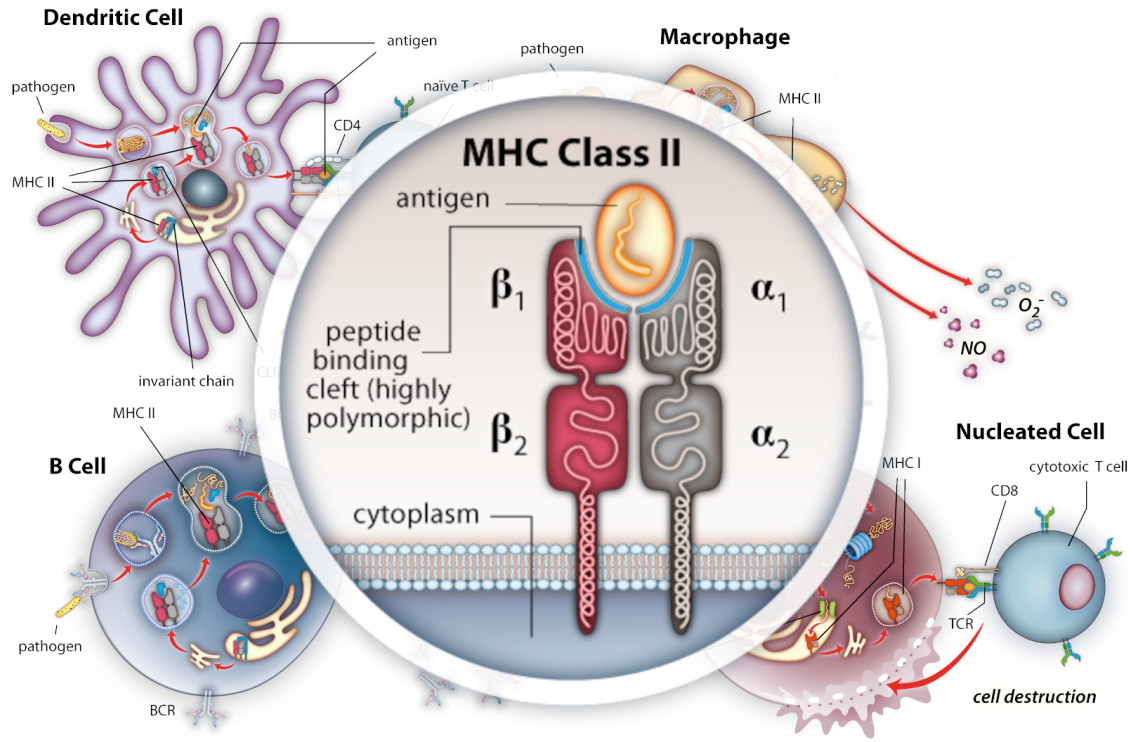
BIOINFORMATIC ANALYSIS OF THE HUMAN LEUCOCYTE ANTIGEN IN INFLAMMATORY BOWEL DISEASE

MAREIKE WENDORFF



Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

Kiel, November 2022

Mareike Wendorff: *Bioinformatic analysis of the Human Leucocyte Antigen in Inflammatory Bowel Disease*, Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel, © November 2022

Bioinformatic analysis of the Human Leucocyte Antigen in Inflammatory Bowel Disease

Dissertation zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät der
Christian-Albrechts-Universität zu Kiel

Vorgelegt von
Mareike Wendorff
Kiel, November 2022

ERSTER GUTACHTER:	Prof. Dr. Andre Franke
ZWEITER GUTACHTER:	Prof. Dr. Tal Dagan
TAG DER MÜNDLICHEN PRÜFUNG:	7. März 2023
ZUM DRUCK GENEHMIGTE VERSION:	22. Mai 2023
DEKAN:	Prof. Dr. Frank Kempken

In rowing you are going backwards and you don't see the finish line, but you know it is there. In science you are rowing like crazy and hope that a finish line will be there.

— Katalin Karikó

An individual is more than the sum of genes – but nothing without them.

— Spontaneous thought

Zeichnet es nicht ein gutes Immunsystem aus, dass man es überhaupt nicht bemerkt?

— Dobby modified (J.K.Rowling)

SUMMARY

Immune mediated chronical diseases, like inflammatory bowel disease (IBD), are lifelong diseases for which curative therapies are not available until now. The current treatment, aiming to reduce the symptoms of the disease, can have severe side effects and sometimes only a limited treatment response. One main reason why the treatment options are so limited, is that the pathogenesis of the disease is still not fully understood.

Over the last decades, different factors associated with increased risks of these diseases have been identified with the human leukocyte antigen (HLA) as one of the strongest associated genetic regions for many different inflammatory traits. The profile of the HLA association differs across diseases. A causative role of the HLA, including the elucidation of disease-specific variation in the HLA (HLA alleles), in chronic inflammations remains to be revealed. One exception is Celiac disease. The identification of gluten and a specific HLA variation as the triggering factor in Celiac disease has improved the live of affected patients drastically as they can manage their symptoms by avoiding gluten contained in wheat. Patients with IBD do not have this possibility as no clear factors related to the HLA association (i.e., a specific HLA-peptide-T-cell interaction) have been identified yet.

Studying the HLA is a challenging task as the classical HLA genes are highly polymorphic, neighboring genes are inherited together (linkage disequilibrium), and the function of the HLA is based on presenting "something of everything" that is present in the body. It can be expected that each HLA allele can present about 2% of all peptides of suitable lengths, and that any HLA-peptide interaction resulting from this, might induce an immune reaction in presence of a specific T helper cell. The UniProt database is a database for all protein sequences of any natural source. It includes more than 10^{10} different peptides by fragmenting the protein sequences into peptides of a length of 15 amino acids using a sliding window approach. One hypothesis is, that a single peptide or a handful of peptides of the even larger pool of peptides that exists in reality (including unknown species, unknown sequences and modified peptides) explain the strong association of the classical HLA genes with ulcerative colitis (UC) and Crohn's disease (CD), the main types of IBD.

The aim of this thesis is to gain as much understanding as possible about the HLA alleles genetically associated with UC, and the source and structure of peptides hypothesized to drive the disease. This thesis includes two papers published in peer-reviewed journals and an additional manuscript which is in progress. Furthermore, publications where I have been involved are listed and cited if suitable. In the first publication, presented in this thesis (Section 6.3), we built an imputation panel that enabled the analysis of IBD and the associated HLA alleles and their corresponding haplotypes across different ancestries. In the second publication (Section 7.3), I analyzed the interaction of peptides with a defined set of HLA alleles associated with UC. This study was the first to use ultra-high density microarray data

for predicting the binding status of HLA alleles and peptides. In the final manuscript (Section 8.3), I studied the genetics of UC in Caucasian individuals. I analyzed, what the genetics, combined with prior knowledge about different genes and their protein function, can tell us about a hypothesized peptide that might be a key player in the pathogenesis of UC.

Next to some improvements in the imputation of HLA genotypes and the binding prediction, this thesis points out first concrete candidate peptides and suggests a path to continue to discover more about the contribution of the HLA in UC.

ZUSAMMENFASSUNG

Immuninduzierte chronische Krankheiten, wie chronisch entzündliche Darmerkrankungen, sind lebenslange Erkrankungen, für die keine heilenden Therapien verfügbar sind. Aktuelle Therapien, die darauf abzielen Symptome zu lindern, haben oft schwere Nebenwirkungen und nur eingeschränkten Behandlungserfolg. Ein Hauptgrund, warum die Behandlungsoptionen so eingeschränkt sind ist, dass die Pathogenese der Erkrankung immer noch nicht vollständig bekannt ist.

Über die letzten Jahrzehnte wurden verschiedene Faktoren identifiziert, die mit einem erhöhten Risiko für diese Krankheiten assoziiert sind. Für viele Entzündungskrankheiten ist die Region der humanen Leukozyten Antigene (HLA) dabei einer der am stärksten assoziierten genetischen Orte. Das Profil der HLA assoziierten Allele unterscheidet sich dabei zwischen den Erkrankungen. Eine ursächliche Rolle des HLA, einschließlich einer Erklärung für die krankheitsspezifischen Assoziationen der HLA Allele in chronischen Entzündungen, muss noch identifiziert werden. Eine Ausnahme ist Zöliakie. Die Identifikation des Glutens und einer spezifischen HLA Variante als auslösende Faktoren in der Zöliakie hat das Leben von Betroffenen drastisch verbessert. Die Patienten können durch den Verzicht auf Gluten, welches in Weizen enthalten ist, ihre Symptome in den Griff bekommen. Patienten mit chronisch entzündlichen Darmerkrankungen haben diese Möglichkeit nicht, da bislang kein eindeutiger Faktor in Bezug auf die HLA Assoziation gefunden wurde (also eine spezielle Interaktion zwischen HLA, Peptid und T-Zelle).

Die Forschung zu HLA ist eine herausfordernde Aufgabe, da die klassischen HLA Gene sehr polymorph sind, benachbarte Gene zusammen vererbt werden (Kopplungsungleichgewicht) und die Funktion der HLA Proteine darauf beruht, von allem, was im Körper zu finden ist, etwas zu präsentieren. Es kann davon ausgegangen werden, dass jedes HLA Protein etwa 2% aller möglichen Peptide in einem festen Längenbereich präsentieren kann und jede HLA-Peptid Interkation, die daraus resultiert, könnte eine Immunreaktion auslösen, wenn eine spezifische T-Helferzelle präsent ist. In der Datenbank UniProt, einer Datenbank für alle Proteinsequenzen natürlichen Ursprungs, sind über 10^{10} verschiedene Peptide verfügbar, wenn die Aminosäuresequenzen mit einem "Sliding Window" in Peptide mit einer Länge von 15 Aminosäuren zerlegt werden. In der Realität kann davon ausgegangen werden, dass sogar noch mehr Peptide natürlich vorkommen, durch bislang nicht sequenzierte Arten und modifizierte Sequenzen. Eine Hypothese ist, dass ein einzelnes Peptid oder eine handvoll Peptide aus diesem gigantischen Pool an Peptiden die Erklärung für die starke Assoziation der klassischen HLA Gene mit Colitis Ulcerosa und Morbus Crohn, den beiden Hauptformen chronisch entzündlicher Darmerkrankung, liefert.

Das Ziel dieser Doktorarbeit war es, die HLA Allele zu erforschen, die mit Colitis Ulcerosa assoziiert sind, und soviel wie möglich über den Ursprung

und die Struktur von Peptiden, von denen hypothetisch angenommen wird die Erkrankung auszulösen, herauszufinden.

Diese Arbeit beinhaltet zwei Publikation, die in wissenschaftlichen Fachzeitschriften veröffentlicht worden sind, und ein weiteres Manuskript, dessen Veröffentlichung noch aussteht. Weitere Publikationen, bei denen ich involviert war, werden in dieser Arbeit aufgelistet und an gegebenen Stellen zitiert. In der ersten Publikation, die hier präsentiert wird (PAPER A, Abschnitt 6.3) haben wir eine Imputationsreferenz gebaut, die die Analyse von chronisch entzündlichen Darmerkrankungen und den damit assoziierten HLA Allelen in verschiedenen Populationen ermöglicht. In der zweiten Publikation (PAPER B, Abschnitt 7.3) habe ich die Interaktion von Peptiden mit einem definierten Satz von HLA Allelen, die mit Colitis Ulcerosa assoziiert sind, analysiert. Diese Studie war die erste, die Daten von einem Mikroarray mit einer ultrahohen Dichte verwendet hat, um die Bindung zwischen HLA Allelen und Peptiden vorherzusagen. Im finalen Manuskript (PAPER C, Abschnitt 8.3) habe ich die Genetik von Colitis Ulcerosa in Kaukasiern untersucht und analysiert. Außerdem habe ich erforscht, was uns die Genetik in Kombination mit bereits vorhandenem Wissen zu verschiedenen Genen und deren Funktion über die hypothetischen Peptide verrät, die womöglich eine Schlüsselrolle in der Colitis Ulcerosa haben.

Neben Verbesserungen in der Imputation von HLA Genotypen und der Vorhersage der Bindung von HLA Allelen und Peptiden beschreibt diese Arbeit erste konkrete Kandidatenpeptide und schlägt einen Weg vor, wie mehr über die Rolle der HLA Gene bei Krankheiten herausgefunden werden könnte.

CURRICULUM VITÆ

MAREIKE WENDORFF

PERSONAL INFORMATION

Date of birth	November 9th, 1988
Place of birth	Warburg, Germany
Nationality	German
Email	wendorff@leibniz-ipn.de

EDUCATION

09/2015 – today	PhD student University Kiel
10/2011 – 03/2014	M.Sc. Environmental Modeling Carl von Ossietzky Universität, Oldenburg Final Degree: 1.27 Thesis: „Untersuchung und Modellierung der thermischen Kapazität von Wohngebäuden zur Speicherung erneuerbarer Energien“
10/2008 – 01/2012	B.Sc. Environmental Science Carl von Ossietzky University, Oldenburg Final Degree: 1.45 Thesis: “Transport through a tidal inlet in the East Frisian Wadden Sea (Otzumer Balje)”
08/1999 – 06/2008	University-entrance diploma (“allgemeine Hochschulreife”) Hüffertgymnasium, Warburg Final Degree: 2.0 Advanced courses: Biology and chemistry

WORKING EXPERIENCE

06/2022 – today	Leibniz Institute for Science and Mathematics Education, Kiel
09/2015 – 08/2022	Institute of Clinical Molecular Biology, Kiel, PhD candidate
05/2013 – 11/2014	Fraunhofer UMSICHT, Oberhausen, Research Assistant
10/2009 – 10/2012	Carl von Ossietzky University, Oldenburg, Student Assistant

PUBLICATIONS

MAIN PUBLICATIONS PRESENTED IN THIS THESIS

Frauke Degenhardt, MAREIKE WENDORFF, Michael Wittig, Eva Ellinghaus, Lisa W. Datta, John Schembri, Siew C. Ng, Elisa Rosati, Matthias Hübenthal, David Ellinghaus, Eun Suk Jung, Wolfgang Lieb, et al. "Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles." In: *Human molecular genetics* 28.12 (2019), pp. 2078–2092. ISSN: 1460-2083. DOI: 10.1093/hmg/ddy443

MAREIKE WENDORFF, Heli M. Garcia Alvarez, Thomas Østerbye, Hesham ElAbd, Elisa Rosati, Frauke Degenhardt, Søren Buus, Andre Franke, and Morten Nielsen. "Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction." In: *Frontiers in Immunology* 11.August (2020), pp. 1–8. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.01705

MAREIKE WENDORFF, Hesham ElAbd, Frauke Degenhardt, Marc Höppner, Florian Uellendahl-Werth, Eike M Wacker, Lars Wienbrandt, Simonas Juzenas, Regeneron Genetic Center, Tomas Koudelka, David Ellinghaus, Petra Bacher, et al. "Genome-wide analysis of individual coding variants and HLA-II-associated self-immunopeptidomes in ulcerative colitis." In: *medRxiv* (Jan. 2023). DOI: 10.1101/2023.03.22.23286498. URL: http://medrxiv.org/content/early/2023/05/17/2023.03.22.23286498.abstract

LIST OF FURTHER PUBLICATIONS

Peer reviewed manuscripts

Y. Bejaoui, M. Witte, M. Abdelhady, M. Eldarouti, N.M.A. Abdallah, A.A. Elghzaly, Z. Tawhid, M.A. Gaballah, H. Busch, M. Munz, M. WENDORFF, E. Ellinghaus, et al. "Genome-wide association study of psoriasis in an Egyptian population." In: *Experimental Dermatology* 28.5 (May 2019), pp. 623–627. ISSN: 1600-0625. DOI: 10.1111/exd.13926. URL: http://www.ncbi.nlm.nih.gov/pubmed/30921485

Frauke Degenhardt, Gabriele Mayr, MAREIKE WENDORFF, Gabrielle Boucher, Eva Ellinghaus, David Ellinghaus, Hesham ElAbd, Elisa Rosati, Matthias Hübenthal, Simonas Juzenas, Shifteh Abedian, Homayon Vahedi, et al. "Transethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals not only shared but also ethnicity-specific disease associations." In: *Human molecular genetics* 30.5 (Apr. 2021), pp. 356–369. ISSN: 1460-2083. DOI: 10.1093/hmg/ddab017. URL: http://www.ncbi.nlm.nih.gov/pubmed/33555323

Frauke Degenhardt, David Ellinghaus, Simonas Juzenas, Jon Lerga-Jaso, MAREIKE WENDORFF, Douglas Maya-Miles, Florian Uellendahl-Werth, Hesham ElAbd, Malte C Rühlemann, Jatin Arora, Onur Özer, Ole Bernt Lenning, et al. "Detailed stratified GWAS analysis for severe COVID-19 in four European populations." In: *Human Molecular Genetics* Epub ahead (July 2022), n.n. ISSN: 0964-

6906. DOI: 10.1093/hmg/ddac158. URL: https://academic.oup.com/hmg/advance-article/doi/10.1093/hmg/ddac158/6644888

COVID-19 Host Genetics Initiative. "Mapping the human genetic architecture of COVID-19." In: *Nature* 600.7889 (Dec. 2021), pp. 472–477. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03767-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/34237774

Hesham ElAbd, Yana Bromberg, Adrienne Hoarfrost, Tobias Lenz, Andre Franke, and MAREIKE WENDORFF. "Amino acid encoding for deep learning applications." In: *BMC Bioinformatics* 21.1 (2020), pp. 1–14. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03546-x

Hesham ElAbd, Frauke Degenhardt, Tomas Koudelka, Ann-Kristin Kamps, Andreas Tholey, Petra Bacher, Tobias L Lenz, Andre Franke, and MAREIKE WENDORFF. "Immunopeptidomics toolkit library (IPTK): a python-based modular toolbox for analyzing immunopeptidomics data." In: *BMC Bioinformatics* 22.1 (Dec. 2021), p. 405. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04315-0. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04315-0

D. Ellinghaus, F. Degenhardt, L. Bujanda, M. Buti, A. Albillos, P. Invernizzi, J. Fernández, D. Prati, G. Baselli, R. Asselta, M.M. Grimsrud, C. Milani, F. Aziz, J. Kässens, S. May, M. WENDORFF, L. Wienbrandt, et al. "Genomewide Association Study of Severe Covid-19 with Respiratory Failure." In: *New England Journal of Medicine* 383.16 (Oct. 2020), pp. 1522–1534. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2020283. URL: http://www.nejm.org/doi/10.1056/NEJMoa2020283

Quirin Hammer, Josefine Dunst, Wanda Christ, Francesca Picarazzi, MAREIKE WENDORFF, Pouria Momayyezi, Oisín Huhn, Herman K. Netskar, Kimia T. Maleki, Marina García, Takuya Sekine, Ebba Sohlberg, et al. "SARS-CoV-2 Nsp13 encodes for an HLA-E-stabilizing peptide that abrogates inhibition of NKG2A-expressing NK cells." In: *Cell Reports* 38.10 (2022). ISSN: 2211-1247. DOI: 10.1016/j.celrep.2022.110503

Tim Hollstein, Dominik M. Schulte, Juliane Schulz, Andreas Glück, Anette G. Ziegler, Ezio Bonifacio, MAREIKE WENDORFF, Andre Franke, Stefan Schreiber, Stefan R. Bornstein, and Matthias Laudes. "Autoantibody-negative insulin-dependent diabetes mellitus after SARS-CoV-2 infection: a case report." In: *Nature Metabolism* 2.10 (2020), pp. 1021–1024. ISSN: 2522-5812. DOI: 10.1038/s42255-020-00281-8. URL: http://dx.doi.org/10.1038/s42255-020-00281-8

Elisa Rosati, Gabriela Rios Martini, Mikhail V Pogorelyy, Anastasia A Minervina, Frauke Degenhardt, MAREIKE WENDORFF, Soner Sari, Gabriele Mayr, Antonella Fazio, Christel Marie Dowds, Charlotte Hauser, Florian Tran, et al. "A novel unconventional T cell population enriched in Crohn's disease." In: *Gut* Published (2022), gutjnl–2021–325373. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2021-325373

Christine Strippel, Marisol Herrera-Rivero, MAREIKE WENDORFF, Anja K Tietz, Frauke Degenhardt, Anika Witten, Christina Schroeter, Christopher Nelke, Kristin S Golombeck, Marie Madlener, Theodor Rüber, Leon Ernst, et al. "A genome-wide association study in autoimmune neurological syndromes with anti-GAD65 autoantibodies." In: *Brain* Online ahe (Mar. 2022), pp. 1–6. ISSN: 0006-

8950. DOI: 10.1093/brain/awac119. URL: http://www.ncbi.nlm.nih.gov/pubmed/35348614

Christian P Kratz, Dmitrii Smirnov, Robert Autry, Natalie Jäger, Sebastian M Waszak, Anika Großhennig, Riccardo Berutti, MAREIKE WENDORFF, Pierre Hainaut, Stefan M Pfister, Holger Prokisch, Tim Ripperger, et al. "Heterozygous BRCA1 and BRCA2 and Mismatch Repair Gene Pathogenic Variants in Children and Adolescents With Cancer." In: *JNCI: Journal of the National Cancer Institute* 114 (2022), pp. 1523–1532. ISSN: 0027-8874. DOI: 10.1093/jnci/djac151

Manuscripts under Review

Hesham ElAbd, Frauke Degenhardt, Tobias L. Lenz, Andre Franke, and MAREIKE WENDORFF. "VCF2Prot: An Efficient and Parallel Tool for Generating Personalized Proteomes from VCF Files." In: *bioRxiv* (2022), p. 2022.01.21.477084. DOI: https://doi.org/10.1101/2022.01.21.477084. URL: https://www.biorxiv.org/content/10.1101/2022.01.21.477084v1.abstract

Hesham ElAbd, MAREIKE WENDORFF, Tomas Koudelka, Christian Hentschker, Ann-Kristin Kamps, Christoph Prieß, Lars Wienbrandt, Frauke Degenhardt, Tim A Steiert, Petra Bacher, David Ellinghaus, Uwe Völker, et al. "Predicting Peptide HLA-II Presentation Using Immunopeptidomics , Transcriptomics and Deep Multimodal Learning." In: *bioRxiv* (2022). DOI: https://doi.org/10.1101/2022.09.20.508681

Eun Suk Jung, David Ellinghaus, Frauke Degenhardt, Akira Meguro, Khor Seik-Soon, Sören Mucha, MAREIKE WENDORFF, Michael Wittig, Simonas Juzenas, Elaine F Remmers, Ahmet Gül, Nobuhisa Mizuki, et al. "Genome-wide association analysis reveals HLA-B*46:01 as a risk factor for intestinal involvement in patients with Behçet's disease." In: *Gut (under review)* (2022)

ACKNOWLEDGEMENTS

I cannot express enough thanks for all the amazing people around me. This thesis would not have been possible without contributing colleagues and a supportive network.

First, I want to thank Prof. Andre Franke. He enabled this project and have had the time to discuss my topics. He was always approachable, and gave me the room to find my own way. Furthermore, he supported and initialized international collaborations that supported me in extending my knowledge.

Based on this I have had the fortune to stay a month in Sydney together with Prof. Shoba Ranganathan who was patient in giving me a basic understanding of the HLA when I was just starting. Unfortunately, we soon figured out, that her expertise was not the way to go to approach my research question in the first place, even though it might be now, after the analysis performed within this thesis is done.

Next, I have to thank Prof. Morten Nielsen, who is the expert on HLA-peptide binding prediction based on machine learning. He welcomed me in his lab for a visit, was open for a collaboration, and actively supervised the study and writing process related to PAPER B.

I want to thank Prof. Tobias L. Lenz for all the deep discussions we have had about my research. He opened me all doors to be a part of his group and join discussions and social activities.

You do not have to be a Professor to be the most important figure in another ones PhD journey. Therefore, I want to thank Frauke Degenhardt as she was there for all my concerns - from the on-boarding in the first days to the finalization of the manuscript - from twelve hour-a-day working weeks to walks through the botanical garden - from crying to laughing - from integrating me into her project, which resulted in PAPER A to her proofreading my thesis.

I want to thank Elisa Rosati for initializing the immunogenetics meetings and keeping them alive. Those meetings were a great opportunity to reflect, discuss, and redirect the own work.

Additionally, I want to thank all my other colleagues including Lars Wienbrandt, Matthias Hübenthal, Sören Mucha, Louise Thingholm, Malte Rühlemann, Michael Forster, Simonas Juzenas, Jan Kässens, and Hesham ElAbd with whom I shared my office for some time. We have had nice discussions in coffee breaks and visits of the canteen. Also the container-live with Florian Uellendahl-Werth, Eike Wacker, and David Ellinghaus, as addition to some of the previously mentioned, included all kinds of discussions.

There have been also other members of the working group, who were essential for my work. Some solved concrete content related problems, others are maintaining the system as a whole and keep everything running. Among those Marc Höppner, the IT-department, including Iacopo Torres and Georg Hemmrich-Stanisak, the computational center of the university with Michael Kisiela who is in parallel part of our group, and the adminis-

tration, with among others Eike Zell, Natalie Tepling, Tosca Heinrich, and Christiane Wolf-Schwerin.

Other scientists were around outside our working group, giving great opportunities for an exchange. Two networks to be mentioned here as they have had a positive influence on me and my career are the RTG1743 and the PMI.

All those amazing people I get to know as colleagues but hopefully some stay friends.

Other individuals I directly met as friends. While all of my friends have had an important role in my live. Some have had a direct influence on this thesis.

Marika who, actively as a science communicater, and unwillingly as a victim of IBD, made me see the meaning of my work. With the contrast between the pure vital energy, next to the suffering and early death, she highlighted the importance of live and the aims we are following.

Judith helped me to come into the writing process and therefore, I want to thank her for the company.

Thanks to my family as they have been there for me every single day of my life. A special thanks to my mum who listened to my daily reports when I needed it most, and my brother who always has had time to help me with my informatics related questions.

There are also two contributions to this thesis I want to thank for: First, thanks to Michael Lange for sending me the English transcript of his interview with Katalin Karikó. Second, thanks to Renate Nikolaus who designed some of the graphics used within this thesis, including the graphic the cover is based on.

Even though, some were already mentioned, many thanks go to Frauke Degenhardt, Hesham ElAbd, Audrey, my mother, and the whole LaTeX-community for support, an open eye and ideas regarding the finetuning of this thesis.

Most likely I missed to mention the one or the other person who actively contributed to this work or who earned to receive special thanks for cheering me up and celebrating my successes. Herewith, I want to say: Thanks!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

1KGP	1000 Genomes Project
3'NT	3' non translated region
A	adenine
APC	antigen presenting cell
B2M	beta-2 microglobulin
BA	binding affinity
C	cytosine
CD	Crohn's disease
CLIP	class II-associated invariant chain peptide
CY	cytoplasmic region
DC	dendritic cell
DEG	differentially expressed gene
DL	deep learning
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide
EL	eluted ligands
eQTL	expression Quantitative Trait Loci
ER	endoplasmic reticulum
G	guanine
GRC	genome reference consortium
GSA	Illumina Global Screening Array
GTEx	Genotype-Tissue Expression project
GWAS	genome wide association study
HWE	Hardy Weinberg equilibrium
HLA	human leukocyte antigen
IBD	inflammatory bowel disease
IEDB	immune epitope database and analysis resource
InDel	insertion or deletion

LD	linkage disequilibrium
LoF	loss of function
MAF	minor allele frequency
MHC	major histocompatibility complex
MIC	MHC class I chain-related
ML	machine learning
MS	mass spectrometry
NGS	next generation sequencing
NOD2	nucleotide-binding oligomerization domain 2
OR	odds ratio
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PepWAS	peptidome wide association study
PTM	post translational modification
QC	quality control
RNA	ribonucleic acid
SKAT	sequence-based kernel association test
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SP	signal peptide
T	thymine
TAP	transporter associated with antigen processing
TCR	T cell receptor
TM	transmembrane region
TNF	tumor necrosis factor
TS	targeted sequencing
UC	ulcerative colitis
WES	whole exome sequencing
WGS	whole genome sequencing

Part I

GENETICS OF INFLAMMATORY BOWEL DISEASE

Genetic studies are an important angle for the study of complex diseases like inflammatory bowel diseases. Inflammatory bowel diseases involve chronic inflammations of the gastrointestinal tract, caused by genetic factors and environmental factors. Within this part, I am giving some background about genetics and genetic studies of diseases. Furthermore, I am introducing inflammatory bowel diseases and give a brief insight into the current knowledge about the genetics in inflammatory bowel diseases.

1

INTRODUCTION

Over the last decades, a huge effort has been invested into understanding the role of genetics in the etiology of various diseases. The developments in genotyping and the tremendous success of GWASs, as a means to analyzing genetic associations in disease, has resulted in the discovery of an increasing amount of genetic risk factors for complex diseases over time. For some diseases, including IBD, several hundred independent genetic risk factors are known today. For many of those associated genes, no conclusive connection or no connection at all to the disease pathophysiology has been discovered to date. Pharmacological studies have shown that drug candidates have a higher success rate if their target is related to genes associated with the disease [92]. Overall, a disease with an understood cause and pathophysiology, might offer more defined treatment and therapy opportunities. For many complex diseases, we still have only a fragmentary knowledge of the cause of disease, as many factors have an impact on the phenotype. Currently, the scientific community aims to understand the identified risk factors of diseases following the path of correlation to uncovering causation (Figure 1) [103].

One genetic locus associated with many immune-related diseases is the HLA. The classical HLA proteins are important players in adaptive immunity. Through the presentation of potential antigens (antibody generators), in

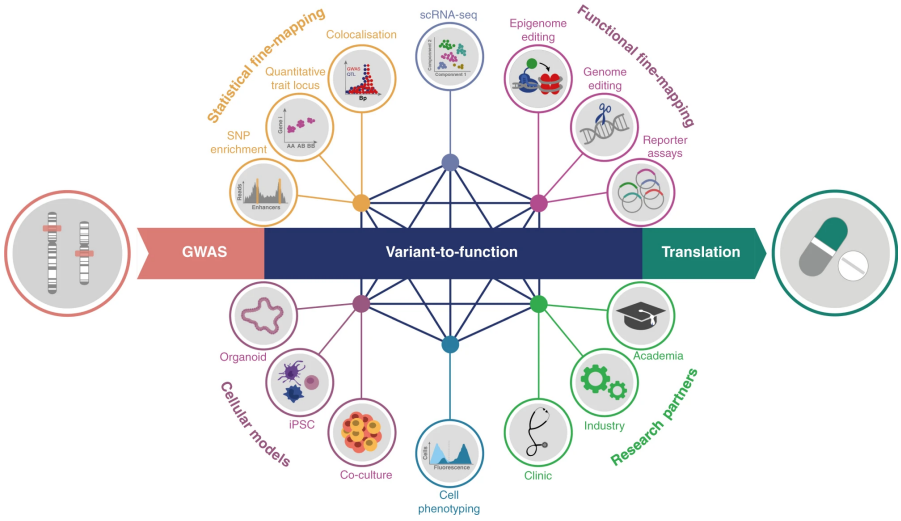


Figure 1: Variant to function approach for interpreting GWAS hits. The interpretation of GWAS signals is advancing with more tools and models arising to study the impact of genetic variations from different angles. Figure taken from Lichou and Trynka [103].

the form of peptides to T cells, allow the immune system to identify and eliminate sources of suspicious antigens from the body. Only antigens that are presented by the HLA (i.e. not floating around freely in the cytosol or cell environment) may lead to the generation of antibodies. Many different peptides derived from non-immunogenic sources as well and are presented by the HLA. The role of the HLA is to provide a little of everything to the immune system. Different peptides bind to different HLA proteins. The HLA proteins vary in the chemico-physical properties of the HLA-peptide binding groove and bind peptides with different affinities and selectivity. The immune system then decides which sources need to be eliminated. This selectivity is then joined by the T cell receptor (TCR). The limited but still broad spectrum of potential antigens, induced by the different HLA proteins based on the genetic variability in this region, is expected to be the reason for the association of this genetic region with a multitude of different immune-mediated diseases. The generality of the HLA makes the identification of the disease specific role of the HLA a challenging task.

Within this thesis, I will analyze the HLA using bioinformatical techniques. The aim of the studies, presented in this thesis, is to reveal what the human genetics tell us about the function of the HLA in IBD. A special focus is on UC, as the genetic variation within the HLA has a greater impact in UC than in the other main form of IBD, CD. Therefore, I will initially focus on the determination of HLA proteins based on genetic information, described in PAPER A (Section 6.3). In a second step, I will analyze the presentation profile of HLA alleles associated with UC, published in PAPER B (Section 7.3). Finally, I will investigate the genetic variability in UC patients and healthy controls with a focus on the HLA genes and how other mutations might impact the repertoire of presented peptides as shown in PAPER C (Section 8.3).

As the human genetics are a fundamental factor for this study, and previous findings on genetic associations are the motivation for this study, Part i focuses on genetics and IBD. First, the fundamentals of the human genetics and genetic associations are introduced (Chapter 2). Afterwards, a short introduction of IBD is given in Chapter 3. This is followed by an overview of the genetic associations in IBD as described in the literature in Chapter 4.

Part ii then focuses on the HLA in general. After an introduction into the HLA (Chapter 5), the methods for differentiating the HLA types based on the genetics are described (Chapter 6). This section includes PAPER A, which introduces a multi-ethnic reference panel for the computational inference (imputation) of HLA alleles. The last chapter in this part moves away from the genetic perspective and concentrates on the description of the presentation of peptides by the HLA molecules (Chapter 7). As part of this section, PAPER B (Section 7.3) is included, which presents the first prediction tool for HLA-peptide interaction that is based on high-density peptide microarray data.

The last part (Part iii), as a connection to the preceding chapters, deals with the role of the HLA in IBD (Chapter 8). The chapter starts with an overview of the associations of the HLA with IBD as described in the literature (Section 8.1) including a trans-ethnic study, where I supported the analysis as co-author [44]. Then, the chapter discusses the different hypothesis for the role of the HLA in the pathogenesis of IBD (Section 8.2) and

finishes with my own analysis of the HLA in UC based on the human genetics (PAPER C, Section 8.3). The second chapter of this part (Chapter 9), complements the thesis by discussing different aspects affecting the presented studies and giving an outlook on future work that could be conducted to identify the disease specific role of the HLA.

2

GENETICS

The genome of an organism is the basis of all appearance and the starting point for the investigation of most diseases. Susceptibility, immunity, missing and mislead functionality are based on genetics.

The human genome is divided into 22 autosomal chromosome pairs and two sex chromosomes, XX in case of female and XY in case of male individuals [200]. Each chromosome consists of two deoxyribonucleic acid (DNA) strands. Each strand is made of DNA nucleotides composed of a sugar molecule (deoxyribose), a phosphate molecule, and one of the organic base adenine (A), thymine (T), cytosine (C), and guanine (G) [65, 123].

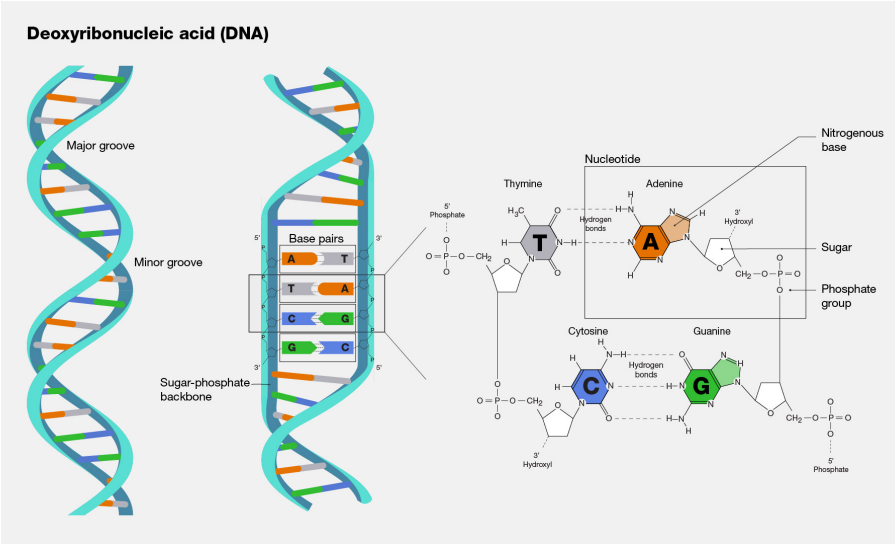


Figure 2: Structure of the DNA. The DNA double helix is constructed out of nucleotides. The nucleotides form a sugar-phosphate backbone along the strands and the nucleotides between the bands interact via hydrogenic bonds. Figure taken from the National Human Genome Research Institute [123].

Each base of one strand pairs together with the corresponding base of the other strand via hydrogen bonds – Adenine (A) pairs with thymine (T) and guanine (G) with cytosine (C) [65].

DNA strands are read from the 5' (phosphate) to the 3' (hydroxyl) end [65] and are also commonly referred to as plus (+) and minus (–) strand [65]. The plus (+) strand is defined as the strand with the 5'-end at the telomer of the short arm of the chromosome [129]. Telomers are the tip regions of the chromosomes, while the central part dividing the two arms of the chromosome is called centromere [65].

The human genome contains about 20 000 protein coding genes [52] i.e., genes that are transcribed from DNA to messenger RNA (mRNA) (transcription) and then translated info functional proteins (translation). The exonic regions within a gene define protein sequences of the translated protein. Intronic regions are nucleotide sequences that are most commonly not incorporated into the protein but may contain regulatory elements or carry non-informative nucleotide sequences [59, 155]. Variants of the same gene are produced by splicing events, which are called transcripts of a gene. During splicing, nucleotide sequences are cut from an immature mRNA containing the whole sequence read from the DNA to form a mature mRNA. By splicing, whole sections across both exons and introns may be omitted. [33, 203, 210]

More than 99% of the genome are identical in all people [57]. Parts of the remaining 1% of the remaining genome has been linked with basic phenotypic differences across individuals and different diseases.

While some diseases are caused by either chromosomal abnormalities, e.g., Down syndrome, or based on the disfunction of a single gene, e.g., hemophilia A. Other complex traits e.g., Alzheimer's disease or IBD, are influenced by variation across multiple genes at different locations across the genome [57]. Those variants often only have a small statistical effect size and can only be detected by studying large sample sizes in GWAS (Section 2.3).

2.1 REFERENCE AND VARIATIONS OF THE GENOME

2.1.1 *Genome builds*

To define a specific nucleotide variation and to be able to communicate it, a common nomenclature is necessary. With this goal in mind, the genome reference consortium (GRC) began to define genetic assemblies based on long, high quality sequences, as early as 2004. GRCh38, also called hg38, is the current major assembly for the human genome, which was primarily released in December 2013 [126]. The previous reference assembly GRCh37, also called hg19, is still in use [125]. Next to the 25 main assemblies for the different chromosomes (1-22, X, Y) and the mitochondrial DNA, additional alternative contigs (a continuous sequence generated from overlapping sequences from a single genetic source) were included in the reference. In the initial hg19 reference were 9 alternative contigs, 7 of those were for the HLA region and the remaining two for the genes *MAPT* and *UGT2B17* [66]. The number of loci with alternative contigs increased to 207 with 261 alternative contigs, including 8 alternative HLA sequences in GRCh38 [67].

Data with different genome builds may be brought together to the same build by a liftover as the UCSC implemented [73].

2.1.2 *Genetic variations*

Different versions of a whole gene or a genetic location are called an allele [122]. If there are only two variants a locus is called biallelic or otherwise multiallelic. A replacement of one base pair is called single nucleotide variant

(SNV). If the frequency of the minor (less frequent) allele is above 1% it is called single nucleotide polymorphism (SNP). Alternative variability originates from InDels (insertions or deletions) or copy number variations.

Genetic variants are differentiated based on the influence on the resulting protein sequence (Figure 3). Synonymous mutations are the most common mutations in the coding region, as they do not lead to a change in the amino acid sequence. A missense mutation induces the replacement of one single amino acid by a different one. InDel mutations are caused by insertions or deletions of nucleotides within the DNA sequence. If the number of inserted nucleotides is a multiple of 3 within a coding DNA sequence, the InDel mutation is called an in-frame mutation. In-frame mutations lead to the addition or removal of one or a few amino acids to/from the protein, possible paired with a replacement of the initial amino acid. In other cases, SNVs or InDels can lead to a more severe change in the protein structure, e.g., start loss or gain of a stop codon, leading to the premature termination of the amino acid sequence (truncated protein) or an InDel induces the shift of the reading frame (frameshift) which results in a different follow up sequence. Most of the latter mutations cause a loss of function (LoF) of the protein [97]. Compared to the reference genome, the genome of an individual carries approximately 3 to 4 million SNVs and 0.4-0.5 million InDels, whereof about 100 have a protein truncating effect [97].

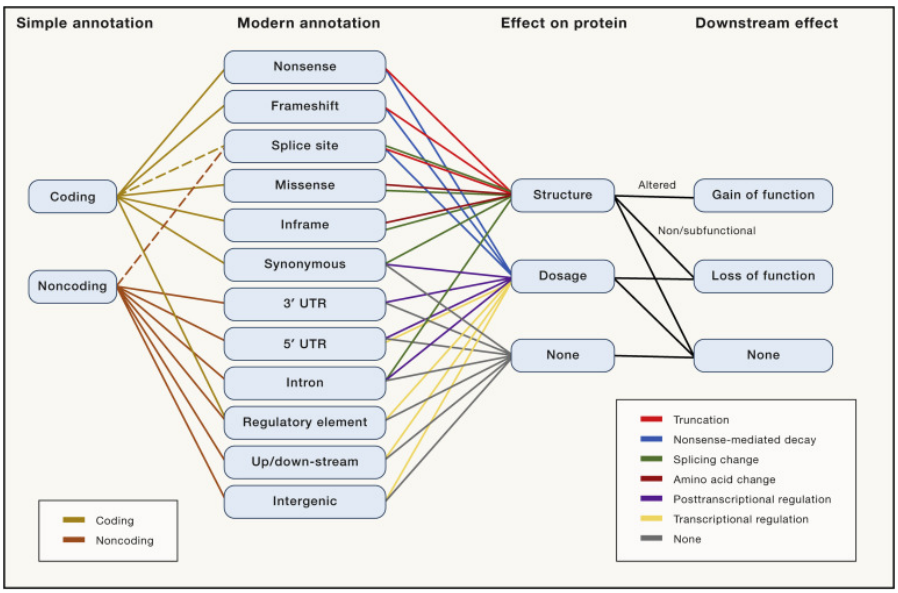


Figure 3: Functional annotation and downstream consequences of SNVs and Small InDels. Figure taken from Lappalainen et al. [97].

2.1.3 *Linkage disequilibrium*

A specific (sub-)sequence of nucleotides on one single chromosome is called haplotype. The frequency of haplotypes across a population is influenced by the linkage disequilibrium (LD). The genome of an individual is shaped amongst others by heritability, recombination of chromosomes via cross-over and de-novo mutations i.e., mutations not inherited from a parent. A group of neighboring genetic variants on the same haplotype are often

inherited together. As a result, the frequency of one variant is dependent on the frequency of another variant. This effect is called linkage disequilibrium (LD). [166]

About 20 years before the very first GWAS (Section 2.3) Lander and Botstein [96] proposed that screening the whole genome for disease loci might be possible by making use of the LD. Though further improvements, especially in genotyping (Section 2.2), were needed before GWAS could be performed and impacted the knowledge about many disease and phenotypic traits.

2.2 GENOTYPING

It took until 1972 for determining the sequence of a whole protein-coding gene after uncovering the general structure of DNA in 1953 [72]. Since then, different methods have been developed to define genetic sequences and individual genetic variation. The technologies make use of different biological, physical and chemical properties of the DNA. Those technologies differ in their throughput, accuracy, pricing, sequencing length, and de-novo applicability [74]. Consequently, different methods are used for different scientific questions.

Sequencing technologies are used to uncover the genome of a new species or rare, potentially individual mutations (Section 2.2.1). For GWAS (Section 2.3) typically genotyping arrays (Section 2.2.2) are used, as those enable the genetic characterization of huge sample sets for a selected set of known variants with comparably low cost and high accuracy [178]. The variants are usually selected to be representative for different haplotypes and enable an imputation of further variants (Section 2.2.3).

2.2.1 *Sequencing*

The first sequencing technology used over a long-time span was the dideoxy chain-termination method, also called Sanger sequencing [72]. It is defined as the standard method due to its accuracy, robustness, and ease of use. Briefly, in each single step of the analysis, four reactions are run parallelly. In each reaction one of the monomers of the DNA, the deoxyribonucleotide (dNTP) is partly replaced by the corresponding radiolabeled dideoxynucleotide (ddNTP) with a lacking 3' hydroxyl group required for the bonding of the next dNTP. Thereby the extension of the sequence is interrupted randomly after the occurrence of a specific base. Using gel electrophoresis, the sequence of the bases can be read. This step is repeated many times to analyze a whole stretch of DNA. The technology was later improved e.g., by changing the labelling to allow the reaction in only one vessel and by using capillary-based electrophoresis. Those improvements contributed to the development of increasingly automated DNA sequencing machines. [72, 74]

DNA sequences analyzed with Sanger sequencing typically have a length of approximately 1 kilobase (kb). For the generation of longer sequences overlapping sequences are generated and combined in silico to a continuous

sequence ('contig') [72]. Furthermore, Sanger sequencing is characterized by high accuracy, low throughput, and high costs. [142]

The first methods of the next generation of DNA sequencers (next generation sequencing (NGS)) officially appeared in 2005 [112, 116, 163]. Compared to Sanger sequencing, NGS comprised several different methods with a dramatic increase in throughput [116]. Two approaches described here are pyrosequencing as the first commercially available NGS technology and the nowadays most important NGS technology Solexa, later acquired by Illumina [72].

Pyrosequencing does not use radio- or fluorescently-labelled molecules, instead a luminescent method for measuring pyrophosphate synthesis is applied. In brief, a single dNTP is added during each reaction cycle. The inorganic phosphate that is naturally released during DNA strand synthesis induces a light signal based on the conversion of luciferin to oxyluciferin [72]. This in turn results in the emission of light, that can be measured [69, 72]. This technology uses the natural nucleotides and the resulting sequence can be measured in real time during synthesis [72].

The Solexa method starts with a library preparation where the DNA is fragmented, and adaptors are ligated to the fragments. The adapter-modified, single stranded DNA is then added to a flow cell. A flow cell contains anchor sequences (oligonucleotides) that are complementary to the adaptors. The fragmented DNA is thus immobilized by hybridization to the flow cell [58]. The DNA templates are then amplified by "bridge" polymerase chain reaction (PCR), relying on captured DNA strands "arching" over and hybridizing to an adjacent oligonucleotide [58, 192]. This process is also referred to as clustering, as due to the immobilized DNA sequences, clusters of sequences are generated [58]. The DNA fragments are sequenced in a sequence-by-synthesis manner using fluorescent 'reversible-terminator' dNTPs [72]. The fluorophore occupies the 3' hydroxyl position of the dNTP and must be cleaved off before the polymerization can continue. This allows synchronous sequencing (Figure 4) [58, 72, 192]. Light emitted by the fluorophores is captured by a charged couple device and then translated into a nucleotide sequence [72]. Typically read length of 150 base pairs are generated [77].

With application of NGS the costs and time for sequencing a whole genome decreased massively [74, 142]. This method is also named second-generation sequencing or short-read sequencing, in contrast to the first-generation sequencing (Sanger-sequencing) and the third-generation sequencing or long-read sequencing [74].

Third-generation sequencing technologies provide advantages over second-generation sequencing, but for a long time lacked in accuracy [74]. PacBio machines, which are single molecule real time machines, enable the measurement of kinetic data next to the sequencing information and can produce very long reads. Accordingly, they are suitable for the generation of de novo genome assemblies. Nanopore sequencing, as another example, allows sequencing of longer reads and the sequencing can be done in the field [72]. Until recently this method had an error rate of around 14%. Optimizations during the last years reduced this lack of quality massively [74].

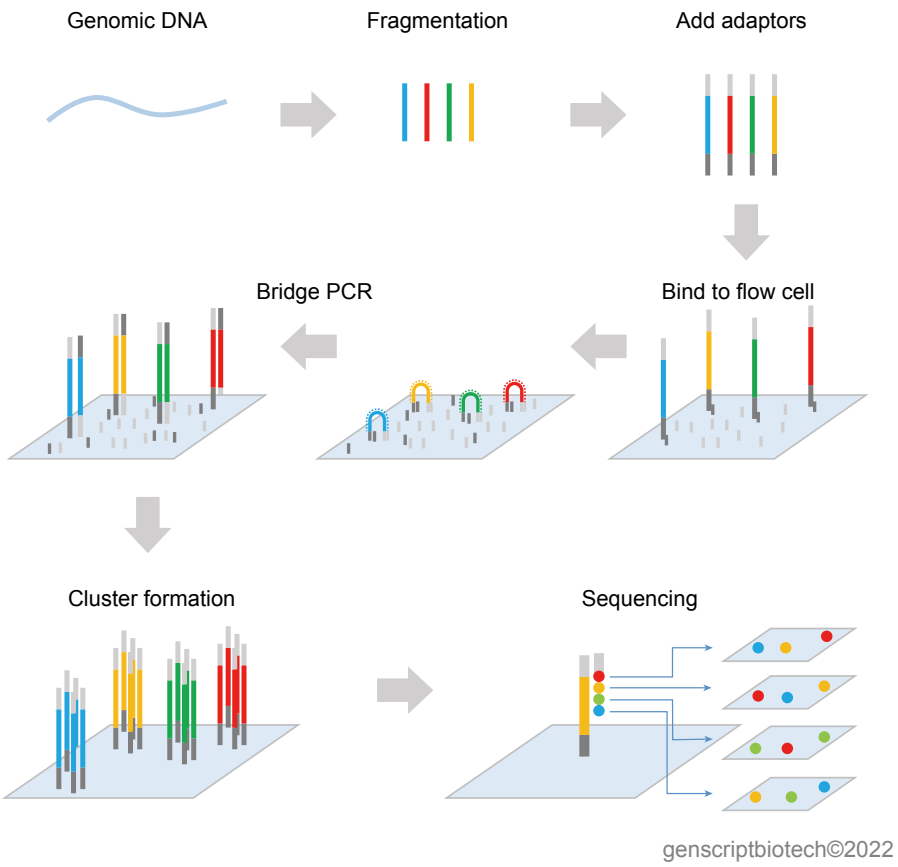


Figure 4: NGS workflow in the Solexa method as used by Illumina. [58]

Depending on the sequencing template, DNA sequencing includes whole genome sequencing (WGS), whole exome sequencing (WES), epigenome sequencing and targeted sequencing (TS) [74]. Template preparation can be based on the PCR (commonly used for TS) or hybridization capture-based approaches (commonly used for WES and TS).

In TS, opposed to WGS, targeted regions are sequenced to reduce cost and time of the experiment. TS can focus on single genes or a group of genes or even on whole exomes (all exons within a genome) (WES). The human exome represents about 2% of the genome. WES is designed to cover the entire exome. WES is unbiased, as no preliminary assumptions about relevant genes needs to be made. Most mutations with large effects on disease-related traits are expected to be covered by WES. [115, 142]

2.2.2 *SNP arrays*

In the early 2000s, high throughput genotyping was introduced, enabling genetic association studies (Section 2.3) in big datasets. This led to the identification of genetic variants with small statistical effects on different disease traits. [142]

In most cases those association studies are based on genotyping information generated by SNP arrays which enable the measurement of specific predefined sets of genetic variation.

Different SNP arrays are based on different chemistries but always call for the hybridization of fragmented single-stranded DNA to labeled oligonucleotide probes. Each array contains different unique nucleotide probe sequences. A signal intensity associated with each probe and its target after hybridization is measured and translated into specific genotype. Widely used SNP arrays are manufactured by the companies Illumina and Affymetrix and have been developed over time.

SNP arrays designed by Illumina use silica microbeats, hence they are called Illumina Bead Arrays, coated with multiple copies of a nucleotide probe. [78]

A variation of selection criteria offering different advantages and application possibilities are used to design SNP arrays. The first SNP arrays mainly included common variants. Nowadays a greater density of variants with a wider range of allele frequencies is included [178]. Variants covered by various arrays increased from a few 1000s to nearly a million over the years [95]. Whole-genome SNP arrays are typically designed to include variants with the highest information content on the genetic variation (utilizing LD) [95, 111]. Other SNP arrays have a special focus e.g.:

- On the analysis of protein-coding regions (exome chips). [178]
- On loci important for a specific research question (custom arrays like the Illumina ImmunoBeadChip also called Immunochip, designed to cover genetic variation related to the immune system) [37].
- On the ancestry of individuals [76, 111].

Multi-ethnic arrays are designed to cover genetic variation across different ethnicities (e.g., the multi-ethnic Illumina Global Screening Array (GSA)) [76].

Two genotyping arrays suited for the analysis of IBD and the HLA are the GSA and the Immunochip both designed by Illumina. The Infinium Global Screening Array-24 BeadChip (short GSA) is a whole-genome array that was designed to cover clinical research variants as well as allowing high imputation accuracy (Section 2.2.3) of nucleotide variants at a minor allele frequency (MAF) larger than 1% in all populations of the 1000 Genomes Project (1KGP) [6]. Version 3 of the GSA includes 654 027 fixed markers plus additional capacity for up to 100 000 custom markers. [76]

The Immunochip as a custom-array was designed to enable cost-effective analysis in the major autoimmune and seronegative diseases. The array contains 196 524 nucleotide polymorphisms. All of them were identified by GWAS on one of the considered diseases, including UC and CD. The chip also contains a dense set of SNPs in the HLA region. The cost of this chip was lower than other genome-wide chips, due to the comparably lower number of measured SNPs and high production numbers. However, the Immunochip was designed using information from Caucasian samples and might be less representative of variation observed in other populations and marker selection was based on previous GWAS. [37]

According to company specifications genotype call rates and reproducibility of the SNP arrays are >99.7% for both types [178]. Compared to Sequencing technologies, SNP arrays are relatively inexpensive (about 40US$

per sample), highly accurate and the analytical pipelines for GWAS analysis are well established for array data. Though the acquired data is mainly restricted to common and low-frequency variants with a bias towards well-studied or sequenced populations [178].

2.2.3 *Phasing and Imputation of Genotypes*

As mentioned above, SNP arrays (Section 2.2.2) are commonly used for genetic analysis. They consider only a predefined set of genetic variants without phase information. Non-genotyped variants might, however, be of biological interest or relevant to compare studies that were e.g., measured on different genotyping platforms. This data can be generated to some degree from SNP array data by imputation, using reference genomes Figure 5. Imputation methods take advantage of local LD patterns that enable the inference of SNP genotypes within a region from genotyped SNPs that are representative for a haplotype. These representative SNP are called tag-SNPs and are chosen as part of the SNP array design.

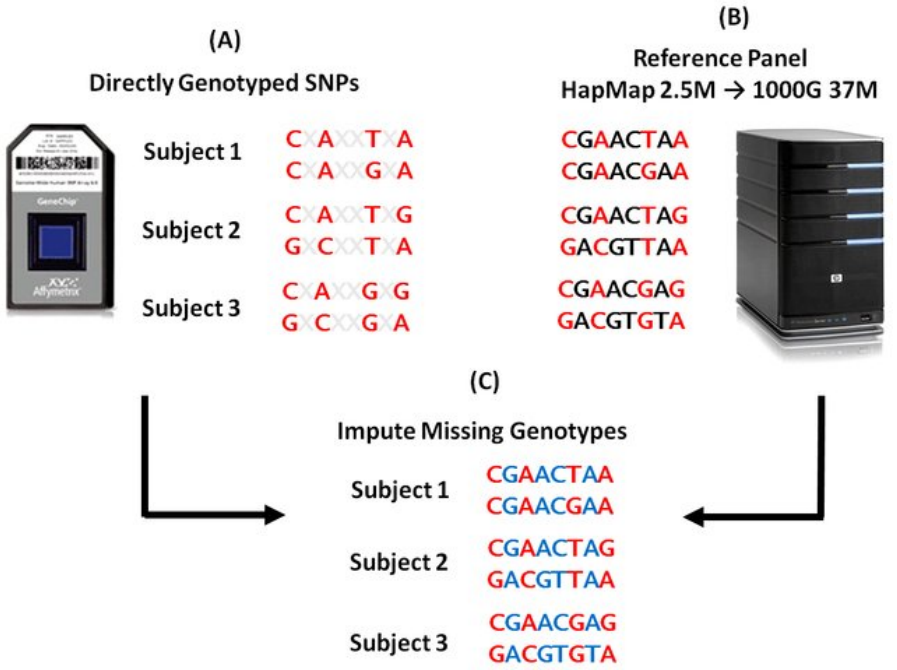


Figure 5: Schematic overview of imputation. (A) Genotyped SNPs are phased to estimate haplotypes that alleles reside on. (B) Publicly available reference haplotypes with dense nucleotide sequence coverage ate downloaded from projects including the International HapMap Project and the 1KGP. (C) Phased haplotypes from the study are assessed and compared to the publicly available haplotypes. The most likely genotypes are imputed based on the alleles found in the reference haplotype panel. Figure from Wood et al. [202].

Imputation is typically performed in two steps, phasing of the genotypes and the imputation. Theoretically, phasing and imputation could be performed in one combined step, but the separated approach is common practice for computational efficiency [171], though slightly less accurate. Both processes are based on high quality reference data commonly produced by large consortia. These are used to infer the phase of the data and to infer

missing genotype information. Alternative approaches, in which only the study genotypes themselves are used as a reference for both phasing and imputation, are uncommon but also possible. The size of the reference panel is directly and inversely related to the imputable allele frequency [178]. The number of imputable variations is only expected to increase further in the coming years.

The most commonly and widely used reference panels are the 1KGP panel (number of individuals (n) = 2 504 in phase 3), the Haplotype Reference Consortium panel (HRC, n = 32 470) and the most recently released TOPMed panel (n = 97 256) [121, 178].

The 1KGP first released sequencing data of the initial 1 092 (phase 1) whole genomes in 2012 [6] and up until today it is the only fully open source WGS data set consented for public distribution of raw sequence data without access restriction [25]. The phase 3 data include 88 million phased variants (84.7 million SNPs, 3.6 million InDels and 60 000 structural variants) from 2 504 individuals from 26 different populations. This resource includes more than 99% of SNP variants with a frequency of >1% for a variety of ancestries. The dataset is generated out of a mixture of low-coverage WES, deep WES, and dense microarray genotyping [15]. Recently, an updated version was published recently including a higher coverage of the genome resulting in more variants for the sample set and 602 related parent-child trio samples [26].

The first version of the Haplotype Reference Consortium (HRC) reference panel initially described 2016 by McCarthy et al. [113], comprised 64 976 haplotypes from 32 488 samples and provides accurate genotype imputation at MAFs as low as 0.1%. The dataset includes the phase 3 data of the 1KGP. It comprises a set of 39 235 157 variants and does not include any InDels. The majority of these samples is of Caucasian ancestry. [113]

The largest imputation reference panel so far is the 2021 presented Trans-Omics for Precision Medicine (TOPMed). It includes 97 256 individuals of diverse ancestries and 308 107 085 SNV as well as InDels based exclusively on deep WGS data [177]. As stated in Taliun et al. [177], this reference can be used to accurately impute ($r^2 > 0.3$) variants with minor allele frequencies above 0.002-0.003% in individuals of European and African ancestry.

Even though the TOPMed dataset is not publicly available, it can be used for imputation and phasing via publicly available imputation servers: the TOPMed Imputation Server and the Michigan Imputation Server. [39]

Phasing algorithms are typically based on a Hidden Markov Model [171]. Briefly, probabilities for a specific allele or haplotype are calculated based on given variants. The probabilities are defined based on the haplotype information in the reference dataset. Some of the most common phasing tools are Eagle (Eagle2.4.1 [107]), SHAPEIT (SHAPEIT4 [45]), and Beagle (Beagle5 [20]). For imputation, the commonly used tools are Minimac (Minimac3 [39], Minimac4 [118]), IMPUTE (IMPUTE5 [156]) and Beagle (Beagle 5.4 [21]). All those tools have been improved continuously over the years, often with the focus on computational time. A combination of Eagle v.2.4 and Minimac4 can be run online on the TOPMed imputation Server [121] with the TOPMed reference or the Michigan Imputation server [39, 117] using different references: 1KGP, CAAPA (a reference panel mainly for

African Americans), HRC, or TOPMed [39]. According to Stahl, Gola, and König [171] all those tools are comparable in performance.

2.3 GENOME WIDE ASSOCIATION STUDIES

The aim of genome wide association study (GWAS) is the identification of variants that are statistically more prevalent in individuals with a certain trait [95]. The identified variants are expected to lead to disease predisposition or inherited together with another variant, leading to disease predisposition [95]. It can be discussed if the first GWAS was presented by Ozaki et al. [141] in 2002 on myocardial infections and identified a candidate locus on chromosome 6p21 including the HLA through gene-based genotyped SNPs, or if the first GWAS was conducted in 2005 by Klein et al. [93] as they applied Bonferroni-correction to define the significance. They presented the first Manhattan-Plot plotting on all SNPs vs their $-\log_{10}(\mathrm{P})$ association and presented loci of interest as high "skyscrapers" (even though it was a barplot, where nowadays a scatterplot is used). Most GWAS make use of an autosomal additive model [178]. Therefore, a logistic regression model is applied, fitting the phenotype (typically cases vs. control) in dependency of the genotype dosages of a variant. Genotypic dosages are most commonly coded as 0 (homozygous for the major allele), 1 (heterozygous), and 2 (homozygous for the minor allele).

More than 50 000 genome-wide associated variants for different traits have been reported since the first GWAS was performed [178]. The NHGRI-EBI GWAS Catalog, a publicly available database that catalogues variants observed in different GWAS, comprised 71 673 variant-trait associations from 3 567 publications in September 2018 [24]. According to Visscher et al. [190] the number of observed risk variants is predicted to increase for all traits with the ever-growing size of discovery data sets, enabling the identification of low-frequency variation [178, 190]. Some of the associated variants have led to a better understanding of the disease susceptibility and allowed the development of new drug targets, disease biomarkers or personalized treatments [95, 178]. The study of Nelson et al. [128] shows that the drug development using genetically supported targets has a higher success rate.

As in GWAS many statistical tests are performed simultaneously, a correction for multiple testing is necessary to filter out false positive observations. A classical approach would be the application of the Bonferroni correction. Ergo, the accepted significance threshold of $\alpha = 0.05$ is divided by the number of conducted statistical tests. Due to the LD (Section 2.1.3) not all variants observed across the genome are independent of each other, in consequence a commonly set genome-wide significance threshold is defined as 5.0×10^{-8}, based on the approximately 1 million independent variants in the HapMap Phase II [55] data set [86, 142]. To reach those significance levels it needs large sample sizes, high allele frequencies, and large effect sizes.

2.3.1 *Quality Control*

An appropriate and stringent quality control (QC) is a key factor for the generation of reliable and replicable findings in GWAS. Biases might be introduced into a study by the sampling procedure or technical issues of the genotyping experiment. Even though a complete removal of those biases post data generation is impossible, a carefully conducted QC can help to reduce biases and may uncover potential problems with genotyping data.

Typical QC steps are:

1. Remove variants with a high genotyping missingness.
2. Remove samples with a high genotyping missingness.
3. A principal component analysis (PCA) to control for population stratification and perform the exclusion of ancestral outliers.
4. Control variants for the Hardy Weinberg equilibrium (HWE) to remove variants with poor genotyping quality.
5. Assess the degree of relatedness amongst the samples.
6. Analysis of sex for consistency between reported and genetic sex of the samples.
7. Eventually analysis of batch effects, caused by i.e., measurements of batches of samples at different timepoints or labs.

2.3.1.1 *Missingness*

A high missingness of variants within a sample or overall missingness for a single variant is an indicator for problems either caused from sample processing or genotyping. For example, the DNA that is used for a sample might have been of bad quality or contaminated. For a variant, the capturing of this genetic variant might have failed systematically. Therefore, variants and samples with high missingness are removed in the quality control. The cutoff value defining "high" missingness changed over the years with improving sequencing procedures and depending on the sequencing technology. Kässens, Wienbrandt, and Ellinghaus [89] defined a default threshold of 0.02 for a variant or 0.02 for a sample.

2.3.1.2 *Principal component analysis (PCA)*

By conducting a PCA, the similarity between different samples can be analyzed. Hence, genotyping data are represented in reduced dimensions and general patterns within the genetic data can be reflected. Based on the preferences of the user, several principal components (PCs), typically 10, are calculated. Each PC captures some variability of the data, with the first and second PC containing the highest degree of information. Since PCA should be conducted on uncorrelated data, genomic data are pruned, meaning that only independent variants, based on the LD are chosen for analysis. Regions with generally high LD, such as the HLA region, are excluded.

For the identification of differences in genetic background, the data of interest is normally mapped onto the PCs generated by data with diverse and known ethnical background, for example the 1KGP dataset [15]. Outliers within the dataset are then removed as big trait independent allele variations are expected for those samples [9]. Further, the first principal components are often used as covariates in the association analysis to account for remaining trait independent patterns.

2.3.1.3 *Hardy Weinberg equilibrium (HWE)*

In general variants are expected to follow the law of Hardy-Weinberg [172]. A biallelic variant with a MAF f is expected to be heterozygous in $2 \times f \times (1 - f)$ of the samples and homozygous with the minor allele in f^2 of the samples. Variants which fail this rule indicate a genotyping or genotype calling error [9]. Deviations from HWE may also be indicative of selection, therefore variants related to a disease might fail HWE in case samples. In conclusion, during QC the evidence of deviation from the HWE is calculated [9, 89]. For this only control samples are used [9, 89]. Samples with a HWE p-value less than a predefined threshold are then removed [9, 89].

2.3.1.4 *Sex checks*

Information about the sex of an individual is often available. The sex can also be defined based on the genotyping information on the sex chromosomes. As males only have one X chromosome, they are expected to be homozygous on all variants of the X chromosome and have non-null calls for the Y chromosome. Females in contrast with two X chromosomes are expected to be heterozygous for some variants on the X chromosome but have only null calls for the Y chromosome. A comparison of the reported sex and the genotyping sex can help to identify plating errors or possible errors in metadata. Samples with discordant sex should be further investigated or be removed from the analysis. [9]

3

IBD

IBD is a complex disease. Complex diseases are believed to be caused by the interplay of a variety of genetic variations and the environment. IBD has a higher prevalence in industrialized countries in comparison to non-industrialized countries. IBD can be categorized into two main forms UC and CD. Typical symptoms of this diseases are abdominal pain, diarrhea, fatigue, weight loss, rectal bleeding, and bloody stool. Additional symptoms in severe cases, are blood and mucus in the stool, fistula, fissure, and anemia (Figure 6) [181]. Between 6% and 47% of IBD patients reported additional extraintestinal manifestations [187] i.e., manifestations outside of the gastrointestinal tract, e.g., ocular manifestations [176].

