metadata and data dictionary

This section includes the description of the tables and views that were used in the implementation of the AFND.

Tables and Views

Table B 1: Tables and views in the AFND

Table	Description	Type	Records
alldesig	Catalogue of alleles for mapping old and new nomenclature	Catalogue	5,140
allefreq	Allele frequencies	Frequency data	92,506
alleles	Catalogue of alleles	Catalogue	6,891
allexlab	Rare alleles submitted by laboratories	Catalogue	1,182
allidend	Details of identical alleles	Catalogue	591
allidenr	Relationships of identical alleles	Catalogue	3,110
allident	Group of identical alleles	Catalogue	164
allseq	Amino acid sequences of alleles	Catalogue	6,186
allseqd	Details of amino acid sequences	Catalogue	21,135
allsynon	Synonymous alleles	Catalogue	15,114
celline	Catalogue of cell lines	Catalogue	132
codons	Catalogue of codons	Catalogue	64
continen	Catalogue of continents	Catalogue	7
country	Catalogue of countries	Catalogue	243
couxgeo	Countries by geographical region	Catalogue	243
cytofreq	Cytokine frequencies	Frequency data	3,603
cytokine	Catalogue of cytokines	Catalogue	189
ethnalle	Alleles by ethnic groups	Catalogue	3,193
ethnic	Catalogue of ethnic groups	Catalogue	24
ethxcou	Ethnic origins by country	View / Report	151
ethxgeo	Ethnic origins by geographical region	View / Report	53
event	Event monitor	Settings	70,393
ext_ashi_rares	Allele confirmations by ASHI	External data	764
ext_imgt_confirm	Allele confirmations by IMGT/HLA	External data	61,039
ext_ipd_kir	Allele confirmations by IPD-KIR	External data	1,645
ext_labs_rares	Allele confirmations by other laboratories	External data	585
ext_nmdp_rares	Allele confirmations by NMDP	External data	16,107
extratbl	Additional tables	Catalogue	110
genofred	Detail of genotype frequencies	Frequency data	41,776
genofreq	Genotype frequencies	Frequency data	2,600
geogreg	Catalogue of geographical regions	Catalogue	10
gracolor	Temporary table for colour gradients	Temporary table	40
gradient	Temporary table for general gradients	Temporary table	11
hallele	List of existing hla alleles	View / Report	4,108

Table B 2: Tables and views in the AFND (Continued)

Table	Description	Type	Records
hallele2	HLA alleles dictionary at different level of resolution	Catalogue	6,738
hapldet	Detail of haplotype frequencies	Frequency data	25,414
haplfreq	Haplotype frequencies	Frequency data	9,173
hladict	HLA dictionary	Catalogue	2,130
hlaprot	Coding sequence of HLA alleles	Catalogue	5,010
kallele	List of existing KIR alleles	View / Report	263
kallele2	KIR allele dictionary at different resolution	Catalogue	644
kirgeno	Catalogue of KIR genotypes	Catalogue	438
locus	Catalogue of locus	Catalogue	87
mallele	List of existing MIC alleles	View / Report	76
mallele2	MIC allele dictionary at different resolution	Catalogue	122
metadata	Catalogue of metadata	Catalogue	38
polyreg	Catalogue of polymorphic regions	Catalogue	8
poplocus	Locus typed by population	View / Report	2,600
popshla	HLA alleles by population	View / Report	836
popskir	KIR alleles by population	View / Report	188
populat	Catalogue of populations	Catalogue	1,212
programs	Catalogue of programs	Settings	49
rareall	Rare allele releases	Catalogue	1,700
releases	Catalogue of releases from external websites	Catalogue	32
resolut	Catalogue of alleles at low-resolution	Catalogue	7,914
securacc	User grants	Settings	26
statshla	HLA breakdowns	View / Report	34
statskir	KIR breakdowns	View / Report	18
statsmic	MIC breakdowns	View / Report	2
statspop	Population breakdowns	View / Report	4
tblallele	Kardex of allele frequencies	Kardex	92,488
tblcytokines	Kardex of cytokine frequencies	Kardex	3,603
tblhaplotypes	Kardex of haplotype frequencies	Kardex	9,173
tmpmatrix	Temporary table for differences in amino acids	Temporary table	120
tmpmatrix2	Temporary table for differences in alleles	Temporary table	1,000
tmpmatrix3	Temporary table for differences in populations	Temporary table	100
tmptab02	Temporary table for frequency entries	Temporary table	6,890
tsession	User sessions	Settings	1
typehla	HLA typing data	Catalogue	1,132
users	Catalogue of users	Catalogue	5,074
usrconf	User settings by program	Settings	8
websetup	Website settings	Settings	1