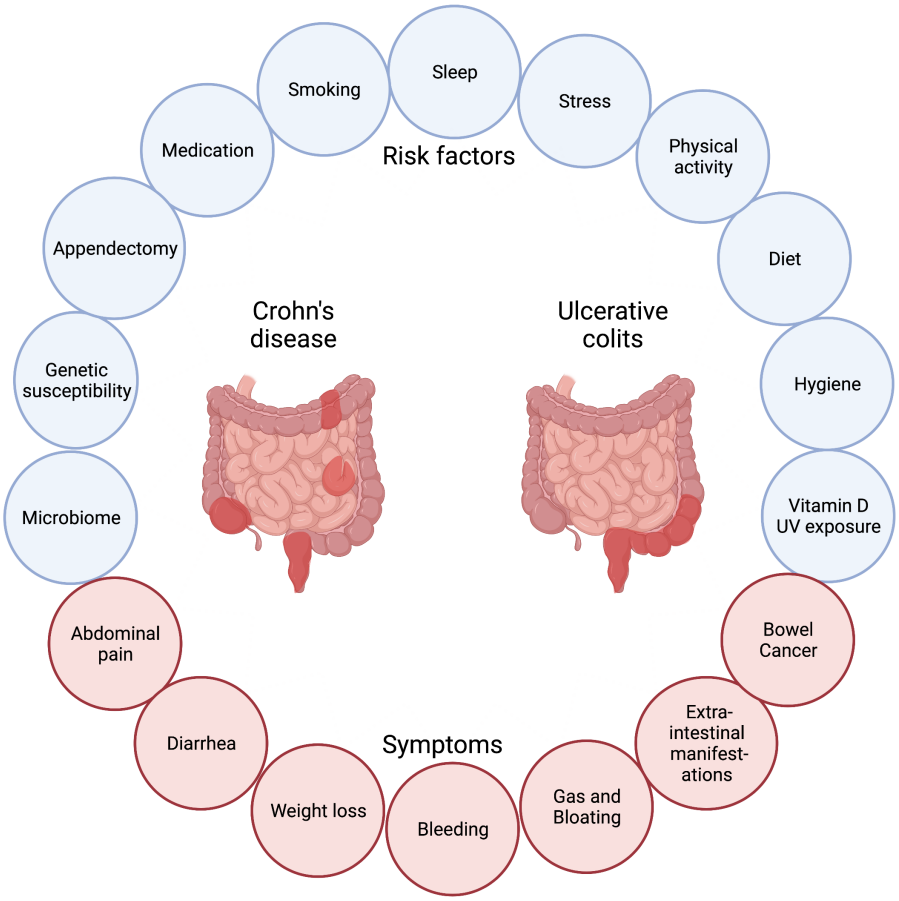


Figure 6: Risk factors, symptoms, and inflammation pattern in IBD.

In the 20th century, the prevalence and incidence of IBD in industrialized countries increased steadily [130]. During the last decades, the incidence in industrialized regions like North America and Europe stagnated with a

prevalence above 0.3% [130]. In other parts of the world, the progression of the disease seems to be in an earlier stage and the incidence is still increasing [130].

IBD is characterized by a remittent or progressive inflammation of the bowel [88]. CD can affect the whole gastrointestinal tract, while UC effects mainly the colon. While CD often shows a patchy pattern of inflammation, UC spreads continuously from the rectum along to the colon (Figure 6) [206]. Histologically, in both diseases the epithelial layer of the bowel is disrupted. In CD the whole epithelial layer and the lamina propria may be affected [206]. In UC the inflammation is characteristically limited to the mucosal surface and the tight junctions are defect, which also increases the permeability of the epithelium [140].

More recent research has reported a dysbiosis of bacteria in the gut in both CD and UC when compared to healthy individuals [170]. The diversity of bacteria in IBD patients is decreased (Figure 6). The totality of bacteria observed at a specific site in the body is also referred to as the microbiome. One common hypothesis in IBD research is that the disruption of the mucosal barrier (Figure 6) allows bacteria to enter the tissue, which in turn causes inflammation. This is accompanied by the presence of activated T cells with an increased amount of T helper cells and a decreased amount of regulatory T cells (Figure 7) [170]. However, it is unclear whether the impaired barrier and inflammation are cause or consequence of the disease. Overall, the disease etiology of IBD is only partly understood. It is, as stated above, multifactorial, including: a genetic predisposition (Chapter 4), the gut microbiome composition, dis-regulated immune responses and environmental factors [184, 206]. Exemplary environmental risk factors associated with IBD are stress, hygiene, diet, and smoking (Figure 6) [88, 205].

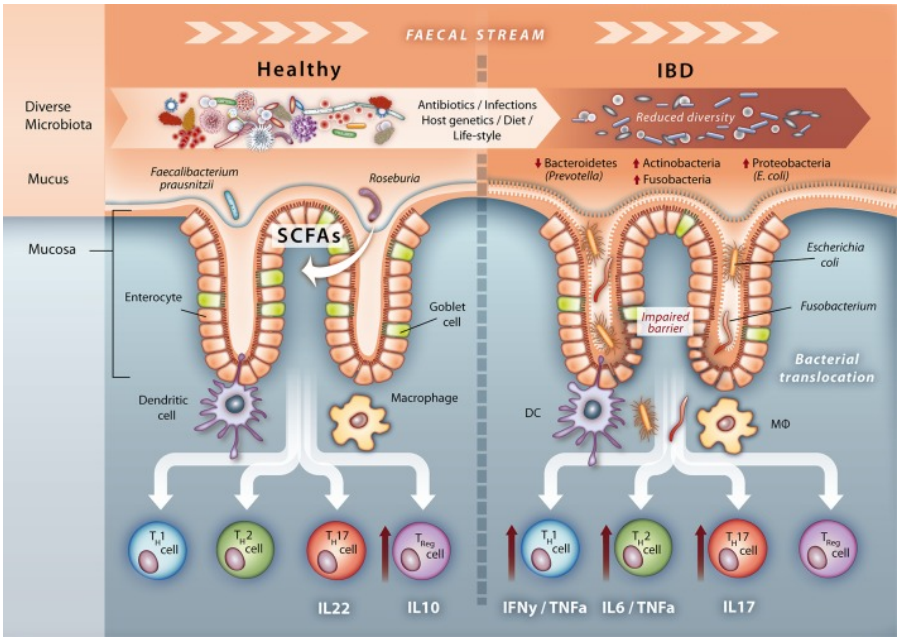


Figure 7: Microbiome signatures of a healthy gut and in IBD. Figure taken from Sommer et al. [170].

Due to the complex background of the disease, there are different classes of treatments: either only affecting the symptoms or influencing one par-

ticular biological pathway [137]. None of the drug treatments have been reported to lead to a curation of the disease, only a remission of the acute inflammation and a procrastination of another flare of the disease may be reached [137]. Currently, the knowledge on which treatment is optimal for which patient is limited and many patients change their medication during treatment several times [137]. Other IDEAs, in the large treatment spectrum for IBD, include curative surgeries in which parts of the bowel are resected. This is usually considered as the last option. A curation of UC might be possible but often accompanied with severe side effects [184]. For a more detailed overview across treatment options and drugs in IBD see Yeshi et al. [205].

4

GENETICS IN IBD

To date, about 240 genetic risk loci have been associated to IBD [11, 40, 82]. Associations of these loci to the disease are based on statistical analyses, such as linkage analysis in IBD families or genome-wide association analyses in large cohorts of diseased and healthy unrelated individuals. Most risk variants have a small statistical effect size on the case control status of IBD cohorts based on logistic regression (odds ratio (OR)<1.3). The majority of IBD patients carry many of those common variants [42, 50]. About 70 % of the 240 gene loci overlap between UC and CD [206].

Most of the observed association signals are consistent between cohorts of different ethnic backgrounds [105]. Variations in disease association of a locus between different cohorts are predominantly driven by the distinct MAFs of variants and in some cases the effect size of the associations (Figure 8) [105].

In specific ancestries, variants associated with IBD are absent or highly invariable (low MAF), as is the case for variants located in the *NOD2* gene. The strong association of nucleotide-binding oligomerization domain 2 (NOD2) with CD (Figure 8) is based on rare mutations present mainly in Caucasian samples [54, 105] (see also PAPER C, Section 8.3). The most significantly associated CD locus among the Asian population is interestingly therefore not NOD2 but *TNFSF15-TNFSF8* [105].

The association of the HLA locus with IBD, the strongest association in UC, was reported with a different OR between Caucasian and East Asian samples by Liu et al. [106]. A more detailed analysis by Degenhardt et al. [44] showed that the associations within the HLA locus are highly conserved in UC across different ancestries, based on which HLA alleles are associated and concur with Liu et al. [105] on differences in effect sizes. The potential role of HLA in IBD is reviewed and analyzed in more detail in Part iii.

For many associations, the causal disease driving gene or variant remains unknown [42, 185] and up until now only a few genes have been well-studied and linked to IBD [206]. The translational results and the direct consequence of identified variants on the disease phenotype are still poorly understood [11].

However, the genes identified to be associated with IBD can be categorized broadly in three groups: Genes involved in pathogen recognition, genes involved in pathogen clearance by innate and cell-mediated immunities, and genes hindering pathogen invasion through the intestinal mucosal barrier [206].

Graham and Xavier [64] categorized IBD associated genes into seven categories and pathways related to the three groups described by Younis,

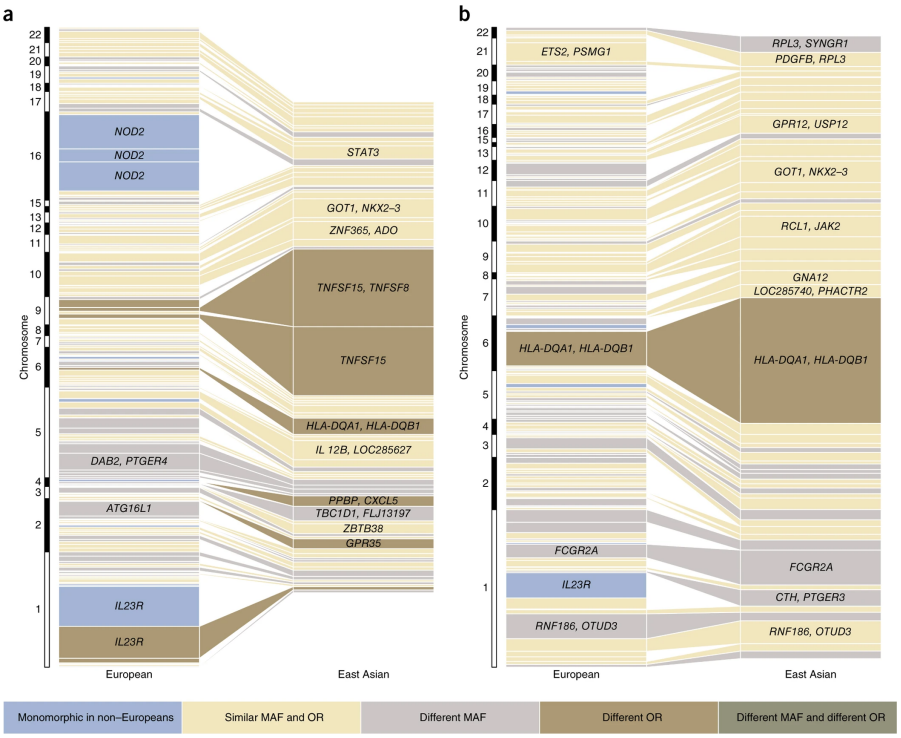


Figure 8: Variance explained per risk variant for CD (**a**) and UC (**b**) in European and East Asian samples. This figure is taken from Liu et al. [105]. In this first trans-ancestry association study of IBD the authors analyzed the variance in disease liability explained by the single independently associated SNPs. Each box in the figure represents one of the associated variants. The box size is relative to the variance in the disease liability. The coloring, as shown in the legend, highlights whether variants differ between the European and East Asian cohort in MAF or OR or both.

Zarif, and Mahfouz [206] (Figure 9a-g). The first category encompasses "microbe-sensing and effector pathways" (Figure 9a) [64]. The group of genes in this category include the aforementioned *NOD2* gene. The gene product of *NOD2* is an intracellular pattern recognition receptor, that can recognize bacterial peptidoglycans and induce an immune reaction [54, 105, 159]. The second category describes effects for the integrity of the intestinal barrier (Figure 9b) [64]. The third category describes the adaptive immunity (Figure 9c) [64]. This pathway involves the HLA locus introduced in Part ii. Briefly, the HLA molecules present small peptides from different sources to T cells and can thereby induce an immune reaction [105, 206]. The fourth category described by Graham and Xavier [64] includes genes involved in inflammation and fibrosis (Figure 9d). This pathway and the associated genes might lead to a pathological inflammation-healing [64]. The fifth category leads to the cell death induced by cell-stress (Figure 9e) [64]. The gene *ATG16L1* is an example for a gene association pointing towards a new pathway, as the role of autophagy in CD was previously not known [178]. The sixth described category deals with the communication between cells by different cytokines (Figure 9f) [64]. The seventh and last category describes the inflammasome signaling (Figure 9g) [64]. This is another microbe-sensing pathway, which is placed in the cytosol of the host cells based on inflammasome complexes [64].

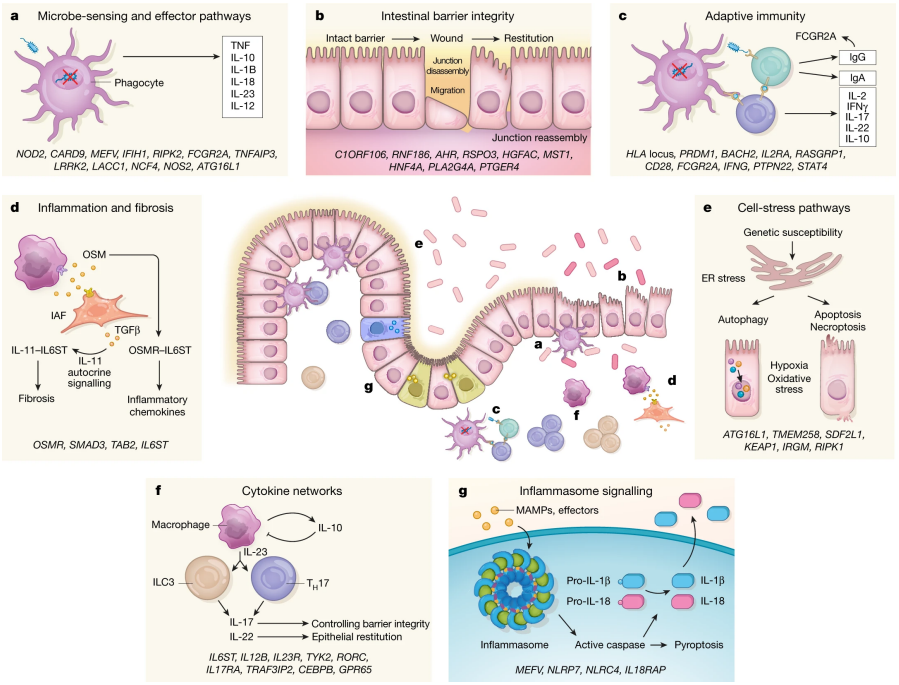


Figure 9: Pathways identified by IBD genetic analysis. Disease associated genes are noted at the bottom of each panel. Figure taken from Graham and Xavier [64].

Although over 240 variants have been identified, only about 7-37 % of phenotype variance (heritability) can be explained by genetic variation based on GWAS data [62, 82, 105]. Twin studies have estimated the heritability to be more than twice as high with 75% in CD and 67% in UC [62]. This observed difference between these two measurements, is also referred to as the effect of "missing heritability" and is based on different factors: The heritability estimated by GWAS might be underestimated, as only variants reaching the significance threshold are considered. Furthermore, GWAS rely on the inference of untyped variants from imputed references based on LD of variants, however knowledge on the LD might be incomplete [62]. On the other hand, the heritability estimated by twin studies might be overestimated. The heritability calculated based on twin studies is grounded on a comparison of monozygotic to heterozygotic twin pairs [62]. The attributed assumption that the impact of environmental factors for monozygotic and heterozygotic twins is are identical fails for different reasons [62]. For example, in contrast to heterozygotic twins, 30% of monozygotic twins share the same placenta [62]. Additionally, heterozygotic twins might have different gender leading to a sex bias. Moreover, behavioral traits, like smoking, are also correlated to genetic factors [62]. The true quantitative heritability might never be known [62].

The genetics of IBD was mainly studied by GWAS based on SNP array genotyping data. This approach is described in more detail in Section 2.2.2 and Section 2.3. It is limited to variants present on the array or in an applied imputation reference [64]. By using WES or WGS to study the genetic background novel and rare variants may be identified [64]. I performed a GWAS based on paired genotyping and exome data in PAPER C (Section 8.3) with a special focus on the HLA gene locus.

Part II

THE HUMAN LEUKOCYTE ANTIGEN

The human leukocyte antigen is a key factor in our adaptive immune system. It has been identified as an important risk factor, not only for inflammatory bowel disease but also for other immune-related diseases. In the following, I introduce the function, structure, and nomenclature of the human leukocyte antigen. Furthermore, I describe how alleles of the human leukocyte antigen can be inferred from genetic data by HLA allele imputation and computational methods can facilitate the prediction of which peptides an HLA allele will bind. In this chapter, Paper A and Paper B will be included. In Paper A, we generated a new multi-ethnic reference panel to impute human leukocyte antigen alleles. In Paper B, we used novel human leukocyte antigen microarray data to predict the binding of peptides to a human leukocyte antigen protein.

5

INTRODUCTION INTO THE HLA

The HLA locus is located on chromosome 6p21.1-p21.3. The region encodes over 140 genes, of which many play a role in the immune system [168]. Often only the three classical HLA class I, six classical HLA class II genes, and their products are considered in scientific studies. All other jawed vertebrates have equivalent genes to the HLA. A broader term for HLA is major histocompatibility complex (MHC) [94]. HLA is specific for humans, while e.g., BoLA Bovine leukocyte antigen is the name of MHC in cattle and histocompatibility system 2 (H-2) in mice [16, 90, 151].

5.1 BIOLOGICAL PATHWAYS OF THE HLA

HLA class I proteins play an important role in adaptive immunity against intracellular pathogens like viruses or cancer (Figure 10) [144]. They are expressed on the surface of all nucleated cells [157] and present peptides of 8-12 amino acids length to $CD8^+$ T cells [17, 27]. These peptides originate from inside the cell, e.g., from a replicating virus. Within the cell, these peptides are cleaved by the proteasome, transported into the endoplasmic reticulum (ER) via the transporter associated with antigen processing (TAP) protein complex and loaded there onto the HLA binding cleft [144]. The loaded HLA is subsequently transported to the cell surface [144]. The interaction between the HLA-peptide complex with a TCR can activate the T cell and induce the destruction of the peptide presenting cell [144]. While only some HLA presented peptides are T cell epitopes, all T cell epitopes need to be presented by an MHC molecule [98].

HLA class II molecules are only expressed on the surface of professional antigen presenting cells (APCs) [7]. These include dendritic cells (DCs), B cells, and macrophages. The HLA class II molecules are synthesized in the ER together with the invariant chain. A small peptide of this molecule is the class II-associated invariant chain peptide (CLIP). It occupies the binding groove to impede other peptides within the ER to bind to the HLA [134]. Antigens enter the cell via phagocytosis, pinocytosis or receptor-mediated endocytosis. They are processed and broken down into smaller peptides in the vesicles of the endocytic pathway [161]. The invariant chain is degraded and the HLA protein is loaded with the antigenic peptide. This process is assisted by HLA-*DM* and HLA-*DO*, two non-classical HLA class II molecules [153]. The HLA-peptide complex is then presented on the cell surface. When it is recognized by a CD4+ T cell different immunological processes can be set into motion [153]. Briefly, these processes encompass [1]:

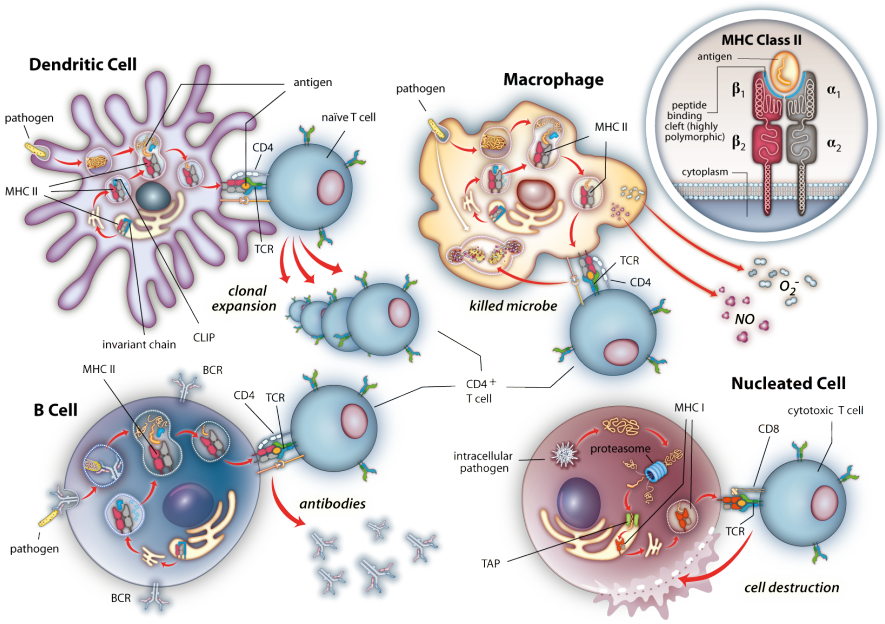


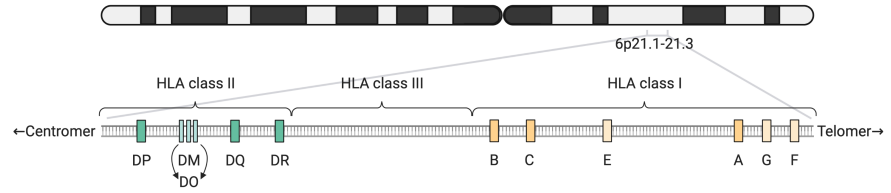
Figure 10: Biology of the HLA for classical HLA class I molecules (nucleated cell, bottom left) and for classical HLA II molecules for the different types of APCs (Dendritic Cell: top right; Macrophage: top left; B cell: bottom right). The figure was generated in cooperation with the designer Renate Nikolaus.

- clonal expansion of the antigen specific T cells,
- activation of B cells involving stimulation of antibody secretion,
- or release of substances like nitric oxide (NO) and superoxide (O_2^-) by macrophages which are important for the elimination of a pathogen.

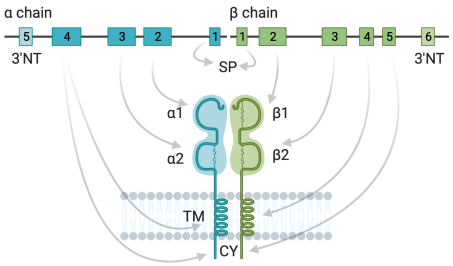
Other than HLA class I molecules, HLA class II molecules are mainly responsible for the defense against extracellular pathogens like bacteria [13, 153]. Peptides are presented by classical HLA proteins. The classical HLA genes comprise the HLA class I genes HLA-*A*, HLA-*B*, and HLA-*C*, and the HLA class II genes HLA-*DPA1*, HLA-*DPB1*, HLA-*DQA1*, HLA-*DQB1*, HLA-*DRA*, and HLA-*DRB1* (Figure 11a) [13]. Some HLA haplotypes encode for an additional HLA-*DRB1* paralog, either HLA-*DRB3*, -*DRB4* or -*DRB5*, with the same structural and functional role [13, 43]. Each human individual has between three and twelve different peptide presenting HLA class II molecules on the surface of its APC depending on the combination of its genotype (Figure 12).

5.2 STRUCTURE OF THE HLA GENE AND PROTEIN

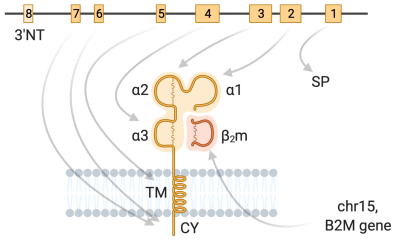
The classical HLA class I and class II proteins, besides their functional differences, also differ in their protein structure. HLA class II proteins are heterodimers built by an α and a β chain (Figure 11b) [147]. The two chains are encoded by a pair of classical HLA genes, e.g., HLA-*DRA* and HLA-*DRB1* [1]. Both chains contain a cytoplasmic region, a transmembrane region, and two domains of the extracellular domain including the



(a) Chromosomal organization



(b) HLA class II: DNA to protein

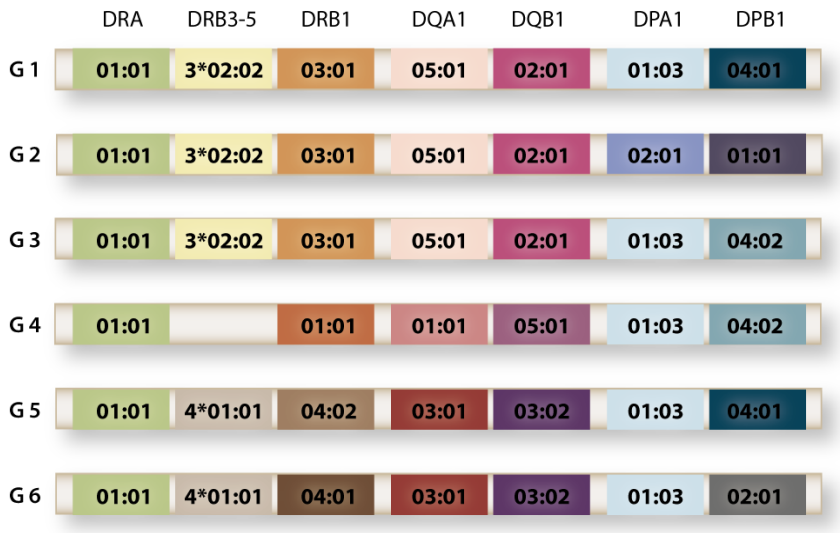


(c) HLA class I: DNA to protein

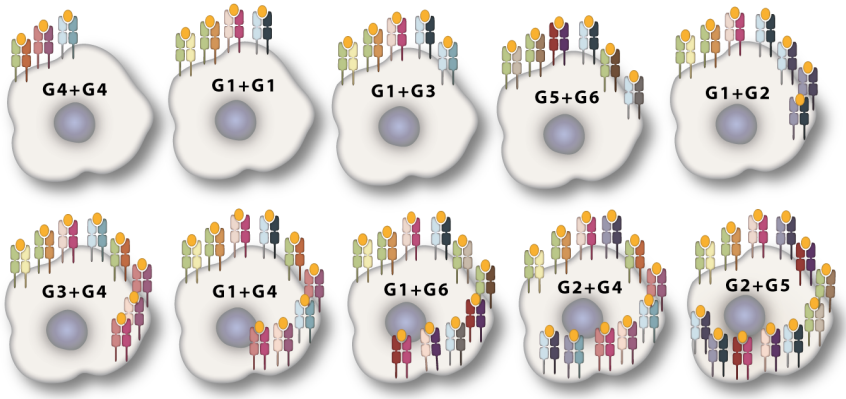
Figure 11: The HLA from the chromosomal location (Figure 11a) to the mature protein (Figure 11b and 11c). Figure 11a shows the location of the HLA genomic region on the short arm of chromosome 6 (6p21.1-21.3) and the organization of the classical HLA loci - DP, DQ, DR, B, C, and A - as well as a small selection of other non-classical HLA genes. Figure 11b shows the exonic structure of the HLA class II α and β genes (direction is turned for graphical reasons) and the resulting mature protein. Exon 1 of the α and β chain code for the signal peptide (SP). Exon 2 of the α and β chain code for the $\alpha 1$ and $\beta 1$ domain, including the highly polymorphic peptide binding cleft of the heteromeric molecule. Exon 3 of both chains encodes the $\alpha 2$ and $\beta 2$ region, respectively, responsible for CD4 binding and structure stabilization. Exon 4 of the α chain and Exon 4 and 5 encode the transmembrane region (TM) and cytoplasmic region (CY) of the protein. Exon 5 of α and exon 6 of β are 3' non translated region (3'NT). Figure 11c shows the exonic structure of the HLA class I genes and the resulting mature protein. The exon 1 codes for the SP. Exon 2 and 3 encode for the $\alpha 1$ and $\alpha 2$ domain including the highly polymorphic peptide binding cleft. Exon 4 encodes for the $\alpha 3$ region responsible for CD8 binding and structure stabilization. Exon 5 codes for the TM and exon 6 and 7 for the CY part of the protein. Exon 8 encodes the 3'NT. The lighter chain of the class I molecule, comprising the $\beta 2 \mathrm{m}$ region is encoded by the beta-2 microglobulin (B2M) gene on chromosome 15. (The figure was created with `BioRender.com`).

peptide binding cleft ($\alpha 1$ and $\beta 1$) and the binding region for the T cell co-receptor CD4 ($\alpha 2$ and $\beta 2$) [1].

HLA class I proteins are also heterodimers, one polypeptide chain is encoded by the HLA gene, the other polypeptide chain is a β_2-microglobulin encoded by the *B2M* gene on chromosome 15 (Figure 11c) [127, 164]. The β_2 microglobulin forms only one extracellular protein domain [1]. A HLA class I gene encodes for three extracellular domains, the $\alpha 1$ and $\alpha 2$ domain forming the binding cleft for the peptide, and the $\alpha 3$ region important for



(a) Exemplary HLA class II haplotyes



(b) Exemplary HLA class II phenotypes

Figure 12: Different genotypes lead to 3-12 different HLA class II proteins on the surface of the APCs. Figure 12a shows six different common HLA class II haplotypes. Figure 12b shows exemplary phenotypes based on combinations of those genotypes with an increasing number of different HLA class II molecules. The figure was generated in cooperation with the designer Renate Nikolaus.

binding of the T cell co-receptor CD8 [1]. Additionally, the HLA class I gene encode for a cytoplasmic and a transmembrane region [1].

All classical HLA genes, except for *DRA*, are highly polymorphic, especially within the binding groove of the molecule. The binding groove is generally formed by a β-sheet with two framing α-helices (Figure 13) [1]. In case of HLA class II, the binding groove is open at both ends, enabling peptides of variable lengths to interact with the HLA molecule [1, 47, 108]. The binding groove of HLA class I molecules is closed, and the interacting peptides are restricted to peptides with lengths of around 9 amino acids [1].

5.3 NOMENCLATURE OF THE HLA

When research into HLA began in the late 1950s to early 1960s [75], HLA proteins were classified into different serotypes, based on serological test-

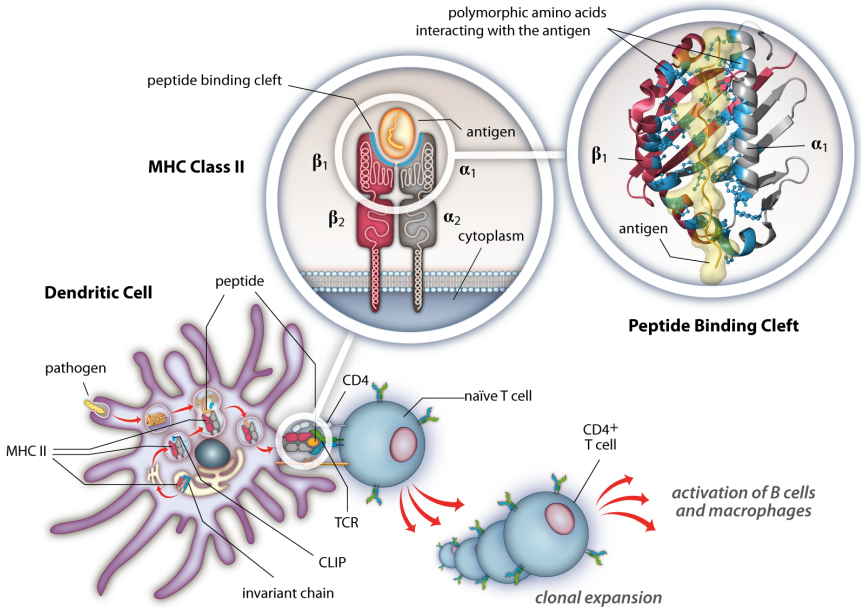


Figure 13: Peptide presentation by HLA class II proteins. Antigens of variable lengths can be loaded into the peptide binding cleft formed by two α-helices on top of a β-sheet. Polymorphic amino acids of the peptide binding cleft are colored in blue. The figure was generated in cooperation with the designer Renate Nikolaus.

ing. Whereas in the beginning only one serological group was know, and subgroups were named by numbering (e.g. HL-A1, HL-A2), it soon became apparent that the observations in the serological testing originated from different HLA loci [75]. Therefore, the naming system was adjusted to account for the genetic origin of the antibodies and newly identified sub-serotypes [75]. With the introduction of sequencing-based methods, the HLA was soon classified based on nucleotide sequence in addition to its serotype (Section 6.1).

Sequence-based analysis allowed for a more detailed distinction of differences in the HLA leading to the discovery of an ever-increasing amount of different nucleotide sequences, also called HLA alleles. The latest HLA nomenclature defines an HLA allele based on a prefix (HLA), the gene name, and four fields separated by colons (Figure 14) [12, 75]. The first field defines the allele group and is in general equivalent with the previously defined serologically defined HLA subtypes. The second field defines the complete protein. The third field includes synonymous mutations within the coding region and the fourth field includes variations within the non-coding region [12].

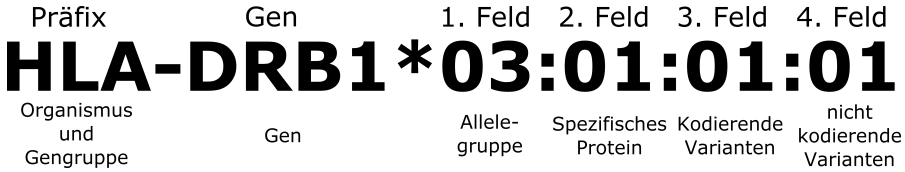


Figure 14: Nomenclature of the HLA.

Previous versions of the nomenclature did not use any separator between the different fields but assumed that differences across each specification could be represented by two digits. In 2010 a new name convention including the separator was introduced, as more than 100 alleles were discovered within distinct HLA groups [12, 75, 138]. As a relic of the previous nomenclature, some papers still refer to 2- and 4-digit alleles, instead of 1- and 2-field alleles, respectively.

Many scientific studies are mainly interested in analyzing differences of the HLA within its peptide binding groove only. This region is encoded by exon 2 of the according gene for both the α and the β chains of the class II proteins (Figure 11b). For class I, the peptide binding groove is encoded by the exons 2 and 3 (Figure 11c). Focusing only on differences in the peptide binding groove results in G-grouping of HLA alleles. All alleles that have identical exon 2 or exon 2/3 sequences are assigned to the same G group. A specific G group is then marked with a 'G' suffix, e.g., A*02:01G. If only coding variants within a G group are considered, the groups can be further collapsed into P groups (indicated by a 'P' suffix) [75]. Unfortunately, many studies lack the clarity to specify the HLA alleles were determined.

In general, other suffixes are used in the HLA nomenclature as well, indicating the expression of the gene. The most important suffix is the 'N' marking a non-expressed allele. [138]

As of today, the IMGT/HLA database includes 34 422 HLA alleles and their sequences (release 3.49.0, 12.07.2022) with currently an increase of several hundred new alleles per quarter year [51, 152].

6

HLA TYPING

The HLA locus is characterized by a high number of polymorphisms, paired with a large amount of homologous sequences between many HLA genes and pseudogenes. The genetic differentiation of specific nucleotide variations is therefore more challenging for HLA alleles than in the rest of the genome. Typically, the best accuracy for measuring an HLA allele is achieved by using sequencing-based methods (Section 6.1). However, sequencing is time and cost intensive in the framework of association analysis, where thousands of individuals are analyzed parallelly. Genotyping arrays (Section 2.2.2), used in GWAS (Section 2.3), do not cover the entirety of the HLA variation. As a consequence, HLA alleles cannot be called based on genotype information derived from these arrays directly. Instead, specifically designed imputation references are used to infer HLA allele information from array-based genotyping (Section 6.2).

6.1 HLA TYPING BASED ON SEROTYPES AND SEQUENCING

The HLA was traditionally typed using the complement dependent cytotoxicity (CDC) assay. This approach relies on sera of known anti-HLA-antibodies that bind to specific HLA allele groups. After adding the complement, cells with the according HLA profile lyse. This dissolution can be visualized under the microscope [19, 196]. In the 80s, DNA-based techniques became popular for HLA type inference [19]. These methods include the application of sequence-specific oligonucleotides (SSO), sequence-specific primers (SSP), and real-time PCR (RT-PCR) [19]. The kits, i.e., specific packages of chemicals and other components, which were used for these techniques, are generally not able to discriminate between closely related HLA alleles as the used primers or oligo probes do not cover all polymorphisms [19].

In addition to this, sequencing technologies of the different generations (see Section 2.2.1) are used. Sanger sequencing typically focuses only on the exons related to the peptide binding groove (exons 2 and 3 for class I and exon 2 for class II, Figure 11) [19]. Sequencing limited to these exons can lead to a high level of ambiguity of the HLA calls [4]. NGS enables sequencing of whole HLA genes in a cost and time effective manner. Many ambiguities can be eliminated with NGS due to sequencing of additional exons [4] and clonal amplification, where the DNA fragments are copied as a whole, and spatially separated [19]. On the other hand, clonal amplification might introduce PCR bias. Clonal amplification and the assembly of short sequences is not necessary any more using third generation long-read sequencing [19].

Based on the type of sequences generated by second and third generation sequencing, different tools can be used to accurately genotype the HLA alleles. Some of them use the IMGT/HLA database [51, 152] as a reference for the HLA allele sequences. Others focus on de novo assembly to build a consensus sequence from read fragments [19]. Exemplary tools used for HLA typing based on NGS data are:

- HLAssign: This open-source tool assigns alleles based on references from the IMGT/HLA database [201]. The tool was, for instance, used in PAPER A (Section 6.3, [43]).
- HLA twin: The commercial HLA twin software from Omixon allows both reference-based and de novo genotyping [139].
- NGSengine: The commercial NGSengine analysis software from GenDx is based on the IMGT/HLA reference. GenDx is compatible with any amplification strategy and different common sequencing platforms [48, 56].

Sequencing data is still not affordable nor available for large cohort studies. Instead, genotyping arrays are used to determine the HLA genotypes (Section 6.2).

6.2 TOOLS FOR THE IMPUTATION OF HLA ALLELES

HLA allele imputation is in its principles similar to SNP imputation (see Section 2.2.3). As in SNP imputation, missing HLA genotypes are inferred from specially designed reference panels. Tools that have specifically been developed for HLA allele imputation include HLA*IMP [46], SNP2HLA [80], HIBAG (HLA Imputation using attribute BAGging) [211, 212] and CookHLA [35].

The imputation of all these tools is based on a reference dataset. HLA*IMP was published with a reference consisting of over 2 500 individuals of European ancestry [46]. SNP2HLA was published with two different reference datasets: First, the HapMap-CEPH panel, including 90 samples of diverse genetic background and second, the T1DGC (Type 1 Diabetes Genetics Consortium) panel, including 5 225 European samples [80]. HIBAG was published with a variety of panels. The single models incorporate data from European, Asian, Hispanic, and/or African samples. Especially for African individuals, the reference size, with around 100 typed individuals, is comparably small to cover the whole diversity of HLA haplotypes. CookHLA validated their algorithm based on different publicly available datasets, e.g., data from the 1KGP or the T1DGC [35].

Jia et al. [80] already pointed out that large population-specific reference panels are needed to achieve high quality HLA imputation. When this research started, all available references for HLA imputation were either of limited size or focusing on Caucasian ancestries only. To analyze the HLA in UC in other populations, as performed in Degenhardt et al. [43], a more diverse reference for HLA imputation was needed. In PAPER A (Section 6.3) we constructed a new multi-ethnic imputation reference based on HIBAG [43]. An updated version of SNP2HLA including a new multi-ancestry reference was recently published on the Michigan Imputation Server [109].

They integrated the HLA imputation in a broader workflow for an association analysis [109].

6.3 PAPER A: MULTI-ETHNIC REFERENCE PANEL

Frauke Degenhardt et al. "Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles." In: *Human molecular genetics* 28.12 (2019), pp. 2078–2092. ISSN: 1460-2083. DOI: 10.1093/hmg/ddy443, p. 2

Aim

Improve the HLA imputation quality in non-Caucasian samples to allow the analysis of disease associated HLA profiles in trans-ethnic cohorts.

Method

Genotypes were measured with Illumina's Immunochip genotyping array for more than 1 300 samples from Germany, Malta, China, India, Iran, Japan, Korea, and samples of African American ancestry. The HLA types of those individuals were established using a high-resolution next generation sequencing approach (HLAssign) for all classical HLA class I and II alleles including HLA-*DRB3/4/5*. Based on this data a new reference panel was trained for HIBAG. We benchmarked the panel by cross-validation and with independent data from the 1KGP.

Results

The benchmark showed high accuracy across the eight populations and the independent data from 1KGP. We identified specific alleles that are challenging to impute in the single populations.

Conclusion

Our new multi-ethnic imputation panel allows accurate 2-field HLA imputation for all classical HLA alleles and HLA-*DRB3/4/5* based on diverse ancestry populations.

Authors contributions

F.D. performed statistical and computational analysis. M.We. performed computational analysis with contributions from M.H.. M.Wi. performed HLA typing with contributions from F.D.. F.D. wrote the manuscript. M.We. revised the manuscript with contributions from E.R., E.E., and D.E.. S.A., B.A., T.B.K., S.R.B., J.H.C., L.W.D., P.E., Y.F., E.S.J., M.K., W.L., R.M, V.M., S.C.N., J.D.R., S.S., A.S., A.T., J.S., S.H.W., and K.Y. were involved in study subject recruitment, contributed genotype data and or/phenotype

data. F.D. and A.F. conceived, designed, and managed the study. All authors reviewed, edited, and approved the final manuscript.

OXFORD

Human Molecular Genetics, 2019, Vol. 28, No. 12 2078–2092

doi: 10.1093/hmg/ddy443
Advance Access Publication Date: 26 December 2018
Bioinformatics Article

BIOINFORMATICS ARTICLE

Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles

Frauke Degenhardt[1,†], Mareike Wendorff[1,†], Michael Wittig[1], Eva Ellinghaus[2], Lisa W. Datta[3], John Schembri[4], Siew C. Ng[5], Elisa Rosati[1], Matthias Hübenthal[1], David Ellinghaus[1], Eun Suk Jung[1,6], Wolfgang Lieb[7], Shifteh Abedian[8,9], Reza Malekzadeh[9], Jae Hee Cheon[6], Pierre Ellul[4], Ajit Sood[10], Vandana Midha[10,11], B.K. Thelma[12], Sunny H. Wong[5], Stefan Schreiber[1,13], Keiko Yamazaki[14,15], Michiaki Kubo[16], Gabrielle Boucher[17], John D. Rioux[17,18], Tobias L. Lenz[19], Steven R. Brant[3,20,21] and Andre Franke[1,*]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [2]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, 0424 Oslo, Norway, [3]Department of Medicine, Meyerhoff Inflammatory Bowel Disease Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, [4]Division of Gastroenterology, Mater Dei Hospital, Msida MSD 2090, Malta, [5]Department of Medicine and Therapeutics, Institute of Digestive Disease, LKS Institute of Health Science, State Key Laboratory of Digestive Disease, The Chinese University of Hong Kong, Hong Kong, China, [6]Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of Medicine, Seoul, 03722, Republic of Korea, [7]Biobank PopGen and Institute of Epidemiology, University Hospital Schleswig-Holstein, Campus Kiel, 24105 Kiel, Germany, [8]Department of Epidemiology, University Medical Center Groningen, 9700 RB Groningen, The Netherlands, [9]Digestive Disease Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, 14117-13135, Tehran, Iran, [10]Department of Gastroenterology, Dayanand Medical College and Hospital, 141001 Ludhiana, Punjab, India, [11]Department of Medicine, Dayanand Medical College and Hospital, 141001 Ludhiana, Punjab, India, [12]Department of Genetics, University of Delhi South Campus, 110021 New Delhi, India, [13]Department of Medicine, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [14]Laboratory for Genotyping Development, Center for Integrative Medical Sciences, RIKEN Yokohama Institute, Yokohama, 230-0045, Japan,

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Received: August 9, 2018. **Revised:** December 17, 2018. **Accepted:** December 18, 2018

Human Molecular Genetics, 2019, Vol. 28, No. 12 | **2079**

[15]Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, 173-8610, Japan, [16]RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan, [17]Montreal Heart Institute, Research Center, Montréal, Québec H1T 1C8, Canada, [18]Université de Montréal Department of Medicine, Montréal, Québec H3C 3J7, Canada, [19]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany, [20]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA, [21]Department of Medicine, Rutgers Robert Wood Johnson Medical School and Department of Genetics, Rutgers University, New Brunswick and Piscataway, NJ 08901, USA

*To whom correspondence should be addressed at: Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Rosalind-Franklin-Street 12, D-24105 Kiel, Germany. Tel: +49 (0) 431/500-15109; Fax: +49 (0) 431/500-15168; E-mail: a.franke@mucosa.de

Abstract

Genotype imputation of the human leukocyte antigen (HLA) region is a cost-effective means to infer classical HLA alleles from inexpensive and dense SNP array data. In the research setting, imputation helps avoid costs for wet lab-based HLA typing and thus renders association analyses of the HLA in large cohorts feasible. Yet, most HLA imputation reference panels target Caucasian ethnicities and multi-ethnic panels are scarce. We compiled a high-quality multi-ethnic reference panel based on genotypes measured with Illumina's Immunochip genotyping array and HLA types established using a high-resolution next generation sequencing approach. Our reference panel includes more than 1,300 samples from Germany, Malta, China, India, Iran, Japan and Korea and samples of African American ancestry for all classical HLA class I and II alleles including *HLA-DRB3/4/5*. Applying extensive cross-validation, we benchmarked the imputation using the HLA imputation tool HIBAG, our multi-ethnic reference and an independent, previously published data set compiled of subpopulations of the 1000 Genomes project. We achieved average imputation accuracies higher than 0.924 for the commonly studied *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* genes across all ethnicities. We investigated allele-specific imputation challenges in regard to geographic origin of the samples using sensitivity and specificity measurements as well as allele frequencies and identified HLA alleles that are challenging to impute for each of the populations separately. In conclusion, our new multi-ethnic reference data set allows for high resolution HLA imputation of genotypes at all classical HLA class I and II genes including the *HLA-DRB3/4/5* loci based on diverse ancestry populations.

Introduction

The major histocompatibility complex, in humans also named human leukocyte antigen (HLA) complex, is a highly variable gene cassette with major functions in the immune system. The HLA region spans ~5 Mb on chromosome 6p21 with genomic positions ranging from 29 Mb to 34 Mb. Genes in this region code for proteins that are involved in many complex functions of the adaptive and innate immune system like the presentation of peptides to the host immune system and also code for proteins that aid peptide presentation or antigen recognition. Results from over 10 years of genome-wide association studies (GWAS) support the HLA as one of the most important disease susceptibility loci for almost every immune-mediated and autoimmune disease. In many cases, the strongest association signals are found within the highly polymorphic classical HLA genes in the class I and II regions, a finding made long before the GWAS era for many of these diseases (1). Therefore, pinpointing the exact genetic variants in the HLA region, which are associated with these diseases, is of utmost importance to disentangle the underlying genetic pathophysiology (2). This is complicated by the highly polymorphic nature of the region, resulting in the need for large disease cohorts to increase statistical power in the detection of genetic association. The costs per sample for Sanger- and next generation sequencing (NGS)-based HLA typing is still at least double that of a genome-wide single nucleotide polymorphism (SNP) array analysis with the new chip platforms. Therefore, imputation methods and reference panels have been developed to provide geneticists with a tool to infer HLA alleles at the classical loci *in silico* using inexpensive and dense SNP array data. These have led to significant advances in fine-mapping of disease relevant genetic variants for many inflammatory and autoimmune diseases (3–5). Published and established HLA imputation tools are amongst others SNP2HLA, HLA Imputation using attribute BAGging (HIBAG) and HLA*IMP (6–8). Imputation of the HLA requires reference panels with high coverage of alleles and genotypes in the region of interest as well as a broad spectrum of samples in order to capture as many different alleles as possible. Additionally, the ancestral background of the reference panel used to impute a data set of interest must be as close as possible to the study population as shown for instance by Jia *et al.* (7). Most HLA imputation reference panels target Caucasian ethnicities and although there has been progress in the development of ancestrally diverse HLA reference panels, studies in which multi-ethnic analyses are performed are still scarce and limited in size (e.g. for chronic inflammatory diseases, (9)). Several imputation references have been published in the past using various genotyping chips and at different resolutions. All reference panels have significantly advanced HLA imputation and analysis conducted with the produced data. However, to date, no full context four-digit multi-ethnic HLA imputation reference panel exists for fine mapping of the HLA region across the totality of the mentioned loci.

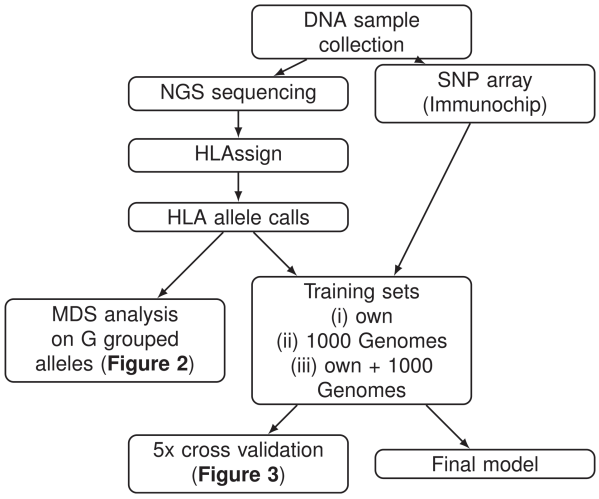


Figure 1. Flowchart of steps taken in preparation and benchmarking of our multi-ethnic reference panel. HLA allele calls were made based on NGS reads. Genotype information was measured using the Illumina Immunochip. These data were combined to train a HIBAG imputation model. Benchmarking was performed using a 5× cross-validation and the independent, previously published, 1000 Genomes data set (24).

With this study, we aimed to create a comprehensive high-quality multi-ethnic HLA reference data set, including *HLA-DPA1*, *-DPB1* and *-DRB3/4/5*, using populations of African American, East Asian (Japan, South Korea and China), European (Germany, Malta) and Middle Eastern (India and Iran) descent.

We generated HLA allele calls from next generation sequencing (NGS) reads for ulcerative colitis (UC) and control individuals of each population, using HLAssign (10) and genotype information using the Illumina Immunochip SNP array [Illumnina, San Diego, CA, USA] (Fig. 1). Using multidimensional scaling (MDS) analysis, we analyzed population structure based on HLA allele frequencies. The combination of called HLA alleles and SNP array genotypes served as training data sets for our new multi-ethnic reference using the HLA imputation tool HIBAG (6). We benchmarked the imputation, applying extensive cross-validation on our multi-ethnic reference panel (Supplementary Material, Fig. S1). The performance of our final model was additionally assessed using the previously published HLA calls of the 1000 Genomes project (11). We also conducted a literature search into the genetic architecture of *HLA-DRB3/4/5* in relation to *HLA-DRB1*, as the presence of the *HLA-DRB3/4/5* are highly dependent on which *HLA-DRB1* allele is carried by an individual. These loci are of particular interest, since they represent a functional variation that has not been considered in many of the previously published reference data sets and hence have been largely excluded in association studies.

Results

MDS-based clustering of reference samples on HLA allele frequencies

Using MDS analysis on relative frequencies of single HLA G grouped alleles across each cohort, we observed distinct clusters for individuals with East Asian, African and European backgrounds (Fig. 2), except for *HLA-DRB3/4/5* and *HLA-DQB1*. The different subpopulations of our multi-ethnic study population cluster well with respective ethnicities of the 1000 Genomes population. For the 1000 Genomes population, exons 2 and 3 (class I) or exon 2 (class II) were typed only for loci *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* but not for *HLA-DPA1*, *-DPB1* and *-DRB3/4/5*. However, to the best of our knowledge no custom G groups were defined (11). Samples did not show population-specific clustering for *HLA-DQB1*, because frequencies of the HLA alleles in European individuals were similar to those in the Yoruban, African American and European individuals of the 1000 Genomes population. We did not detect consistent clusters for the *HLA-DRB3/4/5* genes, possibly because there was not enough variability to allow good clustering results. In our multi-ethnic data set we only observe four, three and six different four-digit alleles for the *HLA-DRB3/4/5* genes, respectively. In addition, these genes also included a high percentage of null alleles (*HLA-DRB3*, 48.45–81.28%; *HLA-DRB4*, 65.78–84.52%; *HLA-DRB5*, 71.28–85.66%; Table 1) that dominate the frequency spectrum and thus the MDS analysis. With 'null allele' we here refer to the absence of a locus in a given individual. These null alleles are named DRB3*00:00, DRB4*00:00 and DRB5*00:00 throughout this paper. In summary, the MDS analysis reveals significant population heterogeneity for the classical HLA genes and thus, imputation tools should be able to account for this heterogeneity by using population-matched and diverse reference panels.

Imputation benchmark

We performed HLA imputation of the HLA class I loci *HLA-A*, *-B*, *-C* and class II loci *HLA-DQA1*, *-DQB1*, *-DPA1*, *-DPB1*, *-DRB1* and *-DRB3/4/5* using HIBAG and three different constellations: (i) our multi-ethnic reference panel in full four-digit context (Fig. 3 and next paragraph), (ii) our multi-ethnic reference panel combined with the 1000 Genomes data set on G group level (Supplementary Material, Fig. S2 and Supplementary Material, Table S1) and (iii) our multi-ethnic reference panel on G group level as a comparison (Supplementary Material, Fig. S3 and Supplementary Material, Table S2). We also used the 1000 Genomes panel to test the performance of our data (Table 2) with special focus on the imputation for the non-European population panels, as one of the main innovations of this work.

Using a cross-validation approach (Supplementary Material, Fig. S1), we divided the data of each specific population into five random subsamples irrespective of case–control status. For each of the subsets, using the remaining 80% of the population, as well as the HLA allele and genotype information of all other populations, we trained a HIBAG model. The HLA alleles were predicted for the 20% of data from the analyzed population that were not used for training. We calculated accuracies for each of the five subsamples of our population of interest and imputation accuracies for unrelated individuals of the 1000 Genomes population. The results of the cross-validation are depicted in Figure 3 and Table 3. Overall accuracies were high with average accuracies ranging from 0.924 in the Chinese to 0.967 in the Maltese populations (Table 3; Supplementary Material, Table S3). More specifically, high overall accuracies were achieved for the *HLA-C*, *HLA-DP* and *HLA-DQ* loci whereas the *HLA-A*, *-B* and *-DRB1* loci were more challenging to impute across all ethnicities with accuracies as low as 0.862 for *HLA-DRB1* in the Iranian panel. This is also reflected in the posterior probability curves depicted in Figure 3b. Posterior probabilities in HIBAG are used as an additional measure to control prediction accuracies and are generated as an average over all classifiers. Low overall posterior probabilities for a locus indicate that the majority of the alleles were challenging to impute. Note, that correct calls, e.g. for rare alleles, also tend to have smaller posterior probabilities,

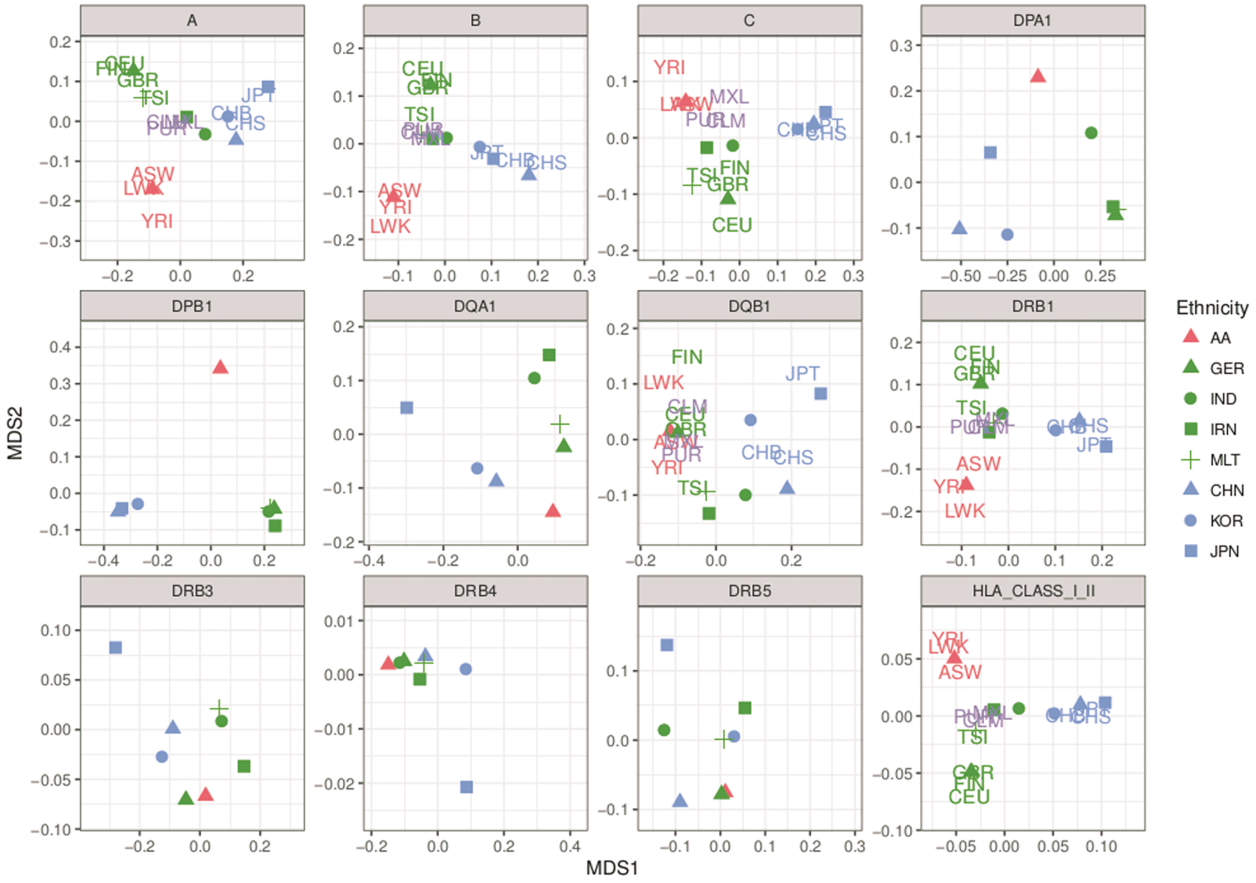


Figure 2. MDS analysis of HLA typed allele data: the MDS analysis was performed using a Euclidean distance measure. Alleles with a frequency <1% were excluded to produce a clustering that is not biased by similarity in low frequency variants. Colors show the origin of the cohort. Red: African American (AA) and African background; Green: European and Middle Eastern background: German (GER), Indian (IND), Iranian (IRN), Maltese (MLT); Blue: Asian background: Hong-Kong Chinese (CHN), South Korean (KOR) and Japanese (JPN); Purple: Non-reference admixed American individuals. Capital acronyms in the panels depict the 1000 Genomes populations as described in Auton *et al.*, (24). The 1000 Genomes populations include Americans of African Ancestry in the Southwest USA (ASW), Africans from Kenya (LWK), Nigeria (YRI), Columbian (CLM), Mexican (MXL) and Puerto Rican (PUR), Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), Japanese in Tokyo (JPT), Finnish (FIN), British (GBR), Tuscan (TSI) and samples with Western European Ancestry collected in the CEPH diversity panel (CEU). For *HLA-DPA1*, *-DPB1*, *-DQA1* and the *-DRB3/4/5* loci no data was available in those panels. For the MDS analysis across all loci (HLA CLASS I_II) we included *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*. Samples of our own cohorts cluster well with the corresponding 1000 Genomes population.

while incorrect calls can have a high posterior probability when haplotypes of two alleles are similar across many classifiers. Therefore, we decided to additionally use other measures such as sensitivity and specificity, and allele specific accuracy to evaluate allele specific results in the following analyses. With 29–55 alleles per population, and 75% (Malta) to 82% (Japan) of the alleles having frequencies of <1% (Supplementary Material, Tables S4 and S5), *HLA-B* presented a particular challenge for imputation. Similarly challenging were *HLA-A* and *-DRB1*, which are discussed further below. The remaining loci were not as variable or had a smaller and more even frequency spectrum (Supplementary Material, Table S5), such that posterior probabilities were higher. *HLA-DPA1* and *-DPB1* had the most "on target" SNPs (30 and 51 SNPs, respectively) (Supplementary Material, Table S6), reflecting the fact, that these loci are least variable and therefore better suited to be captured on a SNP genotyping array. Overall, between 682 (*HLA-DPB1*) and 1,794 (*HLA-A*) SNPs were located within the different gene loci including flanking regions of 500 kb upstream and downstream of each gene. A median of 41.5 (*HLA-DRB5*) to 81 (*HLA-A*) SNPs were used by the single classifiers of HIBAG.

In the following, we show the results of the imputation with our own reference data set divided by ethnic background and also compare our data to previously reported HLA imputation accuracies on published data sets from Dilthey *et al.* (8), Jia *et al.* (7), Okada *et al.* (12), Kim *et al.* (13) and Zheng *et al.* (6) (Table 4). It is of importance to note, that high accuracies for a reference panel using a specific benchmarking panel are best achieved when the benchmarking panel follows the same allele nomenclature and grouping as the panel used for imputation. We could not determine to which extent this was considered in each of the above studies, but we estimate that the effect should not be detrimental if differences only occur between slightly different custom allele groupings (i.e. we assume that the allele that a grouping is based on is also the most frequent allele) and not between different levels of grouping (i.e. full context versus G groups). A summary of these data sets is described in Table 4. The following results are specific to the imputation of HLA alleles into the respective populations using our multi-ethnic four-digit full context reference panel. If not stated otherwise, mean accuracies were compared for four-digit allele imputations of *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*. These are the loci that are

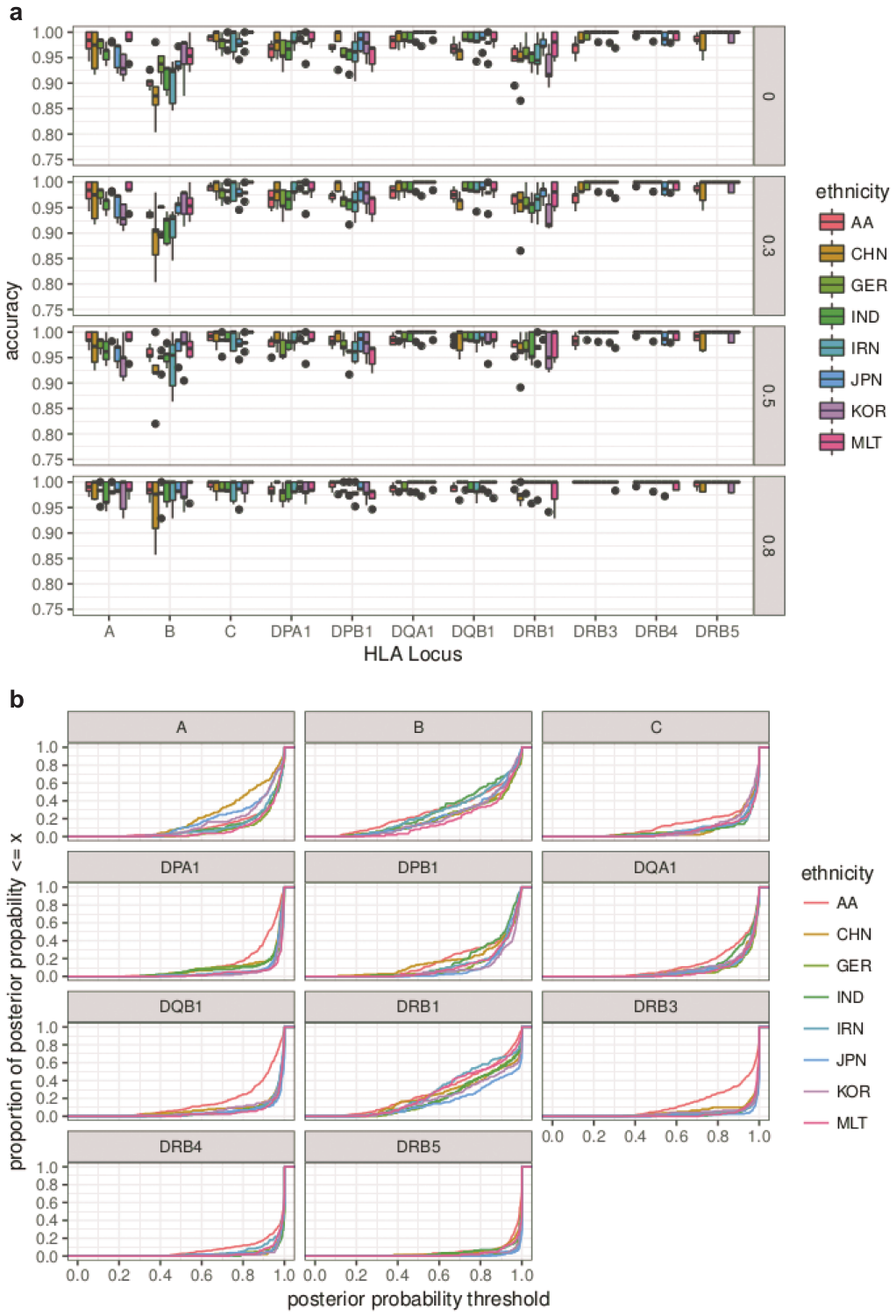


Figure 3. Imputation accuracies employing the multi-ethnic reference panel: accuracies and post-imputation probabilities of HLA imputation with HIBAG using a 5-fold cross-validation scheme and the multi-ethnic data set with full four-digit allele information. 20% of the data with a specific ethnic background were used as the validation set after training a model that used 80% of the remaining data and all data from other ethnic backgrounds. We included 1,360 African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples in total. **(a)** Accuracies are depicted according to post-imputation probabilities with cut-off thresholds at 0 (no confidence filtering), 0.3, 0.5, 0.8 (only high confidence genotypes). Loci are shown according to alphabetical order. Imputation accuracies are especially high for *HLA-C*, *-DPA1*, *-DPB1*, *-DQB1* and the *-DRB3/4/5*. *HLA-DRB1* accuracies are especially lowered by misclassifications of DRB1*04:03, DRB1*04:04 and DRB1*11:04. **(b)** Posterior probabilities are depicted as proportion of the number of samples with a posterior probability smaller than a threshold (*x*-axis).

Human Molecular Genetics, 2019, Vol. 28, No. 12 | **2083**

Table 1. Frequencies of HLA-DRB3/4/5 in our multi-ethnic reference panel: frequencies of *HLA-DRB3/4/5* in the typed HLA data for African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) populations at full four-digit context. Null alleles have the highest frequencies. For *HLA-DRB4* mainly one other allele, DRB4*01:03, exists. DRB5*01:01 is the second most abundant of the *HLA-DRB5* alleles in all but the Japanese and Iranian panels, where DRB5*01:02 is seen more often.