Data dictionary

Table B 3: Data dictionary of the AFND

Table	Attribute	Description	Example
alldesig	alld_allele	Allele name according to previous nomenclature	A*01010101
	alld_type	Type of generation (Automatic or manual)	AUT1, AUT2, AUT3, MAN1
	alld_code	Allele name	A*01:01:01:01
	alld_polyreg	Polymorphic region	HLA, KIR, MIC, Cyt
allefreq	allf_popid	Population ID	1212
	allf_polyreg	Polymorphic region	HLA, KIR, MIC, Cyt
	allf_locus	Locus name	A, B, C, DRB1, ...
	allf_allele	Allele name	A*01:01:01:01, ...
	allf_line	Consecutive number	1, 2, 3, ...
	allf_gene_freq	Allele frequency	0.0001, ...
	allf_pheno_freq	Phenotype frequency	18.50, ...
	allf_added	Date added	01/01/2010
	allf_updated	Last update	01/01/2010
	allf_user	User who entered the frequency	...
	allf_status	Status	A=Active, P=Pending for submission
	allf_indiv	Number of individuals who carry the allele	50
	allf_chrom	Number of copies with the allele	20
	allf_sample_size	Sample size	200
allf_notes	Additional notes	...	
alleles	all_name	Allele name	A*01:01:01:01
	all_name2	Name of the allele according to previous nomenclature	A*01010101
	all_region	Polymorphic region	HLA, KIR, MIC, Cyt
	all_locus	Locus name	A, B, C, DRB1, ...
	all_family	Allele family	01, 02, 03, ...
	all_ethnic_origin	Ethnic origin(s) in which the allele was found in	Oriental - Taiwan, Asia, ...
	all_class	HLA Class	Class I, Class II, ...
	all_common	Common allele in USA (ASHI)	C = Common
	all_documented	Well-documented allele (ASHI)	WD = Well-documented
	all_cells	Number of cells/sources that have been sequenced	1, 2, 3, ...
	all_groups	Number of groups which have reported the allele	1, 2, 3, ...
	all_confirmed	Confirmation of the allele by groups (IMGT/HLA)	Confirmed / Unconfirmed
	all_start	Start position of CDS	1
	all_end	End position of CDS	200
	all_partial	CDS confirmed	Partial / Full
	all_updated	Last update	01/01/2010

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
alleles	all_pops	Number of pops in which the allele has been reported	1
	all_total_nmdp	Number of times in NMDP	1
	all_total_labs	Number of times in other laboratories	1
	all_total_identical_cell	Number of cells considering identical alleles	1
	all_total_identical_group	Number of groups considering identical alleles	1
	all_total_identical_pops	Number of times in website including identical alleles	1
	all_total_identical_nmdp	Number of times in NMDP including identical alleles	1
	all_total_identical_labs	Number of times in labs including identical alleles	1
	all_length	Allele length	4, 6, 8, ...
	all_status	Allele status	A=Active, D=Deleted
	all_year	Year in which the allele was first reported	1990, 2000, ...
	all_levels	Level of resolution	1, 2, 3, 4
	all_level1	Value first level	1
	all_level2	Value second level	1
	all_level3	Value third level	1
	all_level4	Value fourth level	1
	all_clevel1	Code at first level	01
	all_clevel2	Code at second level	01
	all_clevel3	Code at third level	01
	all_clevel4	Code at fourth level	01
all_order	Order	1, 2, 3, ...	
allexlab	axlab_allele	Allele name	A*01:01:01:01
	axlab_indiv_id	ID of individual possessing the allele	HAN-1002, ...
	axlab_user	User who submitted the allele	...
	axlab_ethnicity	Ethnic origin	Caucasoid, ...
	axlab_population	Population name	Brazilian, ...
	axlab_indiv_type	Individual type	Bone Marrow Patient, ...
	axlab_indiv_type_registry	Registry name	Brazilian Bone Marrow, ...
	axlab_indiv_type_other	Specification for other association	Coord blood donor
	axlab_haplotype	Haplotype confirmation	Confirmed by family studies
	axlab_notes	Additional notes	...
	axlab_other_member	Family member with a similar rare allele	Sibling of HAN-1002
	axlab_indiv_same_resol	Individuals typed at the same level of resolution	1, 2, 3, ...
	axlab_methods	Methods used to detect allele	SBT, SSOP, ...
	axlab_sbt	SBT method	Y/N
	axlab_ssp	SSP method	Y/N
axlab_sso	SSO method	Y/N	
axlab_seq	Sequencing method	Y/N	
axlab_oth	Other typing method	Y/N	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
allexlab	axlab_type	Phenotype of the individual	A*01:01:01:02N / 01:01:01:01; B*35:03/44:03:02; DRB1*07/13:64
	axlab_phenotype	Phenotype of the individual validated	A*01:01:01:02N / 01:01:01:01; B*35:03/44:03:02; DRB1*07/13:64
	axlab_haplotype1	First haplotype of the individual	A*01:01:01:02N- B*35:03- DRB1*07
	axlab_haplotype2	Second haplotype of the individual	A*01:01:01:01- B*44:03:02- DRB1*16:64
	axlab_hap1_hla_a	Haplotype 1 HLA-A	A*01:01:01:02N
	axlab_hap2_hla_a	Haplotype 2 HLA-A	A*01:01:01:01
	axlab_hap1_hla_b	Haplotype 1 HLA-B	B*35:03
	axlab_hap2_hla_b	Haplotype 2 HLA-B	B*44:03:02
	axlab_hap1_hla_c	Haplotype 1 HLA-C	
	axlab_hap2_hla_c	Haplotype 2 HLA-C	
	axlab_hap1_hla_dra	Haplotype 1 HLA-DRA	
	axlab_hap2_hla_dra	Haplotype 2 HLA-DRA	
	axlab_hap1_hla_drb1	Haplotype 1 HLA-DRB1	DRB1*07
	axlab_hap2_hla_drb1	Haplotype 2 HLA-DRB1	DRB1*16:64
	axlab_hap1_hla_dqa1	Haplotype 1 HLA-DQA1	
	axlab_hap2_hla_dqa1	Haplotype 2 HLA-DQA1	
	axlab_hap1_hla_dqb1	Haplotype 1 HLA-DQB1	
	axlab_hap2_hla_dqb1	Haplotype 2 HLA-DQB1	
	axlab_hap1_hla_dpa1	Haplotype 1 HLA-DPA1	
	axlab_hap2_hla_dpa1	Haplotype 2 HLA-DPA1	
	axlab_hap1_hla_dpb1	Haplotype 1 HLA-DPB1	
	axlab_hap2_hla_dpb1	Haplotype 2 HLA-DPB1	
	axlab_date_added	Date added	01/01/2010
	axlab_line	Counter	1, 2, 3, ...
	axlab_status	Status	A=Active, P=Pending
	axlab_authors	Contributor	Name
	axlab_add1	Address 1	Address
axlab_add2	Address 2	Address	
axlab_city	City of contributor	City	
axlab_county	County of contributor	County	
axlab_country	Country of contributor	Country	
axlab_postcode	Postcode	Postcode	
axlab_fax	Fax of contributor	Fax	
axlab_email	Email of contributor	Email	
axlab_ihw	Participant of the 15 IHWS	Y/N	
allidend	idd_code	Internal code for identical alleles	A-G01
	idd_allele	Allele name	A*01:01:01:01
	idd_status	Status of the allele	A=Active
allidenr	idr_allele	Allele name	A*01:01:01:01
	idr_identical	Identical allele	A*01:01:01:02N
allident	ide_code	Internal code for identical alleles	A-G01
	ide_name	Allele name	A*01:01:01:01
	ide_status	Status of the allele	A=Active
allseq	aseq_allele	Allele name	A*01:01:01:01
	aseq_protein	Protein sequence	MAVMAPR....
	aseq_genomic	Genomic sequence	GCCAGGCC....
	aseq_size	Size of the coding sequence	3503, ...