	AA	CHN	GER	IND	IRN	JPN	KOR	MLT
DRB3*00:00	51.61	64.60	59.88	56.74	48.45	81.28	64.34	55.00
DRB3*01:01	11.13	2.55	14.51	5.32	8.53	4.55	11.07	4.69
DRB3*02:02	27.74	19.34	22.53	32.98	37.98	8.82	16.39	33.75
DRB3*02:24	0.00	0.00	0.62	0.00	0.39	0.00	0.00	0.31
DRB3*03:01	9.52	13.50	2.47	4.96	4.65	5.35	8.20	6.25
DRB4*00:00	84.52	75.91	80.25	80.85	75.97	65.78	68.44	75.63
DRB4*01:01	6.77	0.00	2.47	0.35	1.55	0.00	0.00	3.75
DRB4*01:02	0.00	0.00	0.00	0.00	0.39	2.14	0.41	0.00
DRB4*01:03	8.71	24.09	17.28	18.79	22.09	32.09	31.15	20.31
DRB4*03:01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31
DRB5*00:00	81.94	72.63	80.56	71.28	85.66	71.66	82.38	81.56
DRB5*01:01	15.97	21.53	16.67	15.96	5.43	6.42	11.07	10.00
DRB5*01:02	0.32	1.82	0.62	12.77	6.98	20.59	4.51	3.75
DRB5*01:03	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00
DRB5*01:08	0.32	2.19	0.00	0.00	0.00	0.27	0.41	0.00
DRB5*02:02	0.97	0.36	2.16	0.00	1.94	1.07	1.64	4.69
DRB5*02:03	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00
DRB5*02:13	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Imputation accuracies for 1000 Genomes populations: population groups are depicted in **bold** and the subpopulations in *italic* type. African (AFR) samples are divided into Americans of African Ancestry in the Southwest USA (ASW), Africans from Kenya (LWK) and Nigeria (YRI). Admixed American (AMR) samples are split into samples with Columbian (CLM), Mexican (MXL) and Puerto Rican (PUR) ancestry. East Asians (EAS) were collected as Han Chinese in Beijing (CHB), Southern Han Chinese (CHS) and Japanese in Tokyo (JPT). Samples with European Ancestry (EUR) are Finnish (FIN), British (GBR), Tuscan (TSI) and samples with Western European Ancestry collected in the CEPH diversity panel (CEU). Accuracies of *HLA-DRB1** are *HLA-DRB1* measured without DRB1*04:03, DRB1*04:04 and DRB1*11:04, which improved accuracies for all ethnicities. HLA-A* are accuracies measured without A*02:03, which improved accuracies for the Chinese samples. Overall accuracies were highest for EUR samples and lowest for the non-AMR, for which no samples with similar backgrounds are included in our novel imputation reference.

	#samples	A	B	C	DQB1	DRB1	mean	A*	DRB1*
AFR	**162**	0.920	0.833	0.932	0.951	0.886	0.904	0.920	**0.906**
ASW	*41*	0.939	0.805	0.915	0.939	0.902	0.900	0.939	**0.923**
LWK	*75*	0.880	0.853	0.960	0.980	0.893	0.913	0.880	**0.899**
YRI	*46*	0.967	0.826	0.902	0.913	0.859	0.893	0.967	**0.902**
AMR	**193**	0.909	0.756	0.972	0.984	0.710	0.866	0.909	**0.766**
CLM	*67*	0.925	0.709	0.970	0.985	0.687	0.855	0.925	**0.711**
MXL	*56*	0.857	0.688	0.973	0.991	0.598	0.821	0.857	**0.674**
PUR	*70*	0.936	0.857	0.971	0.979	0.821	0.913	0.936	**0.888**
EAS	**260**	0.929	0.931	0.975	0.992	0.940	0.953	**0.941**	**0.951**
CHB	*82*	0.939	0.921	0.988	0.994	0.939	0.956	**0.948**	**0.967**
CHS	*92*	0.935	0.924	0.967	0.995	0.935	0.951	**0.963**	**0.944**
JPT	*86*	0.913	0.948	0.971	0.988	0.948	0.953	0.913	0.943
EUR	**322**	0.983	0.944	0.994	0.989	0.890	0.960	0.983	**0.968**
CEU	*52*	0.981	0.922	0.971	1.000	0.865	0.948	0.981	**0.987**
FIN	*95*	0.984	0.974	1.000	0.989	0.926	0.975	0.984	**0.959**
GBR	*86*	0.977	0.959	1.000	0.983	0.884	0.960	0.977	**0.993**
TSI	*89*	0.989	0.910	0.994	0.989	0.871	0.951	0.989	**0.944**

present for all imputation references (Table 4). Within the cross-validation framework, accuracies for a gene were calculated as an average across the different cross-validation runs as it has been done previously (12,13) and enables better comparison of these values between studies. We also report median, minimum and maximum values in Supplementary Material, Table S3. We report accuracies across all imputed alleles in Table 3, Supplementary Material, Tables S1 and S2. A few alleles were especially challenging to impute, both within our as well as in previously published reference panels. These alleles usually have comparably lower sensitivity or specificity scores and similar haplotype structures within the same 2-digit allele groups (Supplementary Material, Tables S7 and S8, Supplementary Material, Tables S5–S8 of Zheng *et al.*, (6)). This is especially important in the context of association analyses where the greatest impact from these issues is seen with higher frequency variants (AF >1%) and thus needs to be considered carefully. Note that this also depends on the ethnicity of the samples

2084 | *Human Molecular Genetics, 2019, Vol. 28, No. 12*

Table 3. Imputation accuracies of the imputation with the multi-ethnic reference panel: 20% of the data with a specific ethnic background were used as validation set after training a model with 80% of the remaining data and all data from other ethnic backgrounds. We included 1,360 African American (AA), Hong-Kong Chinese (CHN), German (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples in total in the imputation reference. Shown are mean accuracies of the HLA imputation with HIBAG using a 5-fold cross-validation scheme and the multi-ethnic data set with full four-digit allele information. The given mean considers only the loci highlighted in **bold**, as these are loci also analyzed in all previous publications. Accuracies of *HLA-DRB1** are *HLA-DRB1* measured without DRB1*04:03, DRB1*04:04 and DRB1*11:04, which improves accuracies for all ethnicities. *HLA-A** are accuracies measured without A*02:03, which improves accuracies for the Chinese samples. Overall, *HLA-B* is the most challenging to impute. Mean accuracies are higher than 0.925 across all cross-validation runs. Best results are achieved for the GER, JPN and MLT populations.

	AA	CHN	GER	IND	IRN	JPN	KOR	MLT
#samples	**312**	**140**	**162**	**143**	**132**	**189**	**122**	**160**
A	0.969	0.900	0.976	0.955	0.973	0.936	0.939	0.984
B	0.877	0.868	0.917	0.875	0.885	0.938	0.934	0.947
C	0.953	0.986	0.975	0.979	0.974	0.973	0.968	0.988
DPA1	0.969	0.979	0.960	0.968	0.985	0.995	0.975	0.988
DPB1	0.925	0.949	0.960	0.944	0.954	0.979	0.963	0.956
DQA1	0.942	0.975	0.975	0.965	0.962	0.968	0.959	0.978
DQB1	0.962	0.964	0.988	0.990	0.981	0.984	0.975	0.984
DRB1	0.925	0.903	0.948	0.924	0.862	0.960	0.918	0.931
DRB3	0.971	1.000	1.000	1.000	1.000	1.000	0.996	0.994
DRB4	0.977	1.000	0.991	0.996	0.996	0.990	1.000	0.988
DRB5	0.987	0.982	1.000	1.000	1.000	1.000	0.992	1.000
mean	0.937	0.924	0.961	0.944	0.935	0.958	0.947	0.967
*A**	0.969	0.954	0.976	0.954	0.973	0.935	0.937	0.984
*DRB1**	0.930	0.904	0.954	0.952	0.956	0.968	0.926	0.971

evaluated. We describe A*02:01/A*02:03, DRB1*11:01/DRB1*11:04 and DRB1*04:03/DRB1*04:04 below for illustration purposes.

African American panel

The imputation of HLA alleles into our own African American data set achieved an average imputation accuracy on full context four-digit level of 0.951 across all analyzed loci and of 0.937 on average for loci *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* only (Table 3). Employing our multi-ethnic reference data set on G group level (ii), we were able to impute alleles of the genes *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* of the 1000 Genomes African ancestry data with a mean accuracy of 0.904 and highest accuracies for the Luhya Kenyan samples alone (0.880–0.980; mean of 0.913; Table 2). In comparison, Zheng *et al.* (6) imputed HLA alleles of random subsets of their African American HLARES data combined with the Yoruba Nigerians (YRI) HapMap samples with a reported mean accuracy of 0.818 using their tool HIBAG (Table 4b). Jia *et al.* (7) imputed the HLA alleles of YRI HapMap samples using their Caucasian Type 1 Diabetes Genome Consortium (T1DGC) reference panel with accuracies between 0.203 (*HLA-DRB1*) and 0.984 (*HLA-C*) across all loci and an overall mean accuracy of 0.750 (Table 4a).

East Asian panel

Employing our multi-ethnic reference data set (i) to impute HLA alleles into our Chinese samples, we achieved accuracies of 0.868 (*HLA-B*) to 1.000 (*HLA-DRB3/4*) and of 0.924 on average for *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*. We imputed HLA alleles into our Japanese samples with accuracies of 0.936 (*HLA-A*) to 1.000 (*HLA-DRB3/5*) and 0.958 on average for *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*. For our Korean samples imputation accuracies of 0.918 (*HLA-DRB1*) to 1.000 (*HLA-DRB4*) were reached, with an average accuracy of 0.947 (Table 3). Additionally, we imputed the HLA alleles of the East Asian 1000 Genomes data on G group level (ii) with mean accuracies higher than 0.953 (Table 2).

In comparison, Okada *et al.* (12), Jia *et al.* (7), Kim *et al.* (13) and Zheng *et al.* (6) reported mean accuracies between 0.77 to 0.922 for *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* (Table 4) for East Asian populations using their respective HLA imputation panels. *HLA-DPA1* or *HLA-DRB3/4/5* is not considered in any of the publications for East Asian ethnicities. For single loci the reported imputation accuracies vary between 0.656 (*HLA-B* with T1DGC reference for Han Chinese in Beijing (CHB) and Japanese samples (JPT); (7)) and 0.984 (HLA-C with a Korean reference panel and the same test population; (13)).

In the cross-validation benchmark the accuracy of locus HLA-A in the Chinese population (Fig. 3a) was decreased due to a misclassification of A*02:03 to A*02:01 in 32% of 37 samples in which this allele occurred. This misclassification is due to the high similarity between these alleles (Supplementary Material, Supplementary Text). When excluding A*02:03 from accuracy calculations for *HLA-A*, accuracies improved for the Chinese subpopulation from 0.900 to 0.954 (Table 3).

Iranian and Indian panels

Overall imputation accuracies for our Indian and Iranian panels over all loci were 0.944 and 0.935, respectively. The accuracies were high for all loci except *HLA-B* (0.875 and 0.885, respectively) and *-DRB1* (0.924 and 0.862, respectively) (Table 3).

The accuracy of the Iranian samples in the cross-validation benchmark (Fig. 3a) at *HLA-DRB1* was low due to a misclassification of DRB1*11:04 to DRB1*11:01 in 39% of the 36 Iranian samples in which this allele occurs (Supplementary Material, Supplementary Text). When excluding the DRB1*11:04 as well as the DRB1*04:04 and DRB1*04:03 alleles (see below) from accuracy calculations for *HLA-DRB1*, the accuracies improved from 0.862

Table 4. Previously reported imputation accuracies: accuracies measured for HLA reference panels, which are mainly based on Caucasian and Asian data, with origin of the publications and cohorts used for training and validation as well as a comparison to accuracies achieved with our own multi-ethnic reference panel (i) in the cross-validation experiment on our own data (see also Table 3) and on the 1000 Genomes cohorts (see also Table 2). Accuracies of the cross-validation (own) framework and of the imputation into the 1000 Genomes population are shown. Mean accuracies are calculated across *HLA-A*, *-B*, *-C*, *-DPB1* and *-DRB1* (loci highlighted in **bold**). Mean accuracies of the listed reference panels are lower compared to our own reference panel in the majority of the cases, especially in the non-European population. **(a)** Accuracies published with SNP2HLA. The international T1DGC reference panel (7) published along with SNP2HLA was used to gain the accuracies on the 1948 British Birth Cohort and the HapMap-CEPH Cohort, two European ancestry panels. The T1DGC panel was further used for imputing the Yoruban Nigerian (YRI), the East Asian Han Chinese from Beijing (CHB) and the Japanese from Tokyo (JPT) samples of the 1000 Genomes data sets. For the East Asian 1000 Genomes panels accuracies reached by later-published ethnic-specific references (12,13) are also listed. **(b)** Accuracies published with HIBAG using the HLARES data from GlaxoSmithKline (GSK) clinical trials of specific ethnic background combined with 1000 Genomes data sets (6). **(c)** Accuracies published with HLA*IMP:02 using different combinations of the Golden Set (GS = 1948 Birth Cohort/ HapMap CEU and CEPH CEU+) and the HLARES data as references (8).

(a) SNP2HLA

Source	**Jia *et al.*** (7)				**Okada *et al.*** (12)	**Kim *et al.*** (13)	
imputation reference	**T1DGC**				**Japanese**	**Korean**	**Korean**
# training samples	5,225				918	330	413
test population	**1948 British Birth Cohort**	**CEPH**	**YRI**	**CHB & JPT**	**JPT**	**random subset**	**CHB & JPT**
# test samples	918	90	not specified	not specified	44	83	61
A	0.981	0.991	0.699	0.981	0.908	0.908	0.91
B	0.968	0.968	0.905	0.656	0.943	0.859	0.893
C	0.969	0.991	0.984	0.688	0.989	0.928	0.984
DPA1	/	/	/	/	/	/	/
DPB1	/	/	/	/	/	0.95	/
DQA1	/	0.985	0.649	0.963	/	/	/
DQB1	0.983	0.991	0.961	0.964	0.894	0.937	0.893
DRB1	0.933	0.969	0.203	0.923	0.843	0.868	0.893
DRB3	/	/	/	/	/	/	/
DRB4	/	/	/	/	/	/	/
DRB5	/	/	/	/	/	/	/
mean	0.967	0.983	0.729	0.864	0.915	0.908	0.915
mean A-C, DQB1, DRB1	0.967	0.982	0.75	0.842	0.915	0.9	0.915
mean A-C, DQB1, DRB1	**own**						
	GER 0.961	GER 0.961	AA 0.937	CHN 0.924	CHN 0.924	CHN 0.924	CHN 0.924
	MLT 0.967	MLT 0.967		JPN 0.958	JPN 0.958	JPN 0.958	JPN 0.958
				KOR 0.947	KOR 0.947	KOR 0.947	KOR 0.947
	1000 Genomes						
	EUR 0.96	EUR 0.96	ASW 0.9	CHB 0.956	CHB 0.956	CHB 0.956	CHB 0.956
			LWK 0.913	CHS 0.951	CHS 0.951	CHS 0.951	CHS 0.951
			YRI 0.893	JPT 0.953	JPT 0.953	JPT 0.953	JPT 0.953

(b) HIBAG

Source	**Zheng *et al.*** (6)			
imputation reference	**HLARES data of Asian ancestry & CHB & JPT**	**HLARES data of Hispanic ancestry**	**African American HLARES data & 60 African YRI**	**HLARES data of European ancestry**
# training samples	720 + 90 (minus test)	439 (minus test)	173 + 60 (minus test)	2668 (minus test)
test population	**random subset**	**random subset**	**random subset**	**random subset**
# test samples	subset	subset	subset	subset
A	0.921	0.934	0.924	0.982
B	0.875	0.75	0.768	0.966
C	0.966	0.962	0.885	0.988
DPA1	/	/	/	/

(Continued).

Table 4. *Continued*

(b) HIBAG

DPB1	0.898	0.931	0.8	0.947
DQA1	0.868	0.938	0.794	0.964
DQB1	0.96	0.957	0.742	0.992
DRB1	0.887	0.82	0.771	0.921
DRB3	/	/	/	/
DRB4	/	/	/	/
DRB5	/	/	/	/
mean	0.911	0.899	0.812	0.966
mean A-C, DQB1, DRB1	0.922	0.885	0.818	0.97

mean A-C, DQB1, DRB1	**own**					
	CHN	0.924	AA	0.937	GER	0.961
	JPN	0.958			MLT	0.967
	KOR	0.947				
	1000 Genomes					
	CHB	0.956	PUR	0.913	ASW	0.9
	CHS	0.951			LWK	0.913
	JPT	0.953			YRI	0.893
					EUR	0.96

(c) HLA*IMP:02

Source	**Dilthey *et al.* (8)**					
imputation reference	**GS**	**HLARES _EU**	**GS & HLARES_ALL**			
# training samples	1,585	1,758	2,055			
test population	**HLARES_EU**	**random subset**	**African Americans of random subset**	**Asians of random subset**	**Europeans of random subset**	**Hispanic of random subset**
# test samples	1,060	872	1,008 (all populations)			
A	0.96	0.97	0.73	0.79	0.96	0.82
B	0.9	0.95	0.73	0.68	0.95	0.63
C	0.96	0.96	0.97	0.82	0.97	0.92
DPA1	/	/	/	/	/	/
DPB1	/	0.90 (2-digit)	/	/	/	/
DQA1	0.87	0.97	1	0.73	0.96	0.93
DQB1	0.98	0.98	0.87	0.83	0.97	0.97
DRB1	0.88	0.91	0.71	0.72	0.9	0.8
DRB3	/	0.94 (2 digit)	/	/	/	/
DRB4	/	0.98 (2 digit)	/	/	/	/
DRB5	/	0.99 (2 digit)	/	/	/	/
mean	0.93	0.95	0.84	0.76	0.95	0.85
mean A-C, DQB1, DRB1	0.94	0.95	0.8	0.77	0.95	0.83

mean A-C, DQB1, DRB1	**own**									
	GER	0.961	GER	0.961	AA	0.937	CHN	0.924	GER	0.961
	MLT	0.967	MLT	0.967			JPN	0.958	MLT	0.967
							KOR	0.947		
	1000 Genomes									
	EUR	0.96	EUR	0.96	ASW	0.9	CHB	0.956	EUR	0.96
					LWK	0.913	CHS	0.951		
					YRI	0.893	JPT	0.953		

											PUR	0.913

to 0.956 (Table 3). Mean sensitivity values for DRB1*11:04 for the cross-validation runs were 0.307 for the Iranian population and 0.208 for the Indian population (Supplementary Material, Table S8). The frequency of this allele was 2.82% and 13.85%, respectively (Supplementary Material, Table S5). The improvement of the overall accuracy by excluding these alleles in the Indian samples (0.924 to 0.952) was not as big as in the Iranian samples because of the lower allele frequency (AF). Previously reported sensitivity values for the DRB1*11 alleles (Supplementary Material, Tables S5–S8 of Zheng *et al.* (6)) range

Human Molecular Genetics, 2019, Vol. 28, No. 12 | 2087

from 0.627 (DRB1*11:04) to 0.993 (DRB1*11:01) in the European population. In this previous study, misclassifications occurred for DRB1*11:04, too, which was called as DRB1*11:01 in 93% of cases when a misclassification occurred in European samples (6). This is in line with our own results.

Imputation for non-reference populations

The Latin American admixed populations of the 1000 Genomes data set (containing Amerindian and European, for Puerto Rico also West African ancestral admixture, here grouped into Mexican, Columbian and Puerto Rican populations) were imputed with mean accuracies ranging from 0.821 for the Mexican, 0.855 for the Columbian to 0.913 for the Puerto Rican population (Table 2). In particular, *HLA-B* and *-DRB1* showed low imputation accuracies (0.688 to 0.857 and 0.598 to 0.821, respectively) while all remaining loci had accuracies higher than 0.857 (Table 2). Overall, the Puerto Rican data set showed highest accuracies and only 40 out of 134 total measured alleles had sensitivity values of lower than 1.000 (Supplementary Material, Table S9). Out of these 40 alleles, 22 have an AF <0.1% in the Puerto Rican panel. Accuracies for loci imputed within the Puerto Rican data set ranged from 0.821 (*HLA-DRB1*) to 0.979 (*HLA-DQB1*) (Table 2).

HLA-DRB3/4/5 haplotypes

Many imputation tools allow the imputation of *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1* but only a few studies have reported on the imputation of the *HLA-DRB3, -DRB4 and -DRB5 (HLA-DRB3/4/5)* loci, such as Dilthey *et al.* (8), who analyzed *HLA-DRB3/4/5* imputation in Caucasian data sets (Table 4c). These genes can be present or absent in an individual depending on the *HLA-DRB1* genotype. For the evaluation of the imputation of these genes and to elucidate which *HLA-DRB3/4/5* loci are known to be located on the same haplotype as a specific *HLA-DRB1*, we conducted an extensive literature review and present the results below. We mainly focus on the information reported by Holdsworth *et al.* (14), Robbins *et al.* (15) and Bontrop *et al.* (16). According to literature, alleles of the *HLA-DRB3/4/5* loci occur within a specific *HLA-DRB1* context, being present in some haplotypes and absent in others. The results of this review are summarized in Figure 4. Haplotypes with *HLA-DRB1* always carry the pseudogene *HLA-DRB9*, which is located downstream of *HLA-DRB1* and that consists of two exons (17). DRB1*01, DRB1*08 and DRB1*10 are not found with any *HLA-DRB3/4/5* allele. Haplotypes with DRB1*03, *11, *12, *13 and *14 are found with *HLA-DRB2* and *-DRB3*. DRB1*04, *07, *09 are found with *HLA-DRB4* as well as *-DRB7* and *-DRB8*. Finally, DRB1*15 and *16 are reported to be located on the same haplotype as *HLA-DRB5*. Exceptions to his rule have been described for DRB1*15 and *16, where especially in African Americans *HLA-DRB5/6* can be missing. DRB1*07 has been reported to occur with a non-expressed form of DRB4*04:01 (15) and DRB1*08 has also been previously identified together with DRB3*03:01 (15).

We investigated our herein-described multi-ethnic data on *HLA-DRB1* and *-DRB3/4/5* for congruence with these previous findings. In short, we determined the *HLA-DRB1* alleles for every sample and checked whether we could also find the expected *HLA-DRB3/4/5* alleles or the absence of these in the same sample. All but four samples followed the haplotype structures depicted in Figure 4. After re-analysis of the remaining four samples we concluded that these samples must have been contaminated, since three or more alleles could plausibly be called for all analyzed loci, with one allele having a smaller number of reads that aligned to it. In further six samples we found one of the exceptions described in the literature. One Maltese sample did not have *HLA-DRB4* while DRB1*07:01 was present and five African American samples did not have *HLA-DRB5* while DRB1*15:03 or DRB1*16:02 was present.

Frequencies of *HLA-DRB3/4/5* are shown in Table 1. Overall, *HLA-DRB3* is the most variable of those genes according to its frequency spectrum, with DRB3*02:02 being the most common non-null allele with an AF ranging from 8.82% in our Japanese panel to 37.98% in our Iranian panel. For *HLA-DRB4*, DRB4*01:03 is the most common non-null allele with frequencies ranging from 8.71% in the African American to 32.09% in the Japanese panel. DRB5*01:01 is the most common non-null allele in all but the Iranian and Japanese panels with frequencies of 5.43% in the Iranian to 21.53% in the Chinese panel, while DRB5*01:02 has a frequency of 20.59% in the Japanese panel and a frequency of 6.98% in the Iranian panel. Our data suggest that DRB1*15:01 is located on the same haplotype as DRB5*01:01, while DRB1*15:02 (which is very common in Japanese samples) is located on the same haplotype as DRB5*01:02 (Supplementary Material, Table S10). Accuracies of the *HLA-DRB3/4/5* imputations are high (>0.971; Table 3 and Fig. 3a). Sensitivity measures for the *HLA-DRB3/4/5* are generally high; however, for low frequency variants (e.g. DRB3*02:24 in the Iranian, Maltese and German panels at frequencies of <0.62%) values as low as 0 were measured. DRB4*01:02 in the Japanese panel, DRB3*01:01 and DRB4*01:01 in the African American panel are common alleles (AF > 1%) classified with mean sensitivity values of lower than 0.800 (0.375, 0.739, 0.690, respectively). We also observed, using the tool Disentangler (18), that the phasing of *HLA-DRB3/4/5* alleles might present a challenge, with many of the null alleles occurring on haplotypes with *HLA-DRB1*, when the respective *HLA-DRB3/4/5* allele is present (Supplementary Material, Fig. S4; *HLA-DRB3/4/5* are excluded here). The analysis of this particular topic, however, is beyond the scope of this paper.

Discussion

We compiled three different imputation panels as pre-trained HIBAG models that can be used for HLA imputation in different ethnicities: (i) a multi-ethnic reference with four-digit full context HLA alleles and (ii) a multi-ethnic reference with four-digit HLA alleles as G groups. Both panels include *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DPA1*, *-DPB1*, *-DRB1* and *-DRB3/4/5* and (iii) a multi-ethnic reference panel combined with the 1000 Genomes data (including data from *HLA-A*, *-B*, *-C*, *-DQB1*, *-DRB1*, *-DPA1*, *-DPB1* at a four-digit G group resolution). Our reference panels have high accuracy values across different ethnicities and subsets of the data and also achieve high accuracies in non-reference ethnicities (Tables 2 and 3). The accuracies in non-reference ethnicities are high, but lower than for our reference data sets, as even though our reference is highly diverse the worldwide diversity of the HLA is still not sufficiently captured. Average accuracies of our multi-ethnic reference are larger than 0.924. Tabulated results describing the accuracy measures of panels (ii) and (iii) are presented in Supplementary Material, Tables S1 and S2. Using our reference data, few alleles remain challenging to impute. This affects alleles of the *HLA-DRB1* locus, like the DRB1*11 and DRB1*04 group, which has already been described as problematic in previous benchmarks of other imputation reference panels (6–8) as well as alleles of the highly diverse *HLA-A* and *-C* genes. We therefore recommend using a two-

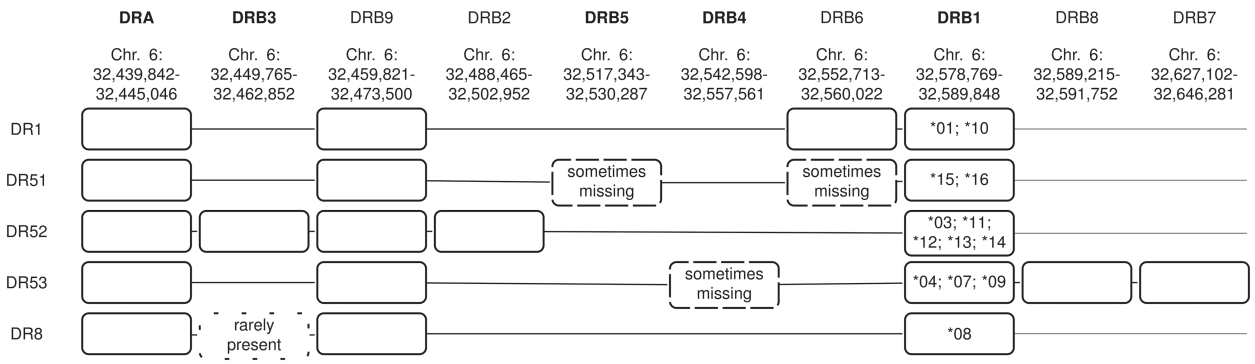


Figure 4. Known architecture of *HLA-DRB3/4/5*: HLA haplotypes that usually contain a specific *HLA-DRB1* allele (*HLA-DRB1* column) are shown. Two-digit alleles are denoted. All loci are depicted in order of their genomic location. *HLA-DRA*, *HLA-DRB1* and *HLA-DRB9* coincide with all haplotypes. The remaining loci are present or absent depending on the haplotype. The most prevalent haplotypes with the known exceptions are shown in the rows below. Exceptions are sometimes seen for *DRB1*08*, DRB1*07, DRB1*15 and DRB1*16. DRB1*08 can occur with *HLA-DRB3*, DRB1*07 can occur without an expressed form of *HLA-DRB4* and DRB1*15 and DRB1*16 can occur without *HLA-DRB5/6*. Loci that usually occur together are joined by a line. The name of the corresponding serotype is shown on the left and haplotypes are ordered by serotype name. Information for this figure was retrieved from Bontrop *et al.*, Holdsworth *et al.* and Robbins *et al.* (14–16).

digit resolution for these alleles and to consider the imputation difficulties in the interpretation of association results for these alleles. We further suggest that the interpretation of specificity and sensitivity measures should be done separately by ethnic background, since measures can vary between ancestries, i.e. haplotypes for an allele that are highly predictive in one ethnicity may not be highly predictive in another ethnicity. We also verified that SNPs missing in the data set for which HLA alleles are imputed—and that exist in the reference—can negatively affect the imputation accuracy. This was the case for DRB1*04:03 and DRB1*04:04, where exclusion of 4.4% of the SNPs used by the HIBAG had a major impact on the imputation accuracy for these alleles (Supplementary Material, Supplementary Text). We therefore suggest, as a general rule, to cautiously investigate the coverage of SNPs used by any imputation reference panel prior to imputation with the respective panel into a data set. Posterior probabilities are often used to improve the quality of the data set. Indeed, we also observe that the accuracies improve when using a posterior probability threshold. However, for some alleles similar haplotype structures can cause incorrect calls despite high posterior probabilities. Especially for rare alleles, correct calls are possible at a very low posterior probability. We therefore suggest using the sensitivity and specificity tables we provide in Supplementary Material, Table S8 to perform data filtering as well as checking the posterior probability.

In summary, imputing HLA alleles into multi-ethnic genome-wide association data sets with our reference panels provides accurate results and can aid HLA fine mapping studies especially in non-Caucasian populations in the future. It allows for HLA imputation using the most recent HLA allele nomenclature at a full context four-digit resolution and a high diversity of different populations.

Nevertheless, larger sample sizes and even more diverse reference panels are needed to adequately cover the existing global HLA polymorphism and frequency spectrum particularly for the ethnicities not included in our panel and also to impute especially rare HLA alleles with high accuracy. DRB1*01:03, for instance, is an allele that has a higher frequency in North American Caucasians (0.9–1.9%) than European Caucasians (~0.6%) (19). As over a million of samples will have been genotyped and whole-genome sequenced in the near future, it is just a matter of warranting global coverage, thus to include representatives from every ethnicity for these efforts. Still, most genetic research focuses on Caucasian ancestry cohorts and neglects large segments of human populations. Decreasing costs of high-resolution NGS-based HLA typing approaches—including phased data sets from long-read technologies—will further fuel the development of more comprehensive and even more accurate imputation reference panels.

Materials and Methods

Resolution of imputation reference panels

Several imputation references have been published in the past using various genotyping chips, allowing for the imputation of different HLA genes at different resolutions, i.e. full context four-digit (two-field), G group and P group resolution (as defined by the IMGT/HLA database) or custom groups (mostly before 2010). Full context four-digit levels provide information on the gene name, their allele group and the protein sequence of the HLA molecule (i.e. A*01:02—Gene: A; allele group: 01; protein: 02). Alleles that are within the same G group have identical nucleotide sequences for exons 2 and 3 (HLA class I) or exon 2 only (HLA class II) and may differ in sequence in the other exons. Alleles that are within the same P group encode for identical amino acid sequences in exons 2 and 3 or exon 2 only. P and G group annotations were introduced in 2010 and a major update in allele naming was conducted (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/Nomenclature_2009.txt), amongst others the separator ':' was introduced and alleles were renamed especially alleles of the *HLA-A*, *-B*, *-C* and *-DPB1* genes. Notably, HLA allele calling conducted before this time, with alleles typed only at exons 2 and 3 or exon 2, may not follow the known G group and P group conventions published by the IMGT/HLA, i.e. HLA alleles might be grouped in custom groups and some of the alleles will carry outdated allele names. This issue should be considered when merging reference panels, such that all included alleles should map to the same allele groups and also in benchmarking studies using external data. G grouping published by the IMGT/HLA database is based on the highest resolution that is recorded for an allele (i.e. eight digits or lower). Note that the post-calling G grouping based on four-digit alleles is problematic for some alleles listed in Supplementary Material, Table S11.

Human Molecular Genetics, 2019, Vol. 28, No. 12 | 2089

Cohorts & data preparation

Multi-ethnic data set. DNA of 96 healthy individuals and 96 UC patients were collected from different studies of Chinese, German, Indian, Iranian, Japanese, Korean and Maltese populations that have been published and described elsewhere (20,21). In short, Chinese samples were collected in and around Hong Kong (Chinese University of Hong Kong), Korean samples in South Korea (Yonsei University College of Medicine and Asan Medical Centre, Seoul), Japanese samples in Tokyo (Institute of Medical Science, University of Tokyo, RIKEN Yokohama Institute and Japan Biobank), Iranian samples were collected in Tehran (Tehran University of Medical Science), Indian samples in North India (Dayanand Medical College and Hospital, Ludhiana), all self-reported North Indian which was consistent with their genetically determined background, German samples in North Germany and Maltese samples in Malta (Department of Gastroenterology, Mater Dei Hospital, Msida, Malta). In addition to the data from the published UC studies, DNA samples were obtained from 192 healthy controls and 192 UC patients, all self-reported as African American, which was consistent with their genetically determined background as each had an admixture of West African and European ancestry (22). These subjects were recruited in the United States of America and Canada by the Johns Hopkins Multicenter African American IBD Study as well as other Genetics Research Centers of the NIDDK IBD Genetics Consortium. We also received 192 (96 healthy, 96 UC) pre-analyzed Japanese samples directly from RIKEN Yokohama Institute.

High density SNP-array data interrogating a wide proportion of the extended HLA region were produced for these samples using the Illumina, Immunochip (all but Malta) with 196,524 markers addressing immune relevant genes or the Illumina Infimum ImmunoArray 24 (Malta only) with 253,702 markers and subjected to strict quality control criteria as described in the Supplementary Material, Supplementary Methods. DNA was isolated and processed as described previously (10) in preparation for sequencing. Sequencing was performed on an Illumina HiSeq2500 (http://systems.illumina.com) with 100 bp or 125 bp paired-end runs on a panel of both case and control data in a pool of 96 libraries per lane. A total of 192 Japanese samples were provided by the RIKEN Yokohama Institute and sequenced using 125 bp paired-end runs on the HiSeq2500 with pools of 94 libraries per lane. Four-digit HLA alleles for all classical HLA I and HLA II genes *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DPA1*, *-DPB1*, *-DRB1* as well as *-DRB3/4/5* were manually curated and called using HLAssign (10). In short, only reads mapping exactly to a reference based on HLA sequences published with the IMGT/HLA database version 3.27.0 (23) were used for calling, taking into consideration evenness of read mapping, read equality and specific read mapping as described by Wittig *et al.* (10). We also cautiously looked at cross-mapping events (reads mapping to multiple HLA loci) and SNP patterns to identify e.g. alleles originating from concatenation of true alleles. In total 1,360 samples were used in this study, having been sequenced and called successfully based on their DNA quality and internal HLAssign measures, i.e. sufficiently large read coverage and also having passed our stringent criteria for the quality control of the Illumina Immunochip array data (Supplementary Material, Supplementary Methods). The *HLA-DRB3/4/5* calls were additionally evaluated for plausibility with respect to the called *HLA-DRB1* genotype. *HLA-DRB3/4/5* alleles, according to reported studies (14–16), occur on certain haplotypes in tight linkage with specific *HLA-DRB1* variants and can either be present or not present at all (i.e. null allele, described as DRB3*00:00, DRB4*00:00 and DRB5*00:00 in the following) or as one functional *HLA-DRB3/4/5* allele in combination with two of the *HLA-DRB3/4/5* null alleles. For a detailed overview we compiled Figure 4. A total of 312 African American (158 Controls, 154 UC cases), 162 German (78 Controls, 84 Cases), 140 Chinese (68 Controls, 72 Cases), 143 Indian (78 Controls, 65 Cases), 132 Iranian (63 Controls, 69 Cases), 189 Japanese (96 Controls and 93 Cases), 122 South Korean (81 Controls and 41 Cases) and 160 Maltese (75 Controls and 85 Cases) samples were available for construction of HLA imputation models with HIBAG.

1000 Genomes data set. Using the Phase 3 [version from 20130502] 1000 Genomes reference data set (24) and Vcftools (version 0.1.12b), we extracted 174,538 phased SNPs that are present in both the Phase 3 data set and on the Illumina Immunochip used for the main part of our trans-ethnic data. We then performed quality control as described in the Supplementary Material, Supplementary Methods leaving out batch and population stratification analyses. HLA data were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_geotypes/. Publicly available data from the 1000 Genomes data set do not include *HLA-DPA1*, *-DPB1*, *-DQA1* and *DRB3/4/5* allele calls. In total 162 samples of African Ancestry, 193 samples of South American Ancestry, 260 samples of East Asian ancestry and 322 samples of European ancestry were available for construction of HLA imputation models with HIBAG. The HapMap data used in other studies (Table 4) are a part of the 1000 Genomes data set.

Calling of HLA-DRB3/4/5 alleles. Data were analyzed visually using HLAssign (10). HLAssign does not calculate phases of the HLA alleles and thus does not make hemizygous calls (i.e. recognize null alleles) such that *HLA-DRB3/4/5* genotypes were edited with respect to the *HLA-DRB1* allele post calling. For consistency with the *HLA-DRB3/4/5* with the literature (Fig. 3), we introduced null alleles DRB3*00:00, DRB4*00:00 or DRB5*00:00 when the *HLA-DRB1* locus was called as DRB1*01, DRB1*08 or DRB1*10, respectively. DRB3*00:00 was assigned if no *HLA-DRB3* was present in the corresponding *HLA-DRB1* haplotype. Equally, DRB4*00:00 and DRB5*00:00 were assigned if haplotypes corresponding to the absence of *HLA-DRB4* or *-DRB5* were called. Samples with inconclusive *HLA-DRB3/4/5* detected during HLAssign analysis were re-analyzed using HLAReporter (25). HLAReporter performs *de novo* assembly on the NGS reads within the investigated HLA locus using the alignment tool TASR (26) and compares these to either G groups or full context alleles known in the IMGT/HLA database with the parameters (-m 50, -o 5, -r 0.7, -u 0, -i 1, -t 0, -e 33, -c 0) for on target reads. Contigs for samples with equal G group predictions were aligned against each other to generate longer overlapping regions using contigs with a coverage higher than 15 and then realigned to the known IMGT/HLA reference alleles.

MDS analysis. Relative allele frequencies were calculated for each allele across the entire multi-ethnic and 1000 Genomes HLA data within the *HLA-A*, *-B*, *-C*, *-DQ* and *-DR* loci. For the MDS analysis alleles with an allele frequency of less than 1% in any subpopulation are excluded to avoid a clustering biased by similarity in low frequency variants. The MDS analysis was performed using R and the stats-Package (cmdscale) with a Euclidean distance measure. For the MDS analysis across all loci we used HLA loci *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*.

HLA imputation benchmark

Training of the reference panel. We performed HLA imputation using the published imputation tool HIBAG (6). This is a machine learning tool implemented in R that employs ensemble classifiers built on bootstrap samples that has been shown to perform with high accuracy in HLA imputation across multi-ethnic data sets (6). In short, a training set with both HLA alleles and SNPs typed in the HLA region on chromosome 6, between 29 and 34 Mb, is used to build several classifiers based on bootstrap samples and a subset of SNPs, similarly to random forest as proposed by Breiman *et al.* (27) that minimize the out-of-bag errors. Once a model is trained, it can be used as reference to predict HLA alleles from unknown samples using their respective SNP genotype information, utilizing the posterior probability as measure of confidence. For the benchmark, we performed a 5× cross-validation using HIBAG (6) and HLA and SNP genotype data from the following two sources: our multi-ethnic cohort described above and the publicly available 1000 Genomes data set (24). The 1000 Genomes data set was typed for *HLA-A*, *-B*, *-C*, *-DPB1* and *-DRB1*, while the multi-ethnic data set contained all classical HLA class I and class II loci and additionally *HLA-DRB3/4/5*. For the 1000 Genomes data set, typed HLA data were available for samples of the following ethnicities: African, South American Ancestry, East Asian and European. We grouped our data into three different data sets: (i) our multi-ethnic reference containing eight different cohorts described above, (ii) the same reference as in (i) with HLA alleles transformed into their respective G groups (G groups combine alleles with identical exon 2 and 3 (HLA Class I) or exon 2 (HLA Class II) nucleotide sequence) using hla_nom_g.txt downloaded from hlaalleles.org date: 2017-07-10, IPD-IMGT/HLA version 3.29.0) and (iii) our multi-ethnic panel and the 1000 Genomes data set combined. In total we used 1,360 samples and 7,428 SNPs within the HLA region for the multi-ethnic reference, as well as 937 samples from the 1000 Genomes data and 7,551 SNPs within the HLA region from the 1000 Genomes data set, with 2,297 samples and 7,126 SNPs for the combined data set as well as their respective HLA calls. For the 1000 Genomes panel, we checked for nomenclature issues, making sure that all of the HLA alleles used in the 1000 Genomes panel mapped to the nomenclature for HLA alleles used since April 2010 (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/Nomenclature_2009.txt). For alleles with unambiguous G groups (Supplementary Material, Table S11), we assigned the lower number allele for reference panels (ii) and (iii). Genotype data were prepared as described in Supplementary Material, Supplementary Methods. Samples with typed HLA information were extracted from each quality-controlled, genotyped data set. The different cohorts were merged and those SNPs with a consistent minor allele frequency (MAF) of <1% (across all cohorts typed for the particular SNP) were excluded. The data were randomly split into five equal parts per cohort with respect to case–control status, thus ensuring that a training set would include both case and control data. Using HIBAG (version.1.8.3), we trained our models using the reference containing the merged subpopulations, excluding 20% of the population of interest and 100 classifiers, as suggested by the authors of the tool (Supplementary Material, Fig. S1).

Validation of the reference panel. The quality-controlled genotype data for each cohort were imputed using Beagle version 4.1 (28) with the cohort itself serving as an internal reference to fill in any remaining missing data. Pretrained HIBAG HLA models (see above) were provided with the respective 20% of the remaining data of each analyzed population (Supplementary Material, Fig. S1), using the genomic position as the identifier. HLA calls were calculated and stored with their respective posterior probabilities. Accuracies and the number of samples to be excluded were calculated for different posterior probability thresholds and compared between the different populations.

Calculation of accuracies. Imputation accuracies were calculated on best-guess alleles compared with the known alleles of the typed data. Accuracies for best-guess alleles were calculated by counting the number of alleles imputed correctly per locus and dividing by the number of samples multiplied by two. Per locus and per allele accuracies were evaluated. We also calculated single allele specificity and sensitivity values if possible. For this we evaluated each allele separately, counting the number of times an allele was predicted correctly as present (True Positive; TP) or absent (True Negative; TN) and the number of times an allele was incorrectly predicted as present (False Positive; FP) or absent (False Negative; FN). We then used the standard definitions to calculate sensitivity and specificity from these values.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

For the calculation of the accuracy, specificity and sensitivity values within the cross-validation, the mean values across the different runs were calculated for each locus or allele, as well as median, minimum and maximum values for comparison. To establish which alleles might have low sensitivity and specificity values in a general setting for (i), we calculated these measures using a model based on the entire population (i).

Imputation reference panels for comparison

A Caucasian reference panel based on genotypes retrieved from the T1DGC (29), as well as a Pan Asian data set (30) using three different Asian populations, were published along with SNP2HLA (7) and are available on request from the SNP2HLA authors. Here, loci *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DPB1* and *-DRB1* were typed (Table 4a). Two additional Asian reference panels based on SNP2HLA were published at a four-digit resolution. First, a Korean reference panel was published in 2014 (13) for the imputation of amino acids and HLA alleles into East Asian populations for *HLA-A*, *-B*, *-C*, *-DQB1*, *-DPB1* and *-DRB1* and second, a Japanese reference data set was published in 2015 by Okada *et al.* (12) with an evaluation of loci *HLA-A*, *-B*, *-C*, *-DQB1* and *-DRB1*. For these two last reference panels, we assume that they were typed at a full context four-digit resolution. This has not been explicitly mentioned in the respective publications (12,13), but we find that the typed alleles best fit to the full four-digit context based on which alleles are present. Pre-trained multi-ethnic HLA models with European, Asian, Hispanic and African ancestry (based on a total of 3,738 samples) are provided with the HLA imputation tool HIBAG (6). The samples used for these models were obtained from HLARES (samples GlaxoSmithKline clinical trials) (6) and the HapMap project. Loci *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DPB1* and *-DRB1* were evaluated at four-digit resolution (Table 4b). The remaining considered reference panels based on HLA*IMP:02 (8) are based on HLARES data and a study specific "Golden Set" (GS) (Table 4c).

Availability of resources

The herein-described reference data sets are available on request from the authors (email contact: f.degenhardt@ikmb.uni-kiel.de) as pretrained HIBAG models and are mapped to IMGT/HLA database version 3.27.0 with G group definitions derived from IMGT/HLA database version 3.29.0. Note that allele names at four-digit levels did not change between these two releases. The training of these models was performed as described above without exclusion of any samples. A script that will estimate the haplotype similarity between alleles based on the genotype positions available in a data set is also available upon request.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We acknowledge the efforts of individuals from Johns Hopkins University and the other MAAIS recruitment centers who contributed to recruitment of the African American samples used in the study as reference (22). Additional African American samples used were also provided by Judy H. Cho (Icahn School of Medicine at Mount Sinai, New York, National Institutes of Health (NIH) grant DK062422), Richard H. Duerr (University of Pittsburgh, NIH grant DK062420) and Mark Silverberg (University of Toronto, NIH grant DK062423). All subjects gave informed consent for their samples to be used for genetics research studies related to inflammatory bowel disease. We also want to acknowledge Garima Juyal (Department of Genetics, University of Delhi South Campus, New Delhi, India), Yuta Fuyuno (Laboratory for Genotyping Development, Center for Integrative Medical Sciences, RIKEN Yokohama Institute, Yokohama, Japan and Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan), Atsushi Takahashi (Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, RIKEN Yokohama Institute, Yokohama, Japan) and Behrooz Alizadeh (University of Groningen, University Medical Centre Groningen, Department Epidemiology, Groningen, the Netherland) for their involvement in this project. We thank Marie Dowds (Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany) and Philip Stuart (Department of Dermatology, University of Michigan Medical School, Ann Arbor, Michigan, USA) for helpful discussions.

Conflict of Interest statement. The authors have no conflict of interest to declare.

Funding

German Research Foundation (DFG) (Research Training Group 1743, 'Genes, Environment and Inflammation' to M.W.); DFG Excellence Cluster No. 306 'Inflammation at Interfaces'; European Union Seventh Framework Programme (FP7-PEOPLE-2013-COFUND) (No. 609020; Scientia Fellows to E.E.); Funding for the Multicenter African American IBD Study (MAAIS) samples was provided by the USA National Institutes of Health (DK062431 to S.R.B.); University Medical Center Groningen, Groningen, The Netherlands (to S.A.); Institute for Digestive System Disease, Tehran University of Medical Sciences, Tehran, Iran (to S.A.); BioBank Japan Project and, in part, by a Grant-in-Aid for Scientific Research (B) (26293180) funded by the Ministry of Education, Culture, Sports, Science, and Technology, Japan; Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI18C0094); Funding for the Indian samples was provided by the Centre of Excellence in Genome Sciences and Predictive Medicine (BT/01/COE/07/UDSC/2008 from the Department of Biotechnology, Government of India); BMBF e:Med research and funding concept (SysInflame grant 01ZX1306A; GB-XMAP grant 01ZX1709); J.D.R. holds a Canada Research Chair and this work was supported by National Institutes of Health grant DK62432. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Rose, N.R. (1978) HLA and disease. *Arch. Intern. Med.*, **138**, 527–528.
2. Dendrou, C.A., Petersen, J., Rossjohn, J. and Fugger, L. (2018) HLA variation and disease. *Nat. Rev. Immunol.*, **18**, 325–339.
3. Okada, Y., Yamazaki, K., Umeno, J., Takahashi, A., Kumasaka, N., Ashikawa, K., Aoi, T., Takazoe, M., Matsui, T., Hirano, A. *et al.* (2011) HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology*, **141**(864–871), e861–865.
4. Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E.S., Annese, V., Hauser, S.L. *et al.* (2015) High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.*, **47**, 172–179.
5. Patsopoulos, N.A., Barcellos, L.F., Hintzen, R.Q., Schaefer, C., van Duijn, C.M., Noble, J.A., Raj, T., IMSGC, ANZgene, Gourraud, P.A. *et al.* (2013) Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS Genet.*, **9**, e1003926.
6. Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R. and Weir, B.S. (2014) HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, **14**, 192–200.
7. Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.M., Concannon, P.J., Rich, S.S., Raychaudhuri, S. and de Bakker, P.I. (2013) Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*, **8**, e64683.
8. Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M.R. and McVean, G. (2013) Multi-population classical HLA type imputation. *PLoS Comput. Biol.*, **9**, e1002877.
9. Franke, A. (2017) Inflammatory bowel disease: a global disease that needs a broader ensemble of populations. *Gastroenterology*, **152**, 14–16.
10. Wittig, M., Anmarkrud, J.A., Kassens, J.C., Koch, S., Forster, M., Ellinghaus, E., Hov, J.R., Sauer, S., Schimmler, M., Ziemann, M. *et al.* (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.*, **43**, e70.
11. Gourraud, P.A., Khankhanian, P., Cereb, N., Yang, S.Y., Feolo, M., Maiers, M., Rioux, J.D., Hauser, S. and Oksenberg, J. (2014) HLA diversity in the 1000 genomes dataset. *PLoS One*, **9**, e97282.

12. Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., Takahashi, A. and Kubo, M. (2015) Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nat. Genet.*, **47**, 798–802.
13. Kim, K., Bang, S.Y., Lee, H.S. and Bae, S.C. (2014) Construction and application of a Korean reference panel for imputing classical alleles and amino acids of human leukocyte antigen genes. *PLoS One*, **9**, e112546.
14. Holdsworth, R., Hurley, C.K., Marsh, S.G., Lau, M., Noreen, H.J., Kempenich, J.H., Setterholm, M. and Maiers, M. (2009) The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens*, **73**, 95–170.
15. Robbins, F., Hurley, C.K., Tang, T., Yao, H., Lin, Y.S., Wade, J., Goeken, N. and Hartzman, R.J. (1997) Diversity associated with the second expressed HLA-DRB locus in the human population. *Immunogenetics*, **46**, 104–110.
16. Bontrop, R.E., Otting, N., de Groot, N.G. and Doxiadis, G.G. (1999) Major histocompatibility complex class II polymorphisms in primates. *Immunol. Rev.*, **167**, 339–350.
17. Gongora, R., Figueroa, F. and Klein, J. (1996) The HLA-DRB9 gene and the origin of HLA-DR haplotypes. *Hum. Immunol.*, **51**, 23–31.
18. Kumasaka, N., Okada, Y., Takahashi, A., Kubo, M., Nakamura, Y. and Kamatani, N. (2011), In 12th International Congress of Human Genetics/61st Annual Meeting of The American Society of Human Genetics, Montreal, Canada, Abstract/Program #708F. http://kumasakanatsuhiko.jp/projects/disentangler/.
19. Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., da Silva, A.L., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R. *et al.* (2015) Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.*, **43**, D784–788.
20. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T. *et al.* (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, **47**, 979–986.
21. Brant, S.R., Okou, D.T., Simpson, C.L., Cutler, D.J., Haritunians, T., Bradfield, J.P., Chopra, P., Prince, J., Begum, F., Kumar, A. *et al.* (2017) Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. *Gastroenterology*, **152**(206–217), e202.
22. Huang, C., Haritunians, T., Okou, D.T., Cutler, D.J., Zwick, M.E., Taylor, K.D., Datta, L.W., Maranville, J.C., Liu, Z., Ellis, S. *et al.* (2015) Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans. *Gastroenterology*, **149**, 1575–1586.
23. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P. and Marsh, S.G. (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.*, **43**, D423–D431.
24. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
25. Huang, Y., Yang, J., Ying, D., Zhang, Y., Shotelersuk, V., Hirankarn, N., Sham, P.C., Lau, Y.L. and Yang, W. (2015) HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med.*, **7**, 25.
26. Warren, R.L. and Holt, R.A. (2011) Targeted assembly of short sequence reads. *PLoS One*, **6**, e19816.
27. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
28. Browning, B.L. and Browning, S.R. (2016) Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.*, **98**, 116–126.
29. Mychaleckyj, J.C., Noble, J.A., Moonsamy, P.V., Carlson, J.A., Varney, M.D., Post, J., Helmberg, W., Pierce, J.J., Bonella, P., Fear, A.L. *et al.* (2010) HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin. Trials*, **7**, 75–87.
30. Pillai, N.E., Okada, Y., Saw, W.Y., Ong, R.T., Wang, X., Tantoso, E., Xu, W., Peterson, T.A., Bielawny, T., Ali, M. *et al.* (2014) Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum. Mol. Genet.*, **23**, 4443–4451.

7

EPITOPES OF THE HLA

7.1 IMMUNE EPITOPES

Nucleotide polymorphisms within the MHC region lead to altered chemicophysical properties of the MHC-peptide binding domain. Therefore, different MHC molecules differ in their peptide binding characteristics (Figure 13). The immune epitope database and analysis resource (IEDB) collects and provides access to experimental results for immune epitope experiments [124, 191]. Epitopes are short sequence patterns to which specific receptors of the adaptive immunity can bind. A T-cell receptor can only recognize a peptide that is presented by MHC molecules but not all MHC presented peptides (ligands) are epitopes [158]. The MHC ligands of the MHC-peptide complex, annotated in the IEDB, have been determined by different methods developed over the last 30 years [83]. These methods differ in their throughput, biological factors included (e.g., splicing), the location of the MHC (cellular or purified), used antibodies, and further conditions. In general, the most relevant lab protocols for the analysis of the MHC-peptide binding can be categorized into three main groups [124]:

1. 3D structure by X-ray crystallography
2. MHC binding assays (binding affinity (BA))
3. MHC ligand elution (eluted ligands (EL))

X-ray crystallography generates a 3-D structure of an MHC-peptide complex. It allows deeper insights into the interaction between the MHC molecule and its ligand [87, 209]. The throughput of the approach is very low.

The MHC-peptide binding assays (BA) include a diverse group of experimental setups. One subgroup of widely used approaches are competitive quantitative MHC binding assays, measuring the half maximal inhibitory concentration (IC50) [124]. The IC50 in MHC binding assays measures the concentration needed to occupy half of the MHC molecules in the presence of a fixed concentration of a reference peptide (typically the CLIP peptide) with a peptide of interest. Two of the main differences of the single protocols are the labelling of the ligand (fluorescently or radio-labeled) and if the MHC proteins are purified or cellular [124]. The throughput of these protocols differs, e.g., purification is a limiting factor [83, 124]. To increase the throughput, the MHC-peptide interactions measurement can be performed on a 96-well microtiter plate [41, 91]. A high throughput multiplex assay can be used to measure hundreds of thousands of MHC-peptide interactions at once. As opposed to the competitive assays, where the inhibitory concentration of a peptide is measured. In these high-throughput assays, an

intensity generated by MHC molecules bound to the fixed peptides is analyzed. This relatively new assay type was used in PAPER B (Section 7.3, [197]) in collaboration with Søren Buus and Thomas Østerbye (Department of Immunology and Microbiology, University of Copenhagen, Copenhagen, Denmark).

In the case of MHC ligand elution, peptides initially bound to the MHC proteins are washed from the MHC proteins [110]. The eluted peptides are then typically determined by mass spectrometry (MS) [110, 124]. Peptide elution experiments result in a set of peptides naturally progressed peptides presented by the MHC complex. In contrast, the other approaches measure the interaction for a defined set of peptides. Therefore, peptide elution renders a big advantage over MHC-peptide analysis using array-technology, as peptides that are previously processed within the cell are presented, thus mirroring the natural conditions more truthfully. On the other hand, the method does not measure which peptides do not bind against the MHC molecules of interest [150]. Furthermore, the identified peptidome often does not exactly represent the *in vivo* peptidome, as the elution is performed in cell lines. Sometimes the cells are genetically modified to express only a single MHC molecule type. Additionally, the eluted peptides are mostly self-peptides (i.e., peptides that originate from the cell itself) or peptides from the medium that is used even if lysates of interest (e.g., containing peptides from pathogens) are added before measurement. Moreover, the identification of peptides based on MS spectra is often based on matching the produced MS spectra against a reference database containing either known spectra or expected amino acid sequences (e.g., the human reference proteome) [183]. Peptides not included in the reference can therefore not be identified. Alternatively, a *de novo* approach can identify peptides directly from the spectral data but if the measured MS spectrum is of moderate quality the complete sequence cannot be extracted [38].

The current research focus in peptide-MHC interaction has shifted more onto peptide elution experiments. They allow for the investigation of more biologically relevant mechanisms than previously used experiments and generate hundreds to thousands of naturally presented peptides in a single experiment [49].

7.2 HLA BINDING PREDICTION

Already in the 1990s, MHC epitope data was collected to perform analysis on MHC-peptide binding including data derived from several experiments [22, 146, 147]. Since then, based on the growing data source, diverse tools have been published to a) summarize the features of peptides binding to a specific MHC molecule and thus create an easily interpretable binding motif of MHC associated epitopes (e.g., SYFPEITHI [146], Seq2Logo [182]) and b) to automatically predict the binding status for chosen peptide-MHC combinations, including novel combinations previously not observed in any experiment (Table 1).

The performance and capabilities of tools have continuously been improving with developments in machine learning algorithms and an increasing data availability. Some tools predict the MHC-peptide binding for spe-

TOOL	COVERED ALLELES	ALGORITHM	INPUT	SOUCE	Year
NetMHCIIpan-1.0	MHC II pan	ML	BA	Nielsen et al. [135]	2008
TEPITOPEpan	HLA-DR pan	static presentations	BA	Zhang et al. [209]	2012
NetMHCIIpan-3.2	MHC II pan	ML	BA	Jensen et al. [79]	2018
MARIA	HLA-DR, -DQ	ML, DL	BA, EL, expression	Chen et al. [29]	2019
MAPTAC	HLA II, mono-allelic	DL	EL	Abelin et al. [2]	2019
MixMHC2pred	HLA II pan	static representation	EL	Racle et al. [145]	2019
MHCAttnNet	HLA I and II	DL	BA	Venkatesh et al. [188]	2020
NetMHCIIpan-4.0	MHC II pan	ML	BA and EL	Reynisson et al. [150]	2020
MHCnuggets	MHC pan	DL	BA and EL	Shao et al. [162]	2020
PIA	4 HLA-DR alleles	DL	BA (microarray)	Wendorff et al. [197]	2020
BERTMHC	MHC II pan	DL	BA and EL	Cheng et al. [32]	2021
PIA-M	HLA II pan	DL	EL, expression, subcellular compartment, glycosylation sites	ElAbd et al. [49]	2022

Table 1: Overview of different tools for MHC class II peptide binding prediction. The prediction is based on three different classes of algorithms: static representation (e.g., position specific scoring matrices, motif representation and evolutionary algorithms), machine learning (ML) and deep learning (DL) (ML algorithms with more layers and potentially more complex structures). The tools are based on binding affinity (BA) and eluted ligands (EL) data. Only a selection of tools is represented in this table.

cific alleles only. These tools were mainly developed in the early times of MHC-peptide binding research like PERUN or NetMHCII-2.0 [22, 133] or developed to make use of data generated by a new lab technique (see Section 7.1) as it was the case for PIA (PAPER B, Section 7.3) or MAPTAC [2, 197]. They were limited in their allele coverage as the availability of MHC-peptide data was limited while the model was constructed but they serve as a proof of principle that the used datatype is suitable for predicting the MHC-peptide binding. Later models are often pan-specific, i.e., the prediction uses a pseudo-sequence of the MHC allele and allows an allele specific prediction independent of the allele specific training data through interpolation [32, 79, 87, 135, 150].
The algorithms developed from static representations, like pocket profiles or positions specific scoring matrixes [23, 47, 136, 146, 149, 165, 174, 208, 209], towards machine learning (ML) [79, 133, 135, 150, 193], to deep learning (DL) algorithms with more than one hidden layer. [2, 29, 32, 49, 143, 145, 188, 197].
Over time, data used for training of the tools in Table 1 have been measured using different lab techniques to determine immune epitopes (see Section 7.1). Early algorithms were mainly trained with BA, later data derived from EL or a combination of both was used for training (see Table 1 for details). The use of EL data for MHC-peptide binding prediction was challenging in the beginning, as the analyzed cells expressed different MHC alleles, hindering the assignment of a presented peptide to the MHC molecule presenting it. From the lab side, this problem was faced by developing cell lines expressing only specific MHC alleles. In silico, this problem could be mitigated by using prediction algorithms based on BA data to determine the MHC allele origin, or by clustering the peptides into separate groups (Gibbs Clustering [10]) [18].
As EL data is influenced by additional biological factors, like peptide processing, Chen et al. [29] included additional gene expression data in their tool MARIA. Our PIA-M model uses expression information as well as information about the subcellular compartment and the glycosylation sites to improve the prediction quality [49].

A review from Chen et al. [31] concluded, investigating and comparing different prediction algorithms, that no algorithm outperformed others. The different algorithms predicted best for particular MHC types and peptide lengths. Wang et al. [194] combined different analysis resources on the IEDB webpage and added a consensus approach to reach the best results. Even though the consensus was continuously updated with new tool versions [124, 191, 195], it currently does not include any of the deep-learning tools or a tool using more than sequence information as input.

One of the most used tools in the field is the NetMHCpan group (the latest published versions for MHC class I is NetMHCpan-4.1 and for class II is NetMHCIIpan-4.0[1] [150]). This tool is one of the pan-specific models discussed above. In comparison to previous versions, the current version of NetMHCIIpan is trained on data derived from EL additional to data derived from BA. NetMHCIIpan-4.0 was used in PAPER C (Section 8.3). The previous version NetMHCIIpan-3.2 [79] was used for comparisons in

1 An updated version NetMHCIIpan-4.1 trained on a larger EL dataset is nowadays available online. [131]

PAPER B (Section 7.3) and next to the DL model PIA, NNAlign [132] was trained with data generated for the purpose of this publication.

7.3 PAPER B: PEPTIDE-HLA CLASS II INTERACTIONS

Mareike Wendorff et al. "Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction." In: *Frontiers in Immunology* 11.August (2020), pp. 1–8. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.01705, p. 2

Aim

Due to the availability of new data and development of suitable algorithms, the quality of predicting MHC-peptide binding has continuously improved over time. Here, we aim to integrate novel high-density microarray data and state of the art machine learning algorithms to perform and evaluate MHC-peptide binding prediction.

Methods

We applied novel high-density peptide microarray technology combined in collaboration with the working group of Søren Buus (Department of Immunology and Microbiology, University of Copenhagen, Copenhagen, Denmark). This gave us unbiased MHC-peptide binding data for over 200 000 defined peptides with four exemplary HLA-II molecules. Two machine learning algorithms were trained with these datasets: NNAlign, a recurrent neural network architecture, that forms the baseline of NetMHCIIpan-3.1, and a novel recurrent neural network model implemented in TensorFlow named PIA (Peptide Immune Annotation). The prediction quality of the models was then compared using the "Frank"s, a statistical score based on the ranks of the predictions.

Results

The identified peptide-MHC binding motifs, generated based on predictions with the high-density peptide microarray, correlated with those generated from NetMHCIIpan-3.1 predictions. PIA reached a significantly higher Pearson's correlation coefficient and Spearman's correlation coefficient that NNAlign in a cross-validation approach with the microarray data. Both tools trained on the microarray data achieved a significantly higher Spearman's correlation coefficient on the microarray data than NetMHCIIpan-3.1. The new models were in general on par with the established NetMHCIIpan-3.1 tool in predicting classical BA and EL data.

Conclusion

The microarray technology is a novel technology to investigate the peptide binding preferences of MHC-II molecules in a large-scale and unbiased manner.

Authors Contributions

ER, TØ, SB, AF, and MW designed the wet lab experiments. TØ performed the wet lab experiments. MW prepared the data for training. HE implemented PIA. HG trained NN-align and prepared the IEDB data and plotted the final figures. MW, HG, FD, and MN performed statistical analysis. MW and MN wrote the paper with input from HG and HE. All authors commented on the final manuscript.

frontiers in Immunology

PERSPECTIVE
published: 05 August 2020
doi: 10.3389/fimmu.2020.01705

Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction

Mareike Wendorff[1], Heli M. Garcia Alvarez[2], Thomas Østerbye[3], Hesham ElAbd[1], Elisa Rosati[1], Frauke Degenhardt[1], Søren Buus[3], Andre Franke[1*†] and Morten Nielsen[2,4†]*

[1] Genetics & Bioinformatics, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany, [2] IIBIO, UNSAM-CONICET, Buenos Aires, Argentina, [3] Department of Immunology and Microbiology, University of Copenhagen, Copenhagen, Denmark, [4] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

OPEN ACCESS

Edited by:
Laura Santambrogio, Weill Cornell Medicine, United States

Reviewed by:
Markus Maeurer, Champalimaud Foundation, Portugal
James Drake, Albany Medical College, United States

***Correspondence:**
Mareike Wendorff
m.wendorff@ikmb.uni-kiel.de
Andre Franke
a.franke@mucosa.de

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Antigen Presenting Cell Biology, a section of the journal Frontiers in Immunology

Received: *21 February 2020*
Accepted: *25 June 2020*
Published: *05 August 2020*

Citation:
Wendorff M, Garcia Alvarez HM, Østerbye T, ElAbd H, Rosati E, Degenhardt F, Buus S, Franke A and Nielsen M (2020) Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction. Front. Immunol. 11:1705. doi: 10.3389/fimmu.2020.01705

Human Leukocyte Antigen class II (HLA-II) molecules present peptides to T lymphocytes and play an important role in adaptive immune responses. Characterizing the binding specificity of single HLA-II molecules has profound impacts for understanding cellular immunity, identifying the cause of autoimmune diseases, for immunotherapeutics, and vaccine development. Here, novel high-density peptide microarray technology combined with machine learning techniques were used to address this task at an unprecedented level of high-throughput. Microarrays with over 200,000 defined peptides were assayed with four exemplary HLA-II molecules. Machine learning was applied to mine the signals. The comparison of identified binding motifs, and power for predicting eluted ligands and CD4+ epitope datasets to that obtained using NetMHCIIpan-3.2, confirmed a high quality of the chip readout. These results suggest that the proposed microarray technology offers a novel and unique platform for large-scale unbiased interrogation of peptide binding preferences of HLA-II molecules.