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
allseqd	aseqd_allele	Allele name	A*01:01:01:01
	aseqd_line	Line number	1,2,3...
	aseqd_type	Type of sequence	EX=Exon, ...
	aseqd_from	Initial position	1
	aseqd_to	End position of CDS	246
	aseqd_order	Order	1, 2, 3, ...
allsynon	asyn_parent	Parent allele	DQB1*03:01:03
	asyn_child	Child allele	DQB1*01:01:01
	asyn_parent_size	Size of the parent allele	3
	asyn_child_size	Size of the child allele	3
cellline	cell_id	Cell line id	1
	cell_dna_no	DNA code of the cell line	AMAI
	cell_ihw_no	Code of the IH Workshop	9010
	cell_ipd_code	Code in IPD-KIR database	10246
	cell_consang	Consanguineous	C=Consanguineous
	cell_2dl1_1	Allele present in the gene	002
	cell_2dl1_2	Allele present in the gene	001v
	cell_2dp1_1	Allele present in the gene	+
	cell_2dl3_1	Allele present in the gene	002
	cell_2dl3_2	Allele present in the gene	005
	cell_2ds4_1	Allele present in the gene	003
	cell_2ds4_2	Allele present in the gene	
	cell_3dl1_1	Allele present in the gene	00101
	cell_3dl1_2	Allele present in the gene	
	cell_3dl1_1_s	Allele present in the gene	
	cell_3dl1_2_s	Allele present in the gene	
	cell_3ds1_1	Allele present in the gene	-
	cell_3ds1_1_s	Allele present in the gene	
	cell_2ds1_1	Allele present in the gene	-
	cell_2ds2_1	Allele present in the gene	
	cell_2ds3_1	Allele present in the gene	-
	cell_2ds3_2	Allele present in the gene	
	cell_2ds5_1	Allele present in the gene	-
	cell_2dl2_1	Allele present in the gene	-
	cell_2dl2_2	Allele present in the gene	
	cell_2dl5a_1	Allele present in the gene	-
	cell_2dl5a_2	Allele present in the gene	
	cell_2dl5a_s	Allele present in the gene	
	cell_2dl5b_1	Allele present in the gene	-
	cell_2dl5b_2	Allele present in the gene	
	cell_2dl5b_s	Allele present in the gene	
	cell_2dl4_1	Allele present in the gene	00202
	cell_2dl4_2	Allele present in the gene	
cell_3dl2_1	Allele present in the gene	001	
cell_3dl2_2	Allele present in the gene		
cell_3dl3_1	Allele present in the gene	+	
cell_3dp1_1	Allele present in the gene	+	
cell_status	Cell line status	A=Active	
codons	cod_id	Codon id	1
	cod_aminoacid	Amino acid name	Phenylalanine
	cod_letter	Amino acid code	F
	cod_3letter	Amino acid three letter code	Phe
	cod_rna	Rna equivalence	UUU
	cod_dna	Dna equivalence	TTT
	cod_polarity	Polarity	Non polar
continen	con_code	Continent code	AFR
	con_name	Continent name	Africa
	con_index	Continent order	1

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
continen	con_status	Continent status	A=Active
country	cou_code	Country code	ARG
	geo_code	Geographical region code	SCAM
	cou_name	Country name	Argentina
	cou_code2	Country code two digits	AR
	cou_number	Country UN number	32
	cou_name2	Country name in caps	ARGENTINA
	cou_tld	Country internet ending	.ar
	cou_timezone	Country time zone	GMT - 3 hrs
	cou_phone_code	Country phone code	54
	cou_population	Country population	40301927
	cou_pop_date	Country population last update	15/10/2008
	cou_pop_source	Country population source	UN estimate
	cou_area	Country area	2780400
	cou_area_date	Country area last update	15/10/2008
	cou_area_source	Country area source	
	cou_capital	Country capital	Buenos Aires
	cou_latitude	Country latitude	34°36 S
	cou_longitude	Country longitude	58°22 W
	cou_status	Country status	A=Active, I=Inactive
	couxgeo	cogr_geo_region_code	Geographical region code
cogr_country_code		Country code	AAE
cogr_geo_region_name		Geographical region name	Asia
cogr_country_name		Country name	Myanmar
cogr_status		Relationship status	A
cytofreq	cytf_popid	Population ID	1366
	cytf_gene	Locus or gene	IL-10/
	cytf_cytokine	Cytokine name	IL-10/ - 1082 AA
	cytf_line	Row number	1, 2, 3...
	cytf_freq	Phenotype frequency	19.4000
	cytf_added	Date added	01/01/2008
	cytf_updated	Last update	01/01/2008
	cytf_user	Username	...
	cytf_order	Order	1085.1000
	cytf_status	Status	A=Active, I=Inactive
	cytf_indiv	Number of individuals who carry the allele	19
	cytf_sample_size	Sample size of the population	100
	cytokine	cyt_name	Cytokine name
cyt_gene		Cytokine gene	AIF-1/
cyt_group		Cytokine group	CC
cyt_updated		Cytokine update	01/02/2008
cyt_index		Cytokine order or index	1200.1
cyt_status		Cytokine status	A
ethnalle	etha_allele	Allele name	A*01:01:01:01
	etha_ethnic_origin	Ethnic origins where the allele has been seen	Caucasoid, ...
	etha_first	First ethnic origin sequenced	Caucasoid
ethnic	eth_code	Ethnic origin code	ASTN
	eth_name	Ethnic origin name	Austronesian
	eth_status	Ethnic origin status	A
	eth_reference	Ethnic origin reference	Pacific Islanders
ethxcou	ethc_country_code	Country code	AAH
	ethc_ethnic_code	Ethnic group code	CAUC
	ethc_country_name	Country name	Serbia
	ethc_ethnic_name	Ethnic group name	Caucasoid
	ethc_status	Relationship status	A

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
ethxgeo	ethg_geo_region_code	Geographical region code	AUST
	ethg_ethnic_code	Ethnic group code	AUST
	ethg_geo_region_name	Geographical region name	Australia
	ethg_ethnic_name	Ethnic group name	Australian Aboriginal
	ethg_status	Relationship status	A
event	eve_id	Monitor id number	19671
	eve_event	Monitor event	LOG
	eve_description	Monitor description	Login to website
	eve_user	Monitor user	...
	eve_program	Monitor program	Main
	eve_date	Monitor date	25/03/2009
	eve_time	Monitor time	102346
	eve_extra	Monitor extra comments	Login...
ext_ashi_rares	ashi_allele	Allele name	A*01:01:01:01
	ashi_release	Release	1.0.0
	ashi_common	Flag to identify a common allele	C=Common
	ashi_documented	Flag to identify a well-documented allele	WD=Well-documented
	ashi_update	Release	25/09/2008
ext_imgt_confirm	imgt_allele	Allele name	B*35:31
	imgt_release	Release	2.28.0
	imgt_cells	Number of cells/sources which have been sequenced the allele	3
	imgt_groups	Number of submitters who have sequenced the allele	2
	imgt_confirmed	Confirmation for the allele	Confirmed
	imgt_start	Nucleotide start position of the sequence compared to the CDS of the reference sequence	1
	imgt_end	Nucleotide end position of the sequence compared to the CDS of the reference sequence	1089
	imgt_partial	CDS sequenced	Full
	imgt_ethnicity	First ethnic origin sequenced	Caucasoid
	imgt_update	Last update	17/02/2010
ext_ipd_kir	ipd_allele	Allele name	2DL1*001
	ipd_release	Release	2.0.0
	ipd_update	Last update	31/07/2008
ext_labs_rares	labs_allele	Allele name	A*01:01:01:02N
	labs_bla	Total for black ethnicity	0
	labs_cau	Total for caucasian ethnicity	0
	labs_mes	Total for mestizo ethnicity	0
	labs_his	Total for hispanic ethnicity	0
	labs_oth	Total for other ethnicity	1
	labs_total	Total alleles	1
ext_nmdp_rares	nmdp_allele	Allele name	B*15:94N
	nmdp_release	Release	2.16.0
	nmdp_afa	African american	0
	nmdp_api	Asian or pacific islanders	0
	nmdp_cau	Caucasian	0
	nmdp_his	Hispanic race	0
	nmdp_nam	Native american	0
	nmdp_oth	Multiple, unknown or decline	0
	nmdp_total	Total	0
nmdp_rare	Rare allele	X	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
ext_nmdp_rares	nmdp_zero	Absent confirmation	X
	nmdp_update	Last update	01/01/2007
extratbl	ext_table	Extra table name	JOU
	ext_code	Extra table code	0001
	ext_value	Extra table value 1	Transplant
	ext_value2	Extra table value 2	Immunology
	ext_status	Extra table status	Additional comment A=Active, I=Inactive
genofred	genfd_popid	Population ID	1618
	genfd_polyreg	Polymorphic region	KIR
	genfd_genotype	Genotype ID	9
	genfd_locus	Locus	2DL1
	genfd_line_h	Row number (header)	6
	genfd_line	Row number	2
	genfd_value	Type	1
	genfd_status	Status	A=Active, I=Inactive
genofreq	genf_popid	Population ID	2444
	genf_polyreg	Polymorphic region	KIR
	genf_genotype	Genotype ID	112
	genf_line	Row number	9
	genf_phenotype	Phenotype code	
	genf_genotype_b2	Genotype (base = 2)	
	genf_genotype_b3	Genotype (base = 3)	
	genf_freq	Genotype frequency	0.6
	genf_added	Date added	06/06/2008
	genf_updated	Last update	06/06/2008
	genf_status	Status	A=Active, I=Inactive
	genf_individuals	Number of individuals who carry the genotype	1
	genf_sample_size	Sample size of the population	166
geogreg	geo_code	Geographical region code	ASIA
	con_code	Continent	ASI
	geo_name	Geographical region name	Asia
	geo_index	Geographical region index	01
	geo_status	Geographical region status	01
gracolor	gcol_gene	Name of gene to draw	2DL1
	gcol_range	Range	60-79
	gcol_color	Name of colour	P
	gcol_color_value	Colour value	#993399
	gcol_bar	Type of bar	6
	gcol_minv	Minimum value	60
	gcol_maxv	Maximum value	80
gradient	gra_value	Gradient value	1
	gra_min	Minimum value	1
	gra_max	Maximum value	10
hallele	hla_allele	Allele Name	A*01:01:01:01
	hla_locus	Locus name	A, B, C, DRB1, ...
	hla_level1	First level	1, 2, 15, ...
	hla_level2	Second level	1, 2, 15, ...
	hla_level3	Third level	1, 2, 15, ...
	hla_level4	Fourth level	1, 2, 15, ...
hallele2	hla_order	Order	1, 2, 3, ...
	hla2_allele	Allele Name	A*01:01:01:01
	hla2_allele2	Last nomenclature	A*01010101
	hla2_locus	Locus name	A, B, C, DRB1, ...
	hla2_length	Level of digits of resolution	1, 2, 3, 4
hla2_source	Source of generation	MAN1, AUT1 ... AUT4	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
hallele2	hla2_resol	Level of resolution	H = High, L = Low
	hla2_level1	First level	1, 2, 15, ...
	hla2_level2	Second level	1, 2, 15, ...
	hla2_level3	Third level	1, 2, 15, ...
	hla2_level4	Fourth level	1, 2, 15, ...
	hla2_order	Order	1, 2, 3, ...
hapldet	hapd_popid	Population ID	1216
	hapd_line	Reference line to haplotypes	1
	hapd_allele	Allele name	A*02:01
	hapd_locus	Locus name	A
	hapd_info	Allele variant	02:01
haplfreq	hapf_popid	Population ID	1481
	hapf_region	Polymorphic region	HLA
	hapf_line	Row number	52
	hapf_haplotype	Haplotype	B*50:01-C*06:02
	hapf_haplotype2	Past nomencl Haplotype	B*5001-Cw*0602
	hapf_freq	Haplotype frequency	2.1000
	hapf_added	Date added	01/02/2008
	hapf_updated	Last update	01/02/2008
	hapf_user	Username	...
	hapf_status	Status	A
hapf_loci	Number of loci	2	
hladict	hlad_allele	Allele name	A*01:01
	hlad_expert	Expert assigned type	A1
	hlad_who	WHO assigned type	A1
	hlad_status	Status	A
hlaprot	hlap_allele	Allele name	A*24:02:31
	hlap_line	Line number	843
	hlap_sequence	Protein sequence	-----V-...
	hlap_reference	Reference sequence	MAVMAPRT...
kallele	kir_allele	Allele name for all resolutions	2DL1*001
kallele2	kir2_allele	Allele name	2DL1*001
	kir2_locus	Locus name	2DL1
	kir2_length	Length of the allele	3
	kir2_source	Source of generation	AUT1
	kir2_resol	Level of resolution	L=Low, H=High
kirgeno	kgen_id	KIR genotype ID	1
	kgen_3dl1	Gene value	1
	kgen_2dl1	Gene value	1
	kgen_2dl3	Gene value	1
	kgen_2ds4	Gene value	1
	kgen_2dl2	Gene value	0
	kgen_2dl5	Gene value	0
	kgen_3ds1	Gene value	0
	kgen_2ds1	Gene value	0
	kgen_2ds2	Gene value	0
	kgen_2ds3	Gene value	0
	kgen_2ds5	Gene value	0
	kgen_2dl4	Gene value	1
	kgen_3dl2	Gene value	1
	kgen_3dl3	Gene value	1
	kgen_2dp1	Gene value	1
	kgen_3dp1	Gene value	1
	kgen_group_a	Belongs to group A	1
	kgen_group_b	Belongs to group B	0
	kgen_type	KIR genotype type	AA
	kgen_order	KIR genotype order	1
	kgen_status	KIR genotype status	A=Active, I=Inactive