Keywords: ultra-high density peptide microarray, MHC class II, HLA, antigen presentation, prediction, peptide binding, high-throughput, machine learning

INTRODUCTION

The highly diverse major histocompatibility complex (MHC) proteins play a major role in the adaptive immune system. MHC class II proteins present peptides of variable lengths mainly derived from extracellular antigens (1). In humans, MHC is called human leukocyte antigen (HLA). The HLA locus is highly polymorphic, resulting in different HLA molecules having a specific peptide binding preference and specific peptidomes. The HLA is an important susceptibility locus in genetic studies of many immune-related diseases, often with multiple HLA alleles playing a role (2–4). However, beyond the suggested association, these studies do not inform about the causes of a disease, i.e., the antigen/epitope that binds to associated HLA proteins and potentially drive the disease onset. To make this link between

HLA and antigen, further studies to characterize the peptidome bound by specific HLAs are necessary (5, 6). To this end, efficient and reliable high-throughput technologies for measuring peptide-HLA interaction are needed. Different assay types may be used to record the interaction between HLA and peptides (7). Classical *in-vitro* assays measure one single interaction of a synthetic peptide and an HLA-molecule in one experiment. Mass spectrometry of HLA eluted peptides considers the whole process of antigen synthesis up to presentation might fail the identification of low abundant peptides or modified peptides. To avoid costs and time delays *in-silico* prediction tools for HLA binding and antigen presentation have been trained on measured assay data (8–17).

Here, we set out to overcome the experimental limitations outlined above by employing our high-density peptide microarray data, a new high-throughput *in-vitro* technology (18, 19), combined with synthetic *in-vitro* generated HLA-II molecules (20) to perform large-scale unbiased characterization of HLA-II allele-specific binding. Earlier work has used peptide microarray for measuring peptide-HLA interaction, but this was limited to thousands of peptides per array (21). Here, the high-density peptide microarray enables the *in-situ* synthesis of over 2 million peptides per array on about 2 cm^2 (18). To this end, we synthesized about 70,000 random peptides in triplicates on one array, allowing us to generate vastly more data points than the combined number of all HLA-DR epitopes registered in the immune epitope database IEDB (www.iedb.org) (7). This technology enables the analysis of whole proteomes of interest in one single experiment and the systematical analysis of post-translational modifications. Our presented technology offers a unique solution to produce large datasets to characterize binding properties of HLA-II molecules and improve the *in-silico* prediction of peptide-HLA interaction while being suitable for hypothesis driven tests.

To prove the quality of the high-density peptide microarray for characterizing peptide-HLA-II interactions, we selected four HLA-DRB1 proteins that are known to be strongly associated with ulcerative colitis, a complex chronic inflammatory bowel disease (2). For DRB1*01:03, DRB1*03:01, DRB1*15:01, and DRB1*15:02, a set of 69,815 random peptides were analyzed. To mine the extracted datasets and to learn predicting peptide-HLA binding, we applied NNAlign (22), as well as a deep learning approach, referred to as PIA (Peptide Immune Annotation). Using the obtained models, we assessed the quality of the chip readout in terms of identified binding motifs, and power to predict publicly available MS data from elution experiments as well as CD4+ epitope datasets in comparison to NetMHCIIpan-3.2 (8).

STATE OF RESEARCH

Peptide-HLA Assays

The IEDB collects all types of immune epitopes. The oldest record is from 1952 (7). From the 90s to 2010, *in-vitro* assays measuring binding of synthetic peptides to HLA molecules (20, 23) were the most common MHC binding assays (www.iedb.org). In the last 5 years, mass-spectrometry (MS) sequencing of HLA eluted peptides (24, 25) became more popular (first records already in 1991).

Both methods have their strengths and weaknesses. *In-vitro* binding studies can measure interaction of individual peptide-HLA combinations. However, this approach is highly cost-intensive (one assay per peptide) and the assay fails to address some events leading up to effective HLA antigen presentation such as antigen processing, the effects of chaperones like HLA-DM, editing of the repertoire of HLA bound peptides (12, 14), and HLA-peptide complex stability. In contrast, recent advances in MS technology have expanded the detectable peptide repertoire presented by HLA molecules (immunopeptidome) by use of liquid chromatography MS. Immunopeptidome data include comprehensive information on the complex HLA ligand presentation (26), and analysis results of such data are a rich source of information for learning about the underlying rules of HLA antigen presentation. However, MS HLA peptide elution data mainly covers self-peptides and is assumed to miss low abundant peptides (26), further post-translational modifications might be identified but misinterpreted (27). Another problem arising with natural cell lines is that they most often present different HLA proteins. To solve this problem, either homozygous cell lines, tagging of a specific HLA allele (14, 28) or algorithms for deconvolution of the HLA proteomes can be employed (16, 17, 29). However, deconvolution has been shown to be of limited success in cases of lowly expressed HLA proteins or cells expressing HLA proteins with overlapping proteome specificity (14).

Peptide-HLA Binding Prediction

Beyond the different experimental approaches developed to specify peptide-HLA interaction, large efforts have been made to develop prediction models capable of accurately predicting peptide-HLA binding. Historically, most *in-silico* methods have been developed based on *in-vitro* binding data and an exemplary state-of-the-art computational method is NetMHCIIpan (8, 9, 30). Recently, prediction methods have been developed from HLA-II elution data (10–15). The results suggest that the inclusion of elution data has a positive impact on the predictive power of *in-silico* methods in particular for the prediction of HLA antigen presentation (10–15). Algorithms can be trained on either *in-vitro* or *in-vivo* data (8, 11, 14, 16), but a benefit from training on the two data types combined has been reported (10, 12, 13, 15, 17). However, currently even the best methods for prediction of HLA-II binding and antigen presentation suffer from an excessive number of false positive predictions.

MATERIALS AND METHODS

Microarray

Peptide microarrays were produced by Schafer-N (Copenhagen, Denmark). Briefly, a Nexterion E microscope slide (Schott, Jena, Germany) were amino functionalized with a 1% w/v linear copolymer (1T0C) of N,N-dimethylacrylamide (Sigma-Aldrich) and aminoethyl methacrylate (Sigma-Aldrich) and used as substrate for solid-phase peptide synthesis. The peptide synthesis

was initiated with the coupling of one unit of epsilon-amino-capronic acid (EACA) followed by the peptide sequences. The 1T0C and EACA unit served as a spacer between the array surface and the peptides allowing the HLA class II molecules to interact and peptides to protrude out of the HLA in both ends. For each experiment the same array-design was chosen. The peptide chips were subdivided into 12 sectors with a marker peptide "PVSKMRMATPLLMQA" of the HLA-II antigen gamma chain (CD74; UniProt: P04233-1: 103-117) placed multiple times in all the sectors corners. 69,815 different random natural 13-mer peptides were placed on the chip in triplicates. The chip does not contain peptides containing more than four poly residues (e.g., RRRR) as poly residues are difficult to synthesize and have a tendency toward unspecific binding.

Peptide microarrays were incubated with different HLA-DR molecules as previously described (31). Briefly, HLA-DR molecules were diluted from a stock (8 M Urea, 25 mM Tris, pH8) to achieve a final concentration of 500 nM HLA-DR in PBS, 0.05% Lutrol F68, 20% Glycerol pH 7.4 and added (overlaid) to the peptide array surface and allowed to fold for 24 h at 18°C before washing and staining with monoclonal mouse anti-HLA-DR (L243) and goat anti-mouse-Cy3. The peptide arrays were scanned with a laser-scanner (InnoScan, Innopsys, France) at a resolution of 1 μm and the amounts of bound HLA-DR were quantified to intensities between 0 and 254 by a proprietary software (Peparray, Schafer-N, Denmark). Larger spots with high values were excluded as noise.

The data was normalized by taking the median intensity of each repeated measurement $\tilde{x}_i$ for each peptide and transformed to fall in the range 0–1 by $\frac{log(\tilde{x}_i+1)}{log(max(\tilde{x})+1)}$. The data was split into one test dataset comprising 10% of the data and a 10-fold cross-validation dataset of the remaining data. To ensure limited data redundancy between subsets, therefore similar peptides [e.g., a 9-mer overlap (underlined amino acids in the following are the same), for example AALITRGLTEMGR and ARTALITRGLTEM, or at least 11 of the 13 amino acids in the same order, for example ADLGSGAGAAGLA and ALGSGAAGAAFGL] were placed into the same subset. For the performance evaluation, the data was back-transformed to the intensity scale.

Consistency Metrics

We evaluated the consistency of the triplicates using the coefficient of variation (CoV) and the Pearson Correlation Coefficient (PCC) between three replicates. The CoV for each repeated peptide measurement was calculated as the standard deviation of intensity divided by mean intensity +1 and the mean CoV over the 69,815 peptides for all four alleles was given.

The PCC between the three replicates was calculated combining the pairwise PCCs R_{12}, R_{13}, and R_{23} as $R_{123}=\sqrt{{R_{12}}^2+{R_{13}}^2+{R_{23}}^2-2^*(R^*_{12}R^*_{13}R_{23})}$ (32).

Epitope and Eluted Ligand Test Datasets

T cell epitopes and HLA ligands obtained from mass spectrometry were downloaded from IEDB and used as independent test data (www.iedb.org, June 18th 2019) (7). Only positive linear peptides with a length between 13 and 19 amino acids were used. Data with an overlapping sequence of at least 9 amino acids with the peptide microarray data or an unknown amino acid were excluded. This resulted in 502 epitopes and 719 ligands for the four alleles (**Supplementary Table 1**).

Negative data (peptides thought not to bind the respective alleles) were added by downloading the sequence of the epitope/ligand source protein as linked by IEDB from NCBI (www.ncbi.nlm.nih.gov), and *in-silico* digesting by a sliding window of the length of the ligand/epitope into overlapping peptides. Peptides with an overlap of 9 amino acids with the peptides used in training or the positive peptides were excluded.

For predicting the binding affinity for a peptide, prediction on all 13-mer subsequences was made and the highest prediction value reported.

Finally, the performance for each epitope/ligand was reported as the Frank value. The Frank value of a binding peptide is the ratio of the number of peptides with a higher predicted binding score in the source protein divided by the overall number of peptides within the protein (8).

NNAlign

NNAlign-2.1 was used on the peptide microarray data (22). NNAlign generates artificial neural network models of receptor-ligand interactions. The program takes as input a set of ligand sequences with target values; it returns a sequence alignment, a binding motif of the interaction, and a model that can be used to scan for the motif in other sequences. Further details of the used parameters can be found in the **Supplementary Methods**. The motifs generation by Seq2Logo (33) is automatically performed by NNAlign.

Deep Learning Model PIA

PIA is a gated recurrent neural network (GRU) based model (34) implemented using Keras (https://keras.io) deep learning framework with TensorFlow (www.tensorflow.org). Further details on the model architecture can be found in the **Supplementary Methods**.

For generating the logos, 500,000 13-mers were randomly selected from the human reference proteome and screened using PIA. The top 1% of peptides were submitted to GibbsCluster-2.0 (29) for motif identification.

RESULTS

High-density peptide microarrays were used to identify large, unbiased peptidome datasets for four HLA-DR molecules. We describe the raw peptide chip readout to quantify data consistency and make comparisons to earlier *in-vitro* binding experimental results. Further, we describe the results of applying two machine-learning frameworks to mine and extract the rules for peptide-HLA binding from the chip data, and we assess the quality of the chip data by comparing the power of the constructed models to that of NetMHCIIpan-3.2 for prediction of HLA ligands and epitopes.

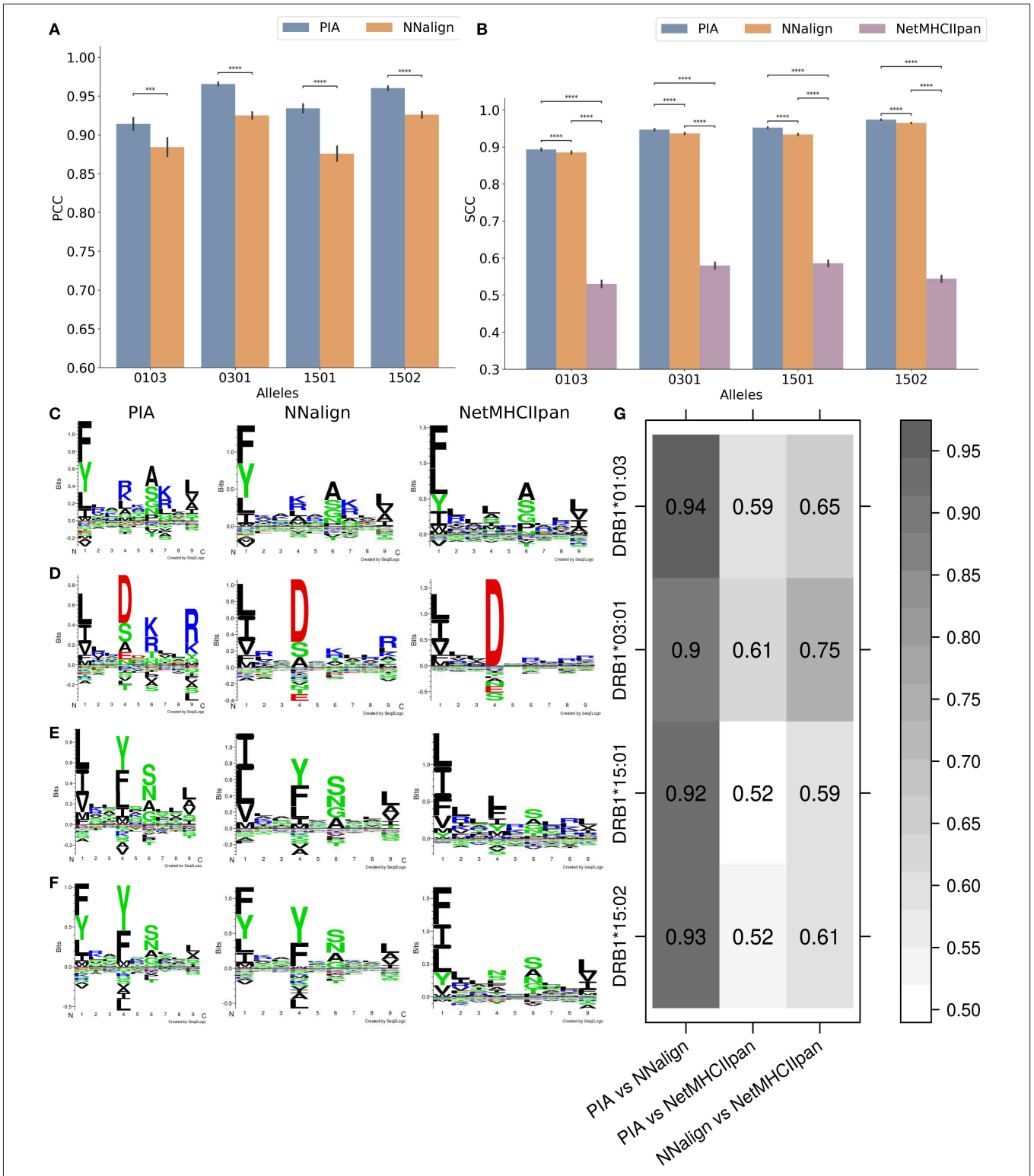


FIGURE 1 | Performance of the models on peptide microarray data and resulting motifs. **(A)** The Pearson correlation coefficient (PCC) and **(B)** the Spearman correlation coefficient (SCC) on the independent test dataset of the peptide microarray are shown. The pairwise *p*-values were calculated using a non-parametric bootstrap hypothesis test with 1,000,000 bootstrap iterations. *$0.01 < p \leq 0.05$, **$0.001 < p \leq 0.01$, ***$0.0001 < p \leq 0.001$, and ****$p \leq 0.0001$. Motif plots of **(C)** DRB1*01:03, **(D)** DRB1*03:01, **(E)** DRB1*15:01, and **(F)** DRB1*15:02 based on the top 1% (from a pool of 100,000 random natural peptide) binding peptides generated with the deep learning model (PIA), NNAlign model and NetMHCIIpan-3.2 (8). **(G)** Pearson correlation coefficient (PCC) of the position specific scoring matrices (PSSM) between the different models.

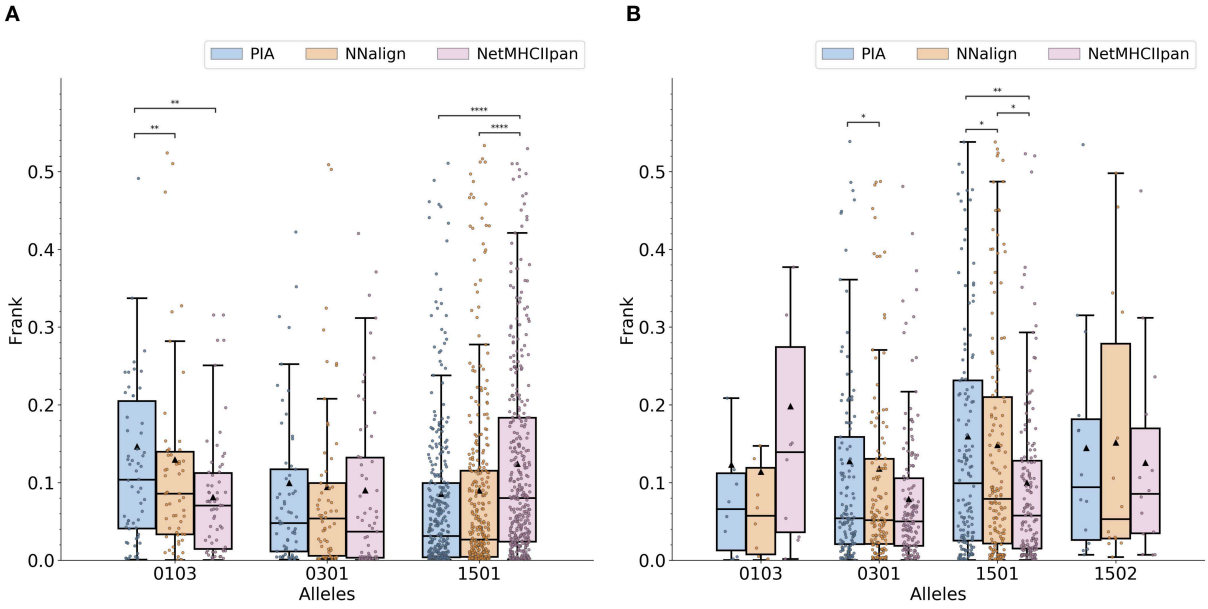


FIGURE 2 | Comparisons of prediction quality on **(A)** MS ligand and **(B)** epitope data. The center line inside the box indicates the median Frank and the triangle shows the mean Frank. The data points available in IEDB are represented using a jitter plot. The colored box covers the interquartile range. The whiskers represent 1.5-fold of the interquartile range. Pairwise *p*-values were calculated using a Wilcoxon signed-rank test (applying Pratt's zero method). *$0.01 < p \leq 0.05$, **$0.001 < p \leq 0.01$, ***$0.0001 < p \leq 0.001$, and ****$p \leq 0.0001$.

Microarray Experiments

The peptide microarray contained 69,815 random 13-mer peptides. An example of the raw readout of the array is shown in **Supplementary Figure 1**, confirming overall clear signals corresponding to discrete peptides. To assess the accuracy and consistency of the array readout, two metrics were used: the CoV and correlation coefficient between the three repeated peptide measurements (for details see **Materials and Methods**). Overall, this analysis demonstrated highly consistent values with a mean CoV over the 69,815 peptides for all four alleles of 0.135 and a correlation coefficient over 0.988 for the single microarrays (**Supplementary Figure 2**).

For further validation of the microarray readout, the amino acid composition of the top 2% peptides with highest signal was compared to the amino acid composition of peptide binders as obtained from the IEDB for the HLA molecules where available. The results of this analysis are shown in **Supplementary Figure 3** and confirmed an overall high consistency between the two with correlation coefficients for HLA-DRB1*03:01 and HLA-DRB1*15:01 above 0.910.

For further analysis, the median of the triplicate was used.

Prediction of Microarray Data

For building the prediction models, the microarray data was log-transformed to reduce the skew and to optimize the range of the data. We trained the NNAlign and the GRU based PIA models using 10-fold CV for each allele. In all cases, PIA outperformed NNAlign. **Figures 1A,B** show the PCC and the Spearman correlation coefficient (SCC) performance values of the two models on the test dataset. Here, PIA outperforms NNAlign in all cases. **Figure 1B** also includes the SCC performance of NetMHCIIpan-3.2, which is trained on *in vitro* IC50 binding values demonstrating at least a SCC of above 0.53 for the different alleles for predicting the chip test data.

To quantify the consistencies between different data types and prediction models, binding motifs were estimated for each HLA molecule and prediction model (**Figures 1C–F**). The binding motifs identified by the peptide microarray based models are close to identical and in most cases similar to those generated with NetMHCIIpan-3.2 using Seq2Logo (8, 33). To compare the motifs obtained by the two microarray-based models, we performed a correlation analysis of the 9 × 20 position specific scoring matrix produced by Seq2Logo defining the predicted binding motif. In all four cases, we obtained PCC values above 0.90 (**Figure 1G**). When comparing the NNAlign and NetMHCIIpan motifs, the correlation values were still very high with 0.59–0.75.

Predict Ligands Measured by Mass Spectrometry

To further assess the predictive power of the developed methods, we performed a benchmark on a set of HLA eluted ligands as obtained from the IEDB (7). Here, the Frank value was used as performance measure (8). In short, Frank is the proportion of peptides within a source protein with a prediction value greater than the given ligand. The Frank is 0 if the ligand is the peptide with the highest binding score and 0.5 for random predictions. To limit the effect of noise and falsely positive assigned data points, only ligands that obtained a Frank value of 0.15 or less for at least one of the included prediction models were included

in the benchmark. As the IEDB currently does not contain any ligands for DRB1*15:02 the molecule was excluded from our analysis (**Supplementary Table 1**). The results (**Figure 2A**) demonstrate an overall comparable performance of the three methods. The microarray-based methods and NetMHCIIpan-3.2 each outperform the other for one dataset (NetMHCIIpan-3.2 performs better for DRB1*01:03, and PIA and NNAlign for DRB1*15:01). No consistent performance difference was observed between the NNAlign and PIA models.

Predict CD4+ Epitopes

The same analysis performed on the HLA eluted ligand data was done on a set of CD4+ T cell epitopes available from the IEDB (**Supplementary Table 1**). The results (**Figure 2B**) show that the microarray-data based models in most cases performed on par with NetMHCIIpan-3.2. For DRB1*15:01, NetMHCIIpan-3.2 significantly outperformed both peptide microarray-based methods. For DRB1*01:03, the microarray-based models showed an increased performance compared to NetMHCIIpan-3.2. This latter difference was, however, not statistically significant due to the limited number of epitopes available in the benchmark. Moreover, the results indicate a slightly improved performance of NNAlign over PIA.

DISCUSSION

Genetic variants in the HLA gene region have been associated with a multitude of diseases, not only autoimmune conditions. Earlier work suggests this to be caused by an intrinsic property of particular HLA variants [for instance different HLA-DQ alleles influencing IL-17 production in T-cells irrespective of the peptide ligand (35)]. However, beyond this and for most HLA's and diseases, the detailed underlying mechanisms and candidate antigens remain unknown. Experimentally testing all possible peptide-HLA combinations to identify the relevant antigens for a given disease is a major undertaking, and with current technologies in most cases not feasible.

To deal with this limitation, we here present a new type of HLA-II antigen interaction assay based on high-density peptide microarrays. This technique allows the assessment of more than 200,000 independent peptide-HLA interaction tests within one single experiment.

We demonstrate how this high-density peptide array serves as a novel, valuable source for high-throughput and high-volume data to accurately characterize the peptidome of HLA-II molecules and its binding specificity. We demonstrated this by quantifying the consistency between internal replica (peptides analyzed multiple times on a given microarray), and by comparing the amino acid composition of the peptidomes as obtained from the peptide microarray to that obtained using conventional *in-vitro* binding assays with solid phase synthesized peptides. We furthered the validation by applying machine learning methods to mine and extract the HLA binding signal from the microarray data and compared the derived binding motif and power of the associated prediction model to state-of-the-art methods trained on conventional *in-vitro* binding data. All comparisons confirmed a high consistency of the microarray data with conventional methods.

In our study, two different machine learning algorithms were applied to mine the large-scale microarray datasets. The first is NNAlign, which is the basis for NetMHCIIpan-3.2 and NetMHCII 2.3 (8, 22) accepted to be among the best available for prediction of peptide binding to HLA-II (30). The second, PIA is based on GRU, a deep learning architecture developed for sequence learning. Both algorithms are able to capture the signal within the peptide microarray data and predict the microarray test dataset with very high performance.

Moreover, the two prediction models trained on the microarray data were benchmarked against NetMHCIIpan-3.2 on independent data of HLA eluted ligand and CD4+ epitope data obtained from the IEDB. Here, all models were found to perform at par, suggesting that the measurements obtained from the microarray are accurately capturing signals of peptide-HLA binding.

The microarray experiments performed here were conducted in the absence of HLA II peptide- loading chaperones such as HLA-DM and HLA-DO earlier demonstrated to play a role in editing the repertoire of HLA class II binding peptides (36, 37). Future work will tell if similar results are obtained in the context of the peptide-microarray technology.

Overall, our results suggest that the described microarray technology for large-scale evaluations of peptide-HLA-II interaction is accurate, precise and highly scalable. We believe this result opens a venue of novel applications addressing challenges and biological problems that can only to a limited extent be addressed using conventional immunoassays. Such applications include mapping the impact of peptide-specific post-translational modifications (such as phosphorylation, deamination, or citrullination, all of these modifications can be added in the peptide synthesis step, i.e., in the array design process) on HLA-II binding and unbiased large-scale screening for HLA-II binding of pathogen proteomes. The herein presented *in-silico* technology data is in our opinion a good addition to immunopeptidome data for the next generation of prediction tools.

DATA AVAILABILITY STATEMENT

The high density peptide raw datasets generated for this study can be downloaded from https://www.ikmb.uni-kiel.de/resources/download-tools/publicly-available-data. All other data are available from the corresponding authors upon request.

AUTHOR CONTRIBUTIONS

ER, TØ, SB, AF, and MW designed the wet lab experiments. TØ performed the wet lab experiments. MW prepared the data for training. HE implemented PIA. HG trained NNAlign and prepared the IEDB data, and plotted the final figures. MW, HG, FD, and MN performed statistical analysis. MW and MN wrote the paper with input from HG and HE. All authors commented on the final manuscript.

FUNDING

MW and HE were funded by the German Research Foundation (DFG) (Research Training Group 1743, Genes, Environment and Inflammation). MN was researcher of CONICET. This project received infrastructure support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2167-390884018, the Agencia Nacional de Promoción Científica y Tecnológica, Argentina (PICT-2016-0089), the Independent Research Fund Denmark award DFF—6110-00644 and the Danish MS Society award A31444.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.01705/full#supplementary-material

REFERENCES

1. Chicz RM, Urban RG, Gorga JC, Vignali DA, Lane WS, Strominger JL. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J Exp Med.* (1993) 178:27–47. doi: 10.1084/jem.178.1.27
2. Goyette P, Boucher G, Mallon D, Ellinghaus E, Jostins L, Huang H, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet.* (2015) 47:172–9. doi: 10.1038/ng.3176
3. Liu H, Irwanto A, Fu X, Yu G, Yu Y, Sun Y, et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. *Nat Genet.* (2015) 47:267–71. doi: 10.1038/ng.3212
4. Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, Doncheva NT, et al. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat Genet.* (2013) 45:670–5. doi: 10.1038/ng.2616
5. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med.* (2017) 9:1–13. doi: 10.1126/scitranslmed.aai8708
6. Miyadera H, Tokunaga K. Associations of human leukocyte antigens with autoimmune diseases: challenges in identifying the mechanism. *J Hum Genet.* (2015) 60:697–702. doi: 10.1038/jhg.2015.100
7. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* (2019) 47:D339–D343. doi: 10.1093/nar/gky1006
8. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology.* (2018) 154:394–406. doi: 10.1111/imm.12889
9. Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, Nielsen M. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics.* (2018) 34:1522–8. doi: 10.1093/bioinformatics/btx820
10. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol.* (2019) 37:1332–43. doi: 10.1038/s41587-019-0280-2
11. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunol Res.* (2020) 8:396–408. doi: 10.1158/2326-6066.CIR-19-0464
12. Barra C, Alvarez B, Paul S, Sette A, Peters B, Andreatta M, et al. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* (2018) 10:84. doi: 10.1186/s13073-018-0594-6
13. Reynisson B, Barra C, Kaabinejadian S, Hildebrand WH, Peters B, Nielsen M. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res.* (2020) 19:2304–15. doi: 10.1101/799882
14. Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity.* (2019) 51:766–79.e17. doi.org/10.1016/j.immuni.2019.08.012
15. Garde C, Ramarathinam SH, Jappe EC, Nielsen M, Kringelum J V., Trolle T, et al. Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics.* (2019) 71:445–54. doi: 10.1007/s00251-019-01122-z
16. Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol.* (2019) 37:1283–6. doi: 10.1038/s41587-019-0289-6
17. Alvarez B, Reynisson B, Barra C, Buus S, Ternette N, Connelley T, et al. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Mol Cell Proteomics.* (2019) 18:2459–77. doi: 10.1074/mcp.TIR119.001658
18. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol Cell Proteomics.* (2012) 11:1790–800. doi: 10.1074/mcp.M112.020800
19. Hansen LB, Buus S, Schafer-Nielsen C. Identification and mapping of linear antibody epitopes in human serum albumin using high-density peptide arrays. *PLoS ONE.* (2013) 8:e68902. doi: 10.1371/journal.pone.0068902
20. Justesen S, Harndahl M, Lamberth K, Nielsen L-LB, Buus S. Functional recombinant MHC class II molecules and high-throughput peptide-binding assays. *Immunome Res.* (2009) 5:2. doi: 10.1186/1745-7580-5-2
21. Gaseitsiwe S, Valentini D, Ahmed R, Mahdavifar S, Magalhaes I, Zerweck J, et al. Major histocompatibility complex class II molecule-human immunodeficiency virus peptide analysis using a microarray chip. *Clin Vaccine Immunol.* (2009) 16:567–73. doi: 10.1128/CVI.00441-08
22. Nielsen M, Andreatta M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions. *Nucleic Acids Res.* (2017) 45:2–7. doi: 10.1093/nar/gkx276
23. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, et al. Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr Protoc Immunol.* (2013) Chapter 18:Unit 18.3. doi: 10.1002/0471142735.im1803s100
24. Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol.* (2016) 41:9–17. doi: 10.1016/j.coi.2016.04.005
25. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol Cell Proteomics.* (2015) 14:3105–17. doi: 10.1074/mcp.O115.052431
26. Vaughan K, Xu X, Caron E, Peters B, Sette A. Deciphering the MHC-associated peptidome: a review of naturally processed ligand data. *Expert Rev Proteomics.* (2017) 14:729–36. doi: 10.1080/14789450.2017.1361825
27. Ramarathinam SH, Croft NP, Illing PT, Faridi P, Purcell AW. Employing proteomics in the study of antigen presentation: an update. *Expert Rev Proteomics.* (2018) 15:637–45. doi: 10.1080/14789450.2018.1509000
28. Yang J, Jaramillo A, Shi R, Kwok WW, Mohanakumar T. *In vivo* biotinylation of the major histocompatibility complex (MHC) class II/peptide complex by coexpression of BirA enzyme for the generation of MHC class II/tetramers. *Hum Immunol.* (2004) 65:692–9. doi: 10.1016/j.humimm.2004.04.001
29. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* (2017) 45:W458–63. doi: 10.1093/nar/gkx248
30. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Comput Biol.* (2018) 14:e1006457. doi: 10.1371/journal.pcbi.1006457

31. Osterbye T, Nielsen M, Dudek NL, Ramarathinam SH, Purcell AW, Schafer-Nielsen C, et al. HLA class II specificity assessed by high-density peptide microarray interactions. *J Immunol*. (2020) 205:290–9. doi: 10.1101/2020.02.28.969667
32. Wang J, Zheng N. Measures of correlation for multiple variables. *arXiv*. (2014). 1–20. Available online at: http://arxiv.org/abs/1401.4827v6 (accessed January 27, 2020).
33. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* (2012) 40:281–7. doi: 10.1093/nar/gks469
34. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014).* (2014) Available online at: https://arxiv.org/abs/1406.1078v3 (accessed September 03, 2020).
35. Mangalam AK, Taneja V, David CS. HLA class II molecules influence susceptibility versus protection in inflammatory diseases by determining the cytokine profile. *J Immunol.* (2013) 190:513–19. doi: 10.4049/jimmunol.1201891
36. Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity.* (2019) 51:766–79.e17. doi: 10.1016/j.immuni.2019.08.012
37. van Lith M, McEwen-Smith RM, Benham AM. HLA-DP, HLA-DQ, and HLA-DR have different requirements for invariant chain and HLA-DM. *J Biol Chem.* (2010) 285:40800–8. doi: 10.1074/jbc.M110.148155

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Part III

THE HUMAN LEUKOCYTE ANTIGEN IN INFLAMMATORY BOWEL DISEASES

The human leukocyte antigen is the strongest associated genetic locus for inflammatory bowel disease, although the specific role of the human leukocyte antigen in the pathogenesis remains unknown. In this part, I summarize the current knowledge about the association of the human leukocyte antigen with inflammatory bowel diseases. In PAPER C I study the genetic association and the autoimmune hypothesis in ulcerative colitis, which is one subtype of inflammatory bowel diseases. Finally, I discuss the different approaches used in analyzing the human leukocyte antigen in diseases, what is known and not known about the human leukocyte antigen in inflammatory bowel disease, which possibilities exist in gain additional knowledge and how this knowledge might help patients in the future.

8

HLA IN IBD

The HLA was among the first identified genetic risk factors in IBD using classical linkage analysis. An alternative name used for this region is IBD3, for inflammatory bowel disease locus 3. Other hints for an association of with HLA with IBD were already published in a serological study from 1972 [5, 61, 204]. Though IBD has been continuously associated with the class II genes of the HLA region, in particular *HLA-DRB1* and *-DQ*, it is still unclear which genetic locus is causal in the disease etiology [44, 63]. Other genes, besides the classical HLA genes within the gene rich HLA region, might also play a role in the association with the disease. For example, tumor necrosis factor (TNF)-α and the MHC class I chain-related genes (MICA/B), located in the HLA class III region between HLA class I and class II, are suspects to a separate genetic association as described in Muro, López-Hernández, and Mrowiec [120].

The association of the HLA region with IBD was among those genetic associations suffering a reproducibility crisis [204]. In retrospect, limited power within the single studies due to small sample sizes and differences in populations are mainly responsible for the different effects seen across different studies.

Today, a consistent picture of the association of the most common HLA alleles is available regarding the association of HLA with the two main forms of IBD. Even though, more samples are needed for rare alleles or alleles predominantly present in understudied populations. An overview of the single associated HLA alleles is given in Section 8.1. Hypothesis on the functional role of the HLA in IBD are described in Section 8.2.

8.1 GENETIC ASSOCIATIONS WITHIN THE HLA WITH IBD

The genetic profile of UC and CD patients differs in the HLA region. The associations are stronger within UC. Ahmad, Marshall, and Jewell [5] noted that the sharing of HLA alleles within families suggests that the HLA region contributes 64%-100% of the total genetic risk of ulcerative colitis but only 10%-33% of the total genetic risk of Crohn's disease. Goyette et al. [63] calculated the variance explained by HLA alleles to be 6.2% in UC and 3.1% in CD. Due to the different identified association profiles of the HLA in the tow diseases, they are now described separately.

8.1.1 *HLA associations in Ulcerative Colitis*

The most significant associations within the HLA for UC are in the genes *HLA-DR* and *HLA-DQ* of the class II. Even though the evidence is limited, there are some hints towards *HLA-DR* to be the gene of interest. First, the expression level of *HLA-DR* is in general higher [81]. Second, Goyette et al. [63] found a proxy for all significantly associated *HLA-DQ* genes in *HLA-DR* but not vice versa.

8.1.1.1 *HLA-DRB1 association in Ulcerative Colitis*

Figure 15 presents the *HLA-DRB1* alleles analyzed in three relevant publications of this field. The inner part shows a meta-analysis of Stokkers et al. [173]. This study published in 1999 included 29 studies on UC, CD, or both subtypes. In comparison to more recent studies, the single studies performed until this timepoint included small sample sizes with up to 344 cases. They were often based on serological studies or alternatively on genetic determination with a one-field resolution. Their meta-analysis identified the serotype DR4, today noted as HLA-DRB1*04, as protective. DR9 (HLA-DRB1*09) and DR2 (DRB1*15 and DRB1*16) and its subtype DR15 (DRB1*15) were identified as risk factors for UC.

In 2015, Goyette et al. [63] published an HLA fine-mapping study based on a large Caucasian cohort including more than 32 000 IBD cases. The HLA alleles were determined by imputation (see Section 6.2). The results are presented in the middle ring in Figure 15. The study confirmed the protective effect of HLA-DRB1*04 and the analysis in a 2-field resolution showed the same direction of effect for all alleles of this group with HLA-DRB1*04:01 and *04:04 being statistically significant. Due to the larger power, HLA-DRB1*07 and HLA-DRB1*03 were statistically significant associated as protective with UC. Furthermore, HLA-DRB1*09 was statistically associated as protective by Goyette et al. [63] and therefore contradicts the finding of Stokkers et al. [173]. None of the single studies agglomerated in Stokkers et al. [173] identified a significant association with this allele. Next to HLA-DRB1*15, the 1-field alleles DRB1*01, DRB1*11 and DRB1*12 were statistically significantly associated with a risk in UC by Goyette et al. [63]. For DRB1*01 interestingly only the rare 2-field allele HLA-DRB1*01:03 was statistically significant with the highest OR among all significantly associated HLA alleles with 3.59 (p-value$= 9.47 \times 10^{-121}$). For DRB1*15 the two alleles DRB1*15:01 and DRB1*15:02 are also separately significantly associated, where DRB1*15:02 has the higher OR with 2.21. The HLA-DRB1*13 serogroup seems to be heterogenic regarding the association with UC. In Goyette et al. [63] the 1-field allele was not significantly associated, but the 2-field allele DRB1*13:01 is statistically significantly associated as a risk factor, while the 2-field allele DRB1*13:02 shows a trend towards being protective.

Even though no significant association within HLA-DRB1*13 was identified by the trans-ethnic analysis from Degenhardt et al. [44], the heterogenic picture was present here as well. In general, the analysis from Degenhardt et al. [44] allowed an insight into cross-ethnicity allele association at the HLA loci at 2-field resolution, with differences in allele frequencies

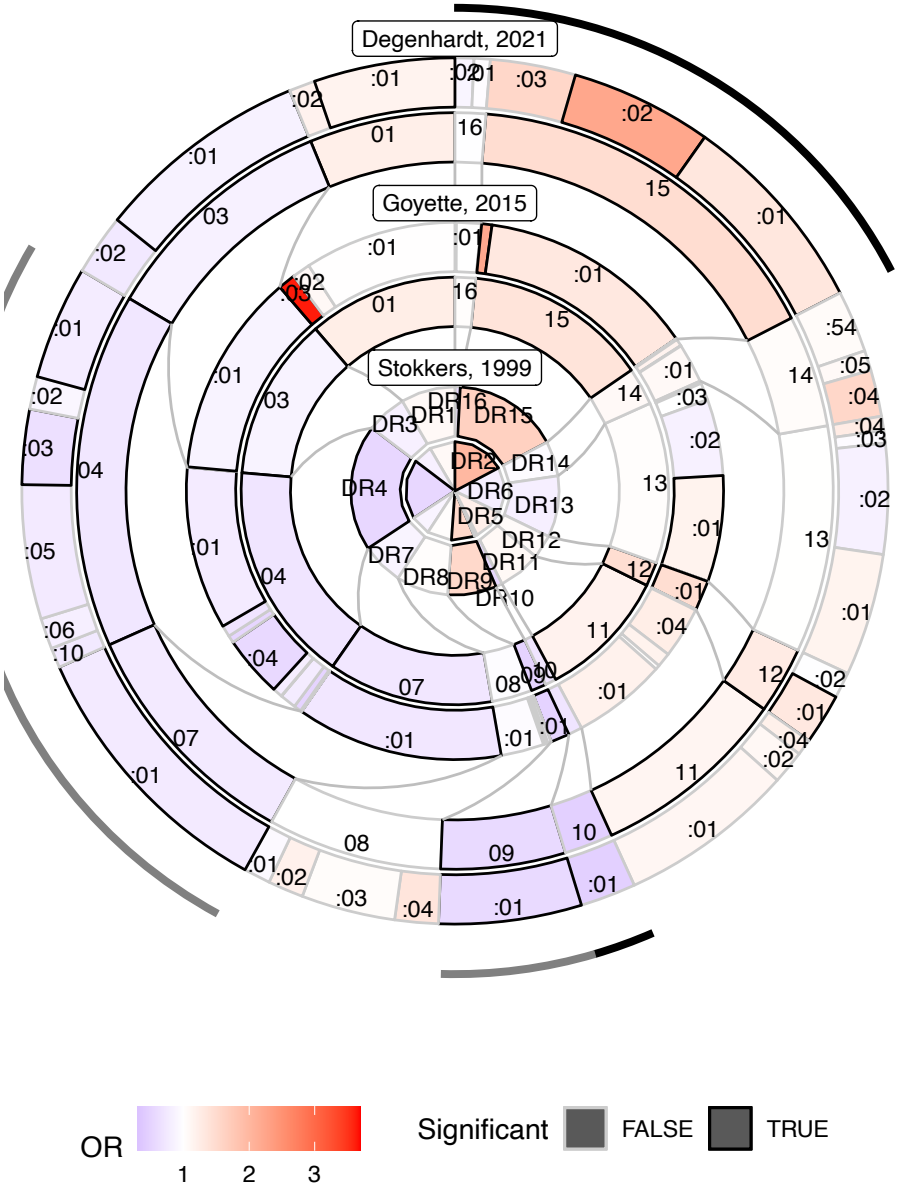


Figure 15: HLA-DRB1 alleles associated with UC. The figure summarizes the results from Stokkers et al. [173], Goyette et al. [63] and Degenhardt et al. [44]. The angle is approximately representing the allele frequency in control samples. The inner ring for each publication presents the lower reported resolution, while the outer ring represents the finest reported resolution. Results from Stokkers et al. [173] were considered significant (marked by a black frame around the allele) if the reported 95% confidence interval does not include an OR of 1, while for Goyette et al. [63] and Degenhardt et al. [44] a p-value of 5×10^{-8} was applied.

observed across the different analyzed populations. The results of the performed meta-analysis are presented on the outer ring of Figure 15. The HLA-DRB1*15 alleles are highlighted here as the different subtypes are especially prevalent in separate global regions. Both the DRB1*15:01 allele in Caucasians and the DRB1*15:02 allele in Asians, were significantly associated with risk in UC by Degenhardt et al. [44] and Goyette et al. [63]. The DRB1*15:03 allele, which is specifically prevalent in people with African ancestry, failed to reach genome wide significance, most probably because of small sample sizes within this ancestry. The analysis from Degenhardt et al. [44] reached genome-wide significance for the HLA-DRB1*01:01 allele, but as explained in the study the strength of this signal may result from a misclassification of DRB1*01:03 as DRB1*01:01 (Supplementary Figure 8 in [44]). All studies concur that there are no significant associations with DRB1*14 and DRB1*08.

Overall, the association of the HLA-DRB1 alleles with UC is stable across different studies. The only exception is HLA-DRB1*09. Except from this, the serogroups DRB1*03, *04 and *07 were found to be associated with a statistically protective effect in UC, while DRB1*01, *11, *12, and *15 were found to be significantly associated with an increased risk of UC. The effect size within the single serogroups varies but even though the sample sizes increased over the years, the power to show all effects on the 2-field level is still limited.

8.1.1.2 *HLA haplotypes and other non-HLA-DRB1 alleles*

Most of the statistically significant associated classical HLA alleles of the other loci are in LD with an *HLA-DRB1* allele. Goyette et al. [63] calculated conditional regression models based on the *HLA-DRB1* signals (Figure 16b). They showed that only four alleles were independently associated with UC when correcting for the *HLA-DRB1* alleles. Those are HLA-A*02:01, HLA-C*12:02, HLA-B*18:01, and HLA-DPB1*03:01. Except for HLA-B1*52:01, located on the same haplotype as HLA-C*12:02, all other HLA alleles of the class I and of *HLA-DP* show only secondary effect in LD with a stronger associated *HLA-DRB1* allele. The haplotype HLA-C*12:02-B*52:01-DRB1*15:02-DQA1*01:03-DQB1*06:01 is the only haplotype described by Goyette et al. [63] where the association signal in the HLA class I region is stronger than in the HLA class II region. This effect is also visible in the non-Caucasian data from Degenhardt et al. [44], who validated the haplotype signals identified by Goyette et al. [63], using a classic phasing approach.

Many *HLA-DQ* alleles were identified with similar effects as the *HLA-DRB1* alleles. Due to the high LD a conclusion which allele is causative for the association remains until know impossible. Among those DQB1*06:02, located on the same haplotype as DRB1*15:01, is the allele with the highest OR and DQB1*03:03 the allele with the lowest OR. In addition to *HLA-DRB1*, Degenhardt et al. [44] included the loci *HLA-DRB3*, *-DRB4* and *-DRB5*. As reviewed in Degenhardt et al. [44] the alleles of these genes are in high LD with the *HLA-DRB1* alleles, they generally show a stronger association based on the P-value, since they have a higher power for detection, which is related to the comparably small number of alleles at these loci.

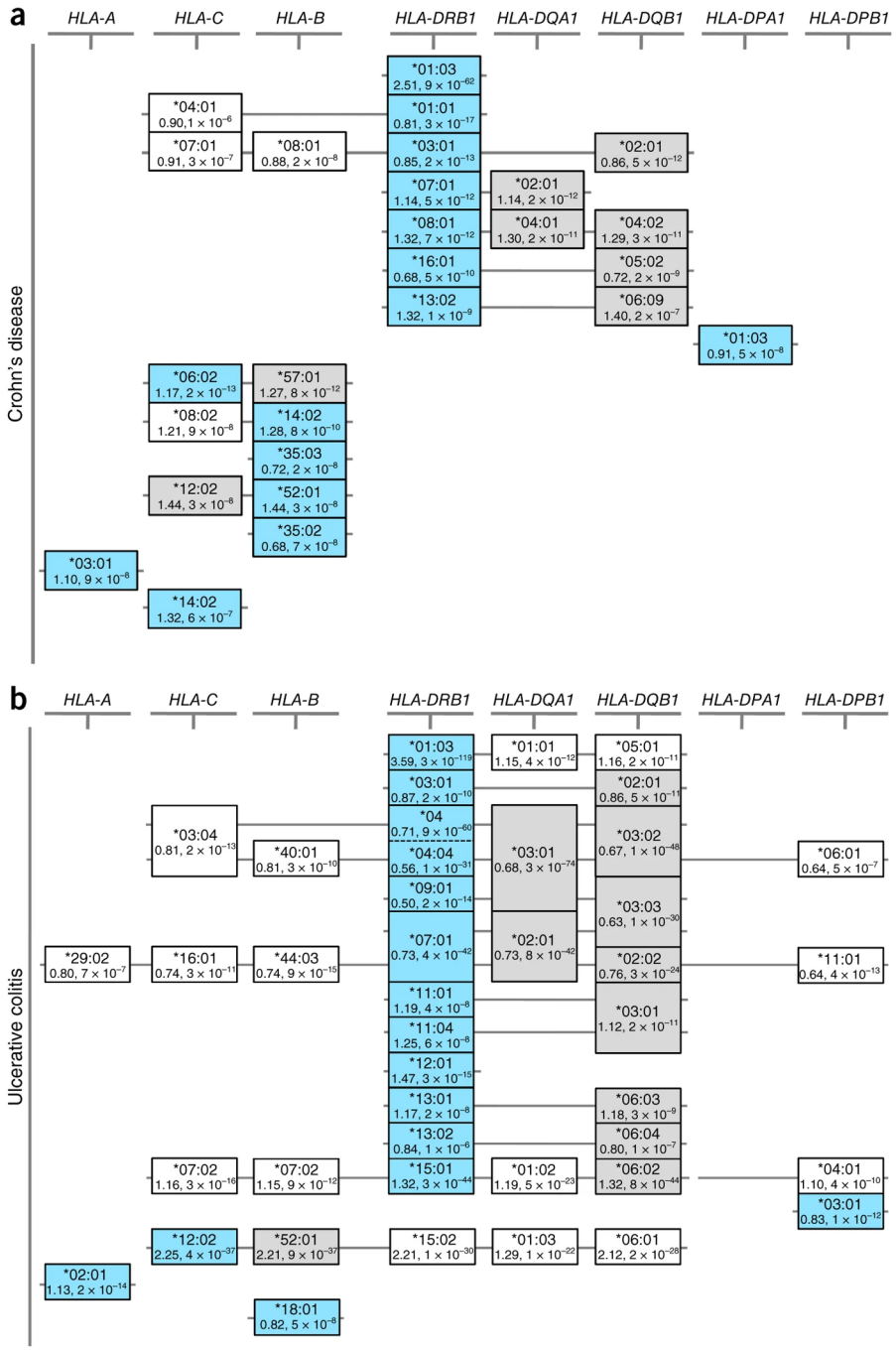


Figure 16: Correlated HLA association signals across the different classical HLA genes. The upper graphic (a) represents the alleles associated with CD, while the lower graphic (b) represents the alleles associated with UC. Each box represents an allele with its corresponding OR (bottom left) and p-value (bottom right). The blue boxes include alleles considered in the model generated by Goyette et al. [63], while connected grey boxes show alternative alleles with a comparable effect and white boxes show alleles with secondary effects. Figure taken from Goyette et al. [63].

8.1.2 *HLA associations in Crohn's disease*

As already mentioned, the HLA association in CD is not as rich as the association in UC, meaning that here is a generally less significant associations of CD with the HLA across both HLA loci. At the HLA allele level, the association of HLA class II is also not that much stronger than the association with the HLA class I comparing both effects and p-values. One exception is the allele DRB1*01:03, which is also the strongest associated HLA allele in UC (compare to Figure 16a and Figure 17). For HLA-DRB1*01:01, the most common allele of the DRB1*01 group, the effect direction identified in [63] was opposite to the one identified in UC and those of DRB1*01:03. Next to HLA-DRB1*01:03 only the DRB1 allele group HLA-DRB1*03 was significantly associated with the same direction of effect in both subtypes of IBD. HLA-DRB1*07 is significantly associated with both diseases but with the opposite direction of effect. Of the other DRB1 alleles, statistically significant associated with CD, HLA-DRB1*08:01 and HLA-DRB1*16:01 were not statistically significant in UC at the p-value of 5×10^{-6}. This cutoff is defined by Goyette et al. [63] to represent statistical significance. Of those alleles HLA-DRB1*16:01 was correlated with a protective effect, while the other two alleles were associated with an increased risk for CD.

When compared to the HLA-DRB1 based model generated by Goyette et al. [63] for UC, the HLA-DRB1 based model for CD included more alleles of other classical HLA loci. Here, the association of HLA-DRB1 covers only a smaller part of the association (Figure 16b). Again, for most haplotypes the association signal cannot be separated between *HLA-DR* and *HLA-DQ* but while all significantly associated *HLA-DQ* alleles can be represented by an *HLA-DRB1* allele, this does not hold true the other way round. There is an independent statistically significant association for the HLA-DPA1*01:03 allele and some HLA class I haplotypes namely: C*06:02-B*57:01, B*14:02, B*35:03, C*12:01-B*52:01, B*35:02, A*03:01, and C*14:02. The analysis on the trans-ethnic data used in Degenhardt et al. [44] has not yet been published for analysis of CD and is therefore not included in the comparison. No DRB1 allele significant in either Stokkers et al. [173] or in Goyette et al. [63] was found to have another direction of effect in the other study, respectively (Figure 17).

8.2 POTENTIAL ROLE OF HLA IN IBD

Ashton et al. [13] summarized potential roles of the HLA in IBD. Their hypotheses are summarized in Figure 18. As described in the previous section, different 1- and 2-field alleles of the HLA are associated with a different risk of IBD. Those alleles mainly differ in their peptide binding region. One hypothesis of the role of the HLA in IBD is that the associated HLA risk alleles present peptides to the host immune system which in turn triggers an immune response. The peptides of interest might either originate from the commensal microbiome in the gut, or from the host itself. Those self-peptides might contain patterns similar to a bacterial peptide (molecular mimicry) and may drive the disease, also referred to as an auto-immunogenic antigen. In combination with triggering of an

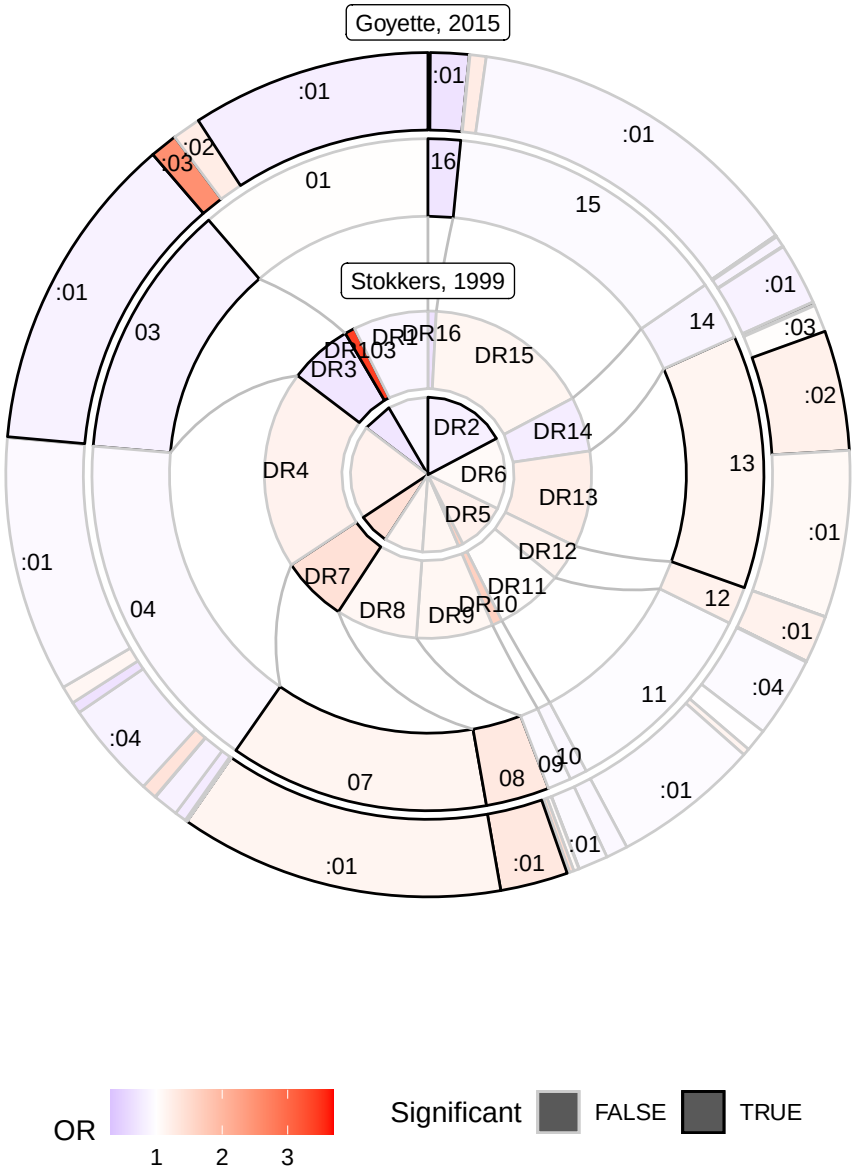


Figure 17: HLA-DRB1 alleles associated with CD. The figure summarizes the results from Stokkers et al. [173] and Goyette et al. [63]. The angle is approximately representing the allele frequency in control samples. The inner ring for each publication presents the lower reported resolution, while the outer ring represents the finest reported resolution. Results from Stokkers et al. [173] were considered significant if the reported 95% confidence interval does not include an OR of 1, while for Goyette et al. [63] a p-value of 5×10^{-8} was applied.

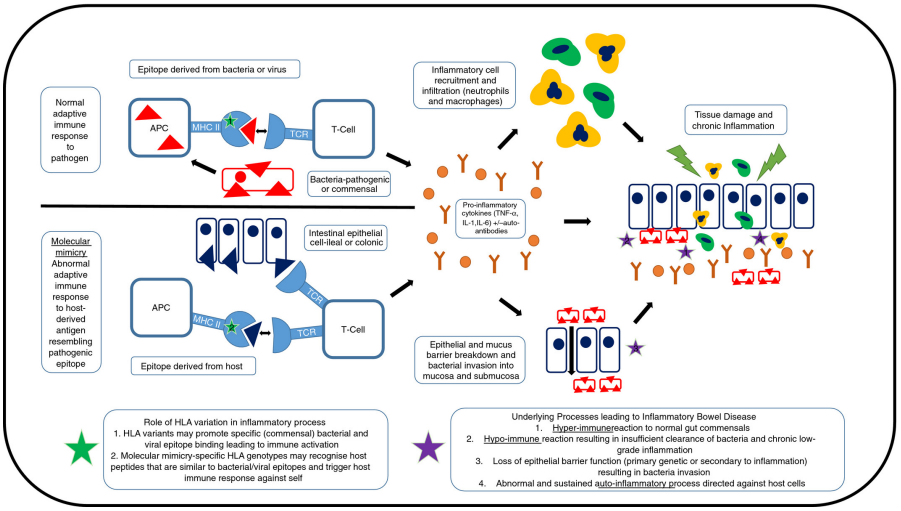


Figure 18: Potential role of HLA in IBD. Different forms of peptide presentation by HLA proteins are hypothesized to initiate an immune reaction leading to IBD (left side with green stars). Different forms of abnormal immune responses as implicated by IBD might be triggered by the HLA (right side with purple stars 1-4). Original figure from Ashton et al. [13].

immune response, inducing the activation of T cells and the production of pro-inflammatory cytokines, other factors may further enhance the disease progression to IBD. Those may, for instance be an overreaction of the host's immune system to the commensal microbiome (hyper-immune reaction), insufficient clearance of bacteria resulting in a chronic low-level immune reaction or an impaired mucosal barrier, either induced by genetic factors or by the immune reaction itself. This in turn may lead to the invasion of bacteria into the tissue. Another factor may be an autoimmune process directed against the host cells.

8.3 PAPER C: ANALYSIS OF THE HLA IN UC

Aim

The HLA is a known risk factor for UC but the concrete mechanisms behind this association are unknown. In this study, we aim to figure out what we can learn from the human genetics in UC patients and controls to enlighten the potential role of HLA in IBD. As the HLA proteins are interacting with various peptides, changes within the protein coding regions are of special interest. Here, we focus on the autoimmune hypothesis and try to identify potential autoimmune peptides driving the disease by their potential to be presented by HLA proteins.

Methods

To reach this aim, we performed a GWAS analysis on imputed SNP array data and exome sequencing data of 863 German UC patients and 4 185 controls. The protein sequences are generated based on WES and used to generate a peptidome including the variation within the sample set. The HLA alleles were imputed using HIBAG. A peptidome wide association study (PepWAS) was conducted on peptides generated candidate protein sequences and the characteristics of the binding peptides was analyzed. Furthermore, the peptides predicted to be differentially presented between cases and controls were analyzed for further supporting attributes like higher expression, overlap with immunopeptidomics data, and the subcellular compartment of the corresponding proteins.

Results

The genetic analysis revealed a novel association with NOD2, a well-known risk gene for CD. The new mutations are located in a different part of the protein. Additionally, mutations were identified in genes by previous GWAS analysis not, or not sufficiently covered (*TAS2R43* and *UNQ6494*). Overall, the genetic association represented the signals as expected based on previous GWAS studies. This is also true for the association of HLA alleles. A further analysis of the binding characteristics showed that HLA alleles associated with a decreased risk of UC have a binding motif including more acidic amino acids. But not all protective alleles show the same binding pattern that distinguishes them from alleles associated with an increased risk of UC. The PepWAS analysis yield 234 significant candidate peptides, whereof one was also present in the immunopeptidomics data in a gene differentially expressed in UC patients. This peptide originates from HLA-DRB1*15 alleles, a group of alleles associated with a comparably high risk of getting the disease. It is the most promising candidate identified.

Conclusion

Here, we identified a few novel genetic associations with UC. Previously, those were most probably missed, because of the used reference and im-

putation panel. Those findings need to be validated with a larger dataset. It suggests that there are still some associations that can be revealed by classical GWAS, using WES, WGS data, or imputation based on the GRCh38 genome build.

The HLA alleles show a constant association with the disease. The allele HLA-DRB1*15:01 having the largest power among the DRB1 alleles in our dataset. The applied PepWAS is one approach to filter down candidate peptides. Whatsoever, as the HLA is not selective on the greater scale, HLA does not give sufficient clues about the protein source. The HLA presentation itself allows to draw conclusions about the pattern of the peptides of interest, as it is based on physico-chemical properties of the peptide. Therefore, follow up analyses need to be conducted to test the 234 candidate peptides from PepWAS for their immunogenicity, especially in relation to the TCR repertoire of UC patients. Furthermore, additional sources of peptides, such as microbial peptides or environmental, should be considered in following analyses.

Authors Contributions

M.W., A.F. conceived and initialized the project. M.W. and H.E. analyzed the data, with help of M.H., F.U.-W., E.M.W., L.W., S.J., X.B., and D.E.. M.W. wrote the manuscript with help of H.E.. M.L., M.Z., B.B., and S.S. collected the samples. M.J. and X.B. generated the WES data. H.E., P.B. and A.T. generated the peptide elution data. T.L.L. and A.F. supervised the project. All authors revised and edited the manuscript for critical content and approved of the final version to be published.

[Title]

Genome-wide analysis of individual coding variants and HLA-II-associated self-immunopeptidomes in ulcerative colitis

[Authors]

Mareike Wendorff[1,2,#], Hesham ElAbd[1], Frauke Degenhardt[1], Marc Höppner[1], Florian Uellendahl-Werth[1], Eike M. Wacker[1], Lars Wienbrandt[1], Simonas Juzenas[1,3], Regeneron Genetic Center[4], Tomas Koudelka[5], David Ellinghaus[1], Petra Bacher[1,6], Andreas Tholey[5], Matthias Laudes[7], Malte Ziemann[8], Bernd Bokemeyer[9], Stefan Schreiber[10], Tobias L. Lenz[11], Andre Franke[1,#]

[1]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany.
[2]IPN-Leibniz Institute for Science and Mathematics Education at Kiel University, 24118 Kiel, Germany.
[3]Institute of Biotechnology, Life Science Centre, Vilnius University, Lithuania.
[4]Regeneron Genetics Center, Tarrytown, NY, USA.
[5]Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine, Kiel University, Kiel, Germany.
[6]Institute of Immunology, Kiel University, Kiel, Schleswig-Holstein, Germany.
[7]Department of Medicine I, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.
[8]Institute of Transfusion Medicine, University Hospital Schleswig-Holstein, Lübeck, Germany.
[9]Interdisciplinary Crohn Colitis Centre Minden, Germany.
[10]Department of Internal Medicine I, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.
[11]Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany.

Corresponding Authors:
Mareike Wendorff
Olshausenstraße 62
24118 Kiel
wendorff@leibniz-ipn.de

Prof. Andre Franke
Rosalind-Franklin-Straße 12
24115 Kiel
a.franke@mucosa.de

[Abstract]

Genome wide association studies contributed to a better understanding of the etiology of inflammatory bowel disease (IBD). While over 240 genetic associations with IBD have since been identified, functional follow-up studies are still in their infancy with the overall pathogenesis of IBD remaining unsolved. E.g., a functional understanding of the genetic association between the human leukocyte antigen (HLA) region and ulcerative colitis (UC) – one subtypes of IBD – is still lacking. Here, we analyzed whether an autoimmune reaction involving the HLA class II proteins HLA-DQ and -DR, both being strongly associated with UC, could be a disease trigger or driver. To this end, genotype data derived from whole exome sequencing and genome-wide SNP array data of 863 German UC patients as well as 4,185 healthy controls were analyzed. Association analyses identified novel variants in the *NOD2* and *SNX20* genes to be linked with UC and confirmed known HLA allele associations. Employing the genetic data, we generated patient-specific self-immunopeptidomes and *in silic*o predicted HLA-peptide binding. Peptidome-wide association analyses of peptide binding preferences in a set of candidate proteins yielded significant associations with 234 specific peptides. Interestingly, none of those peptides showed a differential presence in case and control samples. The disease-associated candidate peptides predicted to be presented by risk HLA proteins contained predominantly aromatic amino acids. In contrast, protective HLA proteins were predicted to bind peptides enriched in acidic amino acids. In summary, we present a proof-of-concept immunogenetic analysis that contributes to a better understanding of the HLA in UC.

[Introduction]

Although about 0.3% of people in the industrialized countries suffer from inflammatory bowel disease (IBD)[1], the etiology of the diseases is still unclear. Different studies have correlated the disease with environmental[2,3] and genetic factors[4–7]. For some of the known factors, a concrete role in the pathogenesis of IBD has been identified, in other cases the impact on the disease remains unknown. Interestingly, the major histocompatibility complex region (MHC), which shows the strongest genetic association with the disease, still has an enigmatic role in IBD[8]. The genetic association at this locus differs between the two main subtypes of IBD, Crohn's disease (CD) and ulcerative colitis (UC)[8,9]. Especially in UC, the MHC class II genes are highly associated with the disease in Caucasians (Goyette *et al.*[9]) and across different ancestries (Degenhardt *et al.*[10]). Though detailed analysis showed a consistent genetic profile, the biology behind the associations remains unsolved. MHC class II proteins mainly present peptides derived from extracellular proteins to CD4-positive T cells, which may then elicit an immune response in the host. The classical MHC class II genes in humans are the human leukocyte antigen (HLA) genes HLA-*DPA1*, -*DPB1*, -*DQA1*, -*DQB1*, -*DRA*, -*DRB1* and depending on the (*DRB1*-related) haplotype possibly one of HLA-*DRB3*, -*DRB4* and -*DRB5*. Except for HLA-*DRA*, which is nearly invariable regarding its protein sequence, all other classical HLA proteins are highly polymorphic. Several different processes of how the HLA may contribute to the inflammation in IBD have been discussed in the literature, most of those are based on the structural differences in the peptide binding pocket and the related differences in antigen recognition[8]. Differences among HLA alleles and their binding preferences have already been discussed in Goyette *et al.*[9] and Degenhardt *et al.*[10]. Here, we dig deeper into the *autoimmune hypothesis* which refers to an immune reaction triggered by HLA-presentation of a host's self-peptide[8].

For this purpose, we analyzed genotype data derived from next generation sequencing (NGS)-based whole exome-sequencing (WES) and genotyping based on Illumina's Global Screening Array (GSA) as well as imputed HLA allele information of 5,048 German individuals for genetic association with UC (**Figure 1, Supplementary Table 1**). This data is unprecedented in its resolution of individual-level coding variation.