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
kirgeno	kgen_phenotype	KIR phenotype	A
	kgen_pops	Pops with this genotype	108
	kgen_individuals	Number of individuals who carry the genotype	3686
	kgen_a_genes	Number of A genes present in the genotype	4
	kgen_b_genes	Number of B genes present in the genotype	0
	kgen_pseudo_genes	Number of Pseudogenes present in the genotype	2
	kgen_total_genes	Total genes in the genotype	9
locus	loc_region	Polymorphic region code	HLA
	loc_name	Locus name	A
	loc_chrom_name	Locus chromosome name	
	loc_chrom_number	Locus chromosome number	6
	loc_chrom_position	Locus position on chromosome	6.23
	loc_chrom_band	Locus chromosome band	
	loc_chrom_subband	Locus chromosome sub-band	
	loc_chrom_ending	Locus chromosome ending	
	loc_hla_class	Locus HLA class	Class I
	loc_hla_type	Locus HLA type	Classical
	loc_kir_domains	Locus KIR domains	2
	loc_kir_cyto_tail	Locus KIR cytoplasmic tail	Large
	loc_kir_gene_number	Locus KIR gene number	1
	loc_kir_type	Locus KIR type	Type I
	loc_kir_code	Locus KIR code	A
	loc_order	Locus order	02
	loc_status	Locus status	A=Active, I=Inactive
mallele	mic_allele	MIC allele name	MICA*001
	mic_locus	Locus name	MICA, MICB
	mic_level1	First level	1, 2, 15, ...
	mic_level2	Second level	1, 2, 15, ...
	mic_level3	Third level	1, 2, 15, ...
	mic_level4	Fourth level	1, 2, 15, ...
	mic_order	Order	1, 2, 3, ...
mallele2	mic2_allele	Allele Name	MICA*001:01
	mic2_allele2	Last nomenclature	MICA*00101
	mic2_locus	Locus name	A, B, C, DRB1, ...
	mic2_length	Level of digits of resolution	1, 2, 3, 4
	mic2_source	Source of generation	MAN1, AUT1 ... AUT4
	mic2_resol	Level of resolution	H = High, L = Low
	mic2_level1	First level	1, 2, 15, ...
	mic2_level2	Second level	1, 2, 15, ...
	mic2_level3	Third level	1, 2, 15, ...
	mic2_level4	Fourth level	1, 2, 15, ...
	mic2_order	Order	1, 2, 3, ...
metadata	met_type	Metadata type	FLD=Field
	met_table	Table Name	kirgeno
	met_number	Entry order	1
	met_name	Metadata name	kgen_id
	met_namedisp	Name to display	Genotype ID
	met_desc	Metadata description	KIR genotype ID
	met_datatype	Field datatype	Numeric
	met_length	Field length	8,0
	met_mask	Field mask	zzzzzzz9
	met_help	Field help	Comments