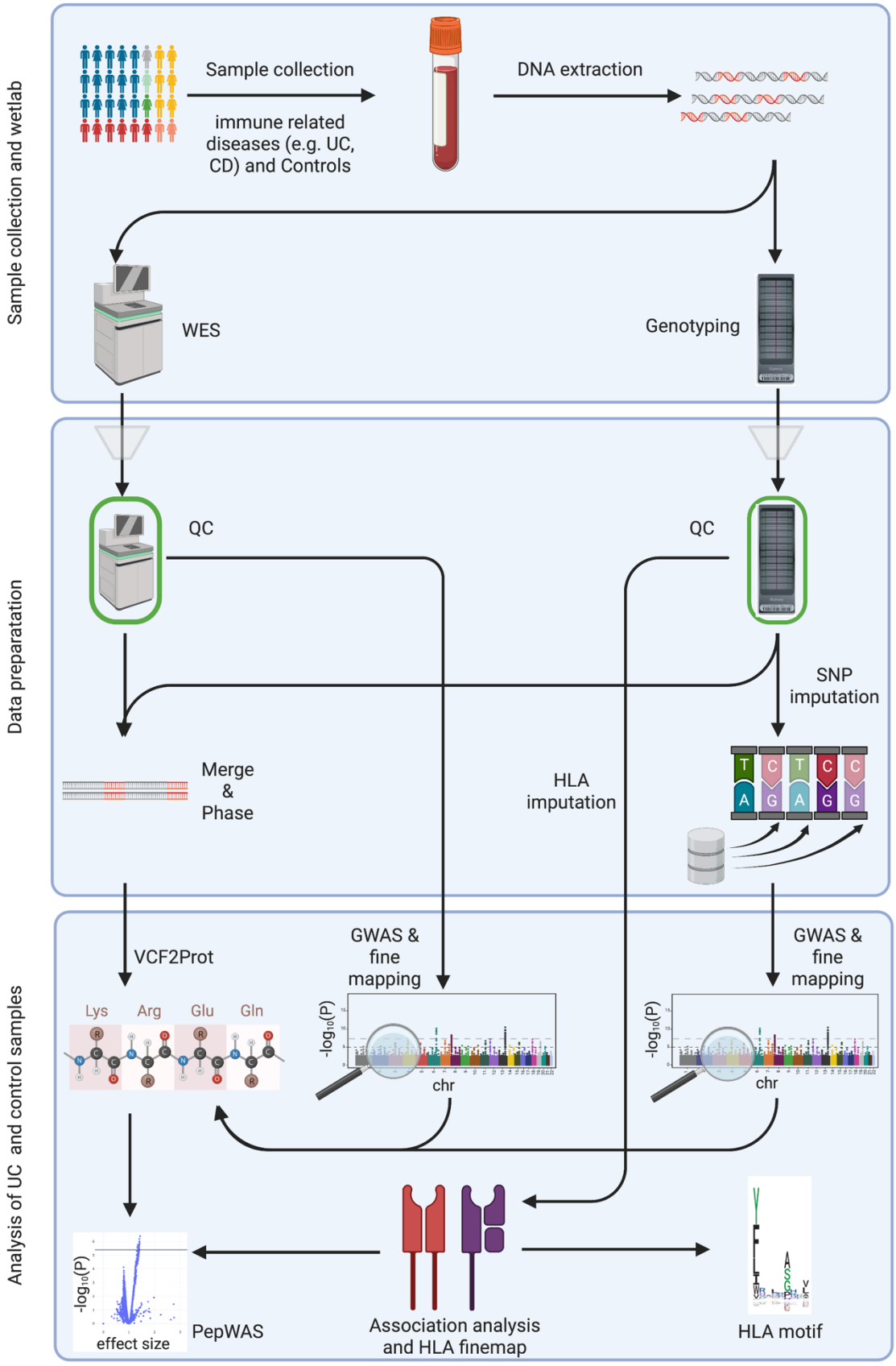


Figure 1: Graphical abstract of the current study. The first box summarizes the sample collection and wetlab part. The second box shows the data preparation. The third box lists the different performed analyses. UC: ulcerative colitis, CD: Crohn's disease, DNA: Deoxyribonucleic acid, QC: quality control, SNP: single-nucleotide

polymorphism, HLA: human leucocyte antigen, GWAS: genome-wide association study, PepWAS: peptidome-wide association study. This Figure was created with BioRender.com.

Based on this exceptionally high-resolution genotype data, we first performed a genome-wide association analysis (GWAS) as well as a fine-mapping of the HLA region. Subsequently, we took advantage of the individual WES data and performed a peptidome-wide association analysis (PepWAS)[11] based on personalized proteomes, focusing on self-peptides from a disease-relevant set of human proteins (**Supplementary Table 2**), to identify candidate self-peptides with a differential likelihood of presentation in patients and controls. The personalized proteomes were generated by translating the per-patient observed nucleotide variations from the WES data into a collection of personalized protein sequences. Further, we checked the peptides identified through the PepWAS for mutations in the coding DNA sequence leading to the according peptide or another amino acid sequence.

[Results]

Comparison of exome data and imputed genotyping data

The exome data is expected to include novel or rarely described coding variants. In a first step we therefore compared how many exome variants are not included in our imputed SNP (single-nucleotide polymorphism) genotyping array dataset. The comparison revealed that half of the variants discovered by exome sequencing are not detected in the imputed data. Most of those variants are very rare (minor allele frequency (MAF)<0.1%). Of the variants with a higher allele frequency about 11% are unique to the exome data. 6.3% of the common exome variants (MAF>5%) are still not imputed.

The fraction of variants specific to the exome data varies not only with the allele frequency but also with the type of mutation. E.g., for missense variants the most common variants were imputed and only 2.8% of variants were specific for the exome dataset but the fraction of overlap is lower for InDels.

GWAS of imputed genotyping data

For the summary statistics of the imputed genotyping data, a genomic inflation factor lambda of 0.995 was calculated based on the data excluding the HLA region (**Supplementary Figure 1**). This means no population stratification is expected to cause false positive hits. The Manhattan plot of the analysis is shown in **Supplementary Figure 2**. Overall, 1,713 of the imputed genotyping markers with a minor allele frequency of at least 1% passed the suggestive significance threshold P-value<10^{-5} and 975 of those reached genome-wide significance. The at least suggestively significant markers were assigned to 26 loci,

considering two variants as belonging to one locus if their physical distance in GRCh38 is below 150 kbp, overlapping largely with results from LD-based clumping (P-value<0.00001) but allowing for a combination of larger association signals such as the HLA class II with 24 clumped signals calculated by PLINK. Of these loci 19 main variants were supported by variants in LD (**Supplementary Table 3**; **Supplementary Figures 3-21**). Three of those loci contained genome wide significant variants: 1p36.13 (*OTUD3,* **Supplementary Figure 5**), 6p21.32 (*HLAII,* **Supplementary Figure 11**), and 5p14.3 (no specific gene, **Supplementary Figure 9**). In agreement with previously published GWAS[7], our dataset shows the main significant genetic association in the HLA region on chromosome 6 with the main peak being in the region of the classical HLA class II genes. In addition, *OTUD3* is one of the strongest associated loci reported previously in the literature for UC[7]. 5p14.3 (**Supplementary Figure 9**) was not reported previously and does not replicate in the publicly available IIBDGC dataset (available through RICOPILI)[30]. The main signal rs2937516 (P-value=1.62×10^{-8}, OR=0.68 [0.60–0.78]) is located on chr5:18748431 between the genes *LINC02100* (chr5:18,514,857-18,746,202) and the pseudogene *UBE2V1P12* (chr5:18,886,622-18,887,095).

Next to the three genome-wide significant loci, we identified 16 suggestively significant loci (**Supplementary Table 3**): 5p13.1 (*PTGER4*, **Supplementary Figure 10***)[7], 10q24.2 (NKX2-3*, **Supplementary Figure 14**)[7] and 22q13.1 (*PDGFB - RPL3*, **Supplementary Figure 21**)[7] as well as 6p21.33 (*HLAI,* **Supplementary Figure 11**) all of which being previously reported to be associated with UC (NHGRI-EBI GWAS catalog[29]). An intriguing novel association signal for UC was detected at 16q12.1 (*NOD2 and SNX20*, **Supplementary Figure 17**), a locus known previously to be associated only with Crohn's disease[7,54]. A detailed follow up-analysis of these variants in the context of UC is described in the paragraph *SNX20/NOD2 association in ulcerative colitis and Crohn's disease.* 12q24.13-q24.21 (*RBM19*, rs3782449, P-value= 8.90×10^{-6}, OR=0.73 [0.64-0.84], **Supplementary Figure 16**) and 4p12 (rs113429955, P-value=5.51×10^{-6}, OR=1.81 [1.40-2.33], **Supplementary Figure 8**) have not been previously described but were replicated using the IIBDGC dataset with replication-P-values of 0.00738 and 0.045, respectively. For 12q24.13-q24.21 (*RBM19*), an association with ocular manifestation in IBD was described by others before[55]. The LD between the T allele rs4766697, the variant associated with ocular manifestation and the A allele rs3782449, our protective lead variant, is with R^2 of 0.0058 very low but the D' is 1. Therefore, all T of rs4766697 (allele frequency of 2.2 %, P-value=0.64, OR=0.92 [0.64-1.32] in our data) are most likely located on the haplotype of the reference allele A of rs3782449 (allele frequency 79.3 %). The two-sided Fisher's Exact test on the contingency table of the dosages yielded a P-value of 1.27×10^{-7}.

The region of the signal at 9q22.2 (*UNQ6494*, rs36147380, P-value=7.4×10^{-6}, OR=0.74 [0.65-0.84], **Supplementary Figure 13**) was not covered in the GRCh37 build and therefore no lookup in the IIBDGC GWAS data was possible. The remaining eight loci did neither replicate in the IIBDGC GWAS dataset nor were they listed in the NHGRI-EBI GWAS catalog, but four of these loci had additional links to UC. Firstly, *TNFRSF8*, also called *CD30L*, at 1p36.22 (rs72641067, P-value=7.4×10^{-3}, OR=1.47 [1.25-1.77], **Supplementary Figure 3**) is an (auto-)immune relevant gene and the gene coding for the corresponding ligand gene *TNFSF8* has been described as associated with CD[7,56]. Secondly, the gene *TPRG1* (3q28, chr3:188,947,214-189,325,304) was found to be associated in the IBD GWAS from de Lange *et al.*[5], but the location of the signal was slightly different. Our lead SNP rs73184427 is located at chr3:189,167,658 (P-value=7.11×10^{-06}; OR=1.56 [1.28–1.89], **Supplementary Figure 7**), while de Lange *et al.* described rs56116661 at chr3:188,683,372, annotated to the *LPP* gene, to be associated with CD (P-value=5.67×10^{-10}; OR=1.14 [1.10–1.18])[5]. The R^2 and the D' between the variants are with respectively 0.00092 and 0.18 low and the association P-value of the previously described variant rs56116661 in UC patients vs. controls is 0.0067 (OR = 0.82 [0.72–0.95]). Thirdly, the gene *CTCF* (16q22.1, rs117327757, chr16:67627037, P-value = 8.6×10^{-6}, OR = 1.8 [1.39-2.33], **Supplementary Figure 18**) is known to influence the expression of TNF[57]. Fourthly, *LYPD5* (19q13.31, rs364691, chr19:43804850, P-value = 1.6e10-6, OR = 1.35 [1.2-1.53]) was reported by Taman *et al.* as upregulated in treatment-naïve UC patients[58].

Associations in the whole exome data

The genomic inflation factor lambda for the exome data was determined to be 1.005 (**Supplementary Figure 22**). Overall, five loci showed at least suggestive associations (**Supplementary Figure 23-28, Supplementary Table 3**). In the whole exome data only the HLA-II region was found as a genome-wide significant signal with LD support (**Supplementary Figure 25**). The intronic *PGAM5* variant rs7973452 at 12p13.2 is LD-supported and reaches suggestive significance (P-value=8.47×10^{-6}; OR=1.34 [1.18-1.53], **Supplementary Figure 27**). PGAM5 is known to regulate antiviral responses[59]. Three additional loci reached suggestive significance, none of these were supported by variants in LD, even when including low frequency variants. One of those is the 1p36.13 locus with *OTUD3,* also identified in the imputed genotyping dataset, with the intronic variant rs773646156 (chr1:19890367; P-value=4.53×10^{-7}; OR=1.33 [1.19–1.48], **Supplementary Figure 24**). Second, the bitter taste receptor gene *TAS2R43*, with the missense variant rs200922417 (chr12:11092088; P-value=6.17×10^{-6}; OR=1.92 [1.45–2.54]; 48L>48V, **Supplementary Figure 26**), is a suggested target for UC treatment as it influences *CLCA1*[60]. This variant is not well covered in the imputed genotyping data and the gene was not distinguished from *TAS2R45* in genome build GRCh37.

The fifth association in the exome data, the third without LD support, is the synonymous variant rs755163625 on 22q11.21 in the gene *SCARF2* (chr22:20429374; P-value=9.2×10^{-6}; OR=2.37 [1.62-3.47], **Supplementary Figure 28**). Coding mutations within *SCARF2* were previously described as responsible for the Van Den Ende-Gupta Syndrome, an extremely rare autosomal-recessive disorder characterized by distinctive craniofacial features[61].

SNX20/NOD2 association in ulcerative colitis and Crohn's disease

NOD2, also formerly known as *IBD1*[4] in linkage studies, was previously described to be a CD-specific disease gene. Here, we identified an association with UC. To investigate this further, we performed additional lookups and calculated the genetic association for this region also for the available CD data (4,097 CD cases and 4,185 controls as in UC analysis).

Our main UC association in the genotyping dataset rs139397276 (P-value=1.3×10^{-6}; OR=3.74 [2.19–6.37]; $MAF_{controls}$=0.9%; MAF_{cases}=2.4%) was located at chr16:50666737 (**Supplementary Figure 17 and 29**). In addition, an association was identified with rs61736932 in the exome dataset (OR=4.22 [2.36, 7.54], P-value=1.15×10^{-6}, MAF=1%, note: MAF is with 0.00990 slightly below 1% therefore it was not described above) at chr16:50711744. Both rs139397276 and rs61736932 are located closer to the centromere as compared to the established CD susceptibility variants rs2066844, rs2066845, and rs2066847 (all of them have an R^2=0 with the lead variant rs139397276). The variant rs139397276 is in the *SNX20* gene and the calculated credible set includes four additional variants with a probability above 1%, including the *NOD2* exome variant rs61736932 (OR=4.47 [2.43, 8.25], P-value=1.60×10^{-6}, MAF=1%). Of the remaining three, two are in introns of the *NOD2* genes and one in an intron of *SNX20.* All five variants reach nominal significance in the association analysis (**Supplementary Table 3**, **Supplementary Figure 29**) and have a frequency around 1%. None of the variants changes the amino acid sequence of the NOD2 protein. While the previously described CD variants have an impact on the leucine rich repeat region (LRR) of the NOD2 protein, our variants are in the nucleotide binding region and in the 3'-UTR region of *SNX20* (**Supplementary Figure 29**). The IIBDGC dataset supported our finding of rs139397276 with a P-value of 2.6×10^{-4} and an OR of 1.36 [1.15-1.60]. We further queried the data of Lesage, 2002[62] and of the IBD Exomes Browser[63] for validation purposes. The data on the IBD Exomes Browser reports an overall P-value of 0.023 and an OR of 1.31 [1.07, 1.60] for rs61736932 in UC versus healthy controls. In the Non-Finnish European batch, the P-value is 6.78×10^{-4} with an OR of 1.93. Lesage reported with 1.9% (6 of 318) in UC patients a lower MAF than in their controls with 2.4% (5 of 206). Therefore, the data of Lesage and colleagues does not support the association we identified.

Gene-based analysis

The gene-based analysis using SAIGE-GENE[35] on the exome data with an AF<1% resulted in no significant gene if correcting for the number of analyzed genes (P-value < 2.61×10^{-6}). We here describe the three genes with the lowest P-values. Those genes are *NOD2*, *KCNK3* and *PSORS1C1*.

NOD2 was also identified in the single variant association of the imputed data. The same variant rs61736932 is responsible for the association signal of the gene-based test (P-value$_{SKAT-O}$ = 3.47×10^{-5}, P-value$_{SKAT}$ = 1.21×10^{-5}, P-value$_{Burden}$ = 0.915). The variant was not identified in the single variant analysis as the allele frequency in the exome data was slightly below 1%. The Burden test result, with very low significance, suggests that the other rare variants have different directions of effect. Of the 13 included variants with an allele frequency between 0.1% and 1% only three other variants were more common in patients than in controls.

The gene *KCNK3* was previously not associated with IBD, of the three performed gene-based tests SKAT has the lowest P-value (P-value$_{SKAT-O}$ = 1.10×10^{-4}, P-value$_{SKAT}$ = 5.17×10^{-5}, P-value$_{Burden}$ = 3.96×10^{-2}). The signal is mainly based on the intronic variant rs926416351 with an allele frequency of 0.0059 (P-value = 4.07×10^{-6}). In the very similar gene *KCNK9* a copy number variation (CNV) associated with UC was identified by Saadati *et al.*[64]. A recent study tested the effect of *KCNK9* and *KCNK3* knockout in a mouse model[65]. They figured out that the absence of one gene increases the expression of the other gene. A *KCNK3* knockout lead to a beneficial outcome in DSS-induced colitis with less mitochondrial damage and apoptosis, while the increased expression of *KCNK3* did not prevent apoptosis after DSS exposure.

The third gene *PSORS1C1* (P-value$_{SKAT-O}$ = 3.04×10^{-5}, P-value$_{SKAT}$ = 4.27×10^{-5}, P-value$_{Burden}$ = 9.49×10^{-5}) is located close to the HLA class I region. The two most common variants included in the gene-based analysis (rs118016068 and rs117114042) were also significantly associated in the publicly available RICOPILI[30] dataset (see also **Supplementary Table 3**).

Power analysis of genome wide associations in ulcerative colitis

Other well-known loci associated with UC like *IL23R*, *LINC02132* and *IL10* failed to show nominal significance but show the trend as reported in previous GWAS studies. The expected statistical power to reproduce those loci with a P-value of $<10^{-5}$ is below 0.75 (**Supplementary Figure 30**). All previously reported loci with a greater suggested power were only reported in studies with samples of Asian populations and therefore could be population-specific findings.

HLA association in UC

We imputed HLA alleles and amino acids for HLA-*A, -B*, *-C*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, *-DRB1*, *-DRB3*, *-DRB4* and *-DRB5* at full 2-field level. Overall, we observed 234 different 2-field alleles in our UC and control data set. For the single loci, between 6 (HLA-*DPA1*) and 62

(HLA-*B*) different alleles were imputed into our samples. In consistency with previous studies of the HLA in UC[9,10] the main association signal was in the locus containing HLA-*DR* and HLA-*DQ* (**Supplementary Figure 31**). Overall, four 2-field alleles were associated at genome-wide significance (*DQB1**06:02 (P-value=2.68×10^{-10}, OR=1.58 [1.37-1.82]), *DRB1**15:01 (OR=1.56 [1.36-1.80], P-value=6.61×10^{-10}), *DRB5**01:01 (OR=1.56 [1.35-1.80], P-value=8.40×10^{-10}), *DRB4**01:03 (OR=0.66 [0.57-0.76], P-value=6.80×10^{-9})), and 3 additional 2-field alleles were nominally significant (5×10^{-8} < P-value < 1×10^{-5}: *C**12:02, *B**52:01, *DQA1**01:02) (**Supplementary Table 4**). The strongest association from a SNP data was observed for the intronic SNP rs6927022 located in HLA-*DQA1* (P-value=2.20×10^{-14}, OR=0.64 [0.58-0.72]). As discussed also previously by Degenhardt *et al.*[10], *DQB1**06:02 is located on the same haplotype as *DRB1**15:01 (**Figure 2**). In our dataset, HLA-*DRB1**15 is the most significantly associated 1-field allele (OR=1.61 [1.40-1.85], P-value=2.07×10^{-11}). Overall, the observed association of HLA alleles is in line with previous studies (**Figure 2**).

The main associated signal described by Goyette *et al.*[9] HLA-*DRB1**01:03 is not nominally significant (cutoff P-value<1×10^{-5}) in our data set due its low frequency in our data, even though its effect size is even larger in our German cohort (OR=5.31 [1.85-15.25], P-value=1.95×10^{-3}, MAF_{cases}=0.41%, $MAF_{controls}$=0.96%) than reported by Goyette *et al.*[9].

Of the 58 alleles previously reported to have a nominally significant association with UC either within the European dataset from Degenhardt *et al.*[10] or the results from Goyette *et al.*[9] only 6 showed an opposite direction of effect in our data. (**Supplementary Table 4** and **Figure 2** and **Supplementary Figure 32**). *DQA1**05:01 showed a risk effect in Goyette *et al.*[9]. However, we suspect that Goyette *et al.* named *DQA1**05:05 as *DQA1**05:01 since both belong to the same g-group of alleles and since the allele *DQA1**05:05 is not reported by Goyette, and the allele frequencies of *DQA1**05:05 (MAF=14.7%) and *DQA1**05:01 (MAF=11.9%) combined are closer to the allele frequency reported by Goyette *et al.* (MAF_{cases}=28.0%, $MAF_{controls}$=25.3%). Further, the associations of the alleles *A**29:02, *C**01:02, *DRB1**13:01, *DQA1**01:03 and *DQB1**06:03 were not supported in our dataset since all of them have an allele frequency below 10% in all cohorts and the lowest P-value in our data is with 0.28 for HLA-*DRB1**13:01 far from significant. Moreover, the three class II genes are located on the same haplotype (**Figure 2**).

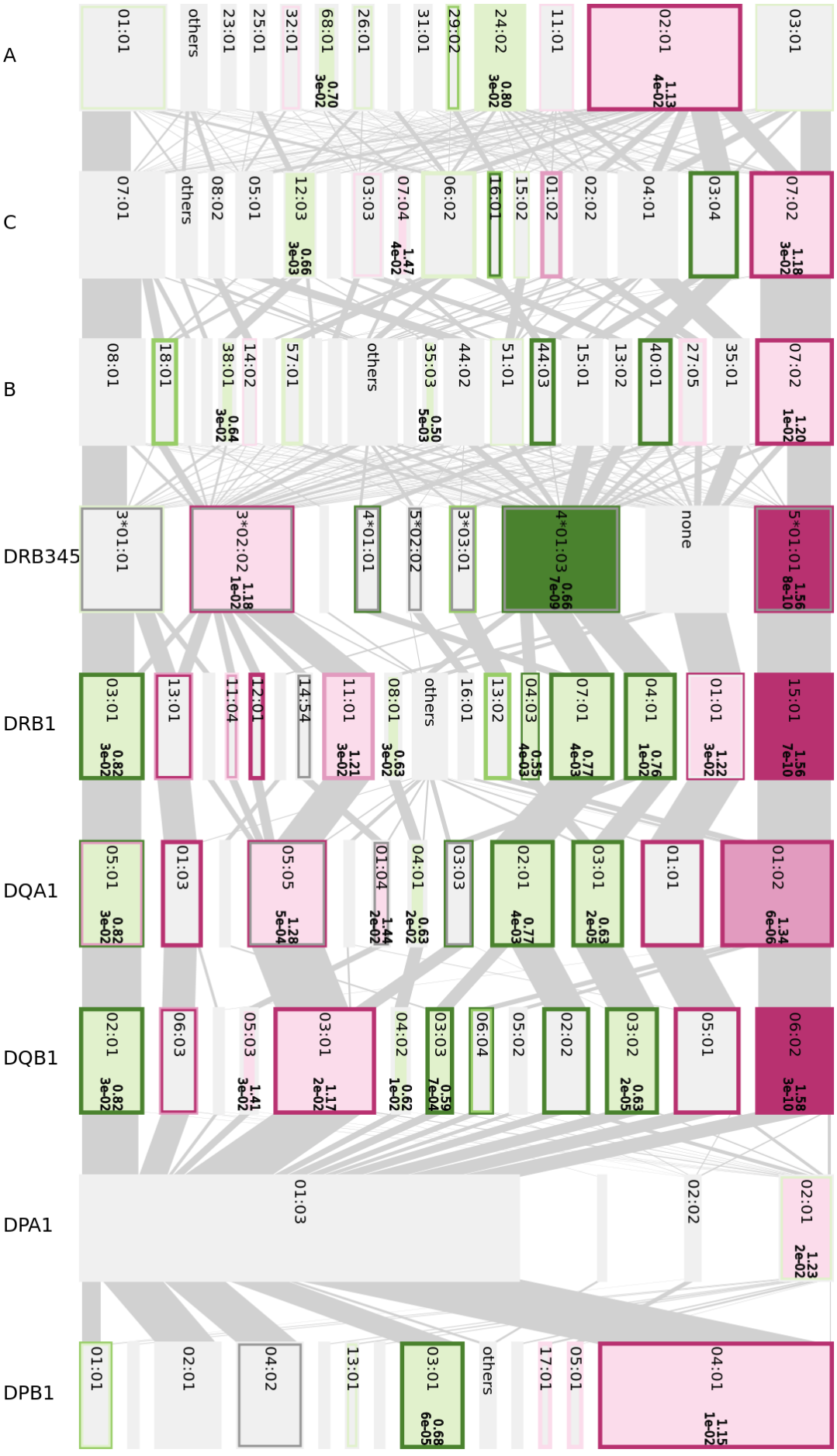

A
C
B
DRB345
DRB1
DQA1
DQB1
DPA1
DPB1
01:01
others
23:01
25:01
32:01
68:01 0.70 3e-02
26:01
31:01
29:02
24:02 0.80 3e-02
11:01
02:01 1.13 4e-02
03:01
07:01
others
08:02
05:01
12:03 0.66 3e-03
03:03
07:04 1.47 4e-02
06:02
16:01
15:02
01:02
02:02
04:01
03:04
07:02 1.18 3e-02
08:01
18:01
14:02
38:01 0.64 3e-02
57:01
others
35:03 0.50 5e-03
44:02
51:01
44:03
15:01
13:02
40:01
27:05
35:01
07:02 1.20 1e-02
3*01:01
3*02:02 1.18 1e-02
4*01:01
5*02:02
3*03:01
4*01:03 0.66 7e-09
none
5*01:01 1.56 8e-10
03:01 0.82 3e-02
13:01
12:01
11:04
14:54
11:01 1.21 3e-02
08:01 0.63 3e-02
others
16:01
13:02
04:03 0.55 4e-03
07:01 0.77 4e-03
04:01 0.76 1e-02
01:01 1.22 3e-02
15:01 1.56 7e-10
05:01 0.82 3e-02
01:03
05:05 1.28 5e-04
01:04 1.44 2e-02
04:01 0.63 2e-02
03:03
02:01 0.77 4e-03
03:01 0.63 2e-05
01:01
01:02 1.34 6e-06
02:01 0.82 3e-02
06:03
05:03 1.41 3e-02
03:01 1.17 2e-02
04:02 0.62 1e-02
03:03 0.59 7e-04
06:04
05:02
02:02
03:02 0.63 2e-05
05:01
06:02 1.58 3e-10
01:03
02:02
02:01 1.23 2e-02
01:01
02:01
04:02
13:01
03:01 0.68 6e-05
others
17:01
05:01
04:01 1.15 1e-02

Figure 2: Disentangler[42] plot that summarizes the HLA haplotype structure in our UC cases and controls. Risk alleles are colored in red (10^{-5}<P-value<=0.05: light-red; 5×10^{-8}<= P-value<10^{-5}: mid-red; P-value<=5×10^{-8}: dark-red), protective alleles in green (10^{-5}<P-value<=0.05 light-green; 5×10^{-8}<= P-value<10^{-5}: mid-green; P-value<=5×10^{-8}: dark-green), dark-grey missing data, light-gray alleles with no effect (P-value>0.05). The inner color of each box is based on our dataset, the inner frame of each box represents the results from Goyette et al.[9] and the outer frame represents the European data from Degenhardt *et al.*[10]. (Note: We suspect that Goyette *et al.* did not distinguish between *DQA1**05:01 and *DQA1**05:05.) For genome-wide-significant alleles the OR and P-value is noted next to the allele names in bold type. The height of the box illustrates the allele frequency. Our results are concordant with the previous analyses by Degenhardt *et al.* and Goyette *et al.* as colors for each box are in most instances the same.

As different alleles share attributes with each other on the protein level based on the amino acid sequence, we further analyzed the data for specificities in the amino acid sequence (**Supplementary Table 4**). Glutamic acid (D) at position 175 of DQA1 is the associated amino acid with the lowest P-value of 3.92×10^{-12} and an OR of 0.64 [0.56-0.72], the amino acid is only present in alleles with a protective direction of effect (e.g., *DQA1**03:01, *DQA1**02:01 and *DQA1**03:02). Alternatively, a Lysine (L) can be present at amino acid position 175, present mainly in the *DQA1**05 alleles, with different directions of effect or glutamine (Q) present mainly in the risk associated alleles and therefore also nominal significant associated (P-value of 3.92×10^{-12} and an OR of 0.64 [0.56-0.72]). The variation is based on rs2308891 (chr6:32642232). The amino acid is in the α2-domain of the protein.

Additionally, nucleotides and amino acids are stronger associated as the single alleles in our data set. The multiallelic variant rs9269955 (chr6:32584361) influences the amino acid 11 of the *DRB1* allele with the nucleotide G as the strongest associated (P-value of 5.08×10^{-12} and an OR of 1.50 [1.33-1.68], MAF=0.2920). Depending on the rs17878703 (chr6:32584360) characteristic either an allele of the *DRB1**15 and *DRB1**16 group with a proline (P) in position 11 (P-value of 1.99×10^{-10} and an OR of 1.54 [1.35-1.76]) or the *DRB1**01 group is present for this risk variant. This amino acid is involved in the interaction with the peptide in binding pocket 6 (peptide-HLA interaction as previously identified in Degenhardt *et al.* 2021[10]).

Other nucleotides and amino acids playing an important role in HLA-peptide interaction with genome-wide significant UC association are at amino acid positions 13 and 71 of the DRB1 protein, and amino acid position 86 of DQB1. All genome-wide significantly associated amino acids and nucleotides are listed in **Supplementary Table 4**.

Binding motifs of associated HLA alleles

To figure out the similarity between the different UC-associated HLA alleles we generated a dendrogram based on the predicted binding peptides. The dendrogram of the HLA-DR-peptide interaction shows three groups of protective and risk alleles each (**Supplementary Figure 32**). The biggest cluster is the HLA-*DRB1**04 group with the main important alleles

*DRB1**04:01 and *DRB1**04:03 but also all other *DRB1**04 alleles have at least a tendency to be protective if reported and except *DRB1**04:02 all cluster closely together. The Second protective cluster combines *DRB1**07:01 and *DRB1**09:01. Both are the only representatives of their 1-field allele group in all three datasets shown. *DRB1**04, *DRB1**07 and *DRB1**09 are the only alleles occurring together with the pseudogenes *DRB7* and *DRB8* and in most cases the protein coding *DRB4* gene. Therefore, this signal might be either based on the similarity between the alleles as all of them evolved most probably from the same ancestor[66] or because of another locus that is in LD, e.g., especially the *DRB4* locus. The third protective cluster includes all present *DRB1**03 alleles (*DRB1**03:01 and *DRB1**03:02).

The risk alleles also cluster based on the 1-field classification in a cluster with *DRB1**01 (present as *DRB1**01:01 and *01:03) and *DRB1**15 (present as *DRB1**15:01, *15:02, *15:03 and only single cases of *15:06), additional *DRB1**12:01 was identified as risk allele but in this case the only other allele of the same serogroup *DRB1**12:02 does not support the same direction of effect but the allele is only present in very low frequencies and so far only Goyette *et al.* reached an association P-value below 0.05 for this allele.

For a closer look into the binding specificities, we generated logos for the genome-wide associated alleles (**Figure 3**). The separate clusters are characterized by individual binding characteristics. *DRB1**03:01 is especially characterized by an enrichment of acidic amino acids in binding pocket 4 and basic anchors in pocket 6 and 9. The *DRB1**04 alleles *DRB1**04:01, *04:03, and *04:04 are characterized by an antigen with a polar or acidic amino acid in position 6 with a residue with one or two carbon atoms. Similar characteristics are also present in anchor position 9 but here the size restriction is not as stringent and additional alanine (A) as a hydrophobic amino acid is present. *DRB1**07:01 and *DRB1**09:01 do not show any enrichment of acidic amino acids but in comparison to the risk alleles (**Supplementary Fig. 33**) they show an enrichment for the polar amino acids serine (S) and threonine (T) in position 4.

The risk alleles *DRB1**01:01 and *DRB1**01:03 are characterized by a small residue in binding pocket 6 (alanine, glycine, and serine) this can be explained by the high-volume amino acids in position 11 (L) and 13 (F) of the beta chain. *DRB1**12:01 peptides are characterized by a hydrophobic amino acid in pocket 4 (leucine or isoleucine) and a big uncharged amino acid in pocket 9 (tyrosine, phenylalanine, leucine, valine), one factor for this characteristic is that the allele has the smallest residue at position 9 (E) and smaller residues at position 39 and 57. The alleles *DRB1**15:01 and *DRB1**15:02 are mainly characterized by preferring aromatic amino acids in pocket 4 (tyrosine, phenylalanine, tryptophan).

Overall, the logo of the *DQ* alleles (**Supplementary Figure 34**) are not as well defined as the motifs of the *DR* alleles. Here, an acidic anchor occurs in position 6 of the risk *DQ* haplotypes

*DQA1**01:01-*DQB1**05:01 and *DQA1**01:02-*DQB1**05:01 and an additional aromatic anchor in position 4 is present. Further, proline plays a more important role in binding peptides, but not in a disease-associated manner.

In summary, a single consistent binding motif for all risk or protective associated alleles cannot be defined. However, the protective alleles do tend to bind more often acidic amino acids, while the risk alleles are characterized by aromatic amino acids based on the binding predictions.

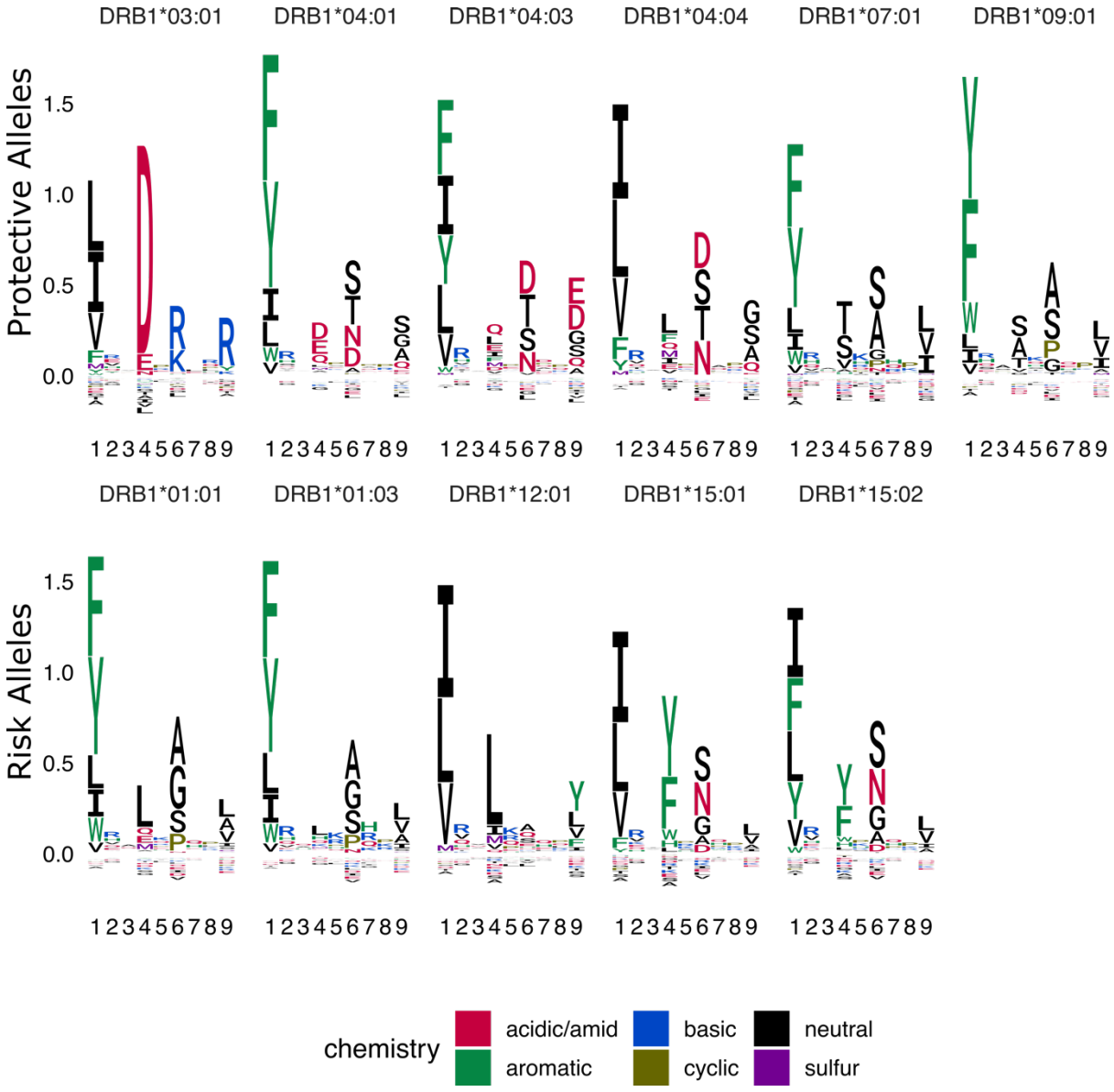


Figure 3: Binding logo plot of associated HLA-DR alleles. The upper row represents the protective associated alleles, the bottom line the logos of the risk alleles. The motifs are based on the NetMHCIIpan-4.0 predictions of the binding cores of all peptides at least annotated as weak binders. The single letters represent the one letter amino acid code colored by the chemical properties of the amino acids.

PepWAS analysis

We performed a PepWAS analysis to investigate which self-peptides are differentially recognized by the UC risk and protective HLA-DRB1 proteins. For computational reasons we limited the analysis to UC susceptibility genes from previous studies and identified within this study testing the hypothesis that chronic inflammation in UC patients is driven by self-peptides encoded by the patient's genome and that are presented by their own HLA molecules. Coding mutations in the genes encoding the peptides may additionally lead to the presence or absence of specific peptides in some patients and controls.

The PepWAS resulted in 234 significant peptides after multiple hypothesis correction (P-value<3.73×10^{-6} based on 13,411 peptides) for HLA-*DRB1* (**Supplementary Figure 35**, **Supplementary Table 5**). Those peptides originate from 46 of the 76 considered GWAS genes (in 155 transcripts). All peptides are specific to one of the genes. Up to 18 peptides identified through the PepWAS are annotated to one single gene. The number of peptides identified within a protein correlates with the length of the amino acid sequence (Pearson's product-moment correlation: R^2=0.27, P-value=1.02×10^{-6}), especially when considering 2 peptides separately, if they have no overlap of 9 amino acids when shifted against each other (Pearson's product-moment correlation: R^2=0.44, P-value=4.71×10^{-11}).

Eight of the proteins containing PepWAS hits are described in the meta-analysis from Linggi *et al.*[50] or in the analysis of treatment naïve patients by Taman *et al.*[51] as differentially expressed in ulcerative colitis patients. Whereof *JAK2*, *KIF21B*, *FCGR2A*, *NOD2,* and the HLA genes *DRB1* and *DQA1* are upregulated and *NXPE1* and *HSD11B2* are downregulated (**Supplementary Table 5**). Of those genes *KIF21B* and *HSD11B2* are neither reported as membrane proteins, nor as extracellular. Peptides of the genes *JAK2, FCGR2A* and *HLA-DRB1* were extracted with the peptidomes described in ElAbd *et al.*[53]. The only hit from the PepWAS analysis having at least nine amino acids overlap to the proteomes is the peptide RRVQPKVTVYPSKTS (P-value=1.52×10^{-6}). This peptide is present in the sequence of alleles of the HLA-*DRB1**15, -DRB1*16, and -*DRB1**10 groups and of HLA-DRB4 alleles. The peptide was identified previously in 5 of 6 samples with a *DRB1**15:01 allele using antibody-based HLA-pulldown and LC-MS analysis[53]. The other sample has a genotype with two *DRB4* alleles, regarding the prediction the genotype of this patient (*DRB1**07:01/*DRB1**09:01) would not present this peptide. In the 9 other samples with a genotype with a DRB4 molecule but no *DRB1**15 allele (one case with 2 haplotypes with *DRB4*), the peptide was not found. Side note: Interestingly, all haplotypes with a copy of the *DRB4* gene (*DRB1**04, *DRB1**07 and *DRB1**09) are associated with a protective effect.

62 mutations are essential to generate one or more of the identified peptides, but thereof 40 would prevent the generation of another identified peptide. In comparison to the 22 missense mutations that lead to the generation of a peptide identified as associated in the PepWAS

analysis, there are 156 missense mutations that lead to a peptide not being identified as significant in the PepWAS.

All PepWAS hits were predicted to bind HLA-DRB1*15:01 and the rare HLA-DRB1*15:06 protein. In most cases also the other tested DRB1*15 alleles (DRB1*15:02 and DRB1*15:03) were predicted to bind. Other alleles binding several hits are DRB1*16:01 and DRB1*16:02, the other alleles of the DR2 serotype and DRB1*01:03. Interestingly no peptide identified in the PepWAS was predicted to bind with DRB1*01:01, even though the allele shows a trend of being a risk factor.

The peptide hits are enriched for isoleucine, asparagine, valine, and tyrosine and have lower frequencies of cysteine, glutamic acid, glycine, and arginine compared to our candidate transcripts.

121 of the peptide hits were present in all individuals within our sample set, 133 are peptides of the reference proteome, of which 19 are not present in all individuals. The remaining 101 peptides require at least one mutation from the reference genome to be encoded (**Supplementary Table 5**). None of the mutations shaping the peptides was statistically significant associated with the disease after correcting for multiple testing and no peptide identified in the PepWAS analysis showed significantly different frequencies in cases and controls.

[Discussion]

Different hypotheses on the underlying processes leading to inflammation in UC including the role of the HLA haplotype are discussed in the literature[8]. Here, we focused on the autoimmune hypothesis in association with autoantigens, excluding microbial peptide candidates from the analysis, because of the excessively larger search space. However, the binding preferences of the HLA alleles are general attributes and therefore parts of our results (and the analysis concept in general) can also be transferred to microbial candidates. While most of the identified genetic associations are not related to HLA antigen recognition, the consistent and strong association of HLA suggests that HLA peptide interaction plays an important role.

Our GWAS results are overall in line with previous studies. The previously not described genome-wide significant hit at 5p14.3 is either specific for our German subpopulation or represents a false positive signal. This intergenic association is well supported by variants in LD but could not be validated in the larger IIBDGC dataset. In addition, no biologically or disease relevant conclusions could be made.

The *NOD2* gene harbors the most prominent genetic associations known for CD. It has also been studied previously in the context of UC[62], however, no clear associations for UC have been described before[7]. The herein determined novel association of rs61736932 with UC is supported by two large publicly available data sets[30,63]. Interestingly, the identified *NOD2* UC variants are in another location of the protein not in LD to those variants described for CD, suggesting a different functional role in UC.

The suggestively protective variant we identified in the *RBM19* gene, rs3782449, and the previously described risk variant with an association in ocular manifestation in IBD rs4766697[55], are located on different haplotypes. The direction of effect for rs4766697 in our data is protective as well, even though far from statistical significance. To our knowledge, there are no other studies linking *RBM19* to IBD, therefore additional studies are needed to clarify the exact role of this candidate gene in UC disease etiology.

Previous GWAS already yielded over 240 genetic variants significantly associated with IBD[67], but still those findings only cover a part of the heritability[68]. Our study cohort was not sufficiently powered to identify low-frequency variants and variants with small effect sizes. However, compared to previous GWAS, we used the genome build GRCh38 (others mainly used build GRCh37) and we employed state-of-the-art imputation reference and exome data sets. This enabled us to analyze previously not, or not properly, covered genetic regions. For example, the signal at 9q22.2 (*UNQ6494*) was not covered in GRCh37, which may explain why no former study identified this locus as associated with IBD. Nevertheless, an independent replication is necessary to validate this signal. Furthermore, for the bitter taste receptor gene *TAS2R43*, where we identified a suggestive hit without LD support in the exome data, the genetic architecture changed between the genome builds GRCh37 and GRCh38 and is still not well covered by imputation. This explains why a potential identification of disease-associated markers at this locus was nearly impossible beforehand but might be improved by further analyses of diverse whole genome sequencing data. It further shows the benefit of improving imputation reference panels and methods.
We identified two suggestive hits located within genes involved in the TNF pathway (*TNFRSF8* and *CTCF)*. TNF levels are typically increased in IBD patients and anti-TNF is a common treatment option for IBD even though the biological role of TNF is much more complex and treatment not effective in all patients[69]. Both signals were not validated in publicly available data sets.

The genetic risk profile of the HLA region in our data is consistent with UC associations described in the literature for Caucasian populations[9]. As shown by Degenhardt *et al.*[10], this

also holds true for different ethnicities in UC, even though some HLA alleles are not observed in Caucasian populations or more frequent in the non-Caucasian population.

The main genetic association signal for UC is located within the HLA-*DR* and HLA-*DQ* genes. Causality of either locus for UC cannot be shown with our data due to the strong LD between the loci. Mechanistic follow-up studies are clearly needed to disentangle this signal. On the computational level, Goyette *et al.*[9] used a gene-based analysis and identified HLA-*DR* as the most likely disease-relevant candidate gene. However, as shown in **Figure 2**, and discussed in Degenhardt *et al.*, HLA-*DQ* cannot be fully ruled out, due to the very strong linkage disequilibrium with HLA-*DRB1*.

The PepWAS analysis ranks peptides based on the HLA profile of the disease-associated HLA alleles and enables the identification of disease-relevant binding motifs. However, the PepWAS approach cannot differentiate disease-relevant peptides from other similar peptides without a prefiltering of peptides of interest, due to peptide similarity among different proteins, statistical restrictions, and the limitation in the specificity of the HLA profile. In our analysis, 46 of the 76 GWAS candidate genes contained PepWAS hits. Longer protein sequences lead to a higher chance of identifying PepWAS hits within this sequence. This influence factor might be increased by the in-silico approach of defining the peptides by a sliding window. All candidates identified by PepWAS are DRB1*15:01 binders, this is associated to the fact that the allele has the largest power due to its frequency and comparably strong effect size. Alternative peptides would need to be predicted as binders for more than one serotype with the same direction of effect to be significant in the analysis.

The only peptide identified by PepWAS that was also present in the 25 immunopeptidomes described in ElAbd *et al.*[53] is a peptide present in DRB1*15. It needs to be considered that the analysis of the immunopeptidomes was based on a reference dataset including only one sequence for each gene, and in case of *DRB1,* the sequence of this gene was DRB1*15:03. But the peptide in the peptidome is the only PepWAS hit in *HLA-DRB1*. The PepWAS was based on the personalized peptidomes of all patients and therefore included sequences of the different alleles in the dataset. Whether this peptide is a strong candidate to play a role in the pathogenesis of UC remains to be elucidated. On the one hand, this peptide is presented especially strong by DRB1*15 proteins, which are related to a higher risk. On the other hand, the peptide is present in the sequence of the protective haplotypes including a *DRB4* allele, and as the peptidomics data show that the peptide can be presented when carrying those haplotypes, even though reduced in comparison to individuals carrying a DRB1*15 allele. If the presentation of this peptide would play a significant role in modulating the disease, a strong

protective effect would be expected e.g., for DRB1*08 alleles, where the peptide is not strongly bound nor present in this haplotype. However, such an effect could not be observed.

The peptide sets predicted to bind the risk alleles are characterized by aromatic amino acids in pocket 4, while the peptides binding the protective alleles are characterized by acidic amino acids in pocket 4. However, both characteristics do not apply to all the significantly associated alleles and no peptide identified in the PepWAS analysis is encoded with significantly different frequencies between our UC patient and control exome data.

Peptides containing mutation sites are less likely to be identified by our PepWAS analysis. One explanation for this might be a bias based on the training dataset used for the prediction algorithm. As the peptidome used for NetMHCIIpan-4.0[46] training was derived mainly from mass-spectrometry data using the human reference proteome to define the peptides, and while the negative peptides were defined by sampling from the UniProt database, the binding peptides in the prediction might be biased towards the human reference proteome. Besides, the version 4.0 of NetMHCIIpan[46] is based on immunopeptidome data and therefore the training data is expected to be closer to the *in vivo* situation than the previous versions that were based on peptide competition assays. Still, the prediction of HLA-peptide interaction does not cover the T cell specificity and therefore presents only a necessary but not complete part of an HLA-induced immune reaction.

In summary, we employed a large exome dataset from UC patients and newly available reference datasets to obtain further insights into the role of the HLA in UC disease etiology. The initial association analysis identified promising genetic signals in *NOD2*, *RBM19*, *TAS2R43*, as well as at the intergenic loci 5p14.3 and 9q22.2. Further, additional suggestive hits related to the TNF pathway were identified. An additional replication of those associations is highly recommended but as most of them either have relatively low frequencies or are not well covered in older genome references, a replication within a large new dataset is necessary. A gene-based analysis on the rare variants revealed by exome-sequencing did not result in any significant results, but as for the three genes with the lowest P-value, an already described connection to the disease could be drawn, it is expected that a more powerful analysis including more samples would result in significant and relevant findings.
Our analysis of the HLA showed again a stable association of UC with multiple HLA alleles. Here, we were also able to highlight some characteristics of the peptides interacting with the HLA even though no specific autoimmune peptide candidates could be identified, where a mutation impacts the peptidome in a disease-specific way. This supports either the hypothesis

of an external/environmental origin of a pathogenic peptide, for example from the microbiome or the diet, or an autoimmune interaction independent from non-synonymous mutations.

[Materials and Methods]

Cohort description

In total 15,877 individuals from studies across Northern Germany, with mainly chronic inflammatory traits (arthritic dermatitis, Crohn's disease (CD), longevity, primary sclerosing cholangitis, psoriasis, sarcoidosis, and ulcerative colitis (UC)) and 4,680 population controls with unknown phenotypes were genotyped and submitted to joint genotyping quality control as described in the section *Array-based genotyping, quality control and imputation* as a resource for the genetic analysis of inflammatory diseases. For reasons of feasibility, this study focuses on the analysis of UC only, since UC exhibits the strongest HLA associations. For this purpose, we included 863 ulcerative colitis patients and 4,185 population controls from the total available quality-controlled genotype cohort. The population control is comprised of 969 individuals recruited within the German Food Chain Plus (FoCus) cohort, previously described in Barbaresko *et al.*, 2020[12] and 3,216 German healthy blood donors as used and described in Degenhardt *et al.* 2022[13]. Of the 863 UC patient samples 424 were previously described by Franke *et al.*, 2008[4] and 439 by Bokemeyer *et al.,* 2016[14].
In total 5,048 individuals (4,185 controls, 863 UC patients) were used for analysis. A detailed overview of sample numbers is shown in **Supplementary Table 1**.

Ethics approval
The study was conducted according to the guidelines laid down in the Declaration of Helsinki and was approved by the Ethic Committee of the Medical Faculty of the University of Kiel (Germany). All participants gave written informed consent, and the recruitment protocols were approved by the ethics committees at the respective recruiting institutions. The following approvals of the project were obtained from the ethics committees: BioColitis samples (D 474/12)[14], German blood donors (A 103/14)[13] and the samples of the FoCus Cohort as well as the samples described in Franke *et al.*, 2008[4] (A 156/03).

Sample preparation/processing
DNA was extracted by the DNA laboratory of the Institute of Clinical Molecular Biology (Kiel University, Kiel, Germany) from whole blood. For a detailed description of the further extraction protocol, we refer to Ellinghaus *et al.*[15].

Array-based genotyping, quality control and imputation

Genotyping was performed using the Global Screening Array (GSA), version 1.0, containing 700,078 variants pre-quality control at the Regeneron Genetics Center, U.S.A. The data were called on genome build GRCh37 (using the cluster file GSAMD-24v1-0-A_4349HNR_Samples.egt). Genotype quality control (QC) was performed as implemented by BigWAS[16]. Briefly, the variants are annotated in a standard way based on a database and filtered on missingness (≥ 0.02 in a single batch or ≥ 0.10 in all batches), and the Hardy-Weinberg equilibrium (false discovery rate (FDR) threshold of 10^{-5} in controls). Then samples with high missingness (≥ 0.02), increased or decreased heterozygosity rates (± 5 standard deviation), relatedness testing (identity by descent ≥ 0.1875) and based on the population structure identified by principal component analysis (PCA) (outside the ± 5*IQR in PC1 and PC2) are removed. Of the final sample set, additional variants are filtered for differential missingness between controls and diseased samples and monomorphic sites.[16] After QC, a total of 5,048 individuals (863 cases/ 4,185 controls) and 579,352 variants remained for analysis (**Supplementary Table 1**). Only variants mapping to the autosomes were used here for association analysis. All genomic positions were lifted to genome build GRCh38 for further analysis on the TOPMed Imputation Server or for the genotyped GSA data within the BigWAS[16] pipeline using the UCSC liftOver tool[17]. To increase the genotyping coverage, SNP imputation was performed using the TOPMed Imputation Server from the NIH using the TOPMed Imputation diverse reference panel (version TOPMed-r2@1.0.0) including 97,256 deeply sequenced human genomes with a post-imputation quality score of R^2 set to 0.1[18–20]. In total 90,406,930 variants had an R^2 larger than or equal to 0.1. In the following analysis 13,780,246 variants with an $R^2 > 0.6$ and a MAF above 1% were analyzed if not noted otherwise.

Exome sequencing, genotyping and quality control

Whole exome sequencing (WES) was performed at the Regeneron Genetics Center, U.S.A. Sample preparation and sequencing as well as the sequence alignment, variant identification and genotype assignment were done as described in Van Hout *et al.* 2020[21]. An extended quality control was conducted with the 'Goldilocks' (GL) filters[21] and additional filters for genotypes, variants and samples. In brief genotypes were set to "no call" based on the WeCall filters allele bias (ABPV <0.009), allele and strand bias (ABPV + SBPV <0.07), bad reads (BR <15), low quality (LQ <10), low mapping quality (MQ <40), quality over depth (QD <15) and strand bias (SBPV <0.01) and further, based on the sequencing depth (DP <7), genotyping quality (GQ >14) or allele balance (AB <0.25). Variants were removed if no reliable sample (AB ≥0.15 or homozygous) remained. For insertions and deletions (InDel) the same filtering was used with different parameters: genotypes with a DP <10 were removed and a reliable

sample to keep a variant was defined by an AB ≥0.20 or homozygosity.[21] Further, single nucleotide variants (SNVs) were removed if the missingness was above 10% (variants that overlapped with an InDel were not considered for missingness filtering) or if the Hardy-Weinberg-Equilibrium P-value in control individuals was below 10^{-5}.

Individuals were removed if they presented with an unusual rate (> 6 standard deviation difference to the mean) of (a) singletons (more than 737) or (b) missingness (equaling a missingness cutoff of 0.21), or (c) heterozygous to homozygous ratio (0.0058 < hethom < 0.0077) or (d) transition/transversion ratio (2.14 < Ti/Tv < 2.43). Additionally, individuals were removed if the reported sex disagreed with the genetic sex. See **Supplementary Table 1** for the sample numbers removed by the different criteria.

Finally, 5,390,149 variants passed exome sequencing filters, summarized in 5,033,063 non-overlapping positions of variation. 19,688 individuals passed QC, of which 17,138 individuals overlapped with the quality-controlled genotyping dataset described above and were used for further analysis.

Genome-wide association analysis

Genome-wide association tests were conducted using SAIGE[22] (0.45.0) on both the variants from exome sequencing and genotype imputation implemented within the BIGWas pipeline that is available at GitHub ikmb/gwas-assoc[16]. In brief, a logistic mixed-effects model was applied on the UC case-control status using genotype dosage (imputed data) or genotype hard-call genotypes (exome data). Genotype dosages were used to appropriately consider imputation uncertainty. The model additionally included the first 10 PCs calculated from the quality-controlled genotypes from the GSA (pre-imputation and post-quality control). We calculated the genomic inflation factor (lamdaGC) with and without excluding the HLA region (chr6, 29Mb-34Mb)[23]. Variants with a P-value of association < $5x10^{-8}$ were defined as genome-wide associated with UC, while variants with a P-value of association < 10^{-5} were considered to have at least a nominal (suggestive) association. Bayesian fine mapping was performed using the tool FINEMAP (version 1.4) with the parameters --n-causal-snps 1 --sss (shotgun stochastic search)[24,25].

Linkage-Disequilibrium-based (LD-based) clumping was calculated using PLINK with a significance threshold for the index SNP (clump-p1) of 0.00001 and a secondary significance for clumped SNPs (clump-p2) of 0.001 in a range of 150 kilo base pairs (kbp).

Gene expression impacts for the associated variants was looked up using the R-package Qtlizer[26] and gtex associated genes with a P-value below 10^{-5} were considered marginally associated.

Replication of suggestive hits

All suggestive variants were investigated extensively using regional association plots created with LocusZoom[27,28]. Variants with insufficient LD-support (less than two additional variants with P-value $<10^{-3}$ in ±150 kbp) and a MAF < 1% were discarded as false positive associations. To validate all remaining hits, we performed a lookup of variants, considering ±150 kbp around our lead SNP in the NHGRI-EBI GWAS Catalog v1.0.2[29] and in the Rapid Imputation for COnsortias PIpeLIne (RICOPILI; dataset IBD_UC_1KG_oct13 dataset[30]) considering variants with linkage-disequilibrium (LD) $R^2 > 0.9$ from our lead-variants[30]. LD was calculated on the exome and TopMED imputed data using PLINK[31]. Variants were considered to replicate if the lead variant itself or variants in high LD ($R^2 > 0.9$) had a P-value of association below 0.05.
Further, a power analysis based on the NHGRI-EBI GWAS Catalog of genome-wide association studies[29] data was performed to investigate replicability of known IBD variants in our dataset at a P-value threshold of $< 10^{-5}$. For this purpose, odds ratios (OR) reported in the GWAS Catalog were used, together with the sample size (n_{cases}=863, $n_{controls}$=4,185) and allele frequencies derived from this study's data. As the power calculated by a single analysis tends to be overestimated (referred to also as the winner's curse)[32–34], we computed a corrected value for power as the median power for the respective reported lead variant within neighborhood of ±5,000 base pairs under exclusion of the highest power value if more than one study was reported in the NHGRI-EBI GWAS Catalog.

Gene based analysis

To include the genetic variation based on rare variants (MAF < 1%), which have limited power in the classical analysis, a gene-based analysis was performed using SAIGE-GENE[35]. SAIGE-GENE performs three different tests: 1) The Burden test analyses the correlation between the number of variants and the disease status and is therefore powerful if the single variants show the same direction of effect. 2) The sequence-based kernel association test (SKAT) aggregates the test statistics of the single variants and is therefore robust against variations in the direction of effect. 3) The SKAT-O is a linear combination of the other two tests. For more details see Lee *et al.* 2012[36] and Zhou *et al.* 2020[35]. The analysis focusses only on the exome sequencing data, as we are focusing here on the genes and as rare variants cannot be imputed in sufficient quality. All variants with an allele-frequency <1% are grouped based on the gene-annotations generated by bcftools/csq[37] using release 100 of the primary assembly of the human proteome from Ensembl[38,39]. A relatednessCutoff of 0.125 and the same 10 PCs as for the GWAS analysis were used. Additionally, the following parameters were applied: minMAC=0.5, LOCO=FALSE, IsSingleVarinGroupTest=TRUE, IsOutputAFinCaseCtrl=TRUE, IsOutputNinCaseCtrl=TRUE,

IsOuputHetHomCountsinCaseCtrl=TRUE. A gene is considered genome-wide significant if the P-value is below 2.62×10^{-6} (based on 19,117 genes).

HLA imputation and fine mapping analysis

Imputation of alleles in the HLA region was performed for the classical HLA class I loci HLA-*A*, -*B*, -*C* and the class II loci HLA-*DQA1*, -*DQB1, -DPA1, -DPB1, -DRB1, -DRB3/4/5* at 2-field full context resolution from quality-controlled SNP genotype data. Here, we extracted pre-imputation SNP genotypes from the extended HLA region (chromosome 6: 29-34Mb) and used them as input for HLA genotype prediction with the random-forest based machine learning tool HIBAG (version 1.20.0)[40] using the multi-ethnic reference model published by Degenhardt *et al.*[41]*,* which was modified to include the variants available on the GSA. We additionally derived amino acid and additional SNP imputation and defined marginal posterior probabilities for single HLA alleles across all predicted HLA genotypes as described in Degenhardt *et al.*[41]*.* Classical HLA alleles with a marginal posterior probability from imputation <0.3 were further excluded. Association analyses were conducted using PLINK's logistic regression framework on the UC case/control status using hard-call HLA alleles, SNPs and amino acids from the imputation and the first 10 PCs calculated on whole-genome genotype information pre-imputation and post-quality control. For a closer look into the HLA haplotype structure, the tool disentangler was used[42].

Generation of personalized proteomes

For subsequent downstream analysis of genome-wide SNP data, a personalized haplotype-aware consequence-caller was used to predict the effect of genetic variants observed in the study cohort at the protein level. These genetic variants were obtained by merging WES data with the quality-controlled genotype data from the GSA (pre-imputation) using the Ensembl human genome FASTA release 100[38,39] as a reference. Four main steps were performed: (1) Phasing of nucleotides from WES with eagle (v2.4.1)[43], (2) calling of the variant effect at the protein level (*i.e.,* the type and if available the effect of a variant on the protein level, for example, a missense variant chr4:85742C>T in ENST00000609518 with the exchange 48P>48S), (3) filtering data for chosen candidate genes (4) translation into FASTA files. (1) To enable the phasing of the sparse WES data the variant files derived from WES and GSA-based genotyping were merged. Further, multiallelic variants were transformed into different biallelic variants with bcftools/norm[44] as eagle is not capable of dealing with multiallelic variants. The phasing was then conducted using eagle (v2.4.1)[43] without an external reference. The resulting file was scanned for contradictory phased variants and rectified

randomly by changing the phase for one of the contradictorily phased genotypes. In detail, as multiallelic variants *e.g.,* ref/alt1/alt2 are split (ref/alt1 + ref/alt2) and phased independently, for the rare case of heterozygous alternative variants (genotype alt1/ alt2) the phasing might lead to the phased genotype ref|alt1 + ref|alt2 which is contradictory as the merging back of the overlapping variants would lead to the ref in one haplotype and the alt1 and alt2 both in the second haplotype. (2) The haplotype aware variant effects were predicted using bcftools/csq[37] on the merged WES/GSA dataset resulting in a VCF file containing 1,353,644 protein changing variants (1,250,354 missense variants, 6,830 inframe-insertions, 17,909 inframe-deletions, 46 inframe-altering variants, 4,904 start-loss variants, 43,201 stop-gain variants, 2,652 stop-loss variants and 43,994 frameshifts[1]). (3) For the following parts the set of genes was filtered to those with an association to UC. Here, we considered genes described previously as associated with UC at a genome-wide significant level in at least 4 different UC studies integrated in the NHGRI-EBI GWAS Catalog or variants not described in the catalog but observed here to have a nominally significant association (for the genotyping data additionally LD support is mandatory) with UC in this study. This resulted in a selection of 76 genes with 322 protein coding transcripts. A list of these genes and transcripts is shown in **Supplementary Table 2**. (4) We utilized VCF2Prot[45] to generate a per-patient personalized version of all proteins transcribed from these 322 candidate transcripts by including the amino acid changes of step (2) into the reference sequences from Ensembl release 100[38,39]. Briefly, the VCFs from step (2) were filtered for records containing the target transcripts into a new VCF file. The newly generated files along with the reference protein sequence of each transcript were fed into VCF2Prot and the results were stored as a FASTA file per patient containing all mutated isoforms.

Peptide binding prediction

The generated personalized protein sequences for the set of candidate transcripts, along with the reference protein sequence were fragmented into 15-mer peptides using a sliding window approach with a step size of 1 and the generated results were stored in an SQL database for downstream analysis.

From all records stored in the database, we next created a set of unique peptides across all transcripts and patients. Additionally, we created a set of unique HLA-*DR* alleles and HLA-*DQ* alleles observed across all patients from the HLA imputation described above. For all combinations of unique peptides and unique HLA alleles from these lists, HLA-peptide binding affinities were predicted using NetMHCIIpan-4.0[46].

[1] The sum of mutations by types are more than the overall number, as the variants include multiallelic variants and mutations might have different effects in different transcripts.

Binding motifs

The predicted binding affinities were used to generate a sequence logo of the predicted bound peptide repertoire for the associated HLA-*DRB1* alleles and HLA-*DQ* (genome-wide significant ($P < 5 \times 10^{-8}$) in Goyette *et al.*, the European dataset of Degenhardt *et al.* or our dataset and no other effect direction in one of the other studies, for *DQA1* and *DQB1* pairings these conditions need to be true for both genes independently). This representation visualizes an enrichment or depletion of amino acids in comparison to the background amino acid frequency. The peptides are truncated to the 9-mer core predicted by NetMHCIIpan.

We generated 2 logos for each allele: (1) The logo generated by including all peptides predicted as weak or strong binder (%rank score <10). (2) Only the peptides that additionally fulfill the criterion of not binding against any of the significant alleles with the opposite direction of effect. The logos present the probability-weighted Kullback-Leibler logos with pseudo counts[47]. The logos were adjusted with pseudo counts based on the BLOSUM 62[48] substitution matrix using a ß of 200, therefore adding 200 artificial peptides reflecting typical evolutionary mutations[47]. The graphical presentation was then performed using the R package *ggseqlogo*[49].

Peptidome-wide association study (PepWAS)

The predicted peptide binding affinities were further used for a PepWAS analysis. Peptides with a percentile rank score below 2% (strong binder) were considered to be presented by an HLA protein. Finally, the set of bound peptides were combined with the patients' phenotypes and HLA genotypes to conduct a PepWAS as described initially by Arora and colleagues[11]. The aim of PepWAS is to identify peptides that might be relevant for immune recognition based on their binding affinity. In brief, PepWAS discriminates peptides based on the predicted binding affinity in patient versus control samples. Therefore, the same logistic regression model as for the genetic association (GWAS) analysis was used including the first 10 PCs.

The proteins and their peptides are analyzed for supporting factors. Those are: (1) Mutations in our sample set influencing the presence of the single peptides. (2) The gene expression signature in ulcerative colitis patients as published by Linggi *et al.*[50] and by Taman *et al.* [51]. (3) The cellular compartment of the genes[52]. The information of the subcellular compartments was exported from https://download.jensenlab.org/human_compartment_knowledge_full.tsv on 25.10.2022, the highest score presented for the gene ontology term GO:0005886 (membrane proteins) and GO:0005576 (extracellular proteins) are considered relevant. (4) A comparison with immunopeptidome data previously published and described by ElAbd *et al.*[53].

[Funding]

M.W. and H.E. were funded by the German Research Foundation (DFG) (Research Training Group 1743, 'Genes, Environment and Inflammation'). T.L.L. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 437857095. The study received infrastructure support from the DFG Cluster of Excellence 2167 "Precision Medicine in Chronic Inflammation (PMI)" (EXC 2167-390884018). The funding agency had neither a role in the design, collection, analysis, and interpretation of data nor in writing the manuscript.

[Acknowledgements]

-

[Contributions]

M.W., A.F. conceived and initialized the project. M.W. and H.E. analyzed the data, with help of M.H., F.U.-W., E.M.W., L.W., S.J., R.G.C., and D.E.. M.W. wrote the manuscript with help of H.E.. M.L., M.Z., B.B., and S.S. collected the samples. R.G.C. generated the WES data. H.E., P.B., T.K., and A.T. generated the peptide elution data. T.L.L. and A.F. supervised the project.

All authors revised and edited the manuscript for critical content and approved of the final version to be published.

[References]

1. Ng, S. C. *et al.* Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet (London, England)* **390**, 2769–2778 (2017).
2. Ananthakrishnan, A. N. Environmental Risk Factors for Inflammatory Bowel Diseases: A Review. *Dig. Dis. Sci.* **60**, 290–298 (2015).
3. Legaki, E. Influence of environmental factors in the development of inflammatory bowel diseases. *World J. Gastrointest. Pharmacol. Ther.* **7**, 112 (2016).
4. Franke, A. *et al.* Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* **40**, 1319–1323 (2008).
5. De Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
6. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
7. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**,

979–986 (2015).

8. Ashton, J. J., Latham, K., Beattie, R. M. & Ennis, S. Review article: the genetics of the human leucocyte antigen region in inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **50**, 885–900 (2019).
9. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.* **47**, 172–179 (2015).
10. Degenhardt, F. *et al.* Transethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals not only shared but also ethnicity-specific disease associations. *Hum. Mol. Genet.* **30**, 356–369 (2021).
11. Arora, J. *et al.* HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 944–949 (2019).
12. Barbaresko, J. *et al.* Dietary patterns associated with inflammatory biomarkers in a Northern German population. *Eur. J. Nutr.* **59**, 1433–1441 (2020).
13. Degenhardt, F. *et al.* Detailed stratified GWAS analysis for severe COVID-19 in four European populations. *Hum. Mol. Genet.* **Epub ahead**, n.n. (2022).
14. Bokemeyer, B. *et al.* P386. Anti-TNF alpha as induction and maintenance therapy in ulcerative colitis patients in the BioColitis Registry in Germany. *J. Crohn's Colitis* **10**, S292–S293 (2016).
15. Ellinghaus, D. *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
16. Kässens, J. C., Wienbrandt, L. & Ellinghaus, D. BIGwas: Single-command quality control and association testing for multi-cohort and biobank-scale GWAS/PheWAS data. *Gigascience* **10**, 1–12 (2021).
17. Kuhn, R. M., Haussler, D. & James Kent, W. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
18. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
19. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
20. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. Minimac2: Faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
21. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
22. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341

(2018).

23. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
24. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
25. Hans, C., Dobra, A. & West, M. Shotgun stochastic search for 'large p' regression. *J. Am. Stat. Assoc.* **102**, 507–516 (2007).
26. Munz, M., Wohlers, I., Simon, E., Reinberger, T. & Busch, H. Qtlizer : comprehensive QTL annotation of GWAS results. *Sci. Rep.* 1–8 (2020) doi:10.1038/s41598-020-75770-7.
27. Boughton, A. P. *et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
28. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).
29. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
30. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).
31. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
32. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
33. Nakaoka, H. & Inoue, I. Meta-analysis of genetic association studies: Methodologies, between-study heterogeneity and winner's curse. *J. Hum. Genet.* **54**, 615–623 (2009).
34. Palmer, C. & Pe'er, I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* **13**, 1–18 (2017).
35. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
36. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
37. Danecek, P. & McCarthy, S. A. BCFtools/csq: Haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
38. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
39. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

40. Zheng, X. *et al.* HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
41. Degenhardt, F. *et al.* Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Hum. Mol. Genet.* **28**, 2078–2092 (2019).
42. Kumasaka, N. *et al.* Disentangler A Visualization Technique for Linkage Disequilibrium Mapping Using Multi-allelic Loci. (2011).
43. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
44. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
45. ElAbd, H., Degenhardt, F., Lenz, T. L., Franke, A. & Wendorff, M. VCF2Prot: An Efficient and Parallel Tool for Generating Personalized Proteomes from VCF Files. *bioRxiv* 2022.01.21.477084 (2022) doi:https://doi.org/10.1101/2022.01.21.477084.
46. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
47. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, 281–287 (2012).
48. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
49. Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
50. Linggi, B. *et al.* Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis. *Sci. Rep.* **11**, 1–12 (2021).
51. Taman, H. *et al.* Transcriptomic landscape of treatment-naïve ulcerative colitis. *J. Crohn's Colitis* **12**, 327–336 (2018).
52. Binder, J. X. *et al.* COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. *Database* **2014**, 1–9 (2014).
53. ElAbd, H. *et al.* Predicting Peptide HLA-II Presentation Using Immunopeptidomics , Transcriptomics and Deep Multimodal Learning. *bioRxiv* (2022) doi:https://doi.org/10.1101/2022.09.20.508681.
54. Franke, A. *et al.* Systematic Association Mapping Identifies NELL1 as a Novel IBD Disease Gene. *PLoS One* **2**, e691 (2007).

55. Taleban, S. *et al.* Ocular manifestations in inflammatory bowel disease are associated with other extra-intestinal manifestations, gender, and genes implicated in other immune-related traits. *J. Crohn's Colitis* **10**, 43–49 (2016).
56. Hong, S. N. *et al.* Deep resequencing of 131 Crohn's disease associated genes in pooled DNA confirmed three reported variants and identified eight novel variants. *Gut* **65**, 788–796 (2016).
57. Watanabe, T. *et al.* Higher-Order Chromatin Regulation and Differential Gene Expression in the Human Tumor Necrosis Factor/Lymphotoxin Locus in Hepatocellular Carcinoma Cells. *Mol. Cell. Biol.* **32**, 1529–1541 (2012).
58. Taman, H. *et al.* Genome-wide DNA methylation in treatment-naïve ulcerative colitis. *J. Crohn's Colitis* **12**, 1338–1347 (2018).
59. Yu, Y. The role of PGAM5 in regulating viral infection and the pathogenesis of intestinal inflammation. (2021).
60. Liszt, K. I. *et al.* Human intestinal bitter taste receptors regulate innate immune responses and metabolic regulators in obesity. *J. Clin. Invest.* (2021) doi:10.1172/jci144828.
61. Anastasio, N. *et al.* Mutations in SCARF2 are responsible for van Den Ende-Gupta syndrome. *Am. J. Hum. Genet.* **87**, 553–559 (2010).
62. Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70**, 845–857 (2002).
63. Broad Institution. Inflammatory Bowel Disease Exomes Browser. https://dmz-ibd.broadinstitute.org/.
64. Saadati, H. R. *et al.* Genome-wide rare copy number variation screening in ulcerative colitis identifies potential susceptibility loci. *BMC Med. Genet.* **17**, 1–10 (2016).
65. Pfeuffer, S. *et al.* Deficiency of the Two-Pore Potassium Channel KCNK9 Impairs Intestinal Epithelial Cell Survival and Aggravates Dextran Sodium Sulfate-Induced Colitis. *Cell. Mol. Gastroenterol. Hepatol.* **14**, 1199–1211 (2022).
66. Doxiadis, G. G. M., Hoof, I., De Groot, N. & Bontrop, R. E. Evolution of HLA-DRB genes. *Mol. Biol. Evol.* **29**, 3843–3853 (2012).
67. Annese, V. Genetics and epigenetics of IBD. *Pharmacol. Res.* **159**, 104892 (2020).
68. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H. & Dermot, P. Host-microbe interactions have shaped the genetic architecture of Inflammatory Bowel Disease. **491**, 119–124 (2012).
69. Perše, M. & Unkovič, A. The Role of TNF in the Pathogenesis of Inflammatory Bowel Disease. in *Biological Therapy for Inflammatory Bowel Disease* 13 (IntechOpen, 2019). doi:10.5772/intechopen.84375.

9

DISCUSSION

9.1 DISCUSSION OF DIFFERENT ASPECTS

IBD is a complex disease, a disease for which the pathogenesis is attributed to many different factors. Over the last decades, different risk factors for IBD have been identified. Among them, 240 genetic risk loci [11, 40, 82], epigenetic modifications [11, 102, 148, 179], differences in the composition of the microbiome [99, 206, 214], and environmental factors [8, 102]. As is typical for complex diseases, like asthma or rheumatoid arthritis, the complete pathogenesis remains unclear and different scientists are working on different puzzle pieces to solve this issue from many different research perspectives. Here, I discuss the limits of the applied methods relevant within this thesis and what we know or suspect about the location of the single puzzle pieces. Afterwards, I will give an outlook on possible future approaches, that might help discover more about the pathogenesis of the disease or might help to further improve the life of patients with IBD.

9.1.1 *Genetic Associations*

Genetic studies are a widely used and accepted method to analyze risk factors that are heritable. Overall, the methods for this analysis have improved over time and biases have been reduced. Still, small biases are hard to avoid due to ethical restrictions within the research projects, technical issues, practical issues (e.g., recruitment strategies), or other unknown factors. Genetic studies are based either on analyzing family trees of individuals affected by a disease or large case-control studies in the context of GWAS (Section 2.3). Whereas the genetic variants of diseased samples are compared against those of healthy individuals. Samples from healthy individuals, such as blood or stool samples, are generally taken from the general population. Blood donors, for instance, are easily accessible and commonly recruited through respective consents, such that many control samples are available through local blood banks. Sometimes, individuals in the social network of the scientists are recruited as well. These groups are not necessarily representative of the general population. They may have e.g., a defined age profile and negative selection for certain diseases. The bias introduced this way might be ruled out by studies replicated with new datasets without the same bias. But as the same controls are often used across different studies, a bias in the sampling of controls would be noticeable in the same genetic association with different traits.

Variants associated with IBD at a high level of significance, like the HLA locus and the NOD2 locus in CD, are well replicated. While some variants among the 240 loci, that have a lower level of significance may be false positives due to statistical noise or the aforementioned biases. **DeLange** reported that 188 of the 232 IBD-associated SNPs of the previous meta-analysis have had at least nominal evidence in their data. Other loci and variants linked with the disease, might not yet been uncovered. Either because of a small effect size, low frequencies, or because of an insufficient representation of the locus in the human reference genome. To also identify variants with a small effect size, even larger datasets will be needed.

As shown by Visscher et al. [190], no saturation of identified associated variants has been reached within the GWAS performed for any trait so far. There is still missing heritability. As already discussed in Chapter 4, the exact heritability of IBD is not known. Estimates of heritability based on GWAS include only the identified genetic factors and therefore underestimate the heritability, while estimates based on twin studies tend to overestimate the heritability [62]. The heritability of complex diseases is expected to be explained partly by common variants and partly by rare to ultra-rare variants [178]. The heritability was so far not completely covered by GWAS studies for different reasons. Ultra-rare variants cannot be detected by classical GWAS studies, which use tag SNPs measured on SNP genotyping assays to extrapolate unknown genotypes from imputation approaches. Accurate prediction of these variants is limited, even when using large reference panels. In addition, some low frequency variants might be poorly captured by the common tag SNPs [37]. Those variants are more likely to be detectable using WES and whole genome-based genotyping. Some of the missing heritability might be in parts of the genome, not covered in classical GWAS studies. As discussed in Chat et al. [28] those variations include copy number variations and structural variants.

In Paper C (Section 8.3), I used WES and SNP array genotyping data of 863 German UC patients and 4 185 control samples to perform a GWAS. In comparison to recent GWAS studies using SNP array data [40, 105], the number of samples in this study was small. Still, I identified novel variants associated suggestively with UC (hits) in regions poorly or not at all covered in previous GWAS studies. Within this study I focused on SNPs and common small InDels. The usage of the new reference genome GRCh38 combined with WES allowed the identification of associations previously undetectable. Likely due to the limited sample size, none of the identified variants though suggestively associated reached genome-wide significant p-values. Hence, a replication is still needed to validate those results. An additional analysis, which focuses on structural variants, might be of additional interest but it is paired with an increasing complexity [28]. The data used in Paper C (Section 8.3), might be insufficient in sequencing depth and read length to accurately identify these more complex variations.

As pointed out in 2015 by Meienberg et al. [115], WES does not cover all possible exons in sufficient depth. However, Chat et al. [28] pointed out that ultra-low coverage whole genome sequencing is a promising approach for cost-effective future GWAS. This approach would lead to lower costs compared to SNP array-based genotyping [28]. It needs to be mentioned, that the quality of sequencing has improved over the last years while the

prices have decreased. Therefore, soon large sequence-based studies for complex diseases are to be expected.

Another approach to uncovering an association of a disease with rare variants within a gene, is to study these variants together. This would base on the hypothesis of a similar effect of these variants within a gene on the disease phenotype. There are different methods developed to jointly analyze rare variants within a gene. These methods, including the burden test and the sequence-based kernel association test (SKAT), have different statistical assumptions.

The burden test analyzes the number of rare variants in a gene, collapsing them in a single metric. This test is powerful if all variants have the same direction of effect but loses power when the variants have different effect directions. The SKAT is more powerful if the variants within one gene have different directions of effect [142]. Both options, i.e., variants having the same effect and variants having different effects, can occur at different loci across the genome for one trait. Often the best fitting model is therefore not known. However, the single tests loose power massively if the respective other hypothesis is true [101]. This problem is addressed by the SKAT-O test, a linear combination of the SKAT and burden test. I performed a gene-based test in PAPER C (Section 8.3), using the tool SAIGE-GENE [213] across all annotated genes. However, no significantly associated genes (p-value$= 2.62 \times 10^{-6}$) were identified. This is most likely due to the limitations based on the number of samples and low frequency of the rare variants in the WES dataset. Nevertheless, the analysis suggests that the same approach on a larger sample set might lead to relevant results.

Other limitations of the GWAS approach include the failure to address epigenetic effects, gene-gene interactions, or gene-environment interactions [37]. While epigenetic effects and gene-environment effects cannot be analyzed on WES or SNP array data alone, one approach to identify gene-gene interactions based on genetic data is a GWAS analysis based on pairs of SNPs. Calculations of SNP×SNP interactions are computationally intensive as the number of tests increase as compared to the single SNP approach in a classical GWAS [199]. The high number of tests conducted, leads to a loss of power of detection of an association after multiple test correction, which leads to the need of even larger sample sets.

A combination of twin or family-based studies together with large population studies might help to identify additional variants associated with IBD. Potentially rare variants of interest are easier to be identified in family trees. Complementary, population studies might help to figure out which variants are of interest on a population level and where a gene is only inherited within a family only by chance together with a disease.

Even if it were possible to uncover the totality of heritability, the functional characterization of the identified variants would remain challenging [178]. Association of a variant with a disease does not per-se imply causality, but merely correlation. To identify causal factors of genetic diseases, typically a hypothesis about the underlying mechanism is needed [178]. For some human genes and their protein products, the function is not known or they might have additional functions, which are yet to be uncovered. Generating a hypothesis for the role of a specific gene with unknown functionality within a complex disease is nearly impossible. Many studies have focused

on mutations affecting the protein structure [97] but also the regulatory effects, like gene expression, can play an important role.

9.1.2 *Analysis of gene expression in IBD*

Non-synonymous mutations within a gene have a direct impact on the protein structure and can influence a protein's functionality, e.g., by leading to a loss of function of the protein or by changing the potential of the protein to interact with other proteins, peptides, and molecules. Mutations can also change the expression level of a gene. Expression in turn, as discussed in more depth in the next chapter, can also be influenced by epigenetic effects, and transcription factors. It varies across different stages within cell differentiation. As the expression is a key factor in the processes of a cell, different types of studies were developed in the last decades to analyze the gene expression dependent on different criteria.

Expression Quantitative Trait Loci (eQTL) are genetic loci that have an impact on the expression of single genes. An eQTL analysis is based on the same statistics as applied in other quantitative GWAS studies, using the expression levels as a phenotype (the dependent variable). Usually, as a secondary step to GWAS analysis of a specific phenotype to identified phenotype-associated variants, eQTL analysis is used to analyze whether identified variants or variants in LD have an impact on the expression of a specific gene. This information can be drawn from publicly available datasets, as for instance, curated by the Genotype-Tissue Expression project (GTEx) [68]. The GTEx Consortium [68] stores pre-analyzed datasets of SNP-expression association analyses, within different tissue types and sometimes also related to specific diseases, e.g., "Blood cells in celiac disease", "Cells - Macrophages", or "Colon - Sigmoid" are noted as description of the tissue.

An alternative way to analyze the impact of expression on a disease, but in this case independent of a genetic variation, is differential expression analysis. Levels of gene expression are compared between patients and control groups. Here, expression levels of genes are measured in samples with a specific phenotype, e.g., IBD, and healthy controls using ribonucleic acid (RNA)-based NGS. The generated data is then analyzed to identify differentially expressed genes (DEGs), using appropriate statistical methods. In case of IBD, tissue samples are typically taken from the gut. Since sampling of gut tissue is much more inversive than sampling from the blood, the sample sizes nowadays used for such analysis, especially for the healthy controls, are limited to tens to a hundred samples [70, 104, 167, 180]. Additionally, the sampling site is not completely homogeneous within and especially across different studies, e.g., as it varies also with the site of inflammation in IBD patients at the timepoint of sample procurement. Those positional effects, but also temporary effects, like daytime, recent activity (e.g., sports, eating, sleeping), or the medication of a patient might impact the genotype expression. It can be expected that those effects are partly indirectly linked to the disease. For example, the fitness might have been impacted by the current disease status.

Currently, gene expression studies in IBD [70, 104, 119, 167, 180] are not reproducing each other well. A biological explanation for this observation, the expression profiles of the disease are not stable over time. More likely, however, the aforementioned sampling-based biases, lab-based biases, and statistical issues lead to problems in reproducibility. Exemplary sources for biases induced in the lab are differences in the RNA-extraction and sequencing protocols, technical differences, or simply the fact that samples may be processed by different lab staff. Another effect that needs to be considered is that RNA sequencing suffers from a high dropout, as stochastically lowly expressed genes are often not detected [30]. More tightly defined and controlled studies will be needed in the future to reduce these biases and increase reproducibility across different studies.

9.1.3 *Potential role of the HLA in IBD*

The HLA proteins are in general well studied and different hypothesis about their function in a disease context exist. So far, no conclusion about the reason for the association of variation in the HLA in IBD could been drawn so far.

The classical HLA alleles are associated with many inflammatory diseases, but the specific role of the HLA with their disease etiology, like for IBD, is unknown for most. One exception is Celiac disease. Celiac disease is triggered by a peptide of the α-gliadin present in wheat. This peptide is presented especially by proteins of the DQ2 and DQ8 haplotype [160]. The connection between wheat and the disease was identified by a Dutch pediatrician who realized that the available diet after the second world war led to the improvement of symptoms in patients with celiac disease [169].

In case of IBD, even though several environmental associations [8, 102] have been found, so far, the "one driver" of the disease could not be identified. One of four possible explanations is most probably the reason why the triggering peptide has not been identified for IBD so far. It might be because nothing comparable to the α-gliadin in Celiac disease exists for IBD. Another explanation would be, that the disease is a collection of more than the two main subforms distinguished in most studies. Further subtypes of the disease are sometimes taken into consideration based on the location of the inflammation [34]. An alternative explanation would be that the triggering peptide is not as easily accessible and its removal not as simple to influence as it is the case for a specific food. It may rather be found in the gut microbiome of the individual patient or even be derived from the individuals own peptidome in form of an autoimmunogenic peptide. And lastly, an initial infection combined with a mimicry of peptides present in the commensal microbiome or self peptidome might be the initial and driving cause of the disease. In this scenario, the initial peptide might be hard to identify, as it is not identifiable anymore when the diagnosis for IBD is made. The driving peptide source alone, might not even be associated with the disease or a strong binding to HLA proteins.

All these hypotheses together lead to a huge search space for the peptide of interest including the microbiome, the food-born peptidome, pathogenic sources, and the human peptidome.

The role of the HLA is to present peptides to T-cells, independent of their source. Therefore, an analysis, focusing on the binding of HLA, leads to the characterization of sequence-based attributes but is not suitable to narrow down the source of the peptide on a bigger scale as shown in PAPER C (Section 8.3).

To identify a specific peptide or a limited group of peptides, that reflect the association pattern of the classical HLA alleles with the disease, additional factors need to be taken into consideration. In PAPER C, additional genetic effects, namely non-synonymous genetic variants, were taken into account but no coding variant influencing the binding of the HLA alleles was found. Other effects might influence which peptides are presented by the HLA in addition to the chemico-physical properties of the HLA proteins coded by the HLA genotype. Those effects include epigenetic factors, that influence the expression of protein sources, and post translational modifications (PTMs) of candidate peptides. All analysis conducted in the framework of PAPER C focus on the auto-immune hypothesis even though the hypothesis of non-human triggers for the disease are equally likely. This would need to be investigated further in the future.

My suggestion for further research into the identification of peptides is a reduction of the search space, including a limitation of candidate sources, e.g., considering only the microbiome species differentially present in IBD patients and controls.

9.1.4 *Peptide-HLA interaction*

Next to the large search space in form of an unknown well of peptides, the analysis of the HLA-peptide interaction, using both experimental and computational approaches, is nowhere near perfect. In the framework of this thesis, the analysis of the HLA-peptide interaction refers to the determination of whether a peptide binds to an HLA protein rather than the physico-chemical properties of potential binding processes. Even though different lab protocols have been developed to study HLA-peptide binding, they all have their weaknesses in representing the peptidome as it would be present *in-vivo* (Chapter 5). As outlined in Chapter 7, experimental protocols for elucidating HLA-peptide binding differ. Protocols determining the BA focus on the affinity of predefined peptides to be bound by an HLA often in comparison to another peptide. They neglect other impact factors of peptide presentation. EL are generated by washing any peptide from HLA molecules as presented on cells.

The most common protocols for BA data achieve limited throughput and the results measured are impacted by different conditions. As the main interest is on positive hits, peptides predicted to bind specific HLA alleles are tested more often by BA experiments. Computational methods, using these data to predict HLA-peptide binding, are limited to the information gained by those measurements and peptides with another characteristic might be missing. This circle might introduce a bias in the data. In addition, different experimental conditions, e.g., pH of the analyte, may influence the binding. All in all, the applied protocols vary in many parameters and no studies, analyzing the effect of the single parameters systematically, have

been performed to date. Nevertheless, different methods measuring BA data repetitively showed similar binding patterns (PAPER B, Section 7.3). This suggests that the results are presenting important characteristics of the HLA binding peptides. BA data include only a small part of the HLA pathway as the molecules are often present without any cellular backbone. This means that the processing of the peptides does not reflect the *in-vivo* situation (see Section 5.1) and that the loading of HLA class II happens in the experiment without the HLA-DM and HLA-DO molecules involved in the loading of the HLA.

While measurements of the BA are mostly hypothesis-driven, i.e., on a range of candidate peptides *ex-vivo*, peptide elution coupled with MS as a method to generate HLA-peptide binding data (EL) can capture any peptide generated *in-vivo* or *in-vitro*. EL obtained with this method represent better biologically relevant and possibly disease-related processes. However, this method inherently lacks the possibility to control for the peptides that are identified. Furthermore, as shown by Bassani-Sternberg and Gfeller [18], MS results can be biased towards the identification of charged peptides.

Since in MS peptides are often identified by comparison of the measured MS spectra to a reference spectrum, either generated by previous MS experiments or a theoretical spectrum generated from a peptidome, identified peptides are often limited to this reference. Peptides that are not included in the reference, may include peptides generated from uncommon genetic variation within single genes or peptides from not included microbial strains. For example, by default only the allele HLA-DRB1*15:03 is included in the reference for *HLA-DRB1* but peptides, which are not in the HLA-DRB1*15:03 molecule, are not present. Another type of peptides not included as such in the reference are discontinuous peptides. Those peptides consist of multiple small fragments of a protein in a non-sequential manner. According to Faridi, Dorvash, and Purcell [53], discontinuous peptides might represent a considerable portion of the HLA-bound peptides at least for HLA class I.

Another disadvantage of peptide elution experiments is that they only deliver information on presented peptides but no knowledge on the peptides that are not presented. To effectively use peptide elution experiments for HLA-peptide binding prediction using machine learning approaches, negatives, i.e., non-binders, need to be generated *in-silico*. Different methods based on different assumptions were developed for this purpose. Non are expected to perfectly reflect the reality [49].

An effect, well captured by the EL is the impact of gene expression on the peptide presentation by an HLA protein. Recent studies from ElAbd et al. [49] and Chen et al. [29] investigated the impact of the expression level of genes on the peptides eluted from HLA molecules by developing HLA-peptide binding prediction algorithms based on EL data coupled with information on the expression of peptide-parent human protein. This leads to a more accurate prediction of self-presented peptides. However, this expression-coupled HLA-peptide binding prediction is currently limited to human proteins, and non-human proteins are hard to include in this model. Therefore, the benefit of this prediction approach is currently limited. Beside the prediction, these studies showed that proteins more highly expressed in an individual's cell are also more likely to be presented by HLA molecules.

PTMs, like glycosylation's, also have an impact on the presentation of peptides by HLA molecules. So far, this effect is hardly studied. Most protocols for the measurement of peptide-HLA binding do not consider any PTMs. To account for this, we integrated the distance to the nearest glycosylation site in PIA-M, the prediction tool for HLA-peptide binding published by ElAbd et al. [49]. So far glycosylation's are not recorded in the IEDB and not considered in the experimental settings in the first place. Nevertheless, methods like the high-density microarray technology used in PAPER B (Section 7.3) would enable the interrogation of the influence of PTM on the binding potential of peptides.

Epigenetic effects might play a role on the expression of genes and therefore indirectly impact the HLA peptidome. Those effects might be induced by environmental factors, like diet or smoking [11]. Differences in the methylation pattern, a common form of epigenetic modifications, was described by Adams et al. [3] for childhood-onset CD.

Another factor that may influence the captured HLA-peptide complex is its stability. HLA-peptide binding is on the one hand characterized by how easily a peptide can be bound and on the other hand by the strength of the HLA-peptide binding. Both factors influence how long a peptide is presented by an HLA protein and peptides presented over a longer timespan are more likely to be detected. While BA data do give a quantitative measurement of the binding, i.e., information on how much of the peptide is bound by HLA molecules, EL measurements are qualitative.

In PAPER C, we identified a candidate peptide using a PepWAS approach, that we could map back to the HLA-DRB1 protein. The potential role of the peptide as driver of UC was supported by gene expression data, peptide elution experiments and the subcellular compartment of the HLA. The HLA is known to present antigens from other HLA alleles. This is especially important in transplantation medicine and the main reason why the selection of a suitable donor is so difficult. The healthy organism is resistant against the own HLA phenotype due to negative T cell selection, i.e., the immunological instance to avoid autoimmunity by avoiding T cells to induce an immune reaction by autoantigens. But as shown in transplanted individuals, the presentation of HLA antigens can in general lead to an immune reaction. Whether this may be also extended as a concept to a chronic inflammation, would need further investigation.

9.1.5 *Impact of variation in the HLA on drug response*

The HLA is not only associated with the prevalence of IBD but as well with the efficiency and side effects of different medications [13, 71, 186]. For example, carriers of the allele DRB1*03:01 were identified to have an increased risk to suffer from nephrotoxicity as a side effect of 5-ASA treatment even though this finding is of limited clinical relevance as this side effect is very rare [71, 186]. Another example is the link between the haplotype DQA1*02:02-DRB1*07:01 and an increased risk of thiopurine-induced pancreatitis [13, 186]. For both associations the functional link is still unknown and might help to get a deeper insight into the role of the HLA in the pathogenesis of IBD. The main aim of pharmacogenetic studies

(studies that investigate the effect of genetic variation on drug response) is to help optimize the treatment on the individual level. The identification of genetic markers linked to drug response can help to reduce side effects and to select the most efficient treatment for a specific patient in the first place [186]. The subcategorization of individuals based on their genetic markers is also called personalized medicine.

9.1.6 *Role of T cells*

Szeto et al. [175] showed that the HLA, next to the direct interaction with the peptide, impacts the stability of the interaction with the TCR. For an immune response to be mounted, the interaction of the TCR with an HLA molecule and the peptide would be necessary (Section 5.1). First approaches were represented by Lee and Meyerson [100] on epitopes of the cytomegalovirus (CMV).

Another effect that needs to be considered when talking about the role of the HLA in disease is T cell selection. The T cells are selected within the thymus in a way that the TCR is functional and does not bind any self-peptides. This should avoid an immune reaction induced by self-peptides. Errors within this process happen and lead to autoimmune diseases. One mechanism against autoimmune diseases are regulatory T cells. As their name suggests, they can suppress an immune response and support self-tolerance.

The importance of the T cells and the TCR in the follow up consequence of a peptide presentation by HLA is striking. Huge knowledge on the role of the HLA in IBD might be gained by considering the T cells. The analysis of TCRs, as well as their interaction with an HLA-peptide complex, is even more challenging than the analysis of HLA-peptide interaction alone because of the high variability of the TCR based on VDJ recombination. While the HLA alleles are "hardcoded" in the genome and can be therefore determined by genome sequencing, the sequences of the TCRs are based on recombination of different V, D, and J gene segments and additional nucleotide editing. Each single T cell presents a different TCR, i.e., the sequences of TCR are highly variable within each individual.

NGS-based TCR profiling can be performed to gain some insights into different TCR sequences. As the majority of the TCR sequences are rare, analysis is challenging because statistically appropriate methods are currently lacking, and only a subset of different TCRs can be captured by sequencing. Single cell analysis of colonic $CD8^+$ T cells [36] and the TCR repertoire [198] in UC revealed differences between patients and controls. Recent analysis of the TCR profile in diseased and in healthy patients led to the identification of TCR sequence patterns, which are more common in patients with CD [154]. Further analysis to explain the pathogenesis of these in the diseases are needed. Also, the impact of the HLA on those profiles needs to be studied.

9.1.7 *Non classical HLA genes in IBD*

There are other immune-related genes not involved in antigen presentation that are also located within the HLA region on chromosome 6. The most studied non-classical HLA genes include TNF (an MHC class III gene) and MICA/B (the MHC class I chain-related A and B genes). So far, only small association studies analyzed the variability within these genes in the IBD-context, some of them describing a genetic association others without any significant results at these loci [5, 60, 120, 204]. Searching the PubMed database, no recent study can be found that had a deeper view into both the classical HLA alleles and the polymorphisms within TNF and MICA/B. The results of my HLA fine-mapping analysis, consistent with other publications [44, 63], revealed no obvious additional peak between the classical HLA class I and HLA class II genes, where those genes are located. However, this might be a consequence of the complex genetic structure of the HLA in general, paired with limited investigations into the genetics of those genes. In fact, TNF is in the scientific focus of IBD research through anti-TNF blockers which are a common treatment option for IBD [14, 120]. MICA has also been implicated in the pathogenesis of IBD by interacting with NKG2D as proposed by Muro, López-Hernández, and Mrowiec [120]. There has been no large, recent study analyzing the alleles in the non-classical HLA genes in IBD, even though the MIC genes are known to be highly polymorphic [204]. Any statistical analysis into the genetics of these non-classical HLA loci should always include the classical HLA alleles in any interpretation, since associations observed for these loci may only result from secondary effects, i.e., because of an LD of these with classical HLA loci. However, potential results might have an impact on the associations of the classical HLA alleles and the resulting functional interpretation. Furthermore, variation, especially within the TNF gene, might have an impact on the response to treatment. First findings in the pharmacogenetics in IBD in relation to the HLA are already published (described in Section 9.1.5).

9.1.8 *Identification of drivers of IBD*

Even though many studies have been performed to figure out more about the factors inducing IBD over the last decades, no coherent picture has been identified. Some pathways have been described (Chapter 4) that keep the chronic inflammatory reaction running but the initiating factors of the disease remain unclear. In general, it is difficult to study an individual pre disease onset or during disease onset. Therefore, only limited knowledge is available for this period. One of the most widely accepted factors for the development of IBD, amongst some environmental factors, is a genetic predisposition, as the genes do not change over time. And those genes point, amongst others, towards the HLA genes and the corresponding antigen presenting proteins. In Section 8.3 (PAPER C), I describe the characteristics of antigens presented in UC patients versus those presented in control individuals. Even though those analyses include uncertainties based on the bioinformatics approach, the overall setting is suitable to limit the peptides to a smaller, though still huge set of candidates, in comparison to the hy-

pothesized peptides relevant to the pathogenesis of IBD. To further reduce this set to a reasonably sized set of candidates, i.e., a number of peptides that can easily be followed up in the lab, further research including the analysis of the interaction of the APC with the T cell is necessary, e.g., analysis of the TCR profile. During the last years, first studies have been published showing a decreased diversity of TCR in IBD patients [85]. For CD an enrichment for a specific sequence pattern was identified by Rosati et al. [154], a study for which I was a coauthor. If those TCR sequences do have an immunogenic role, needs to be further studied.

9.2 OUTLOOK

Many different studies may be conducted to follow up the investigation of the HLA in chronic inflammatory diseases.

The genetic profile of the classical HLA alleles in UC is now well described by Goyette et al. [63], by Degenhardt et al. [44] (with my co-authorship), and in PAPER C (Section 8.3). All three papers focus on imputed 2-field alleles of the classical HLA. Larger cohorts might help to get a clearer picture for rare alleles or alleles with lower effect sizes as well. Another improvement might be achievable by the analysis of HLA alleles at a 3- or 4-field resolution. The genetic variabilities in the third field, accounting for synonymous variation, and fourth field, accounting for intronic variation, have no impact on the protein itself (Section 5.3) but might influence its expression level.

The imputation of HLA alleles is with the available imputation references limited to 2-field resolution. NGS-based or Sanger-sequencing-based HLA data would be needed to train an imputation model using machine learning or, depending on the number of sequenced samples, could be directly applied in studies including higher resolution. As already shown in PAPER A (Section 6.3), accurate imputation panels need to fit the ethnicity of the individuals under study. However, not all ethnicities are sufficiently covered in imputation panels so far.

The impact of genetic factors on the expression of the HLA are also a separate field of interest for further investigations, as genome-wide eQTL studies are expected to have limited success in the complex HLA region.

Furthermore, studies including non-classical genes of interest in the HLA region, e.g., TNF and MICA/B (Section 9.1.7), might help to differentiate the genetic association signal located in this genetic region. It needs to be considered, that based on the current knowledge, an independent association of those genes is uncertain.

The prediction of the HLA-peptide interaction will never be perfect. As already discussed in Section 7.1, the different data types that can be measured have different advantages and disadvantages for determining the binding of peptides by HLA proteins. For the BA, more data especially on rarer HLA alleles could lead to an improvement in HLA-peptide prediction. These data could be generated by the new ultra-high-density microarrays used in PAPER B (Section 7.3). This might be a cost-effective solution but currently the technology has some technical weaknesses. First, in generating the microarray, and second, in the readout, as the signal has a location bias. This

is most probably based either on inequal contact of the array with the used liquids or based on the illumination during the readout (Section 7.3).

Currently, the focus of HLA-peptide research is on peptide elution experiments, to generate data for peptide-HLA interaction. Like for BA there is room for improvement in the protocols used to generate EL data. First, the assignment of the parent-HLA protein to the eluted peptides is challenging in case of cells with a heterozygous combination of targeted HLA alleles. Approaches already developed in order to tackle this problem include the *in-silico* tool MoDec (motif deconvolution) [145] and an HLA tagging approach [2]. The *in-silico* approach adds an additional layer of uncertainty to the data, while the tagging approach is so far not possible in tissue samples. Second, the current protocols sparsely identify non-human peptides. Third, during the last years peptide elution experiments have focused on the elution of peptides from HLA-DR molecules. Consequently, sparse information is available for HLA-DP and -DQ molecules.

Additionally, the current data hardly include any non-linear peptides or peptides including PTMs. The inclusion of those increases the complexity of the lab protocols as well as the peptide-HLA binding prediction. Notably, Faridi, Dorvash, and Purcell [53] and Mei et al. [114] showed the potential relevance of including such information. The analysis of PTMs can be included easily in the high-density microarray protocol (Paper B, Section 7.3) but the biology about the peptide processing needs to be studied separately. First approaches to include PTMs into MS data have already been presented by Yu et al. [207] and Kacen et al. [84].

More focus should also be put on the TCR as well as the peptide presentation by HLA alone is not sufficient to induce an immune reaction. So far, the peptide presentation by HLA and the TCR repertoire are analyzed separately but a connection of those puzzle pieces might provide a major impact to uncovering the pathogenesis of UC. Making this connection might be possible by using tetramer assays, where HLA tetramers (or other HLA multimers) loaded with an antigen are used to detect T cells specific for those complexes. A condition for a successful application of those assays is a reasonable number of antigen candidates.

In Paper C we started to define peptide candidates for UC. Specifically, we focused on the autoimmune hypothesis, i.e. the peptide origin is hypothesized to be within the host itself. A special attention was on proteins expressed from genes previously associated with the disease in GWAS. As the HLA presents "something of everything" to the host-immune system, the HLA binding alone is not eligible to define a set of candidates but needs to be combined with other analyses. Those analyses are for example, TCR analysis as described above, the analysis of the gene expression, or changes within the protein sequences, the last two were included in Paper C. Especially when moving away from the autoimmune hypothesis towards the microbiome or other non-human antigen sources, additional knowledge from other fields need to be included.

9.3 CONCLUSION

The inflammatory processes in IBD are multifactorial [13] and the role of HLA in IBD remains unknown. However, the work performed in the framework of this thesis contributed to the current research by analysis of the HLA class II protein in general and in relation to UC in particular.

The new reference for HLA imputation presented in PAPER A (Section 6.3) improves the accuracy of imputation especially for individuals of middle eastern ancestry. Furthermore, it includes the genes *HLA-DRB3*, *-DRB4*, and *-DRB5*, which are neglected in most other imputation references. Additionally, the study identified some pairs of alleles difficult to distinguish by imputation. In general, HLA imputation, like SNP imputation, has lower accuracy for rare alleles as some alleles might not be present at all in the reference dataset or no tag SNPs are available on the genotyping assay for the discrimination of the corresponding HLA haplotypes.

The imputation panel was, among others, used in the trans-ethnic HLA fine-mapping study in UC, performed by Degenhardt et al. [44]. This fine-mapping study performed by us, replicated, and complemented the picture of associated HLA alleles. Pointing out, that the HLA-DRB1*15 group is a very important risk factor across ethnicities, with different alleles present in different ethnicities. HLA-DRB1*15:01 is also the risk allele identified with the largest power in PAPER C (Section 8.3).

The prediction of HLA-peptide binding is a field developing further with improving lab techniques and machine learning algorithms. As shown in PAPER B (Section 7.3), the ultra-high-density peptide microarrays are a valid data source for training HLA-peptide binding prediction tools. Ultra-high-density peptide microarrays have the big advantage of a high throughput, as hundreds- of thousands of defined peptides can be measured in parallel. On the other hand, the predictions by algorithms trained with these data are limited to the interaction between the HLA and a peptide, and do not reflect any other biological factors, i.e., information on the natural processing of the peptides. For this purpose, EL data is more suitable. We used EL in ElAbd et al. [49]. So far, peptide elution experiments have limited success in the identification of non-human peptides and non-binders are not measured.

The prediction of the interaction between the HLA and a peptide in context of a disease is one way to head from the genetic association of HLA alleles towards a functional understanding. As shown in PAPER C (Section 8.3), the peptides binding HLA-DRB1 proteins that are associated with an increased risk for UC on a genetic level differ from those interacting with alleles associated in a protective manner. This finding supports the hypothesis that the presentation of specific peptides and the related induction of an immune response might play a role in the pathogenesis of UC. Now, the candidate sources for these peptides identified from GWAS, microbiome analysis, and the search for environmental risk factors are so broad that the identification of a proven disease-causing epitope remains elusive. Next to the sheer amount of hypothesis-driven peptide candidates, bioinformatic approaches are currently not able to identify immunologically relevant peptides. As the HLA is supposed to present "something of everything" to the immune system, focusing only on the HLA helps reducing the candidate pep-

tides to some degree but not to prove or disprove any hypotheses regarding the peptide source. To this regard, the inclusion of other fields is necessary, especially the analysis of the T cell repertoire and the antibody reactivity. The first studies showing differences in the T cell repertoire for IBD [154, 198] give hope that an identification of a disease specific HLA-peptide-TCR complex might be possible soon.

To conclude, with the work behind this thesis, I contributed towards a better understanding of the HLA on the genetic and functional level in IBD with a special focus on UC. The findings of the GWAS, based on WES data and using the GRCh38 reference, show that the basic findings on genetic risk factors are not exhausted yet (PAPER C, Section 8.3). But as already pointed out by different authors, it is time to make the link from association to function [64, 185, 189]. The focused analysis of the HLA shows that this holds true for this complex region strongly associated with different diseases as well. It is time to bring together different puzzle pieces to get an even better picture of the disease.

Part IV

APPENDIX

A

SUPPLEMENT OF PAPER A

Supplementary Methods and Figures for "Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles"

SUPPLEMENTARY TEXT

Similarity of some alleles leads to misclassification

We identified a few allele pairs commonly misclassified because of similarity. These alleles can also be identified by a low sensitivity and specificity (**Supplementary Table 8**). The most prevalent allele pairs are described one after another in the following.

The alleles A*02:01:01:01 and A*02:03:01 differ in only three positions across the entire available nucleotide sequences (IMGT/HLA database). This is especially problematic in our Chinese population. A*02:03 had a mean sensitivity of 0.581 across all cross-validation runs of the Chinese population with a median of 0.625 [min=0.364, max=0.750] and a mean specificity of 0.996 and at the same time a frequency of 13.31%, which results in a big impact on the accuracy of *HLA-A* for the Chinese panel (**Supplementary Tables 3, 5, 8)**. A*02:01 had equal haplotypes as A*02:03 in 99 of the 100 classifiers. A median of 1.9% [min=0.0%, max=6.7%] haplotypes do not distinguish between A*02:01 and A*02:03 across the respective classifier, which makes classification of this allele particularly challenging.

The coding sequences of DRB1*11:04 and DRB1*11:01 differ in only one exonic SNP position and in example alleles DRB1*11:01:01:01 and DRB1*11:04:01 differ in only four positions in introns and exons. For the alleles DRB1*11:04 and DRB1*11:01 a median of 4.3% [min=1.4%, max=10.8%] of the haplotypes used for classification in the individual classifiers of our reference panel were overlapping in all of the 100 classifiers, this held true for the alleles DRB1*11:03 and DRB1*11:04 in 94 of 100 classifiers. Here a median of 5.6% [min=0%, max=23.08%] of the haplotypes did not discriminate between the two alleles (**Supplementary Table 7**).

DRB1*04:04 is misclassified in the HLA imputation of the 1000 Genomes samples

Using our multi-ethnic reference panel, we also imputed HLA alleles into the 1000 Genomes population. We observed, that DRB1*04:04 was misclassified in all of the samples of Western European Ancestry, and also in the East Asian and African samples from the 1000 Genomes panel.

The allele frequencies of DRB1*04:03 and DRB1*04:04 were consistently very low in our typed dataset with respective frequencies of 0.31% and 0.62% in our German panel and 2.50% and 0.00% in the Maltese panel (**Supplementary Table 5**). Due to these low allele frequencies only a small number of haplotypes was provided for training in the HIABG model, especially for DRB1*04:04. In the 1000 Genomes population the frequencies

of these alleles were much higher, with respective frequencies of 0.62% and 6.06% in samples of Western European ancestry. Our multi-ethnic dataset and the 1000 Genomes dataset overlapped in 8,417 of a total of 8,803 SNPs (95.6%). Some of the SNPs that were important for the classification of DRB1*04:04 were among the missing 4.4% such that 89 classifiers had a median of 5.2% [min=0%, max=45.6%] of the haplotypes used for allele classification overlap between DRB1*04:03 and DRB1*04:04 (**Supplementary Table 7**). We observed mean HLA allele sensitivity values for HLA-DRB1*04:04 of 0.000 (almost all ancestries) to 0.500 (German) and 0.750 (Korean) in our data (**Supplementary Table 8**). Notably, specificity measures that were reported by Zheng *et al.* (1) were low for DRB1*04, with DRB1*04:03 showing a sensitivity value of 0.150 in the European population and it was classified as DRB1*04:04 in 65% of the cases a misclassification occurred. Overall accuracies of the European ancestry data are high with mean values of 0.961 and 0.967 for the German and Maltese panel respectively based on the *HLA-A, -B, -C, -DQB1, -DRB1* loci with *HLA-B* and *-DRB1* being most challenging to impute.

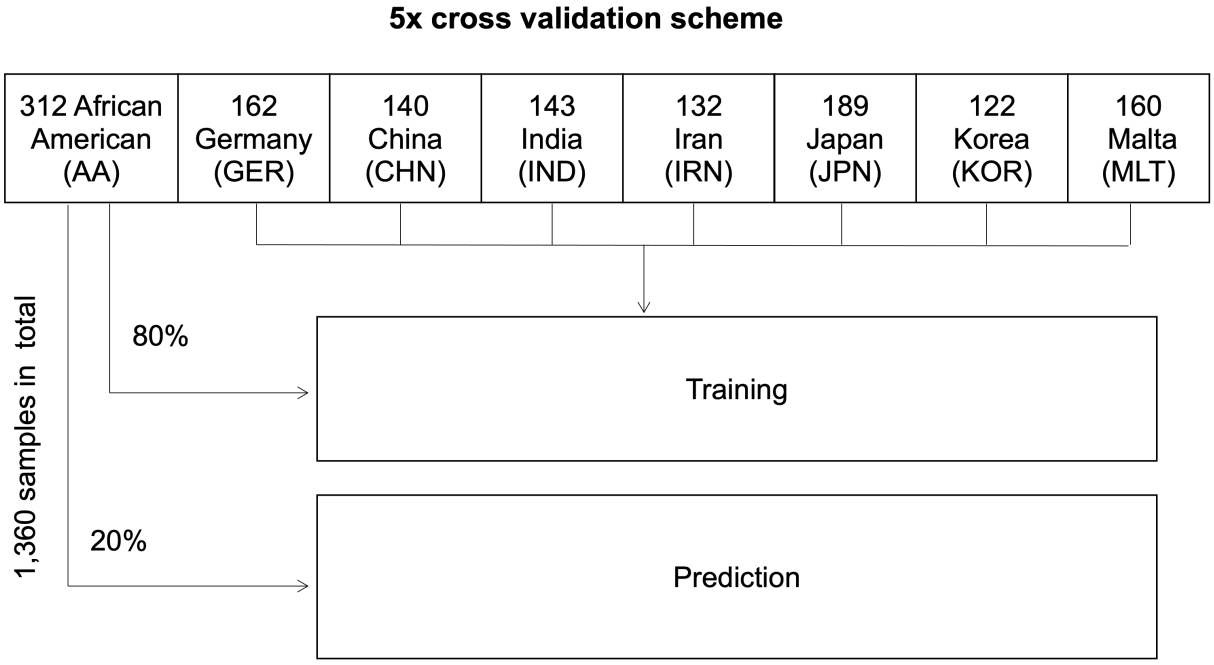
SUPPLEMENTARY METHODS

Quality control of study genotypes

The quality control (QC) was performed in four steps on each the 8 cohorts separately. First, a sample QC was conducted, followed by a SNP QC, the identification population outliers by principal components analysis (PCA) and a cohort-based batch QC. Individuals with missingness > 5%, outlying heterozygosity (not within interval [mean +3·sd, mean -3·sd], sd: standard deviation), with a kinship coefficient (IBD) > 0.185 and those failing gender check were removed from the data set. The kinship analysis was performed as follow: First, possible related samples were identified using PLINK (2, 3) –genome and its method of moments (MOM) estimator. Related samples were then further analyzed with the R-package SNPRelate (version 1.2.0) based on the more computationally intensive calculation of a maximum likelihood estimator (MLE) and related samples were then finally identified using this MLE estimation. The gender check was conducted by counting heterozygous calls on X and Y and plotting the sum. A cluster analysis (based on k-means clustering) was performed on the counts using the R function stats::kmeans (2 centers, 10 iterations, nstart 1) and incorrectly assigned samples (gender and cluster assignment not identical) were removed. SNPs with missingness of > 5% and a deviation from Hardy-Weinberg equilibrium (HWE) $P < 0.00001$ in controls were excluded. No minor allele frequency (MAF) threshold was set. SNPs with differential missingness between cases and controls, differential missingness between batches, and a deviation from HWE within the batches were noted. If a SNP failed one of the batch criteria it was set to missing. For the analysis of population stratification employing PCA, the data were LD-pruned using PLINK's (2, 3) -indep-pairwise 50 5 0.2. Additionally, several regions of high LD as suggested by REF were excluded (chr5:44Mb-51.5Mb, chr6:25Mb-33.5Mb, chr8:8Mb-12Mb, chr11:45Mb-57Mb). The MAF-threshold was set to 0.05 for this analysis. PCs 1 to 10 were calculated. Based on PC1 and PC2 distances of each sample from the “center point” [median(PC1), median(PC2)] were calculated. Outlying samples were defined as those with a Euclidean distance > median(distance to center point) + 3·IQR(distance to center point) were excluded accordingly. For samples of Indian ancestry we could identify 3 distinct subpopulations and chose the largest for further analysis.

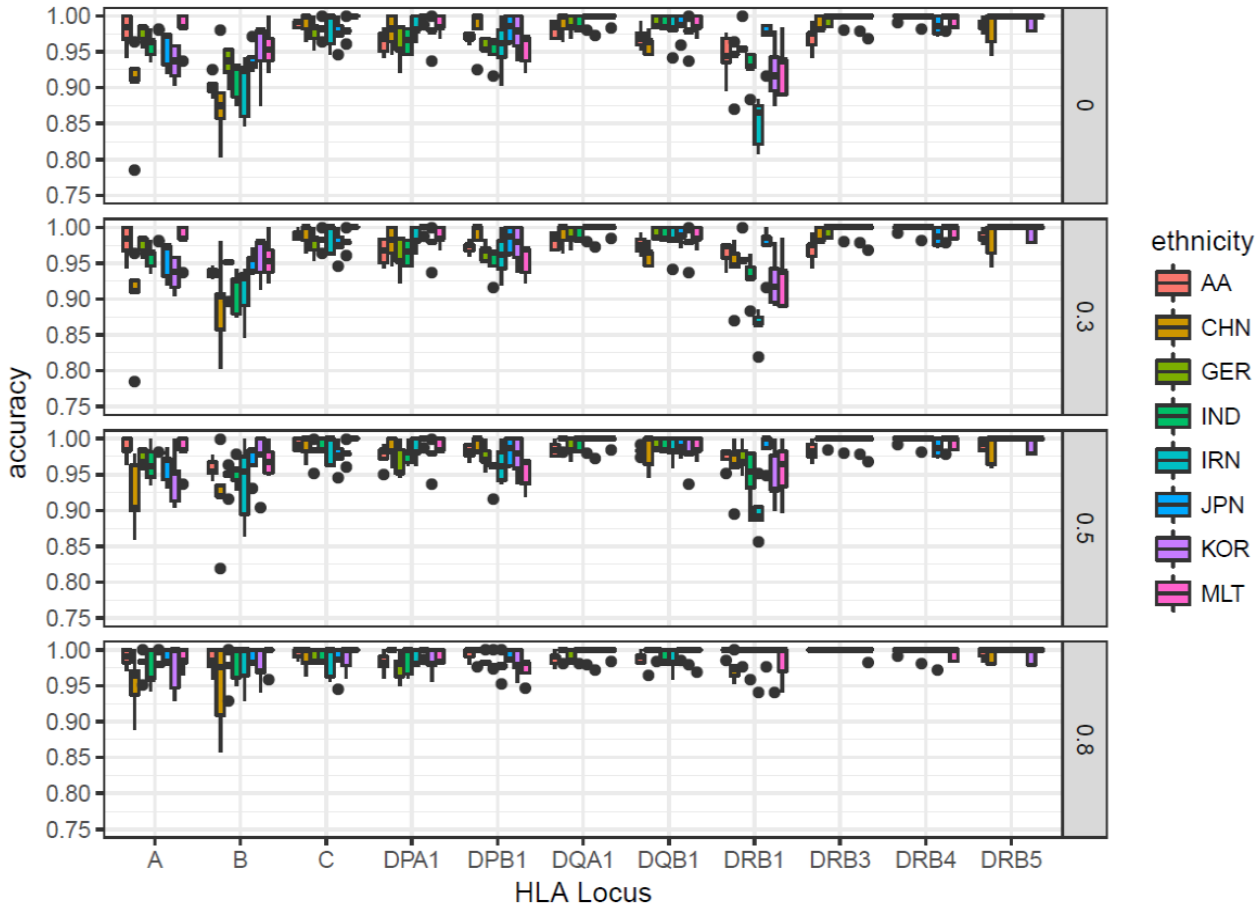
SUPPLEMENTARY FIGURES

Supplementary Figure 1 – Cross validation scheme for the HLA benchmarking: Number of samples used for training of the respective reference using the example of a cross-validation model applied to the African American data panel.

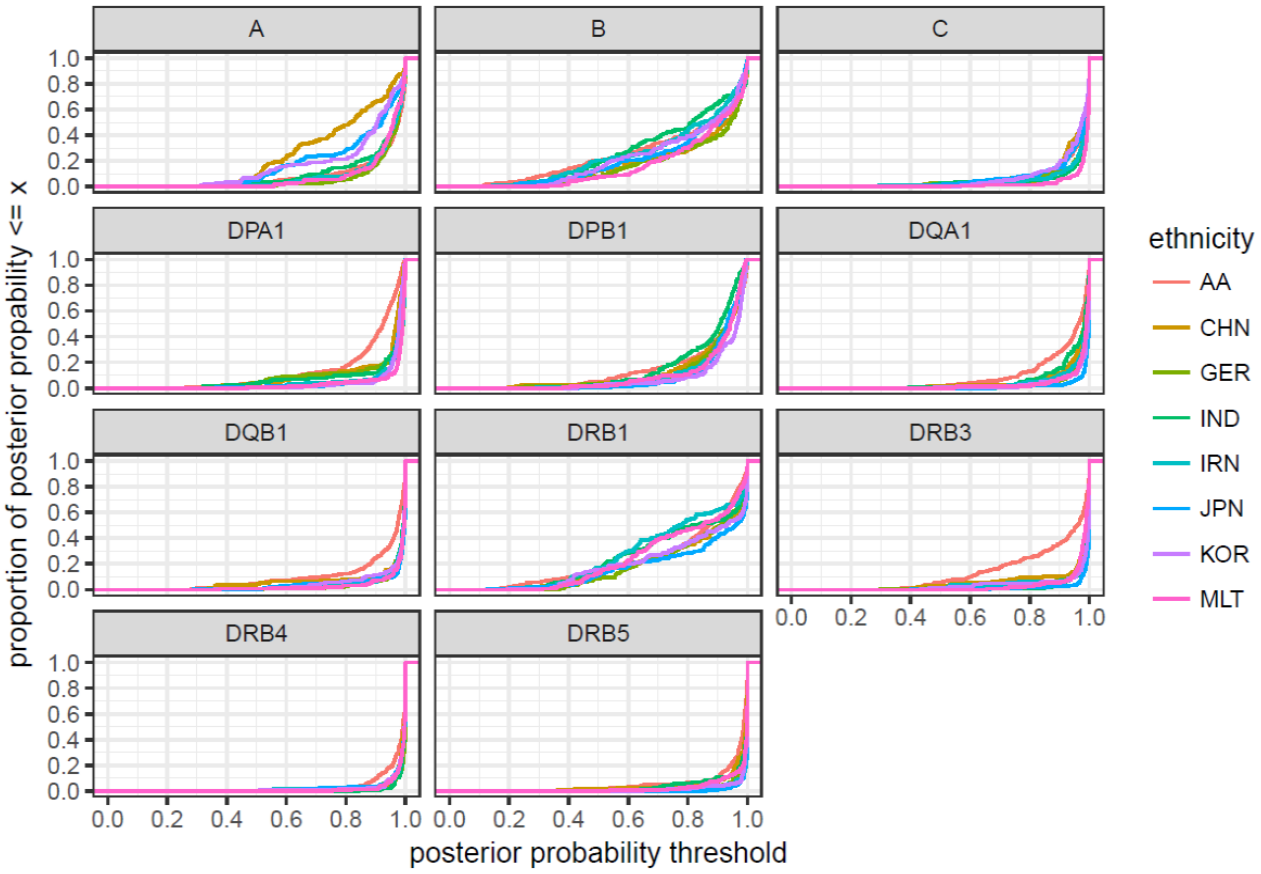


Supplementary Figure 2 – Imputation accuracies employing the multi-ethnic reference combined with the 1000 Genomes dataset: Accuracies and post-imputation probabilities of HLA imputation with HIBAG (1) using a 5x cross validation scheme and the trans-ethnic and 1000 Genomes dataset (4) with 4-digit G group allele information. 20% of the data with a specific ethnical background were used as the validation set after training a model with 80% of the remaining data and all data with other ethnical backgrounds. We used 1,360 African American (AA), Hong-Kong Chinese (CHN), Caucasian (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples and 937 samples from the 1000 Genomes dataset in total. **(a)** Accuracies are depicted according to post-imputation probabilities with cutoff thresholds at 0, 0.3, 0.5 and 0.8. Loci are shown according to alphabetical order. **(b)** Posterior probabilities are depicted as proportion of the number of samples with a posterior probability smaller than a threshold (x-axis).

(a)

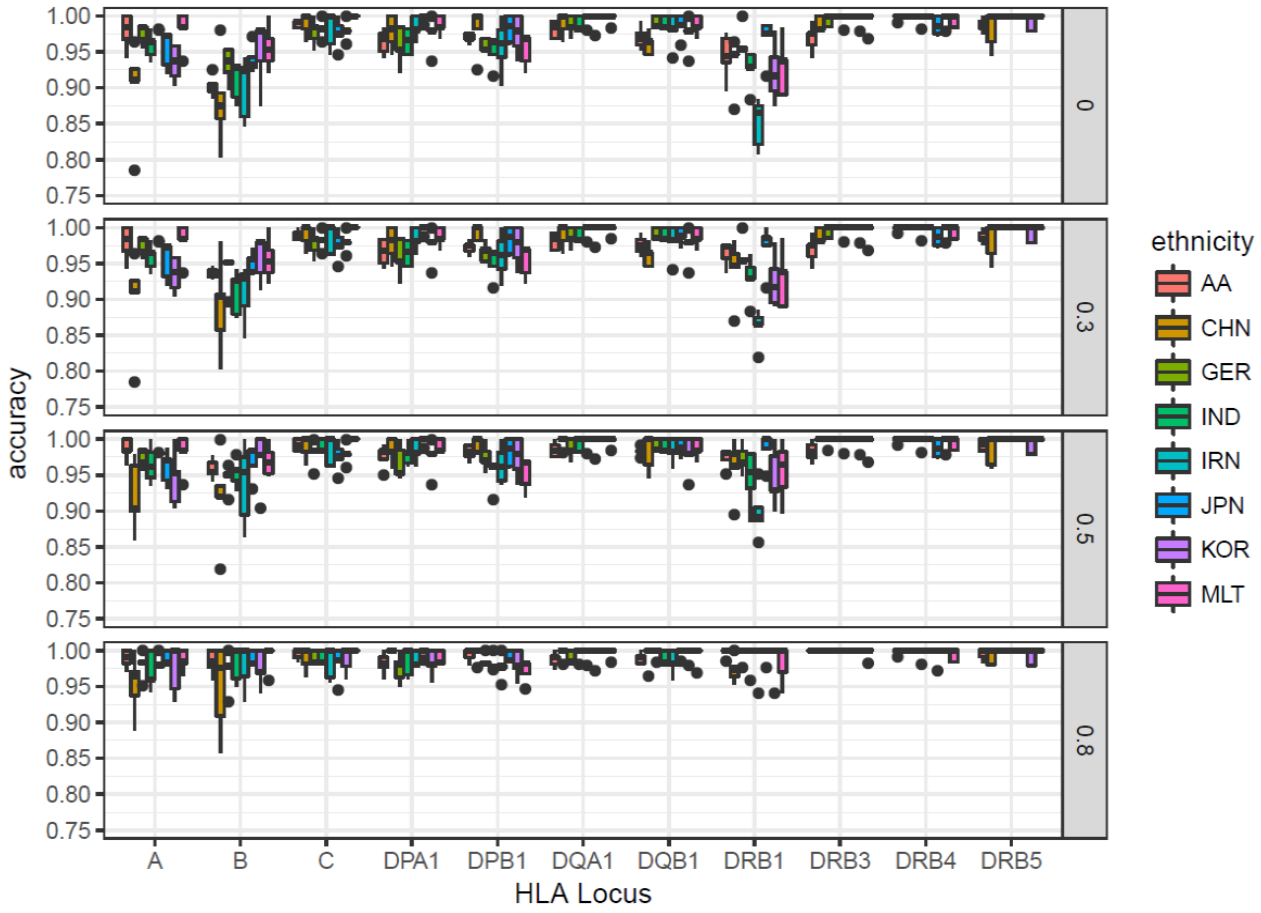


(b)

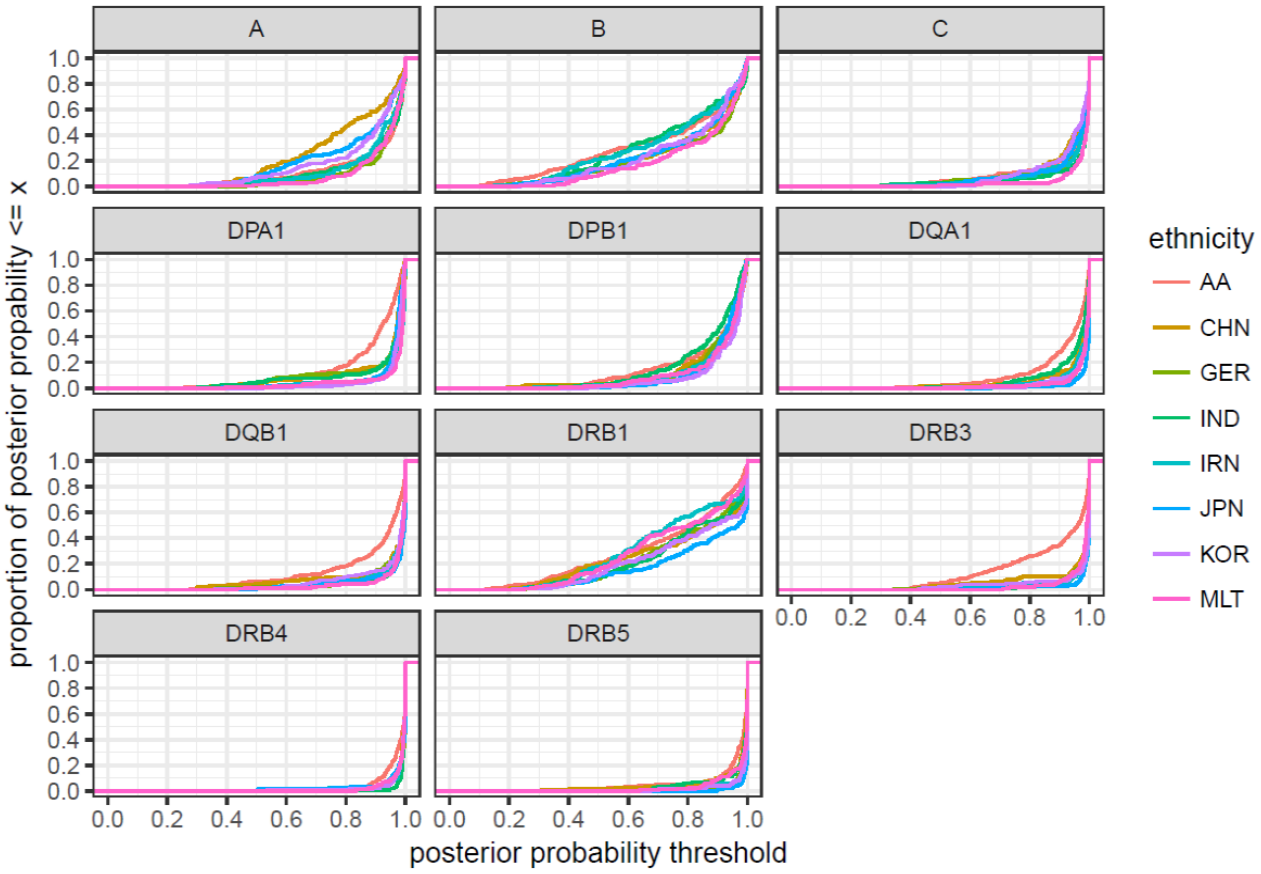

A
B
C
DPA1
DPB1
DQA1
DQB1
DRB1
DRB3
DRB4
DRB5
proportion of posterior propability <= x
posterior probability threshold
0.0
0.2
0.4
0.6
0.8
1.0
ethnicity
AA
CHN
GER
IND
IRN
JPN
KOR
MLT

Supplementary Figure 3 – Imputation accuracies employing the multi-ethnic reference panel in G group context: Accuracies and post-imputation probabilities of HLA imputation with HIBAG using a 5x cross validation scheme and the multi-ethnic dataset with 4-digit G group allele information. 20% of the data with a specific ethnical background were used as the validation set after training a model with 80% of the remaining data and all data with other ethnical backgrounds. We used 1,360 African American (AA), Hong-Kong Chinese (CHN), Caucasian (GER), Indian (IND), Iranian (IRN), Japanese (JPN), South Korean (KOR) and Maltese (MLT) samples in total. **(a)** Accuracies are depicted according to post-imputation probabilities with cutoff thresholds at 0, 0.3, 0.5 and 0.8. Loci are shown according to alphabetical order. **(b)** Posterior probabilities are depicted as proportion of the number of samples with a posterior probability smaller than a threshold (x-axis).