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
metadata	met_common	Display as a common field	Y
	met_upper	Field to uppercase	N
	met_order	Field order	2
	met_status	Field status	A=Active, I=Inactive
polyreg	pol_code	Polymorphic region code	001
	pol_name	Polymorphic region name	HLA
	pol_desc	Polymorphic region description	Human Leukocyte Antigens
	pol_index	Polymorphic region index	01
	pol_status	Polymorphic region status	A
poplocus	popl_poly	Polymorphic region	HLA
	popl_popid	Population id	1212
	popl_locus	Locus name	A
popshla	phla_popid	Population id	1212
	phla_population	Population name	Scotland Orkney
	phla_a	Number of alleles in locus	39
	phla_b	Number of alleles in locus	40
	phla_c	Number of alleles in locus	19
	phla_drb1	Number of alleles in locus	26
	phla_dqa1	Number of alleles in locus	0
	phla_dqb1	Number of alleles in locus	9
	phla_dpa1	Number of alleles in locus	0
	phla_dpb1	Number of alleles in locus	0
	phla_others	Number of alleles in locus	0
	phla_total	Number of alleles in locus	133
	phla_hap	Number of alleles in locus	0
	phla_a_p	Number of alleles in locus	13
	phla_b_p	Number of alleles in locus	20
	phla_c_p	Number of alleles in locus	11
	phla_drb1_p	Number of alleles in locus	26
	phla_dqb1_p	Number of alleles in locus	0
	phla_dqa1_p	Number of alleles in locus	9
	phla_dpa1_p	Number of alleles in locus	0
	phla_dpb1_p	Number of alleles in locus	0
phla_others_p	Number of alleles in locus	0	
phla_total_p	Number of alleles in population	79	
popskir	pkir_popid	Population id	1429
	pkir_population	Population name	Ireland Northern KIR
	pkir_2dl1	Number of alleles in gene	1
	pkir_2dl2	Number of alleles in gene	1
	pkir_2dl3	Number of alleles in gene	7
	pkir_2dl4	Number of alleles in gene	10
	pkir_2dl5	Number of alleles in gene	1
	pkir_2dl5a	Number of alleles in gene	0
	pkir_2dl5b	Number of alleles in gene	0
	pkir_2ds1	Number of alleles in gene	1
	pkir_2ds2	Number of alleles in gene	1
	pkir_2ds3	Number of alleles in gene	1
	pkir_2ds4	Number of alleles in gene	6
	pkir_2ds5	Number of alleles in gene	1
	pkir_3ds1	Number of alleles in gene	6
	pkir_3dl1	Number of alleles in gene	12
	pkir_3dl2	Number of alleles in gene	13
	pkir_3dl3	Number of alleles in gene	7
	pkir_2dp1	Number of alleles in gene	1
pkir_3dp1	Number of alleles in gene	1	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
popskir	pkir_total	Number of alleles in population	64
	pkir_genotype	Number of genotypes in the population	26
populat	pop_id	Population id	2000
	pop_polyreg	Polymorphic region	HLA, KIR, ...
	pop_name	Population name	Scotland Orkney, ...
	pop_name2	Population name	Orkneys, ...
	pop_country	Population country	Scotland
	pop_geo_region	Population geographical region	Western Europe, ...
	pop_subregion	Population sub geographical region	Western Europe, ...
	pop_ethnic_origin	Population ethnic origin	Caucasoid
	pop_linguistic	Population linguistic family	Caucasoid
	pop_center	Population linguistic family	Caucasoid
	pop_origin		
	pop_urban_rural	Population type (Urban/Rural)	Urban, Rural, Urban and Rural
	pop_family	Population family	Parents live at same location, ...
	pop_source	Source of population	Anthropology study, ...
	pop_isolated	Population was isolated	Y/N
	pop_admixture	Population contains admixture	Y/N
	pop_lat1_deg	Latitude degrees	58
	pop_lat1_min	Latitude minutes	59
	pop_lat1_card	Latitude orientation (Cardinal position)	N
	pop_long1_deg	Longitude degrees	3
	pop_long1_min	Longitude minutes	11
	pop_long1_card	Longitude orientation (Cardinal position)	W
	pop_coord_conf	Confirmation of coordinate	Y, N
	pop_size	Population size, if known	10 000
	pop_samplesize	Population sample size	99
	pop_sampleyear	Population sample year	2000
	pop_method_analysis	Type of method used for the estimation	Direct counting, resampling, etc.
	pop_notes	Population notes	Additional comments
	pop_verified	Flag to indicate if population was verified	Y, N
	pop_status	Population status	A = Active, P = Pending
	pop_submitted_date	Population submitted date	04/07/2002
	pop_submitted_by	Population submitted by	Username
	pop_reviewed_date	Population reviewed date	07/10/2008
pop_reviewed_by	Population reviewed by	Username	
pop_update	Last update	07/10/2008	
pop_alleles	Number of alleles in the population	1, 2, 3, ...	
pop_phenotypes	Number of phenotypes in the population	1, 2, 3, ...	
pop_genotypes	Number of genotypes in the population	1, 2, 3, ...	
pop_haplotypes	Number of haplotypes in the population	1, 2, 3, ...	
pop_loci_test_hap	Number of loci tested for haplotypes	1, 2, 3, ...	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example	
populat	pop_main_author	Population author	Bodmer	
	pop_reference1	Population reference 1	Bodmer et al	
	pop_reference2	Population reference 2	E-mail	
	pop_reference3	Population reference 3	Not able to distinguish...	
	pop_dataset	Type of dataset	Literature	
	pop_article	Title of the publication	Title	
	pop_authors	Authors included in the publication	Authors	
	pop_correspondence	Correspondence author	Bodmer	
	pop_email	E-mail corresponding author	E-mail	
	pop_journal	Journal of the publication	Journal name	
	pop_year	Year of publication	2002	
	pop_volume	Volume of Journal	69	
	pop_issue	Issue of Journal	2	
	pop_pages	Pages of the publication	12-17	
	pop_doi	Digital object identifier	Identifier	
	pop_pubmed	Pubmed reference	Reference	
	pop_file_name	Path for electronic version	\\path_file	
	pop_method	Method used	SSOP, SSP, ...	
	pop_ssp	Typed by SSP?	Y, N	
	pop_ssop	Typed by SSOP?	Y, N	
	pop_sscp	Typed by SSCP?	Y, N	
	pop_sbt	Typed by SBT?	Y, N	
	pop_seq	Typed by Sequencing?	Y, N	
	pop_rscs	Typed by RSCA?	Y, N	
	pop_rflp	Typed by RFLP?	Y, N	
	pop_oth	Typed by other method?	Y, N	
	pop_country_lab	Country in which the individuals were typed		
	pop_dna_origin	Source of DNA		
	pop_kit	Kit used for typing		
	pop_certify	Population certification (Y/N)	Y, N	
	pop_reference4	Population reference 4		
	pop_has_alleles	Flag to indicate if pop has alleles	Y, N	
	pop_has_haplotypes	Flag to indicate if pop has haplotypes	Y, N	
	pop_has_cytokines	Flag to indicate if pop has cytokines	Y, N	
	programs	prg_code	Program code	hla0001a
		prg_desc	Program description	HLA allele catalogue
		prg_desc2	Large program description	Program to add...
		prg_desc3	Program keywords	HLA, allele, ...
		prg_desc4	Program description	...
		prg_desc5	Program description	...
prg_type		Program type	CAT	
prg_module		Program module	HLA	
prg_status		Program status	A=Active, I=Inactive	
prg_author		Program author	Username	
prg_created		Program start date	07/01/2008	
prg_last_update		Program last update	07/01/2008	
prg_user_update		Last modification user	Username	
prg_version		Program version	1.0.0	
prg_display		Program display flag	Y	
prg_web		Program display in website flag	U=User, S=Super administrator	
rareall	rar_allele	Allele name	A*01:01:01:02N	
	rar_release	Release	2.22.0	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
rareall	rar_web	Total number of alleles in website	0
	rar_nmdp	Total number of alleles in nmdp	0
	rar_labs	Total number of alleles in laboratories	0
	rar_status	Status	A=Active, I=Inactive
releases	rel_code	Release code	IMGT/HLA
	rel_name	Release name	2.19.0
	rel_date	Release date	01/10/2007
	rel_ftp	Release ftp path source	ftp://alleles.org
	rel_file	Release file source	confirmations_2190.txt
	rel_http	Release http path source	http://www.ebi.ac.uk
	rel_desc	Release description	HLA Alleles
	rel_total_rows	Release total rows updated	3041
resolut	rel_status	Release status	A=Active, I=Inactive
	res_parent	Allele parent	A*01:01:01:01
	res_child	Allele child	A*01:01:01
	res_parent_size	Allele parent size	4
securacc	res_child_size	Allele child size	3
	sec_user	User login	Username
securacc	sec_program	Program code	hla2001a
	sec_password	User password in program	Password
	sec_master_ins	Access to insert in master	Y/N
	sec_master_del	Access to delete in master	Y/N
	sec_master_mod	Access to modify in master	Y/N
	sec_detail_ins	Access to insert in detail	Y/N
	sec_detail_del	Access to delete in detail	Y/N
	sec_detail_mod	Access to modify in detail	Y/N
	sec_print	Access to print	Y/N
	sec_apply	Access to apply updates	Y/N
	sec_undo	Access to undo changes	Y/N
	sec_open	Access to open data	Y/N
	sec_cancel	Access to cancel data	Y/N
	sec_lock	Access to lock record	Y/N
	sec_unlock	Access to unlock record	Y/N
	sec_export	Access to export information	Y/N
	sec_page1	Access to page 1	Y/N
	sec_page2	Access to page 2	Y/N
	sec_page3	Access to page 3	Y/N
	sec_page4	Access to page 4	Y/N
sec_page5	Access to page 5	Y/N	
sec_page6	Access to page 6	Y/N	
sec_page7	Access to page 7	Y/N	
sec_page8	Access to page 8	Y/N	
sec_page9	Access to page 9	Y/N	
sec_page10	Access to page 10	Y/N	
statshla	shla_locus	Locus name	A
	shla_imgt	Total alleles in IMGT/HLA database	1518
	shla_website	Total alleles in Website	193
	shla_entries	Total entries in Website	7498
statskir	skir_locus	Locus name	2DL1
	skir_ipd	Total alleles in IPD-KIR database	37
	skir_website	Total alleles in Website	21
	skir_entries	Total entries in Website	214
statsmic	smic_locus	Locus name	MICA

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example	
statsmic	smic_imgt	Total alleles in IMGT/HLA database	72	
	smic_website	Total alleles in Website	38	
	smic_entries	Total entries in Website	516	
statspop	spop_polyreg	Polymorphic region	HLA	
	spop_pops	Total populations	843	
	spop_alleles	Populations with alleles	709	
	spop_haplotypes	Populations with haplotypes	361	
	spop_genotypes	Populations with genotypes	0	
	spop_family	Populations at family level	503	
	spop_both	Populations with both data	828	
tblallele	tall_popid	Population id	1980, 1981, ...	
	tall_polyreg	Polymorphic region	HLA, KIR, ...	
	tall_locus	Locus name	A, B, C, DRB1, ...	
	tall_allele	Allele name	A*01:01:01:01	
	tall_line	Record identifier	1, 2, 3, ...	
	tall_gene_freq	Gene frequency	0.001	
	tall_status	Status	A	
	tall_user	Username	...	
	tall_pheno_freq	Phenotype frequency	100.0	
	tall_added	Date added	01/01/2010	
	tall_updated	Last update	01/01/2010	
	tall_gene_has_data	Flag to identify gene frequency data	Y, N	
	tall_gene_freq_num	Gene frequency	0.001	
	tall_freq_has_data	Flag to identify pheno frequency data	Y, N	
	tall_pheno_freq_num	Phenotype frequency	100.0	
	tall_has_data	Flag to identify if has data	Y, N	
	tall_indiv	Number of individuals who carry the allele	10	
	tall_chrom	Number of chromosomes with the allele	20	
	tall_sample_size	Sample size of the population	100	
	tall_notes	Notes in the record	any note	
	tblcytokines	tcyt_popid	Population id	2485
		tcyt_gene	Name of the gene	IL-4/
tcyt_cytokine		Cytokine name	IL-4/ - 590 TT	
tcyt_line		Record identifier	3782	
tcyt_genotype		Genotype combination	TT	
tcyt_order		Order	1045.3000	
tcyt_freq		Frequency in the population	3.6	
tcyt_added		Date added	25/11/2008	
tcyt_update		Last update	25/11/2008	
tcyt_status		Status	A=Active, I=Inactive	
tcyt_user		Username	...	
tcyt_indiv		Number of individuals who carry this cytokine	5	
tcyt_sample_size	Sample size of the population	140		
tblhaplotypes	thap_popid	Population id	1216	
	thap_region	Polymorphic region	HLA	
	thap_haplotype	Haplotype	A*02:01-B*15:01	
	thap_line	Record identifier	2450	
	thap_haplotype2	Haplotype last nomenclature	A*0201-B*1501	
	thap_frequency	Frequency in the population	1.4	
	thap_status	Status	A	
	thap_loci	Number of loci included in the haplotype	2	