(a)

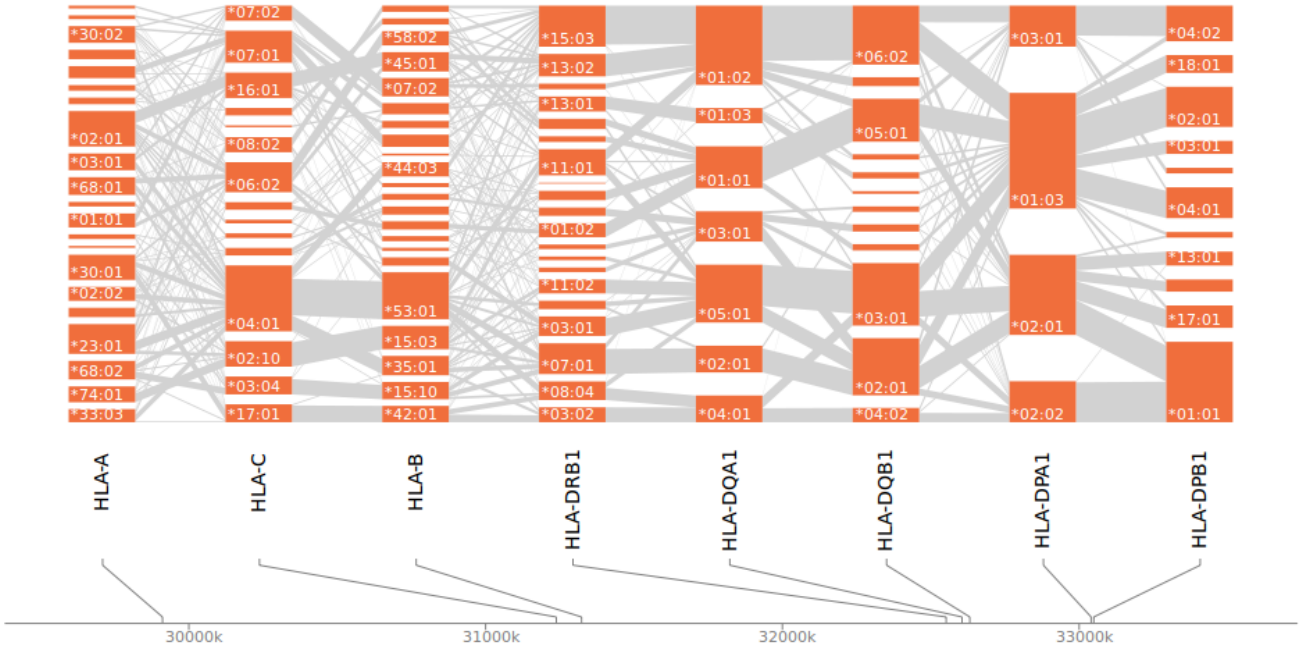


(b)

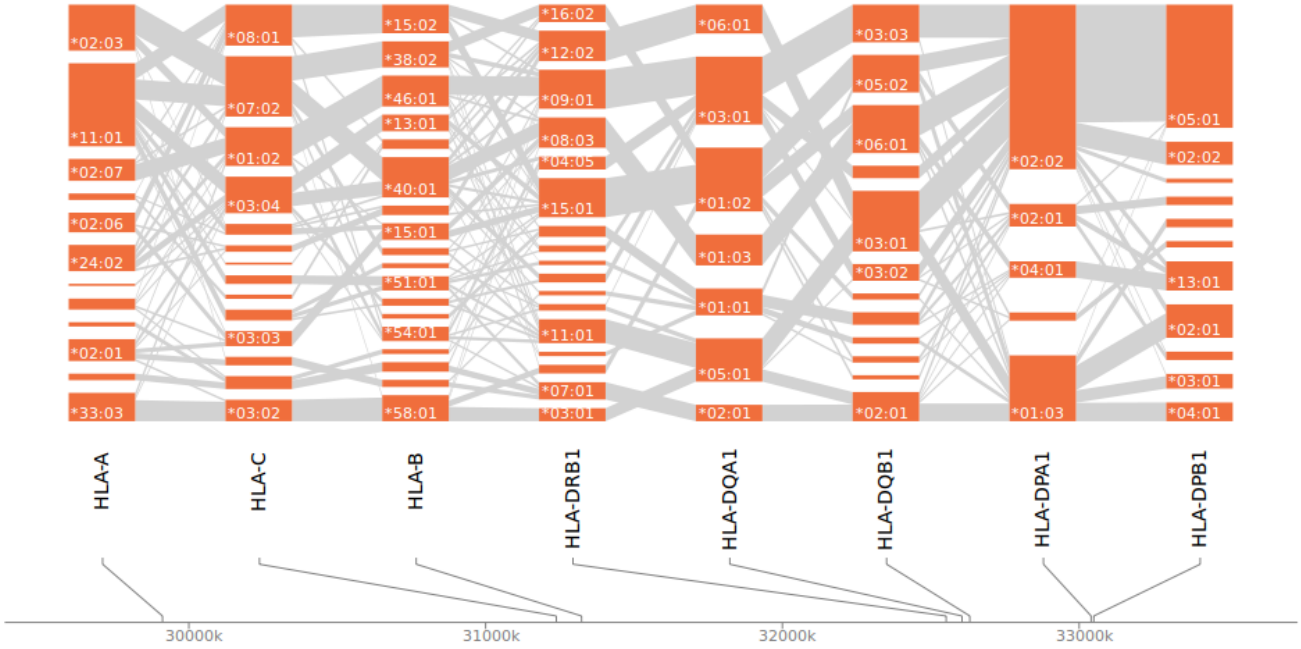

A
B
C
DPA1
DPB1
DQA1
DQB1
DRB1
DRB3
DRB4
DRB5
proportion of posterior propability <= x
posterior probability threshold
0.0
0.2
0.4
0.6
0.8
1.0
ethnicity
AA
CHN
GER
IND
IRN
JPN
KOR
MLT

Supplementary Figure 4 – Disentenglar plots: Disentenglar (5) plot of alleles with a MAF >1% for **(a)** African American (AA), **(b)** Chinese (CHN), **(c)** European (EUR), **(d)** Indian (IND), **(e)** Iranian (IRN), **(f)** South Korean (KOR), **(g)** Japanese (JPN) and **(h)** Maltese (MLT) data. Typing was performed on a 4-digit level using HLAssign (6). Plot shows frequencies as height of the bar and haplotype connections as grey lines. HLA loci are ordered by genomic location.

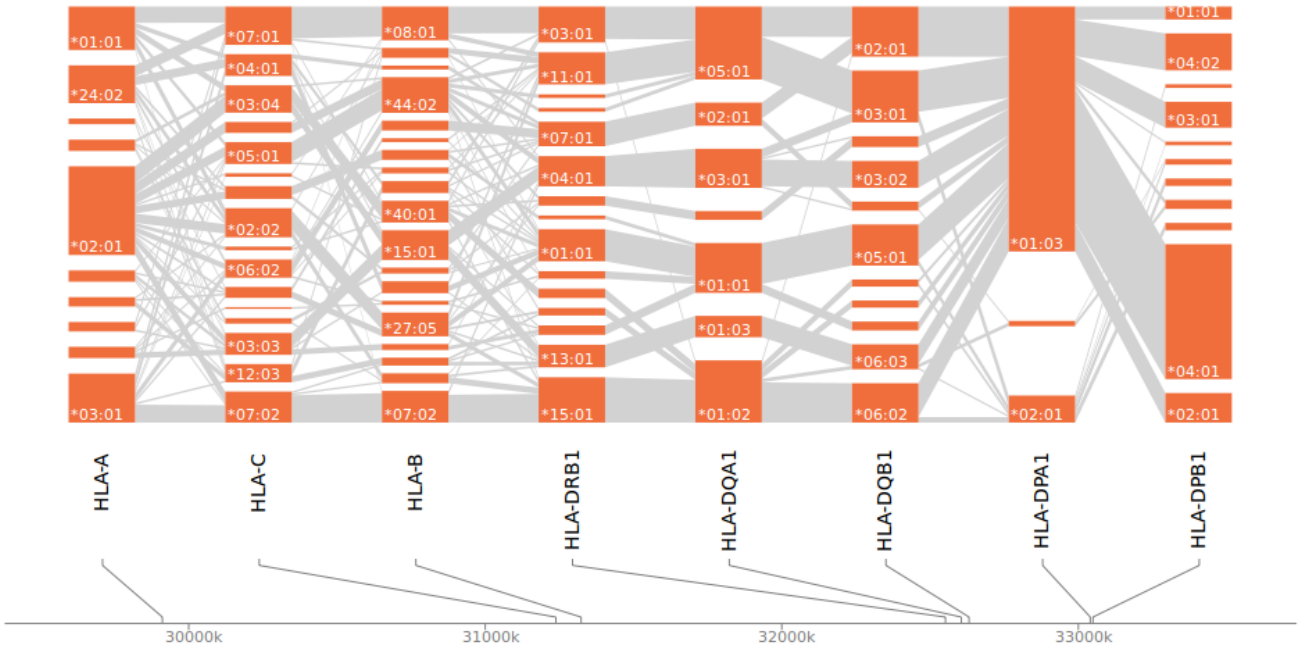
(a) AA



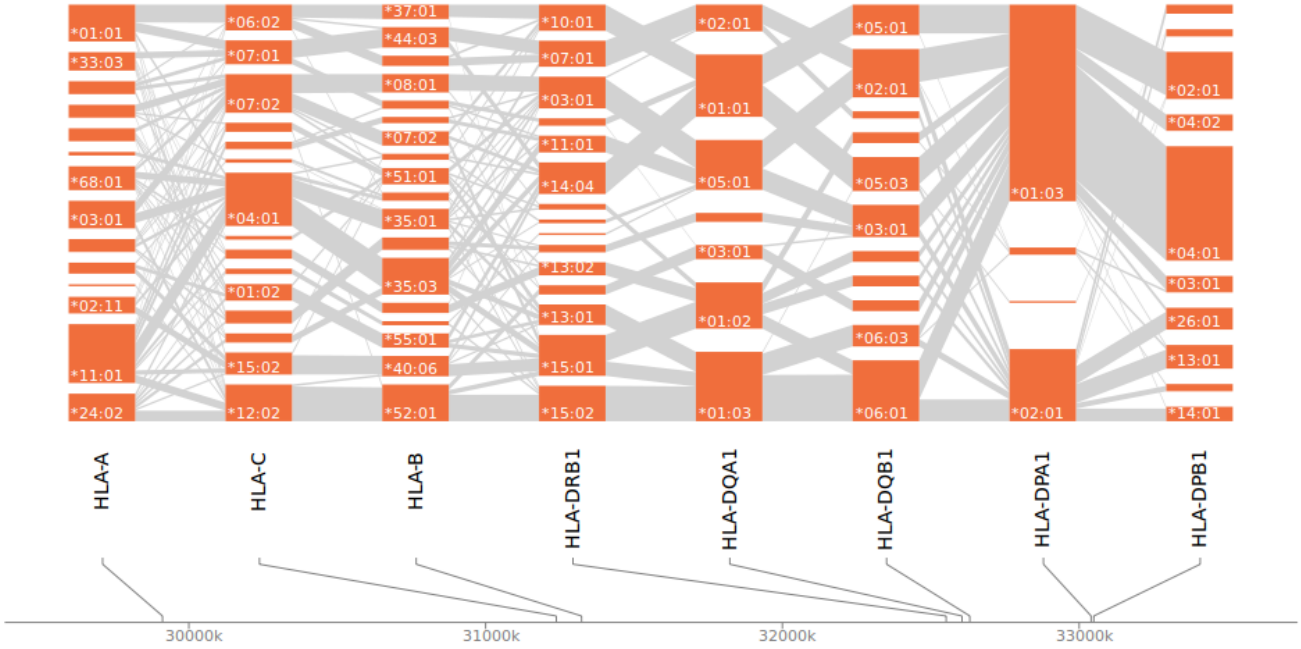
(b) CHN


*02:03
*11:01
*02:07
*02:06
*24:02
*02:01
*33:03
*08:01
*07:02
*01:02
*03:04
*03:03
*03:02
*15:02
*38:02
*46:01
*13:01
*40:01
*15:01
*51:01
*54:01
*58:01
*16:02
*12:02
*09:01
*08:03
*04:05
*15:01
*11:01
*07:01
*03:01
*06:01
*03:01
*01:02
*01:03
*01:01
*05:01
*02:01
*03:03
*05:02
*06:01
*03:01
*03:02
*02:01
*02:02
*02:01
*04:01
*01:03
*05:01
*02:02
*13:01
*02:01
*03:01
*04:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

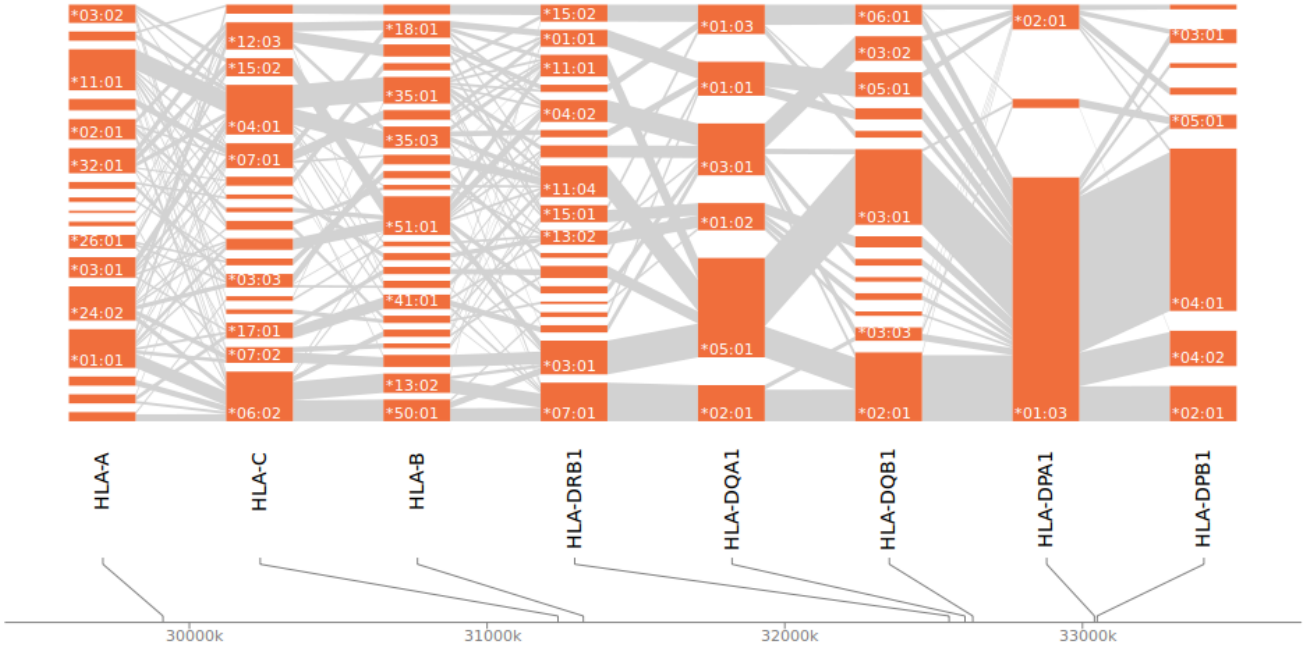
(c) GER


*01:01
*24:02
*02:01
*03:01
*07:01
*04:01
*03:04
*05:01
*02:02
*06:02
*03:03
*12:03
*07:02
*08:01
*44:02
*40:01
*15:01
*27:05
*07:02
*03:01
*11:01
*07:01
*04:01
*01:01
*13:01
*15:01
*05:01
*02:01
*03:01
*01:01
*01:03
*01:02
*02:01
*03:01
*03:02
*05:01
*06:03
*06:02
*01:03
*02:01
*01:01
*04:02
*03:01
*04:01
*02:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

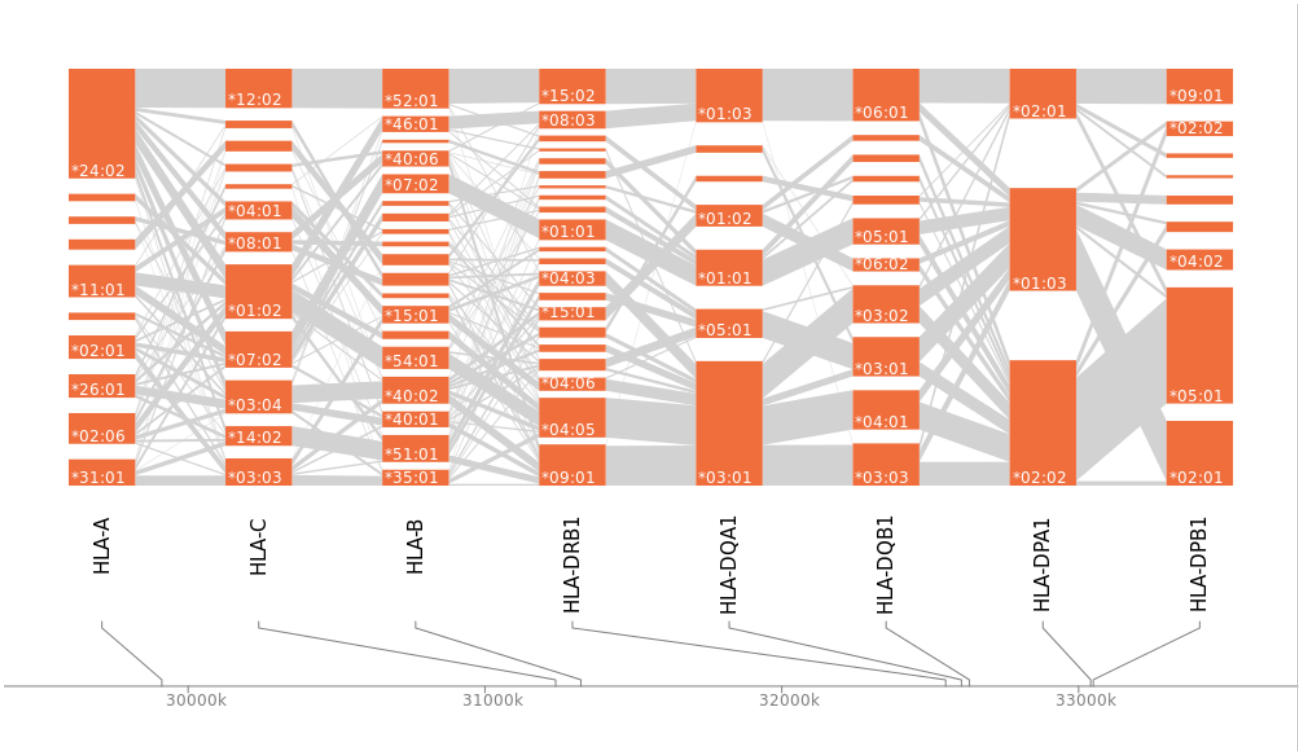
(d) IND


*01:01
*33:03
*68:01
*03:01
*02:11
*11:01
*24:02
*06:02
*07:01
*07:02
*04:01
*01:02
*15:02
*12:02
*37:01
*44:03
*08:01
*07:02
*51:01
*35:01
*35:03
*55:01
*40:06
*52:01
*10:01
*07:01
*03:01
*11:01
*14:04
*13:02
*13:01
*15:01
*15:02
*02:01
*01:01
*05:01
*03:01
*01:02
*01:03
*05:01
*02:01
*05:03
*03:01
*06:03
*06:01
*01:03
*02:01
*02:01
*04:02
*04:01
*03:01
*26:01
*13:01
*14:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

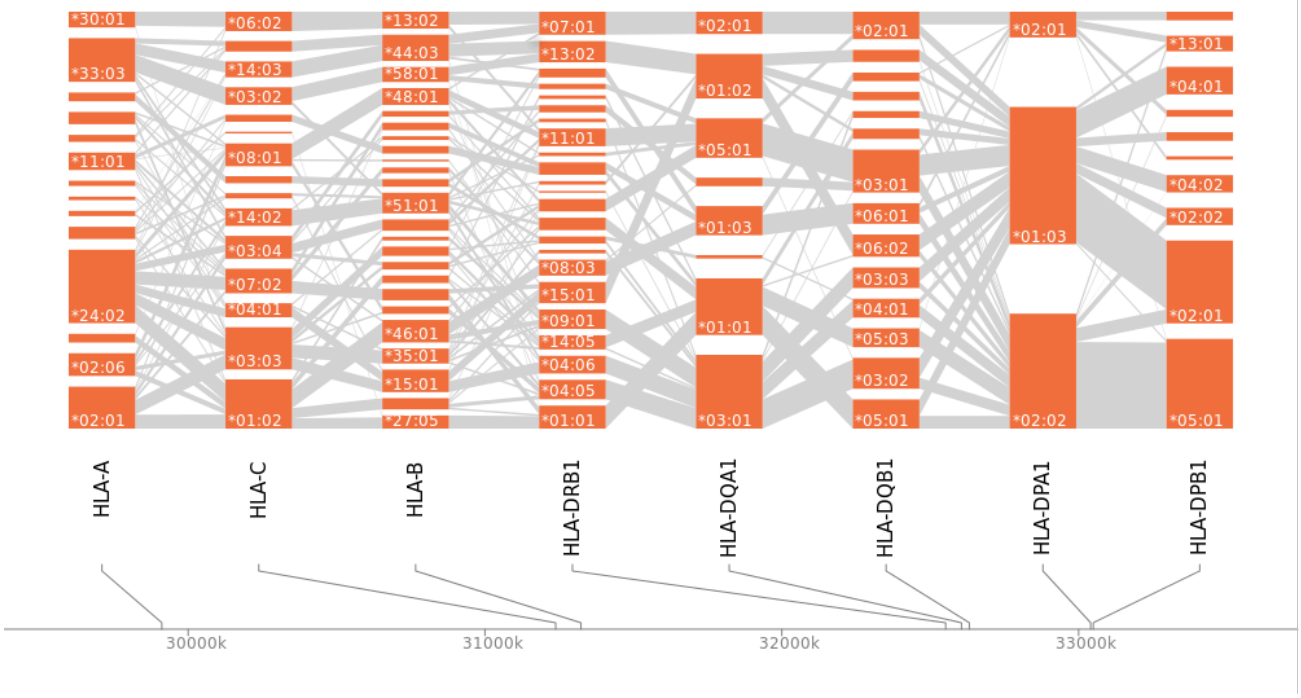
(e) IRN


*03:02
*11:01
*02:01
*32:01
*26:01
*03:01
*24:02
*01:01
*12:03
*15:02
*04:01
*07:01
*03:03
*17:01
*07:02
*06:02
*18:01
*35:01
*35:03
*51:01
*41:01
*13:02
*50:01
*15:02
*01:01
*11:01
*04:02
*11:04
*15:01
*13:02
*03:01
*07:01
*01:03
*01:01
*03:01
*01:02
*05:01
*02:01
*06:01
*03:02
*05:01
*03:01
*03:03
*02:01
*02:01
*01:03
*03:01
*05:01
*04:01
*04:02
*02:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

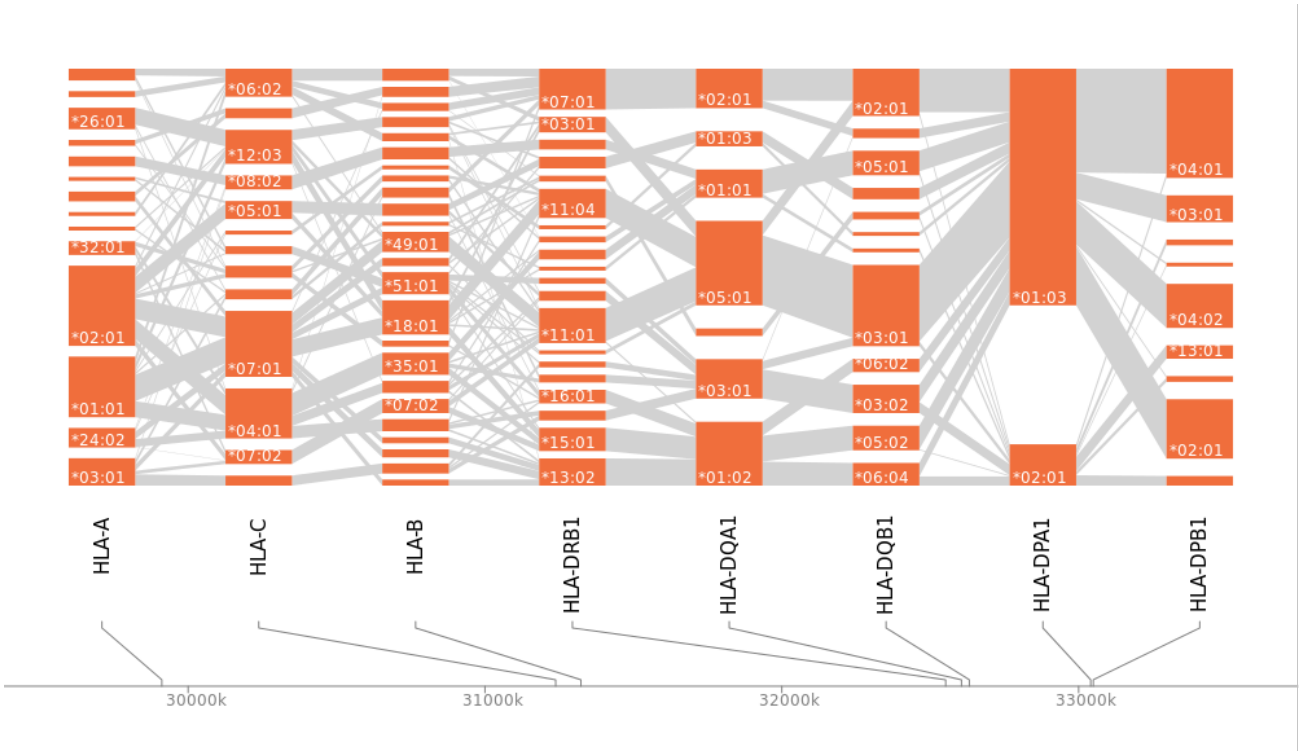
(f) JPN


*24:02
*11:01
*02:01
*26:01
*02:06
*31:01
*12:02
*04:01
*08:01
*01:02
*07:02
*03:04
*14:02
*03:03
*52:01
*46:01
*40:06
*07:02
*15:01
*54:01
*40:02
*40:01
*51:01
*35:01
*15:02
*08:03
*01:01
*04:03
*15:01
*04:06
*04:05
*09:01
*01:03
*01:02
*01:01
*05:01
*03:01
*06:01
*05:01
*06:02
*03:02
*03:01
*04:01
*03:03
*02:01
*01:03
*02:02
*09:01
*02:02
*04:02
*05:01
*02:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

(g) KOR


*30:01
*33:03
*11:01
*24:02
*02:06
*02:01
*06:02
*14:03
*03:02
*08:01
*14:02
*03:04
*07:02
*04:01
*03:03
*01:02
*13:02
*44:03
*58:01
*48:01
*51:01
*46:01
*35:01
*15:01
*27:05
*07:01
*13:02
*11:01
*08:03
*15:01
*09:01
*14:05
*04:06
*04:05
*01:01
*02:01
*01:02
*05:01
*01:03
*01:01
*03:01
*02:01
*03:01
*06:01
*06:02
*03:03
*04:01
*05:03
*03:02
*05:01
*02:01
*01:03
*02:02
*13:01
*04:01
*04:02
*02:02
*02:01
*05:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

(h) MLT


*26:01
*32:01
*02:01
*01:01
*24:02
*03:01
*06:02
*12:03
*08:02
*05:01
*07:01
*04:01
*07:02
*49:01
*51:01
*18:01
*35:01
*07:02
*07:01
*03:01
*11:04
*11:01
*16:01
*15:01
*13:02
*02:01
*01:03
*01:01
*05:01
*03:01
*01:02
*02:01
*05:01
*03:01
*06:02
*03:02
*05:02
*06:04
*01:03
*02:01
*04:01
*03:01
*04:02
*13:01
*02:01
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

REFERENCES

1 Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R. and Weir, B.S. (2014) HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, **14**, 192-200.

2 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559-575.

3 Purcell, S., in press.

4 Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68-74.

5 Kumasaka N., O.Y., Takahashi A. , Kubo M. , Nakamura Y, Kamatani N. (2014), in press.

6 Wittig, M., Anmarkrud, J.A., Kassens, J.C., Koch, S., Forster, M., Ellinghaus, E., Hov, J.R., Sauer, S., Schimmler, M., Ziemann, M. *et al.* (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.*, **43**, e70.

B

SUPPLEMENT OF PAPER B

frontiers

Supplementary Material

1 Supplementary Methods

1.1 NNAlign

The final NNAlign model is an ensemble of 400 networks. The single networks are trained on one of the 10 cross-validation datasets with 10, 20, 40 and 60 hidden neurons and starting with 10 different initial configurations.

Each model was trained for 1200 cycles with a learning rate of 0.10 (eta), without early stopping. A burn-in period of 10 cycles was used where amino acid preference at position 1 is imposed (def: ILVMFYW). The peptide binding core length was set to 9 amino acids.

1.2 Deep Learning Model PIA

PIA is a gated recurrent neural network (GRU) based model (1) that is composite of an embedding layer which projects each amino acid into a continuous representational space of eight dimensions, followed by a GRU layer with 100 units, followed by a batch normalization layer (2) and finally an output neuron giving the prediction score of each peptide. A sigmoid activation function was used for the final layer to restrict the output to be between zero and one.

The model parameters were optimized using Adam (3) to minimize mean absolute error between the predictions and the true labels. Training was carried out for 500 epochs with a batch size of 1024 using an Nvidia Tesla P 100 GPU.

As the models were trained using a 10-fold cross-validation (CV) dataset scheme, a final ensemble of the 10 models, one per CV dataset, with the highest SCC on the validation dataset was constructed. The output of the ensemble was computed as the mean.

2 References

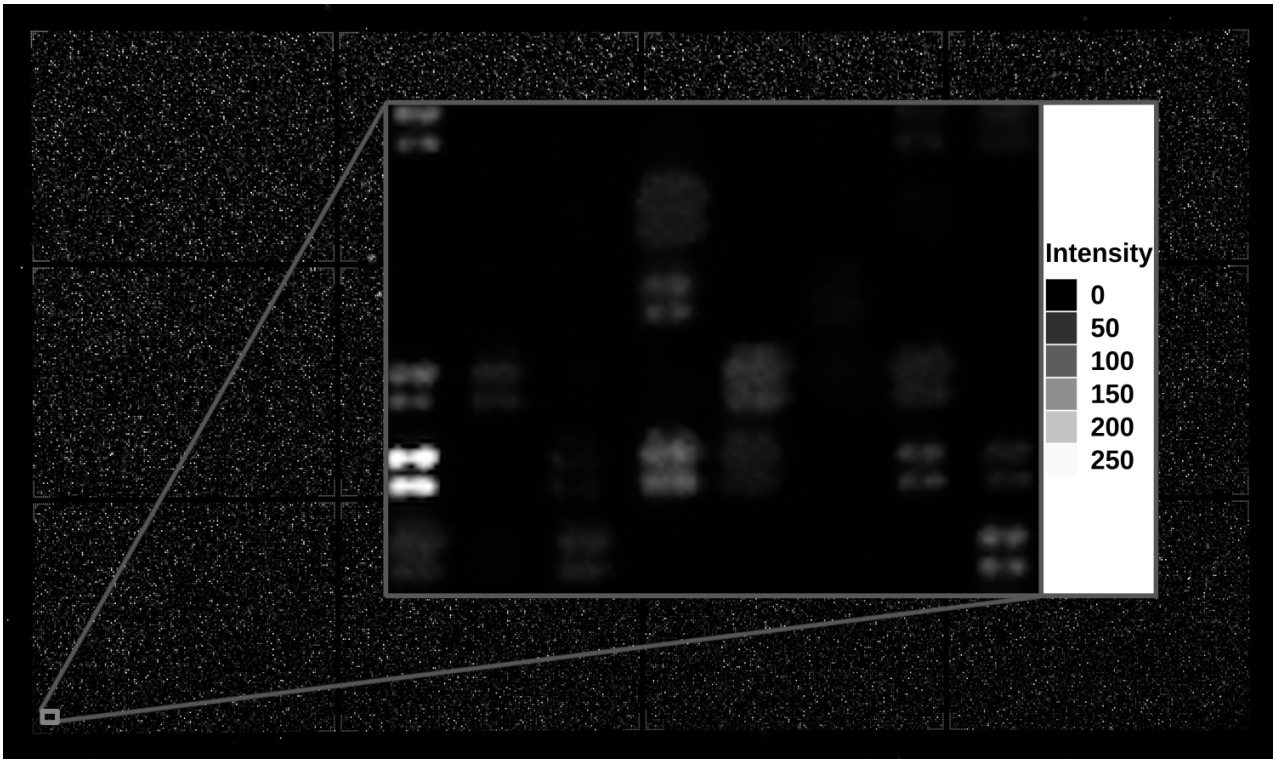
1. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf* (2014) Available online at: https://arxiv.org/abs/1406.1078v3

2. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd Int Conf Mach Learn ICML 2015* (2015) **1**:448–456.

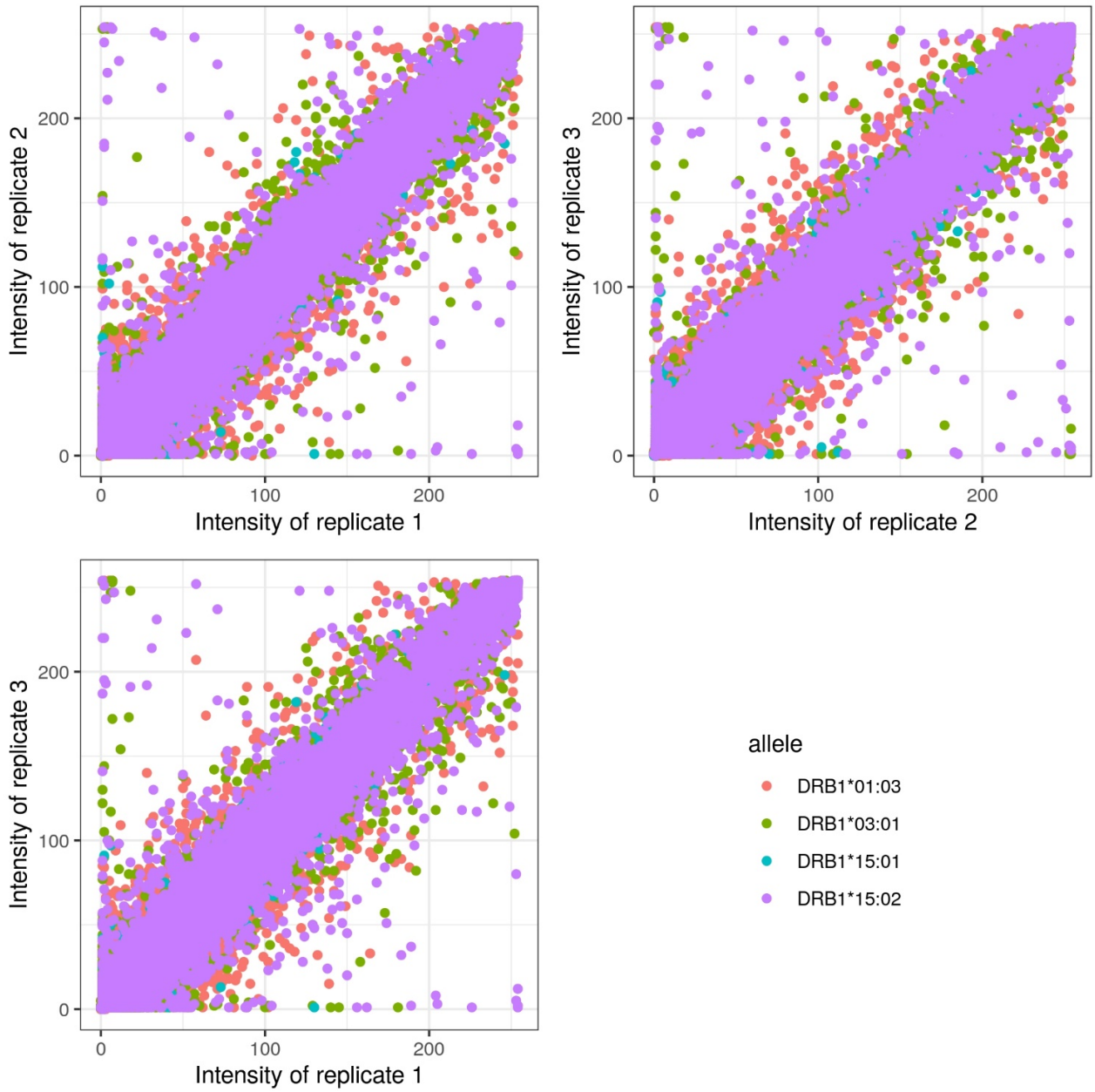
3. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc* (2015)1–15.

3 Supplementary Figures and Table

3.1 Supplementary Figures

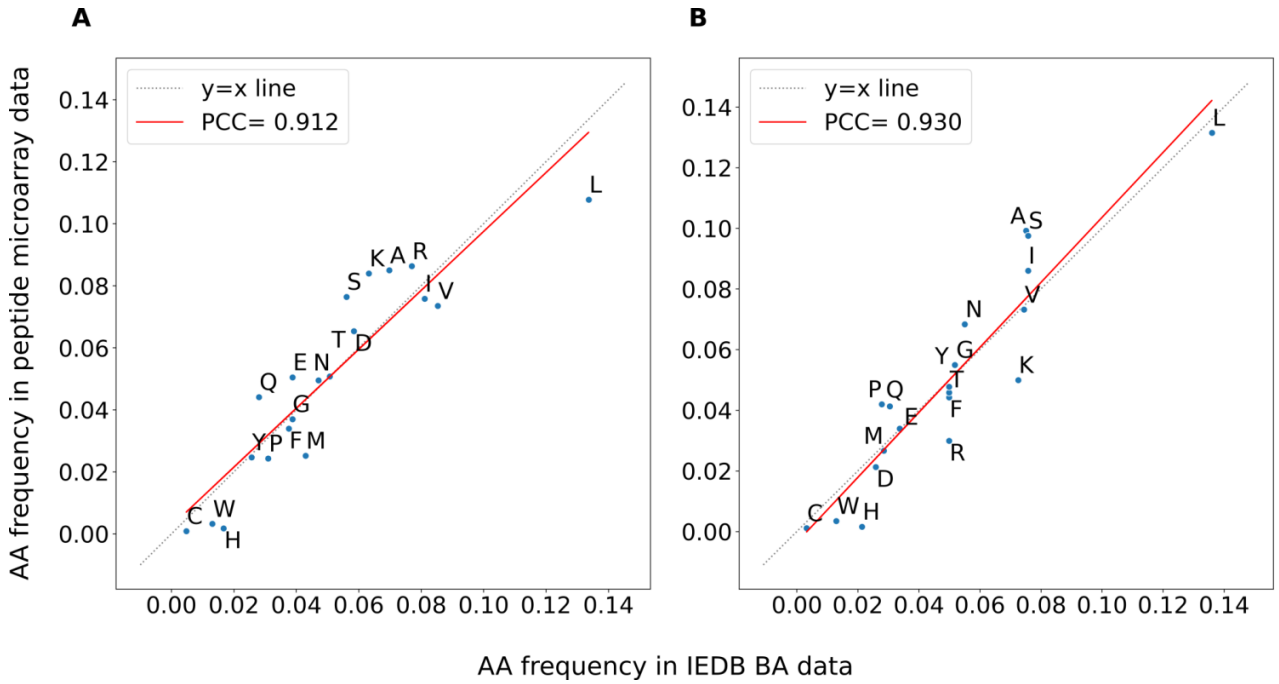


Supplementary Figure 1. Exemplary peptide microarray raw readout. The background figure shows the complete readout with the twelve subfields separated through an empty region (black). The corners of the fields were marked with the CLIP marker peptide. As CLIP is a good binding peptide, the fields are visible as bright angles. Additionally, a zoom-in of a small area is shown. The peptides were synthesized in fields defined by 2 x 2 mirrors and a one-mirror-wide empty region separated the individual peptide fields.



Supplementary Figure 2. Correlation of the three replicates. The measurements get their number of replicate randomly.

Supplementary Material



Supplementary Figure 3. Amino acid frequencies in binding peptides. Comparison of the amino acid frequency of the peptide microarray data against the IEDB binding affinity (BA) data for alleles DRB1*03:01 **(A)** and DRB1*15:01 **(B)**. The grey dotted line represents the identity line (y=x) and the red line is the fitted linear regression line (Pearson correlation coefficient shown in legend). Only the top 2% binding peptides from both data sources are included in the analysis.

3.2 Supplementary Table

Supplementary Table 1. Summary of positive test data downloaded from IEDB. The numbers in brackets are those remaining after applying the Frank filter.

DRB1	*01:03	*03:01	*15:01	*15:02	Σ
Epitopes	11 (10)	224 (145)	242 (152)	25 (14)	502 (321)
MS ligands	126 (62)	114 (65)	479 (373)	0	719 (500)

C

SUPPLEMENT OF PAPER C

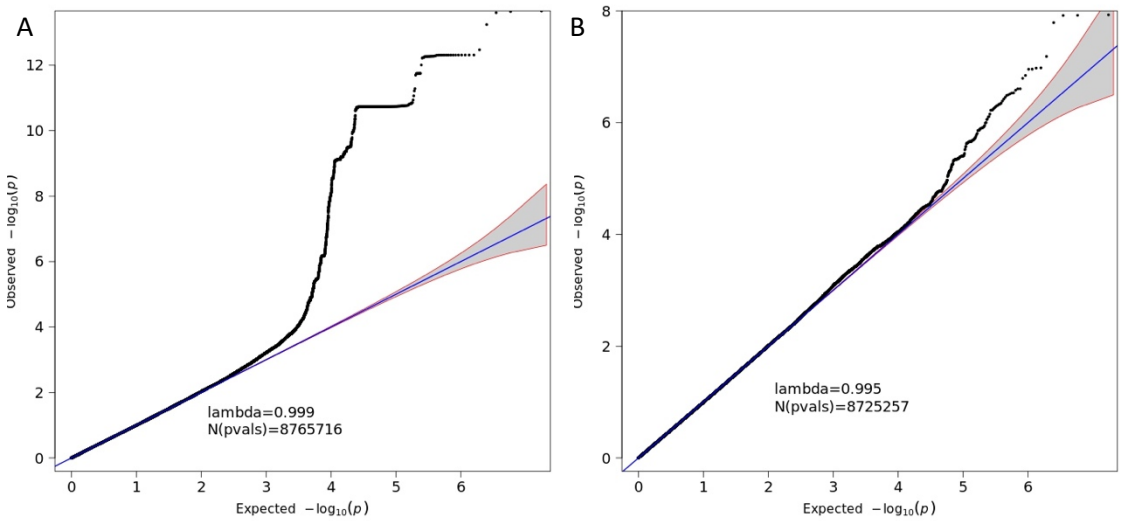
Genome-wide analysis of individual coding variants and HLA-II-associated self-immunopeptidomes in ulcerative colitis – Supplementary Material

Supplementary Figures:

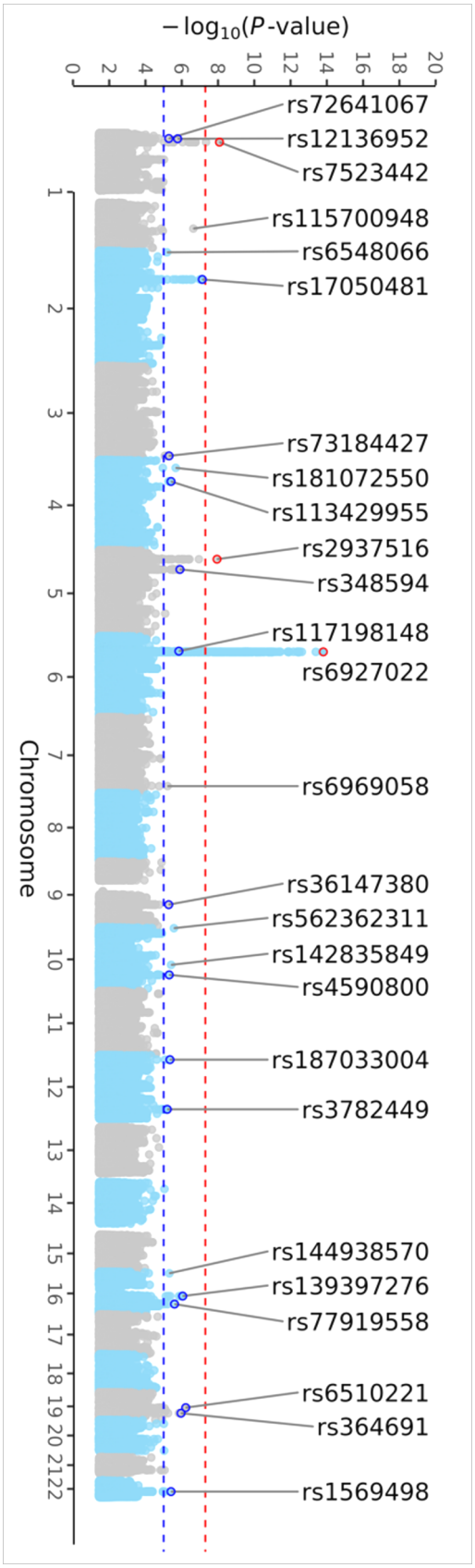
Supplementary Tables:

Supplementary References

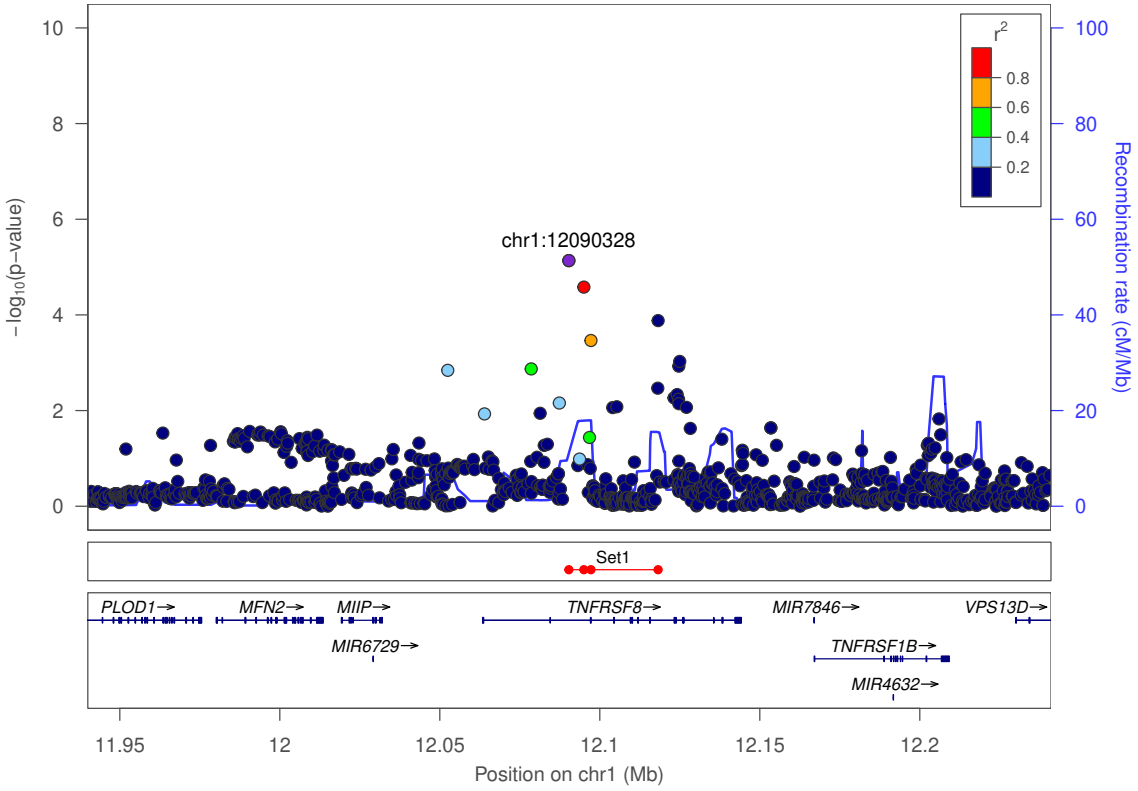
Supplementary Figures



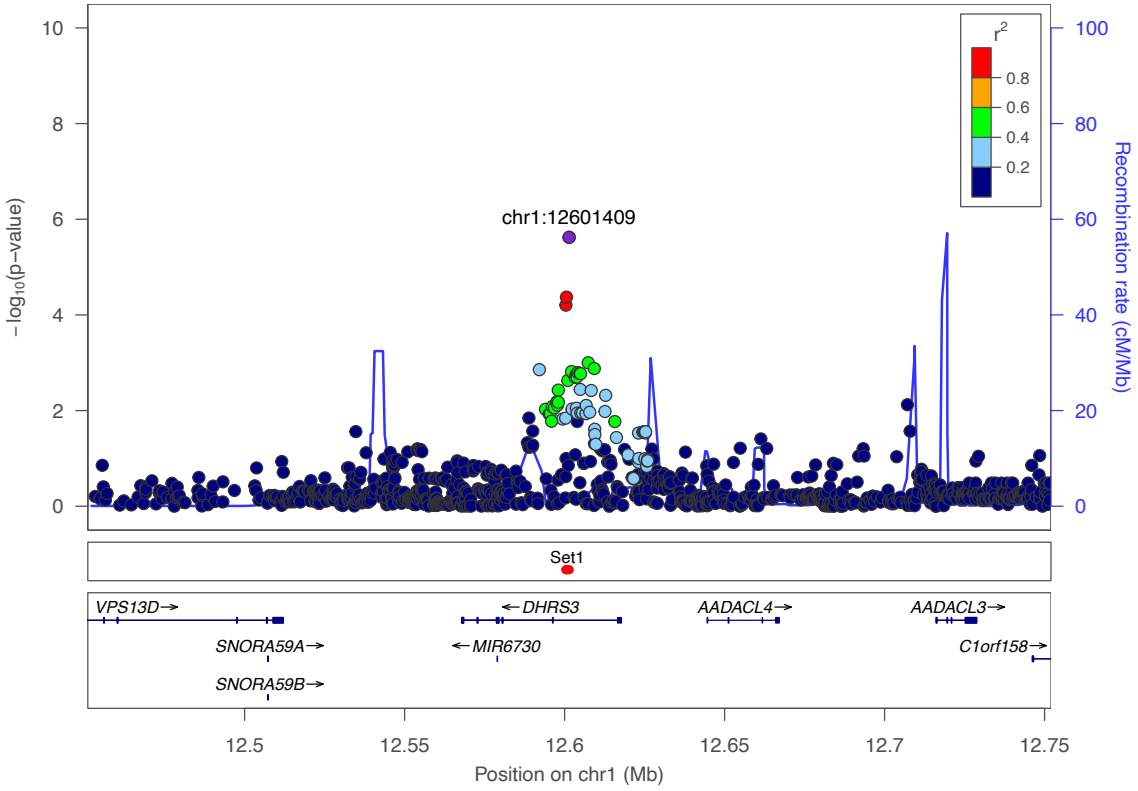
Supplementary Figure 1: *Quantile-quantile plot of association summary statistics of the imputed genotyping data. The 95% concentration band under random sampling is shown in gray. The genomic inflation factor lambda is defined as the ratio of the medians of the sample χ2 test statistics and the 1-df χ2 distribution (0.455).*[1] *Panel* ***(A)*** *includes all 8,765,716 variants with MAF > 1 % and an imputation score* r^2*>0.6. Panel* ***(B)*** *excludes the variants of the HLA-region (chr6:29-34MB).*



Supplementary Figure 2: *Manhattan plot of the imputed genotyping data with a MAF > 1 % and an imputation score r^2>0.6. All loci of at least nominal significance (blue horizontal line; P<1×10^{-5}) are annotated by the SNP-ID. Loci with LD support are highlighted with a blue (nominal significance) or red circle (genome-wide significance, red horizontal line; P<5×10^{8}).*

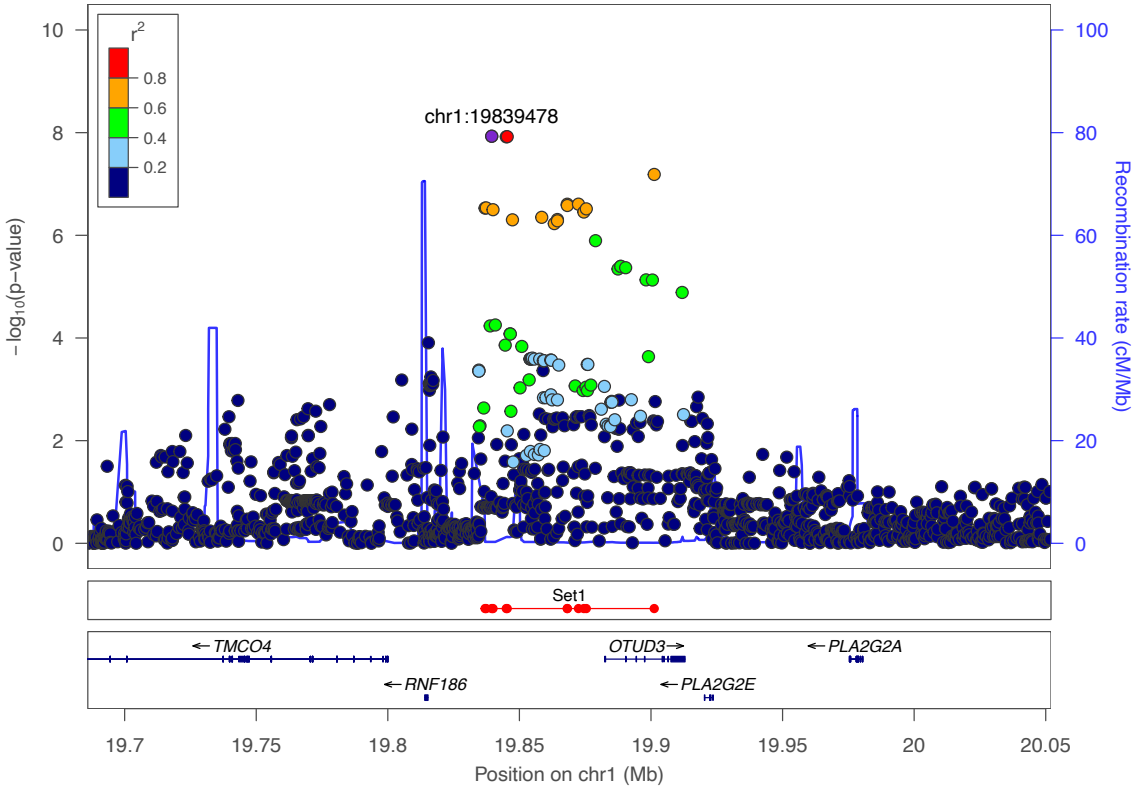


Supplementary Figure 3: *Regional association plot for 1p36.22 with the top hit chr1:12090328 (rs72641067). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*

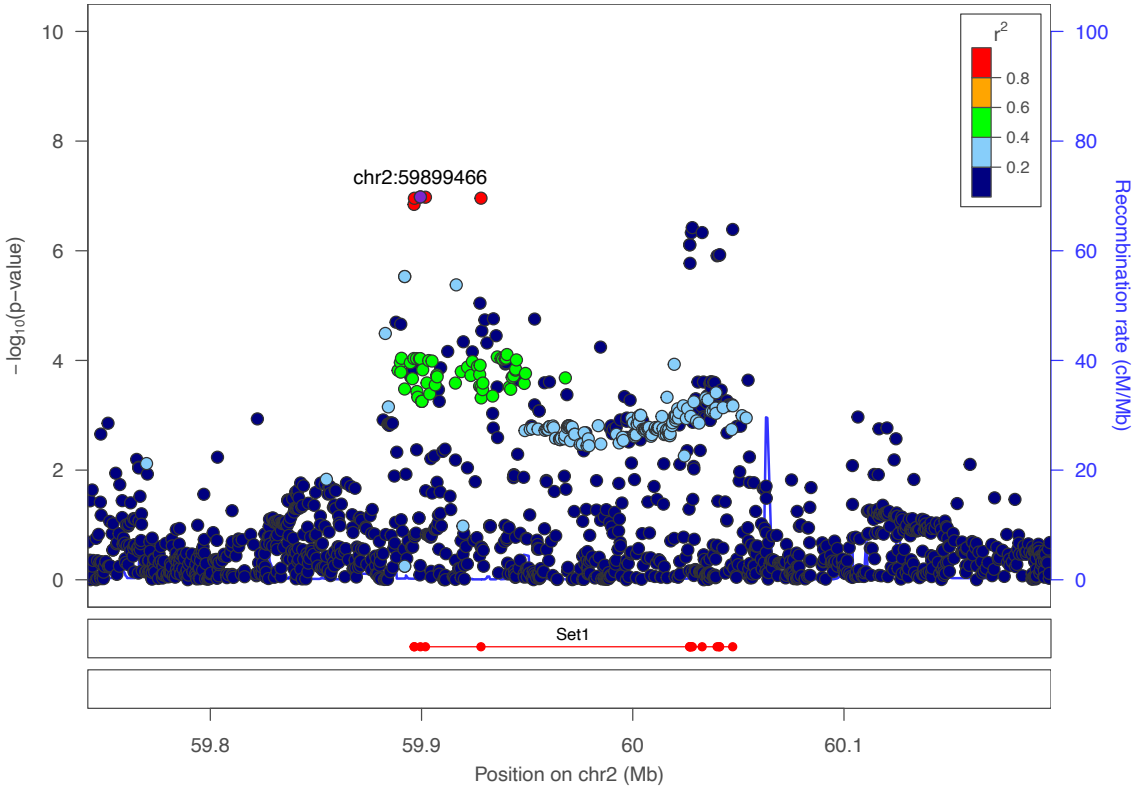


Supplementary Figure 4: *Regional association plot for the locus 1p36.22 with the top hit chr1:12601409(rs12136952). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the*

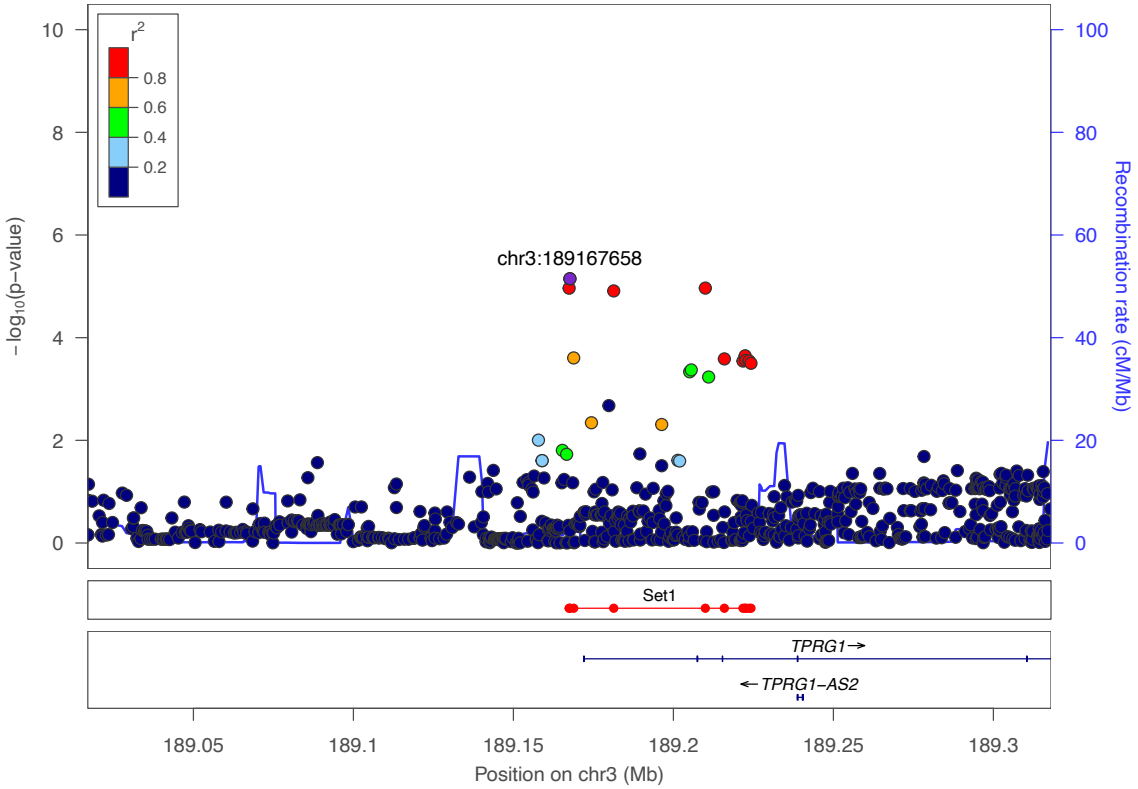
gene. The plot was created using LocusZoom[2].