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
tmpmatrix	tmp_user	User login	...
	tmp_allele	Allele name	DPA1*02:01:03
	tmp_pos	Position of difference in amino acid	42
	tmp_letter	Amino acid	A
	tmp_freq	Frequency in the population	0.3
tmpmatrix2	tmp2_user	User login	...
	tmp2_value	Allele name	A*01:01
	tmp2_position	Order	1
tmpmatrix3	tmp3_user	User login	...
	tmp3_value	Value	2699
	tmp3_position	Order	1
tmptab02	tmp2_popid	Population id	1450
	tmp2_allele	Allele name	C*12:04:01
	tmp2_pfreq	Phenotype frequency	4.20
	tmp2_afreq	Allele frequency	0.021
tsession	tmp2_notes	Additional notes	Additional notes
	ses_user	Session user	...
	ses_last_event	Session last event	main
	ses_status	Session status	A
	ses_date	Session date	28/11/2008
typehla	ses_time	Session time	153420
	thla_user	User login	...
	thla_indiv	Individual id	000001
	thla_locus	Locus name	B
	thla_copy	Copy of the allele	1
	thla_allele	Allele variant	1301
	thla_code	Allele name	B*1301
	thla_mark	Type of record	
users	thla_status	Status	T= Temporary
	usr_login	User login	...
	usr_password	User password	Password
	usr_lastname	User last name	-
	usr_firstname	User first name	-
	usr_title	User title	-
	usr_sex	User sex	M/F
	usr_date_of_birth	User date of birth	-
	usr_email	User email	-
	usr_status	User status	-
	usr_profile	User profile	-
	usr_add1	User address	-
	usr_add2	User address	-
	usr_city	User city	-
	usr_county	User county or state	-
	usr_country	User country	-
	usr_postcode	User postcode	-
	usr_tel1	User telephone 1	-
	usr_tel2	User telephone 2	-
	usr_fax	User fax	-
	usr_mobile	User mobile	-
	usr_url	User url	-
	usr_association	User association or university	-
	usr_interests	User interests	-
	usr_notify	User notifications via email	-
usr_date_added	User date added	-	
usrconf	ucfg_user	Configuration user	...
	ucfg_program	Configuration program	hla2001a
	ucfg_event	Configuration event	Login

Table B.2: Data dictionary of the AFND (Continued)

Table	Attribute	Description	Example
usrconf	ucfg_value	Configuration value	1
websetup	web_name	Website name	AFND
	web_name2	Website name	-
	web_add1	Website address	-
	web_add2	Website address	-
	web_city	Website city	-
	web_country	Website country	-
	web_country	Website country	-
	web_postcode	Website postcode	-
	web_tel1	Website telephone (1)	-
	web_tel2	Website telephone (2)	-
	web_fax	Website fax	-
	web_mobile	Website mobile	-
	web_server	Website server address	-
	web_alias_db	Website database alias	-
	web_image	Website image reference	-
	web_icon	Website icon reference	-
	web_domain	Website domain	-
	web_library	Website library	-
	web_release	Website HLA last release	-
	web_email	Website email contact	-
	web_header	Website page header	-
	web_footer	Website page footer	-
	web_homepage	Website homepage	-
	web_path	Website path	-