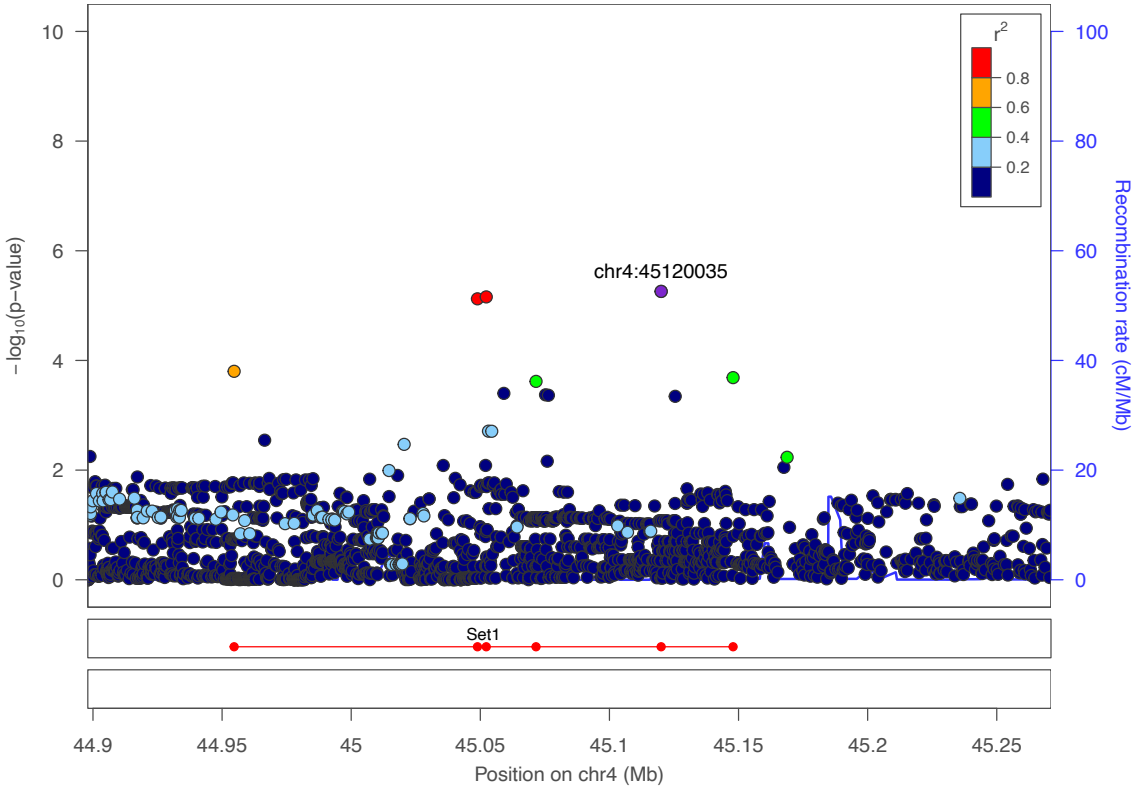
Supplementary Figure 5: *Regional association plot for the locus 1p36.13 with the top hit chr1:19839478 (rs7523442). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



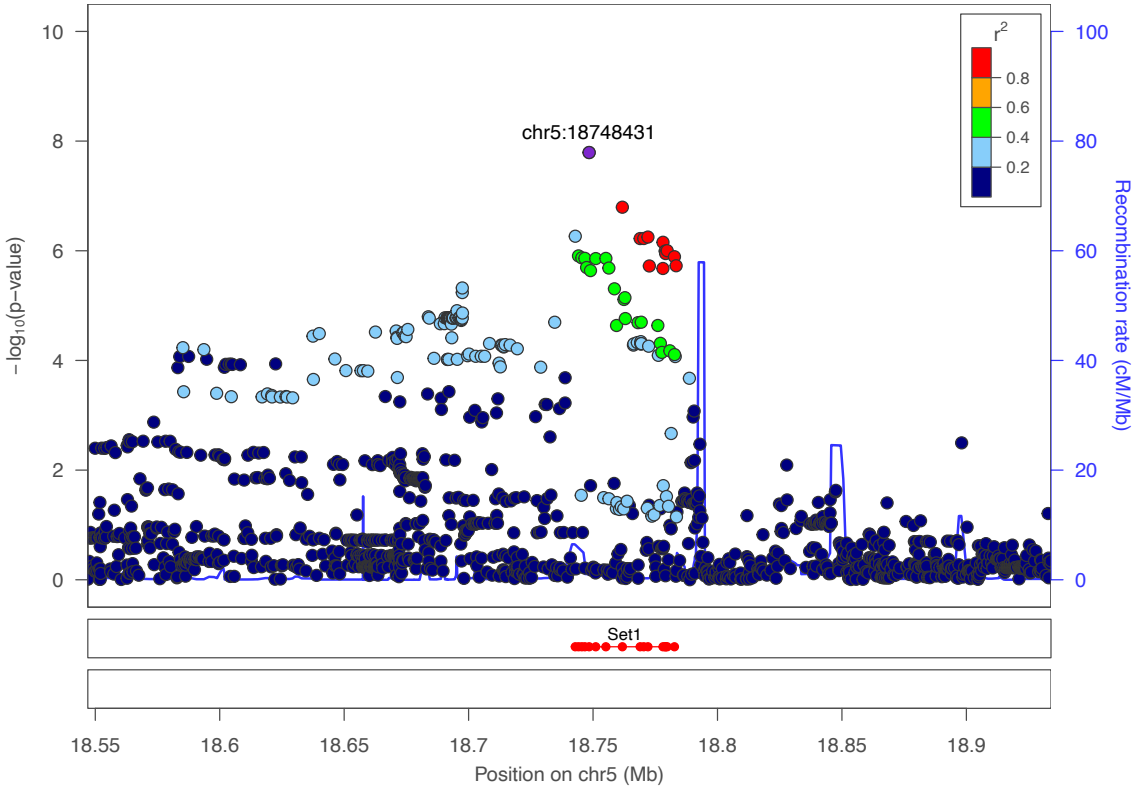
Supplementary Figure 6: *Regional association plot for the locus 2p16.1 with the top hit chr2:59899466 (rs17050481). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



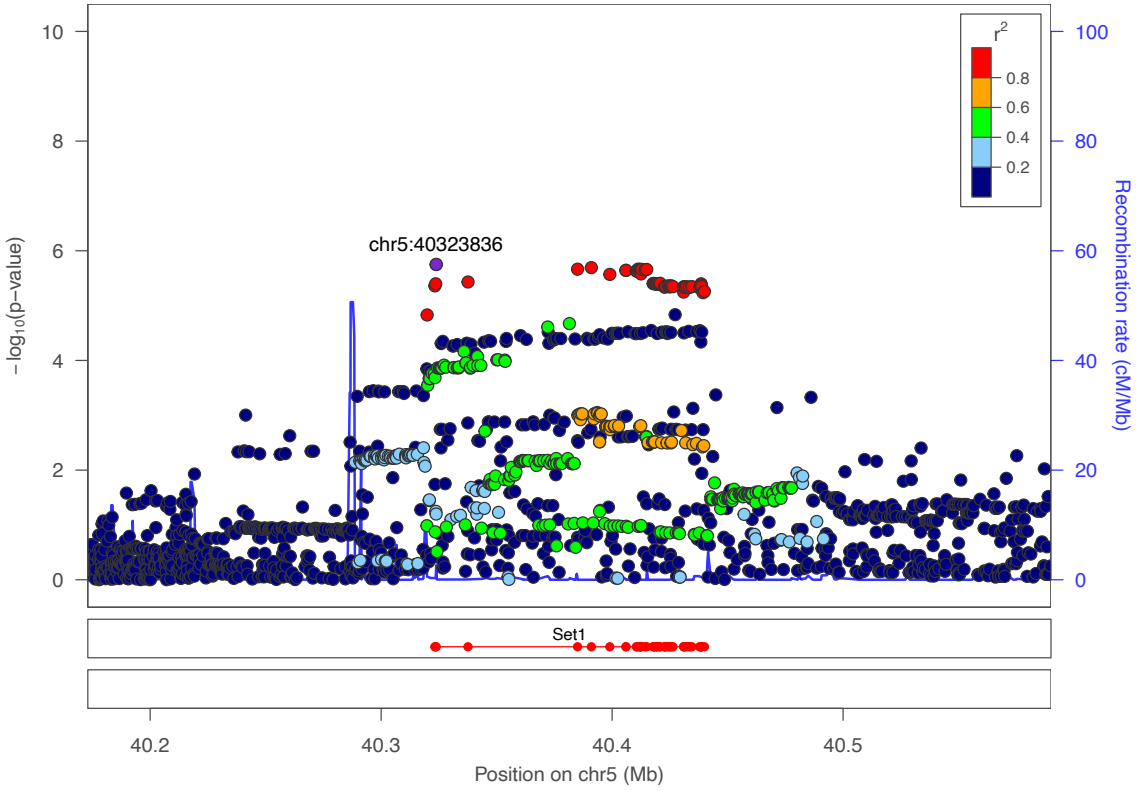
Supplementary Figure 7: *Regional association plot for the locus 3q28 with the top hit chr3:189167658 (rs73184427). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



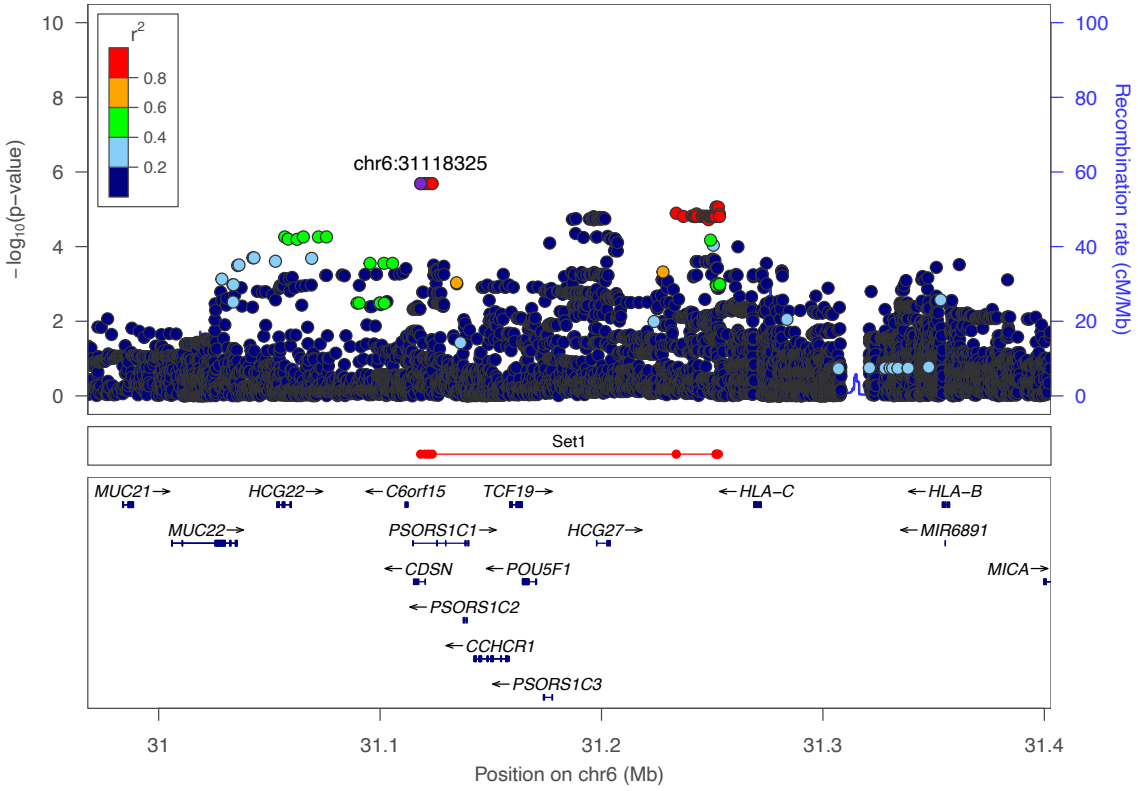
Supplementary Figure 8: *Regional association plot for the locus 4p12 with the top hit chr4:45120035 (rs113429955). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



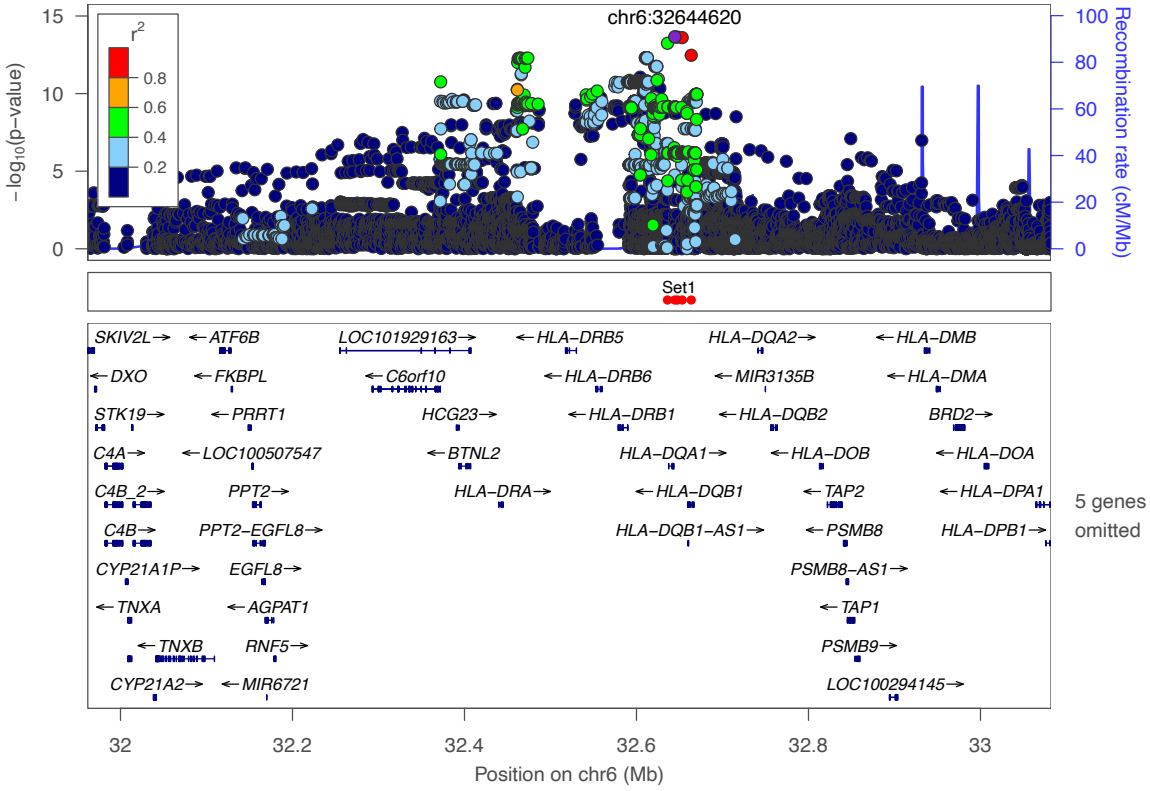
Supplementary Figure 9: *Regional association plot for the locus 5p14.3 with the top hit chr5:18748431 (rs2937516). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



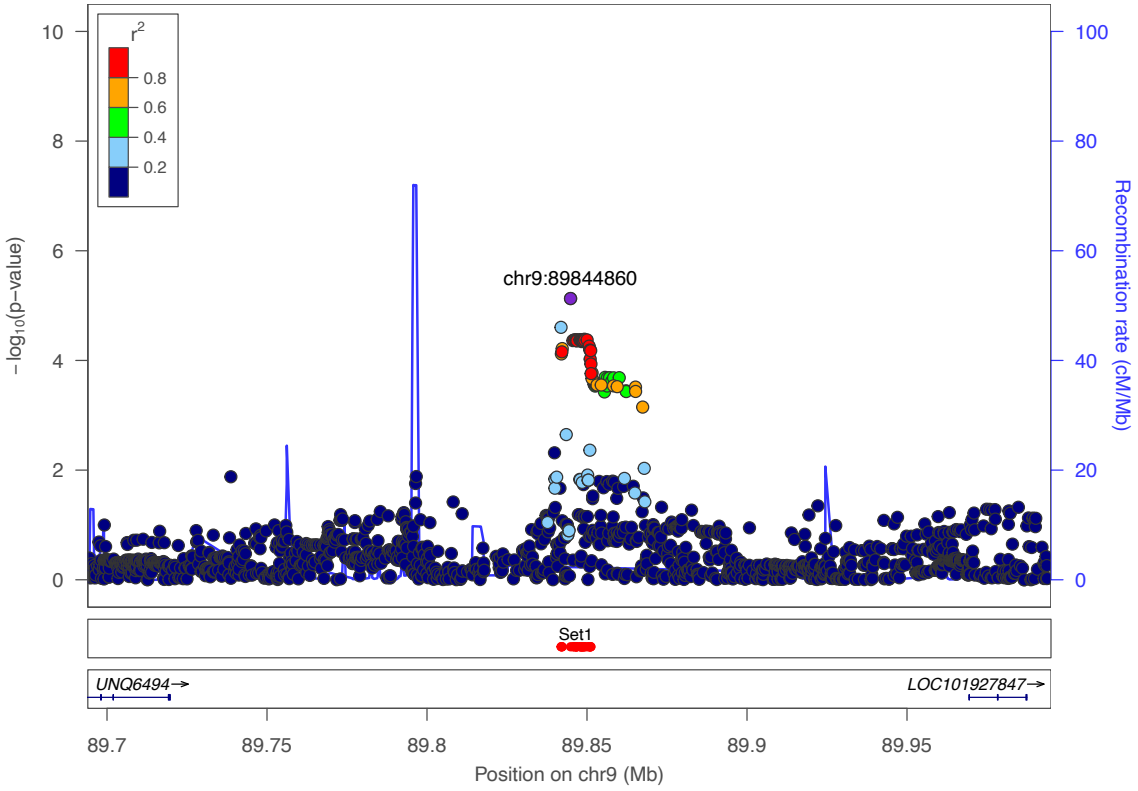
Supplementary Figure 10: *Regional association plot for the locus 5p13.1 with the top hit chr5:40323836 (rs348594). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



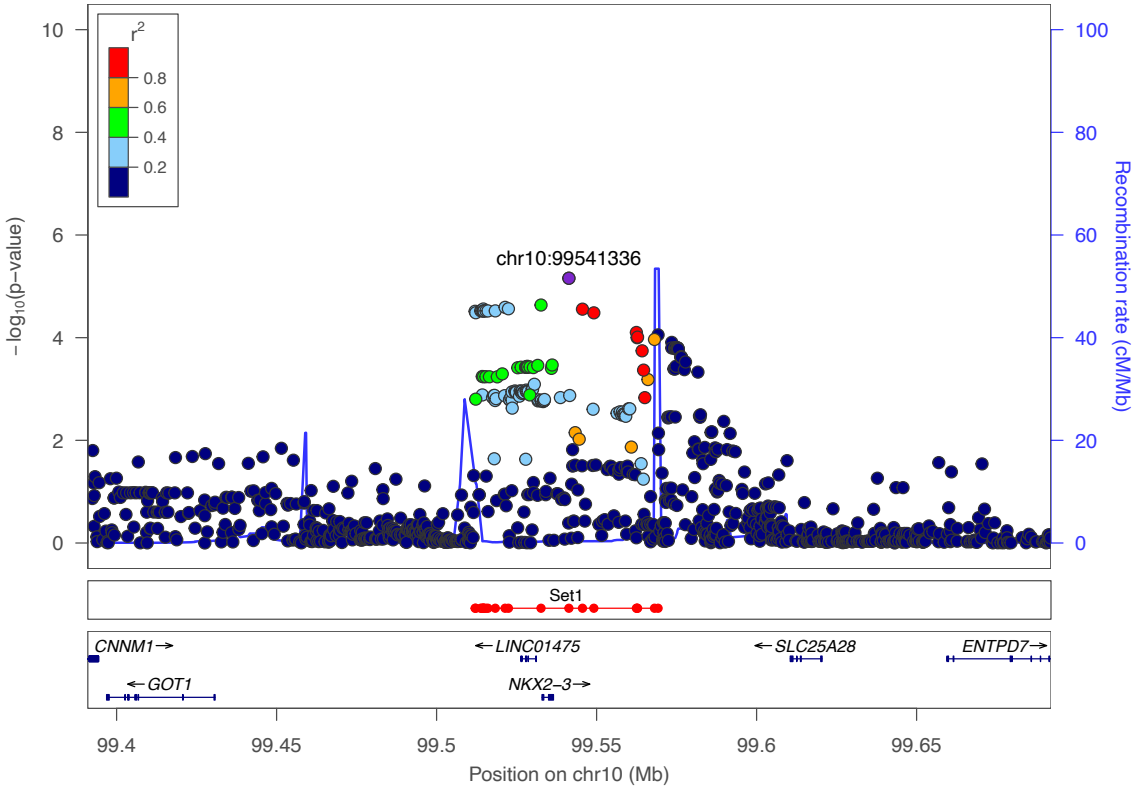
Supplementary Figure 11: *Regional association plot for the locus 6p21.33 with the top hit chr6:31118325 (rs117198148). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



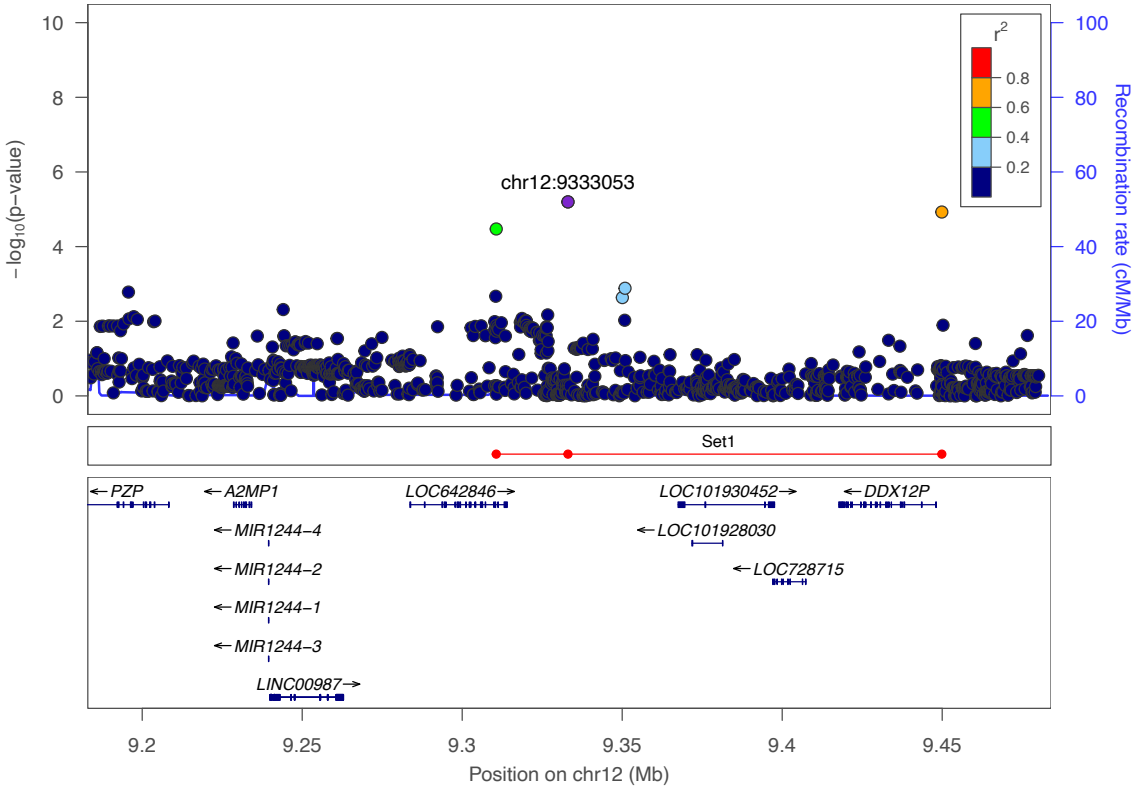
Supplementary Figure 12: *Regional association plot for the locus 6p21.32 with the top hit chr6:32644620 (rs6927022). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



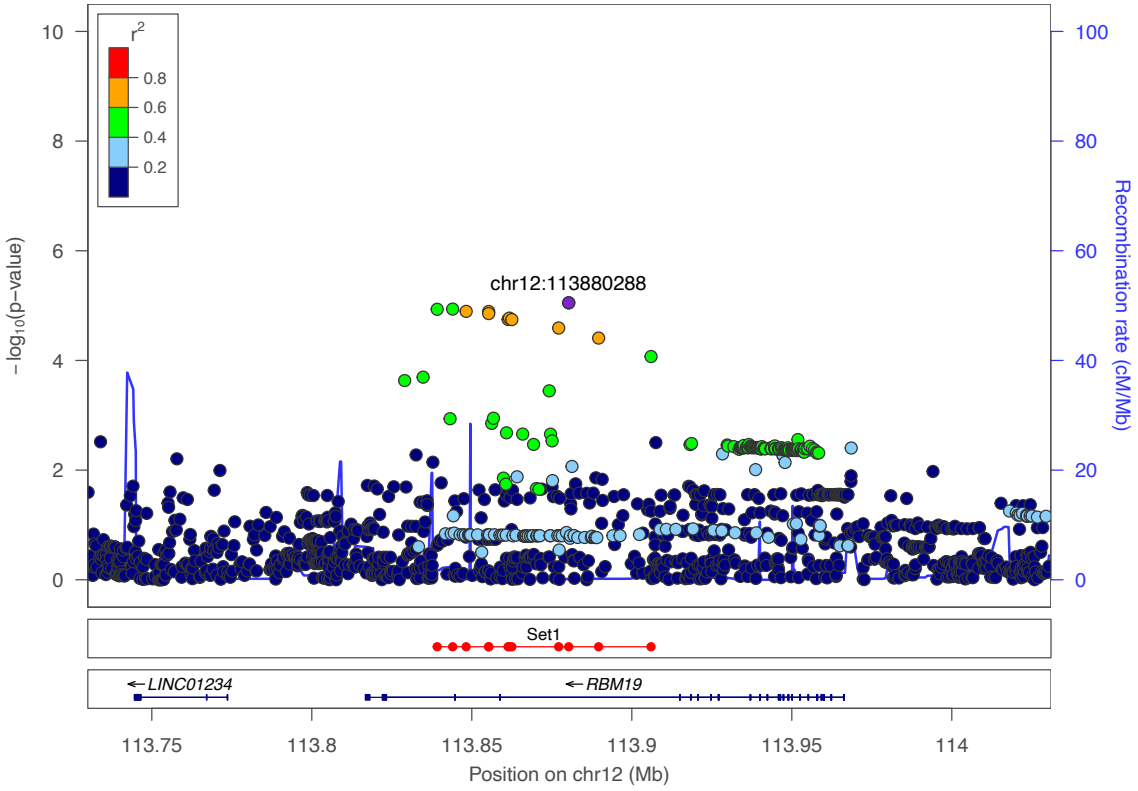
Supplementary Figure 13: *Regional association plot for the locus 9q22.2 with the top hit chr9:89844860 (rs36147380). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



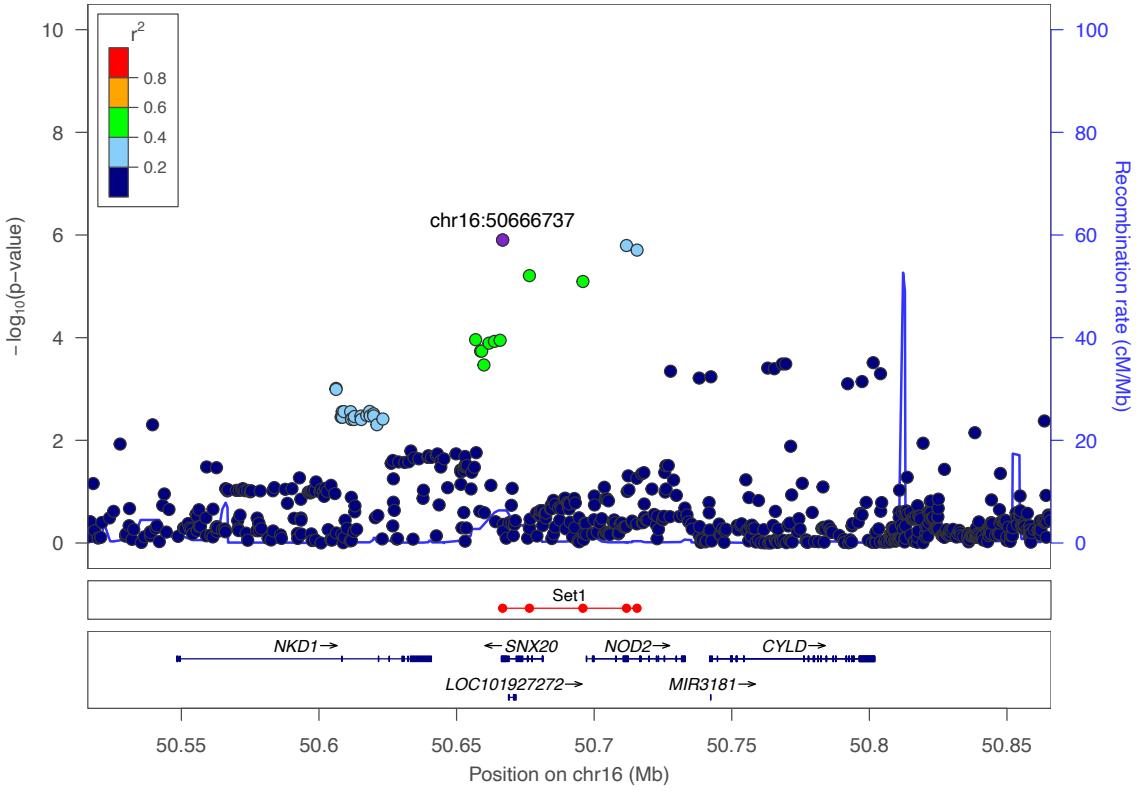
Supplementary Figure 14: *Regional association plot for the locus 10q24.2 with the top hit chr10:99541336 (rs4590800). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



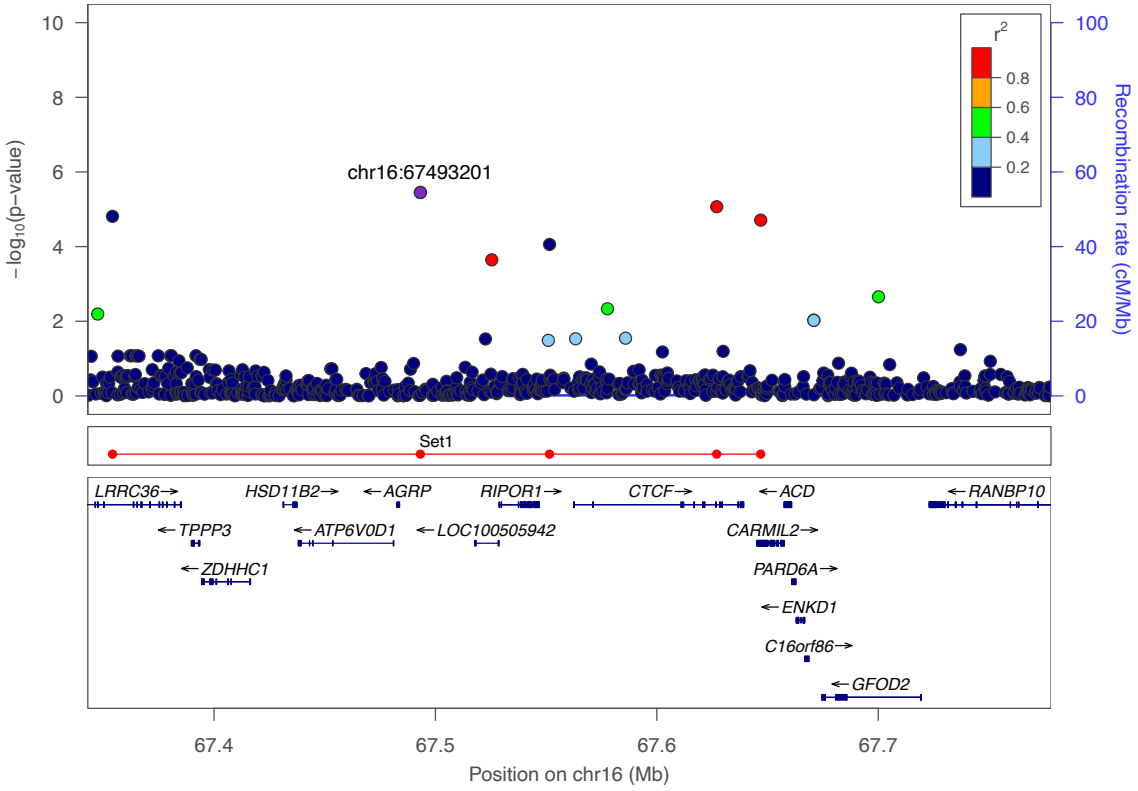
Supplementary Figure 15: *Regional association plot for the locus 12p13.31 with the top hit chr12:9333053 (rs187033004). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



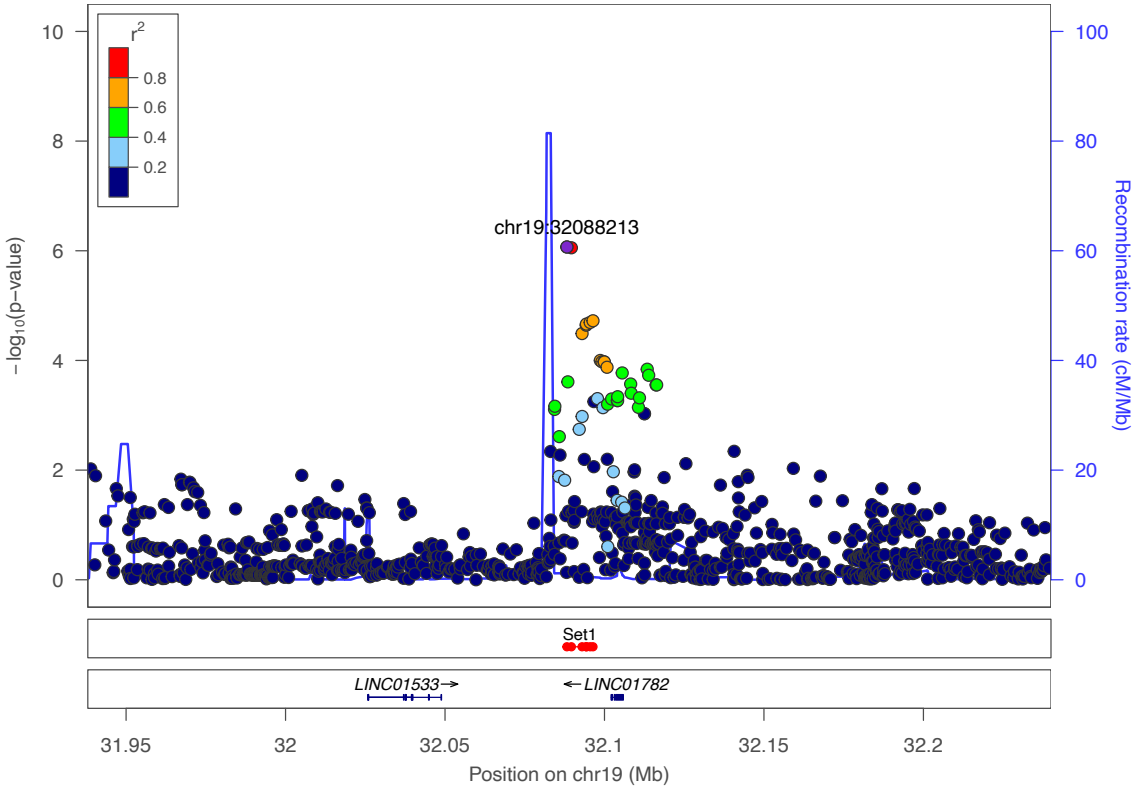
Supplementary Figure 16: *Regional association plot for the locus 12q24.13 with the top hit chr12:113880288 (rs3782449). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



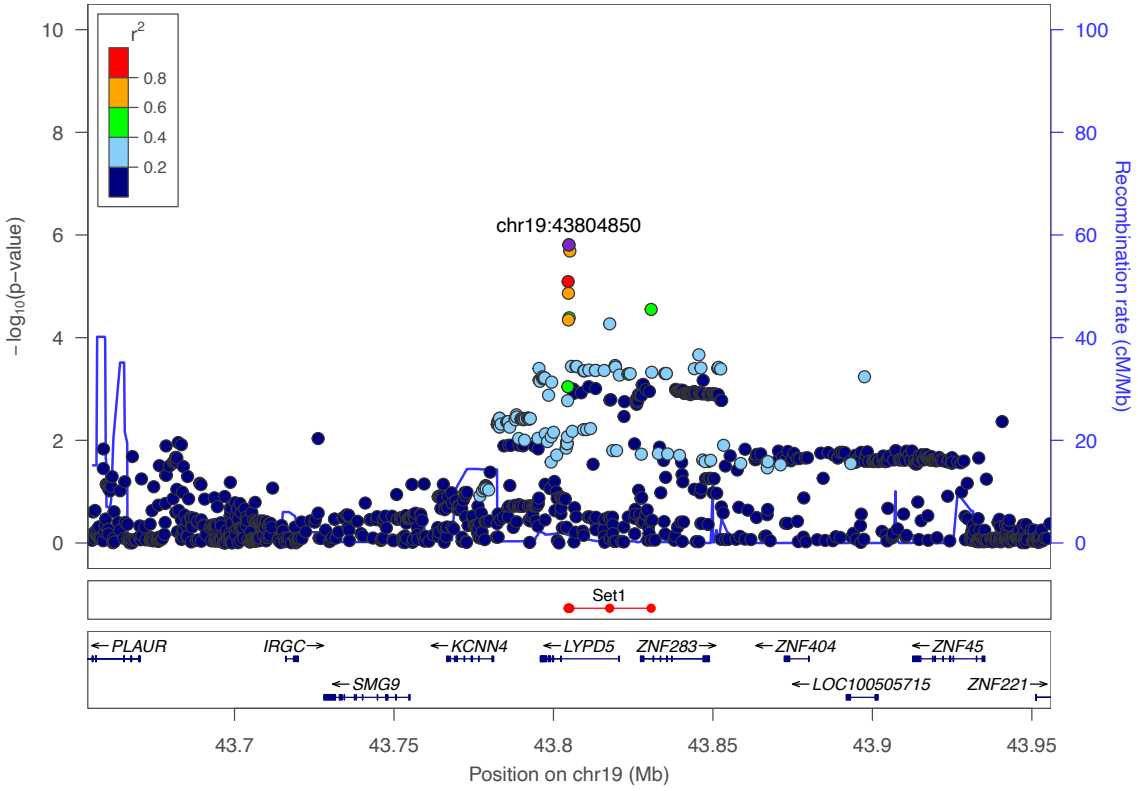
Supplementary Figure 17: *Regional association plot for the locus 16q12.1 with the top hit chr16:50666737 (rs139397276). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



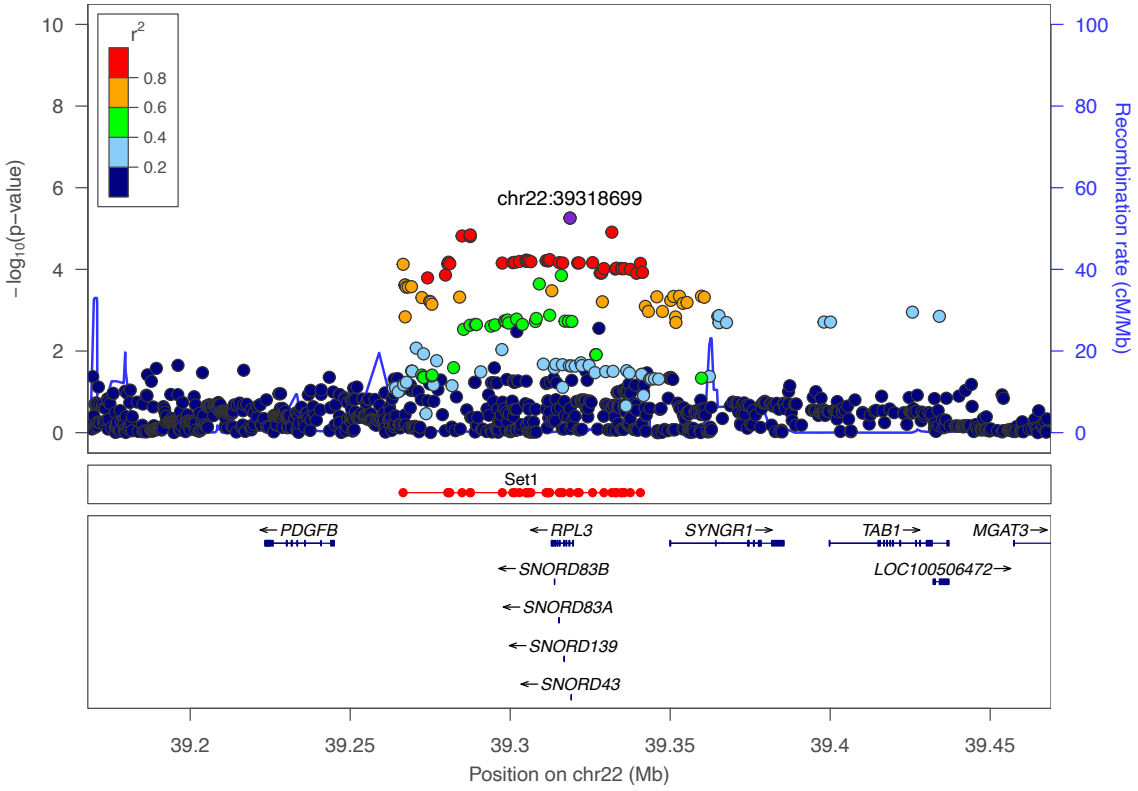
Supplementary Figure 18: *Regional association plot for the locus 16q22.1 with the top hit chr16:67493201 (rs77919558). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



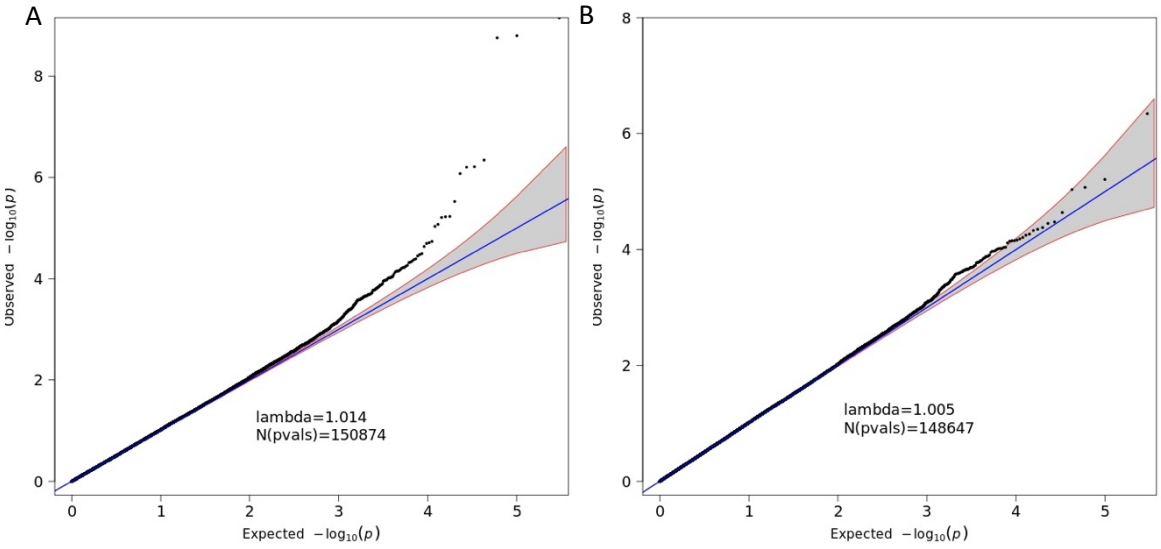
Supplementary Figure 19: *Regional association plot for the locus 19q13.11 with the top hit chr19:32088213 (rs6510221). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



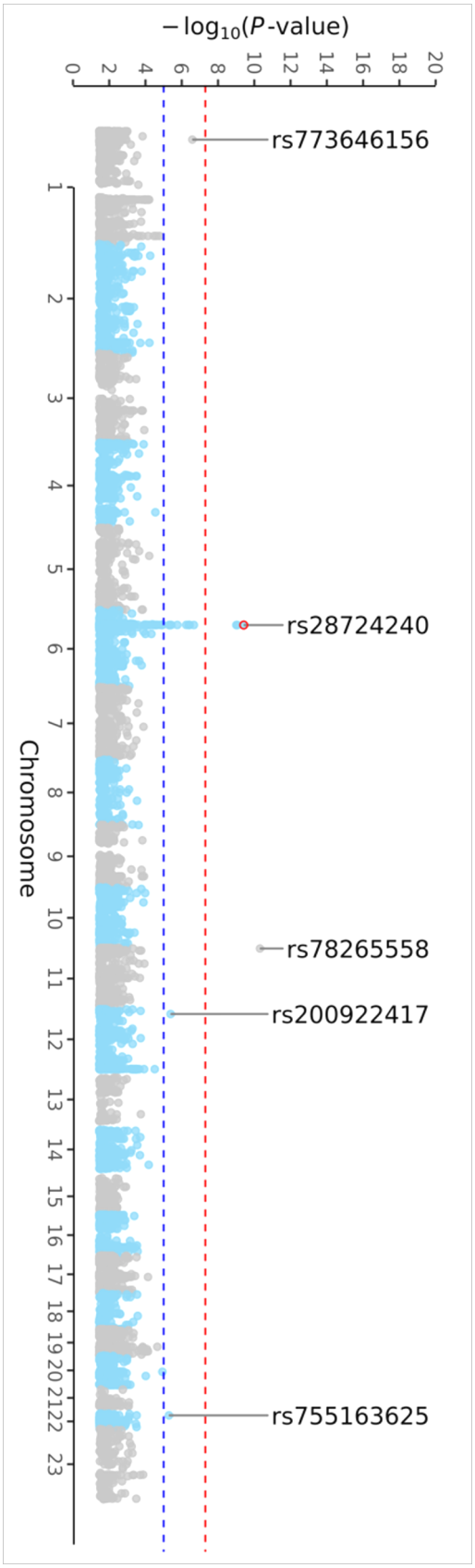
Supplementary Figure 20: *Regional association plot for the locus 19q13.31 with the top hit chr19:43804850 (rs364691). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



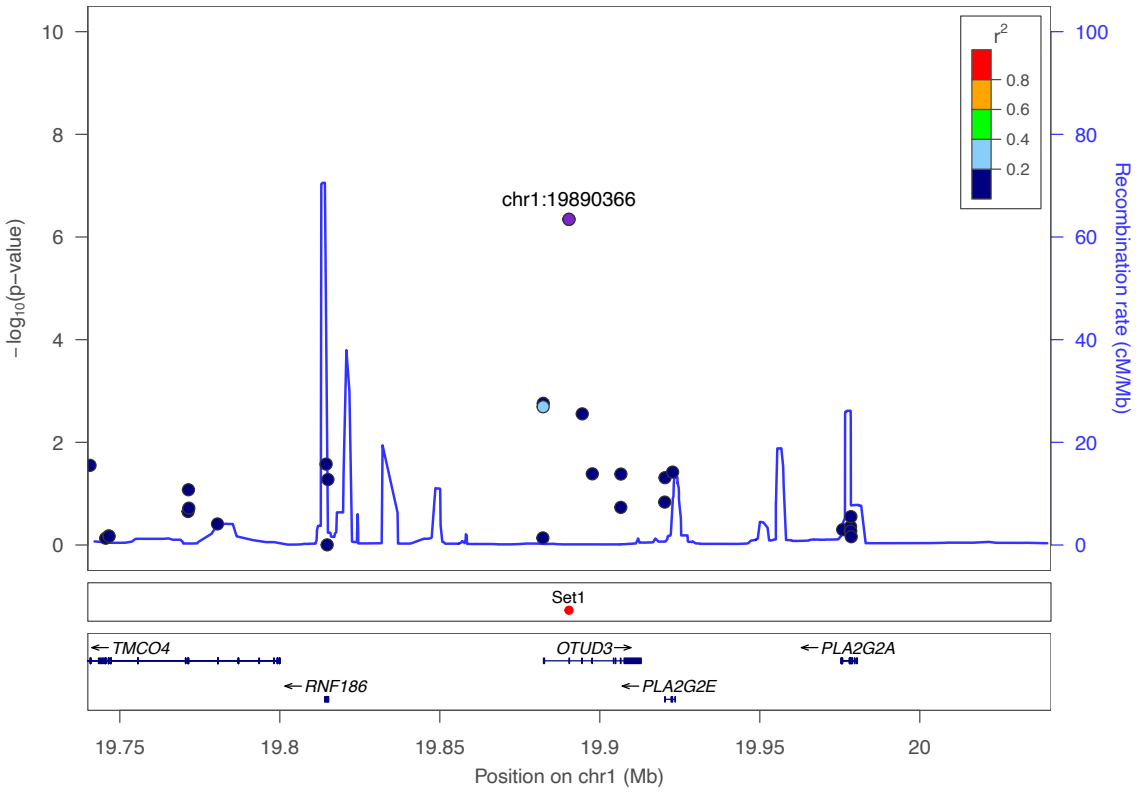
Supplementary Figure 21: *Regional association plot for the locus 22q13.1 with the top hit chr22:39318699 (rs1569498). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



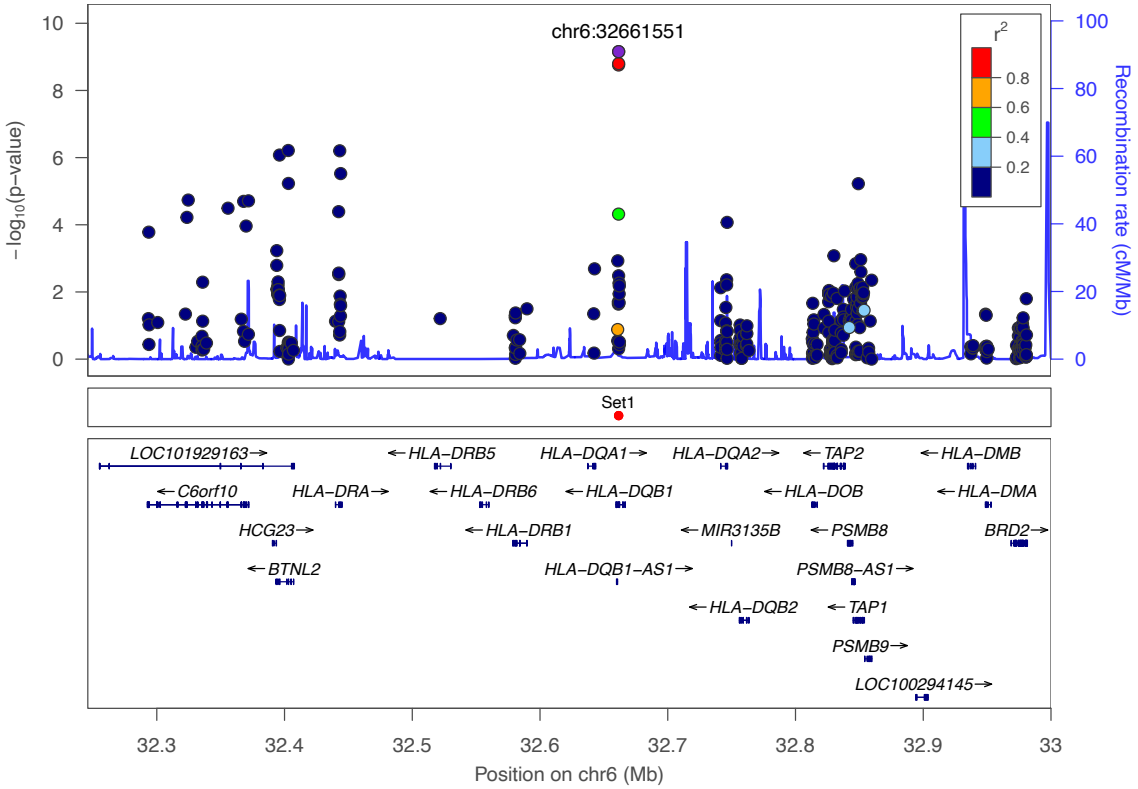
Supplementary Figure 22: *Quantile-quantile plot of association summary statistics of the whole exome data. The 95% concentration band under random sampling is shown in gray. The genomic inflation factor lambda is defined as the ratio of the medians of the sample χ2 test statistics and the 1-df χ2 distribution (0.455).[1] The left figure includes all 150,874 variants with MAF > 1 % and an imputation score r^2>0.6. The right figure excludes the variants of the HLA-region (chr6:29-34MB).*



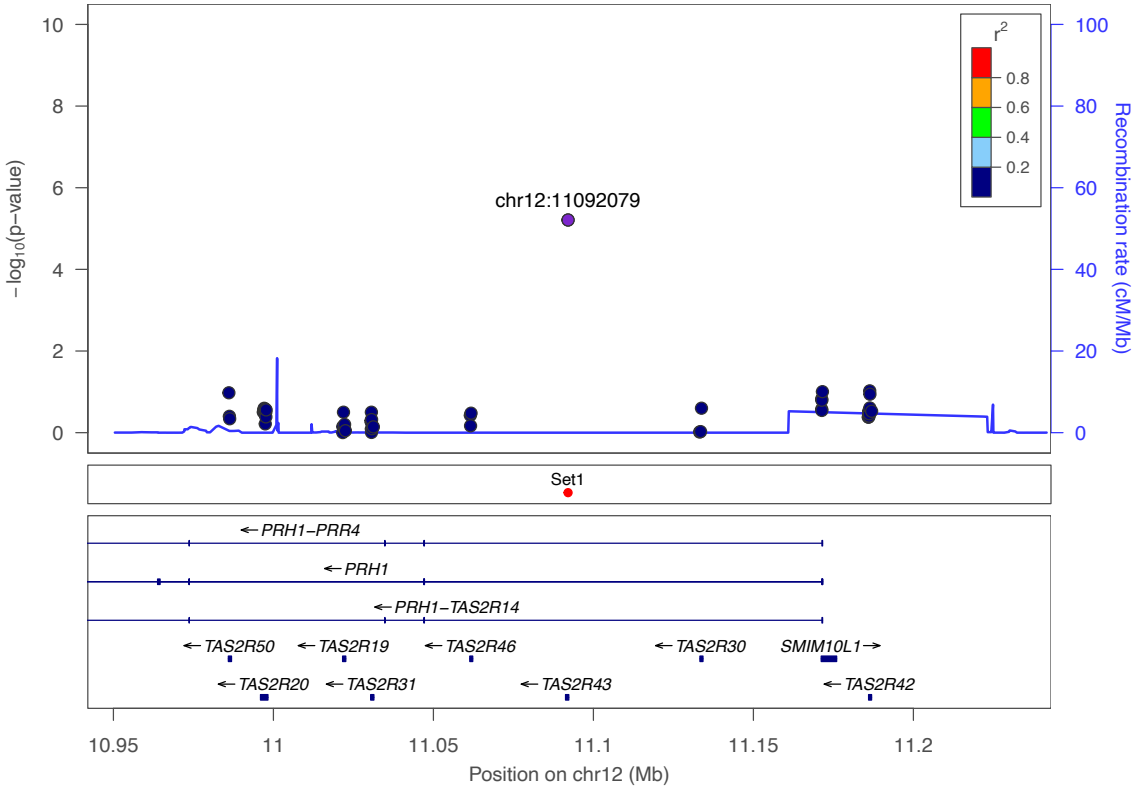
Supplementary Figure 23: *Manhattan plot of the exome data with a MAF >1% and an imputation score r2 >0.6. All loci of at least nominal significance (blue horizontal line; P<1×10-5) are annotated by the SNP-ID. Loci with LD support are highlighted with a blue (nominal significance) or red circle (genome-wide significance, red horizontal line; P<5×10^{8}).*



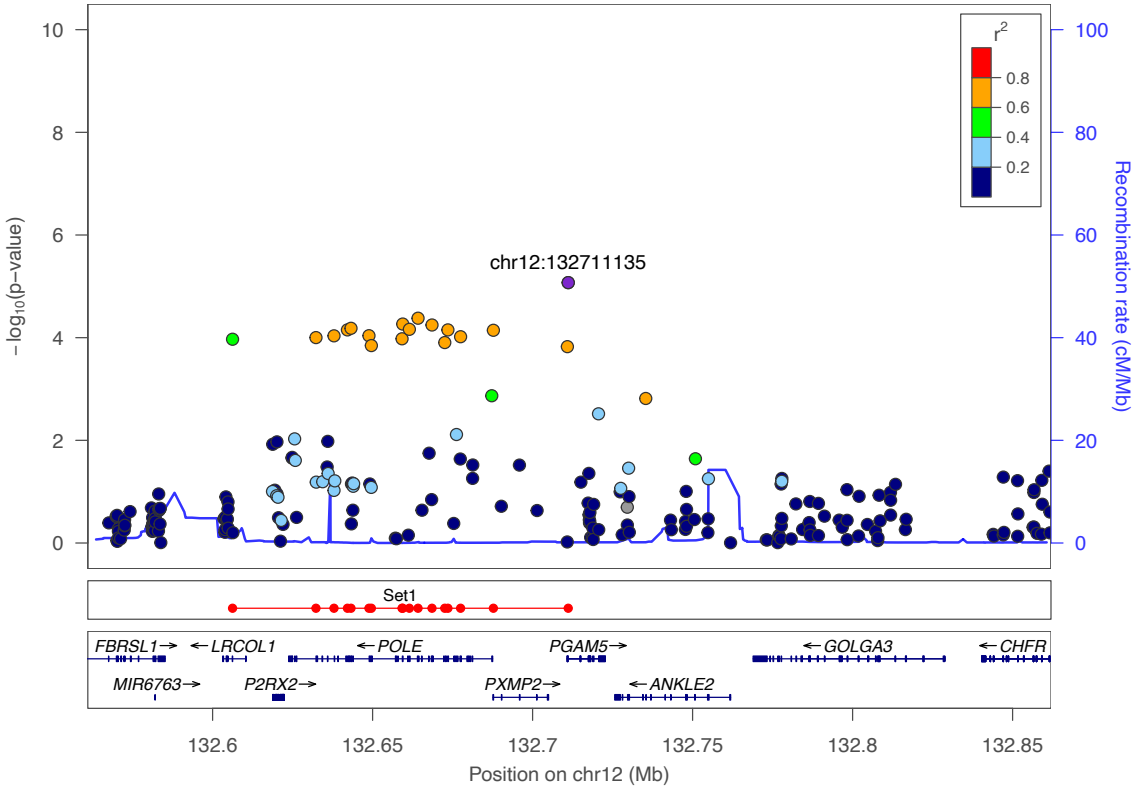
Supplementary Figure 24: *Regional association plot for the locus 1p36.13 in the exome data with the top hit chr1:19890366 (rs7523442). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



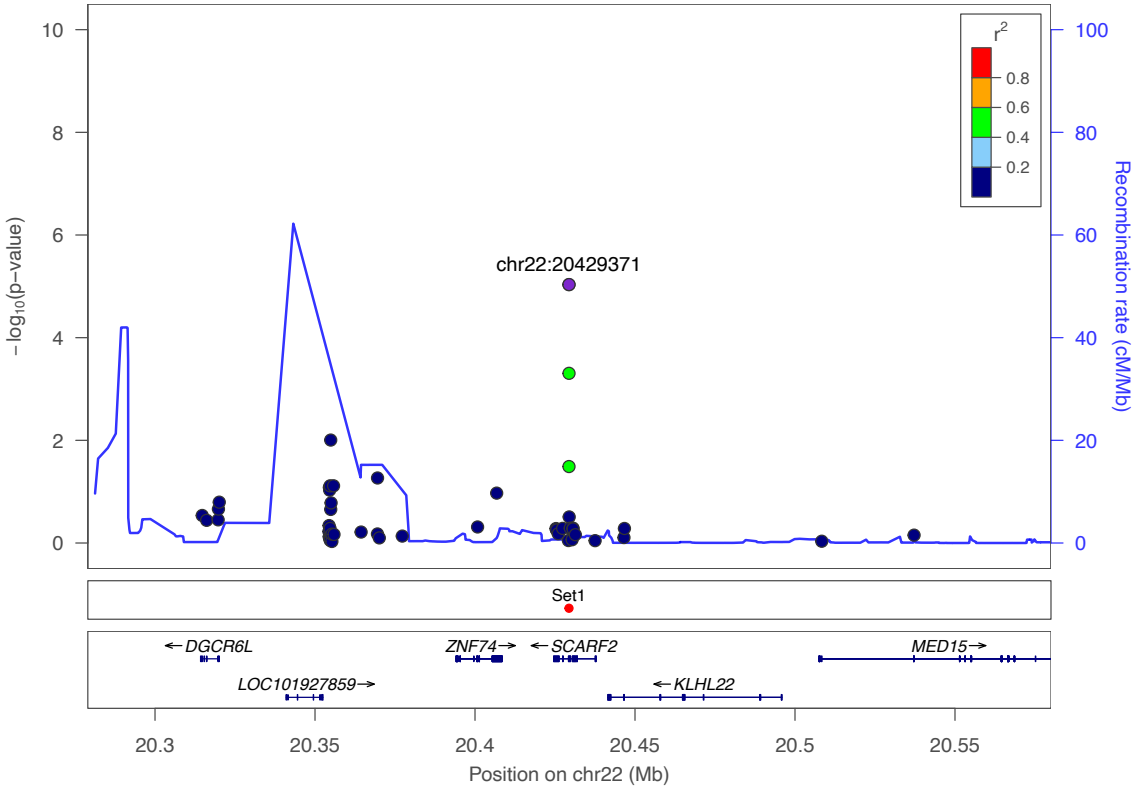
Supplementary Figure 25: *Regional association plot for the locus 6p21.32 in the exome data with the top hit chr6:32661551 (rs28724240). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



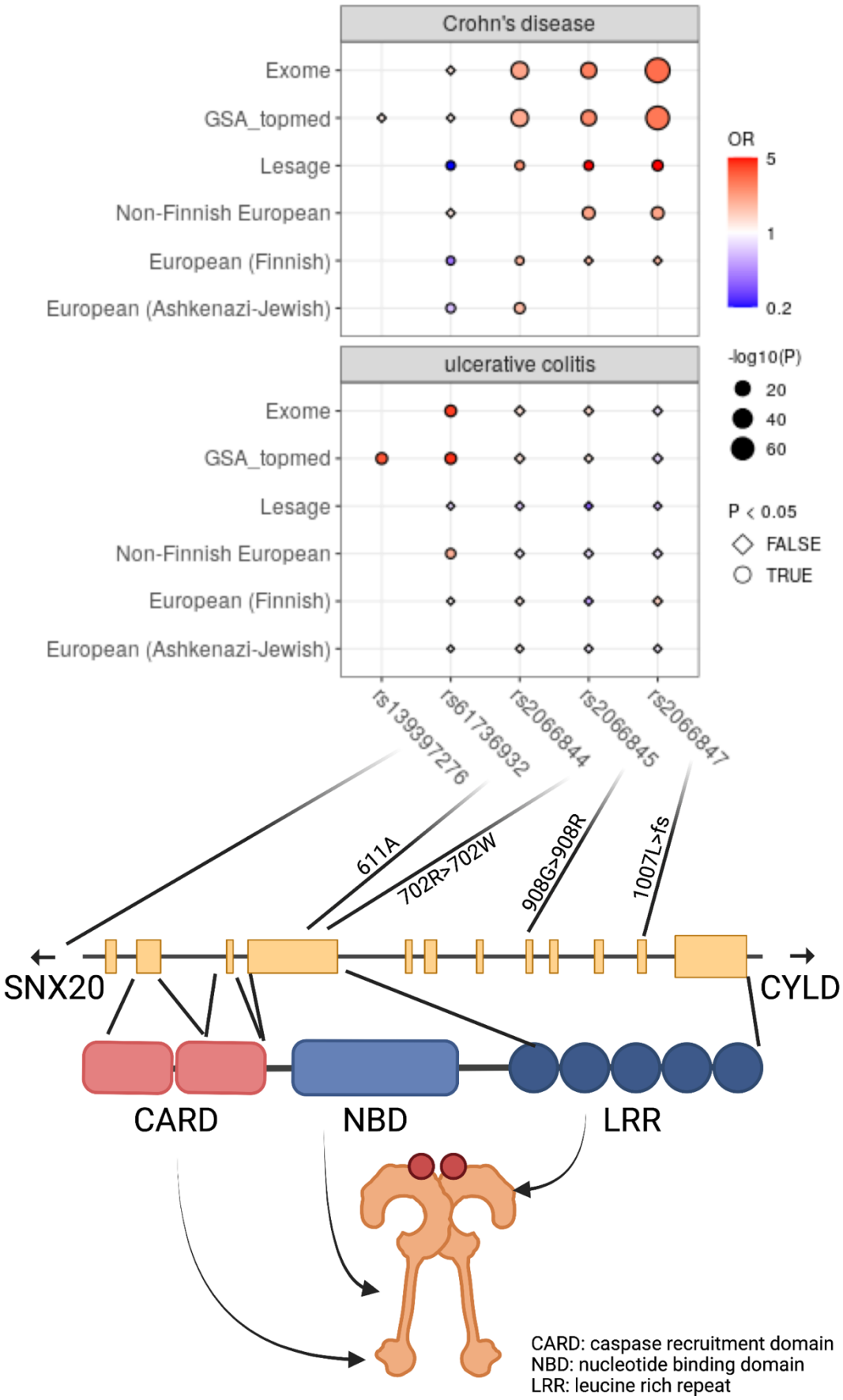
Supplementary Figure 26: *Regional association plot for the locus 12p13.2 in the exome data with the top hit chr12:11092079 (rs113197337). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



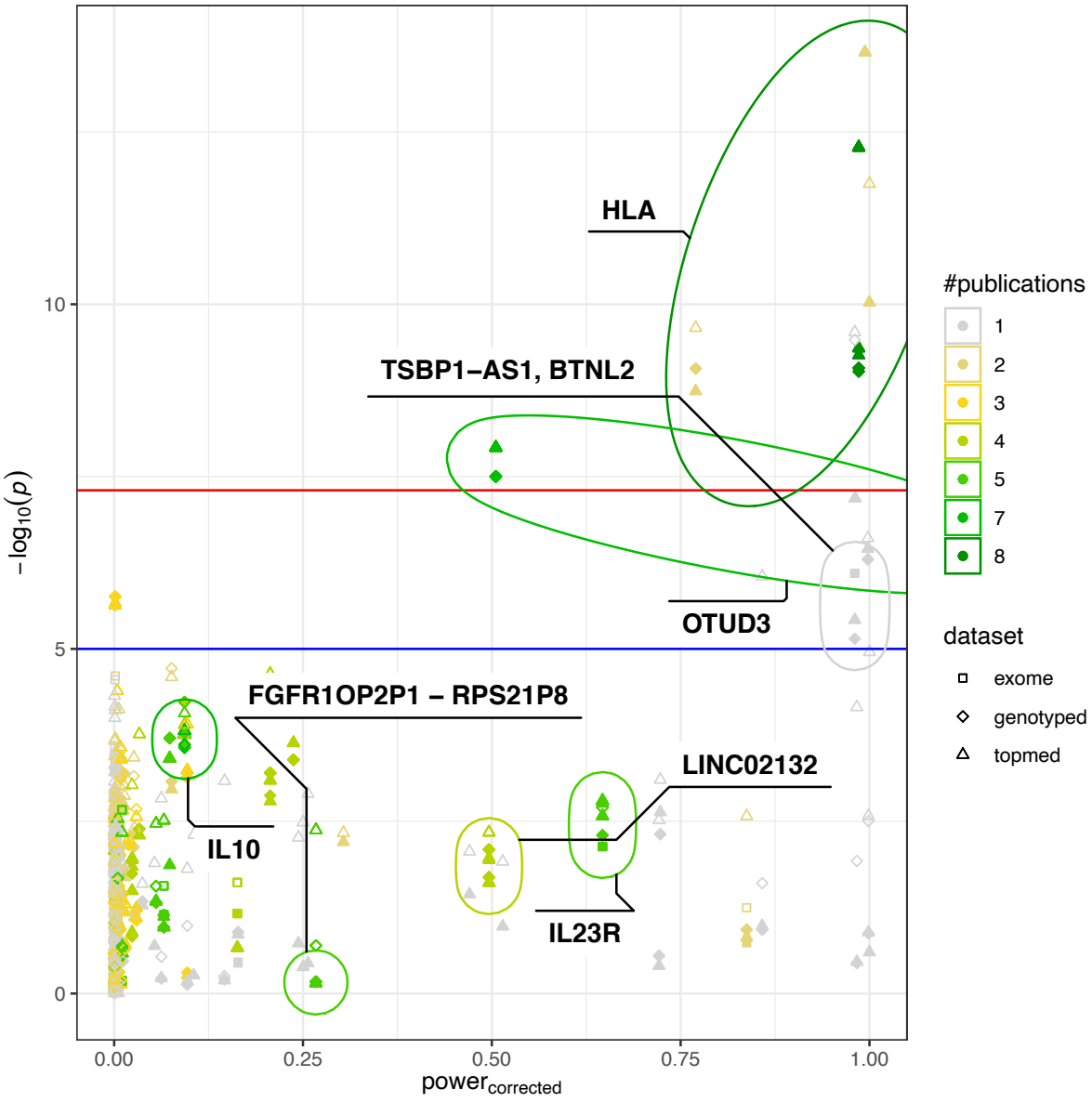
Supplementary Figure 27: *Regional association plot for the locus 12q24.33 in the exome data with the top hit chr12:132711135 (rs7973452). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*



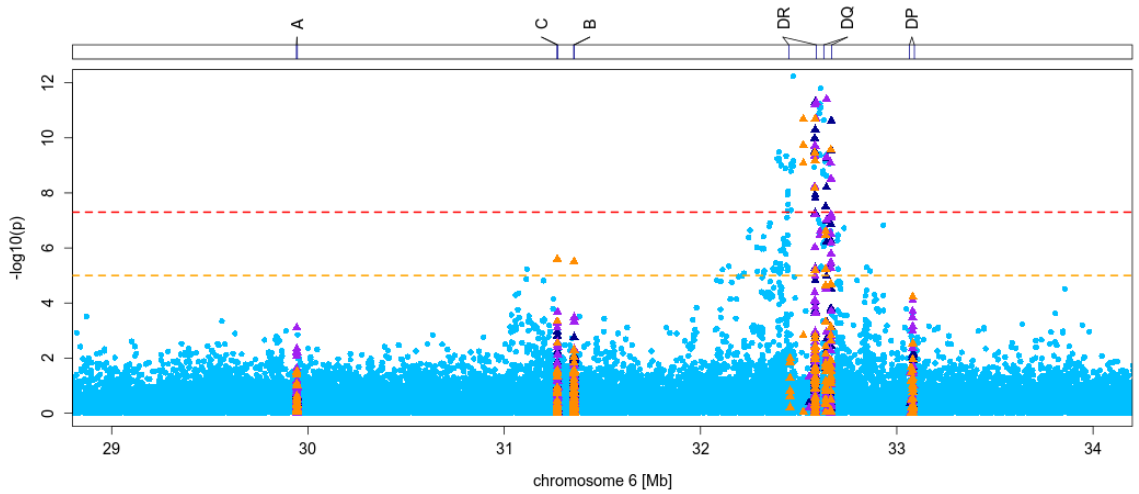
Supplementary Figure 28: *Regional association plot for the locus 22q11.21 in the exome data with the top hit chr22:20429371 (rs755163625). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom[2].*


Crohn's disease
ulcerative colitis
Exome
GSA_topmed
Lesage
Non-Finnish European
European (Finnish)
European (Ashkenazi-Jewish)
OR
5
1
0.2
-log10(P)
20
40
60
P < 0.05
FALSE
TRUE
rs139397276
rs61736932
rs2066844
rs2066845
rs2066847
611A
702R>702W
908G>908R
1007L>fs
SNX20
CYLD
CARD
NBD
LRR
CARD: caspase recruitment domain
NBD: nucleotide binding domain
LRR: leucine rich repeat

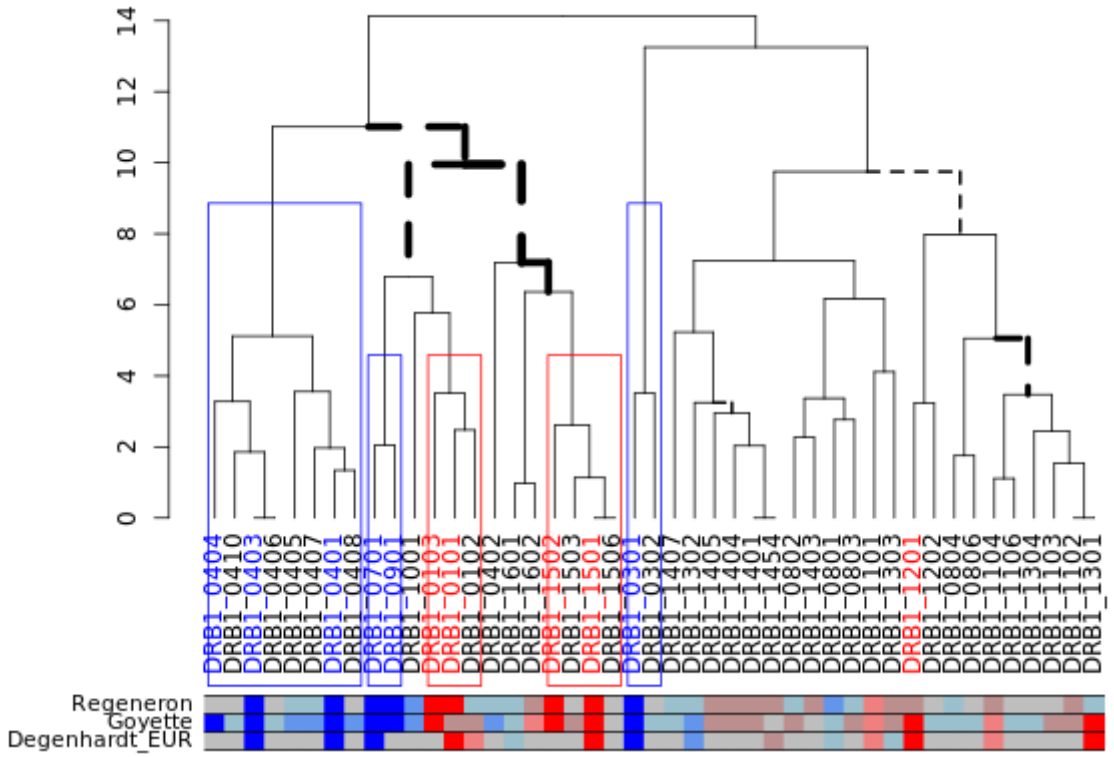
Supplementary Figure 29: *Associations at the NOD2 locus and the influence on the protein level. Rs139397277 is our main signal and rs61736932 the strongest associated variant within the exome data while the other three variants are those previously identified as associated with CD. Created with BiorRender.com.*



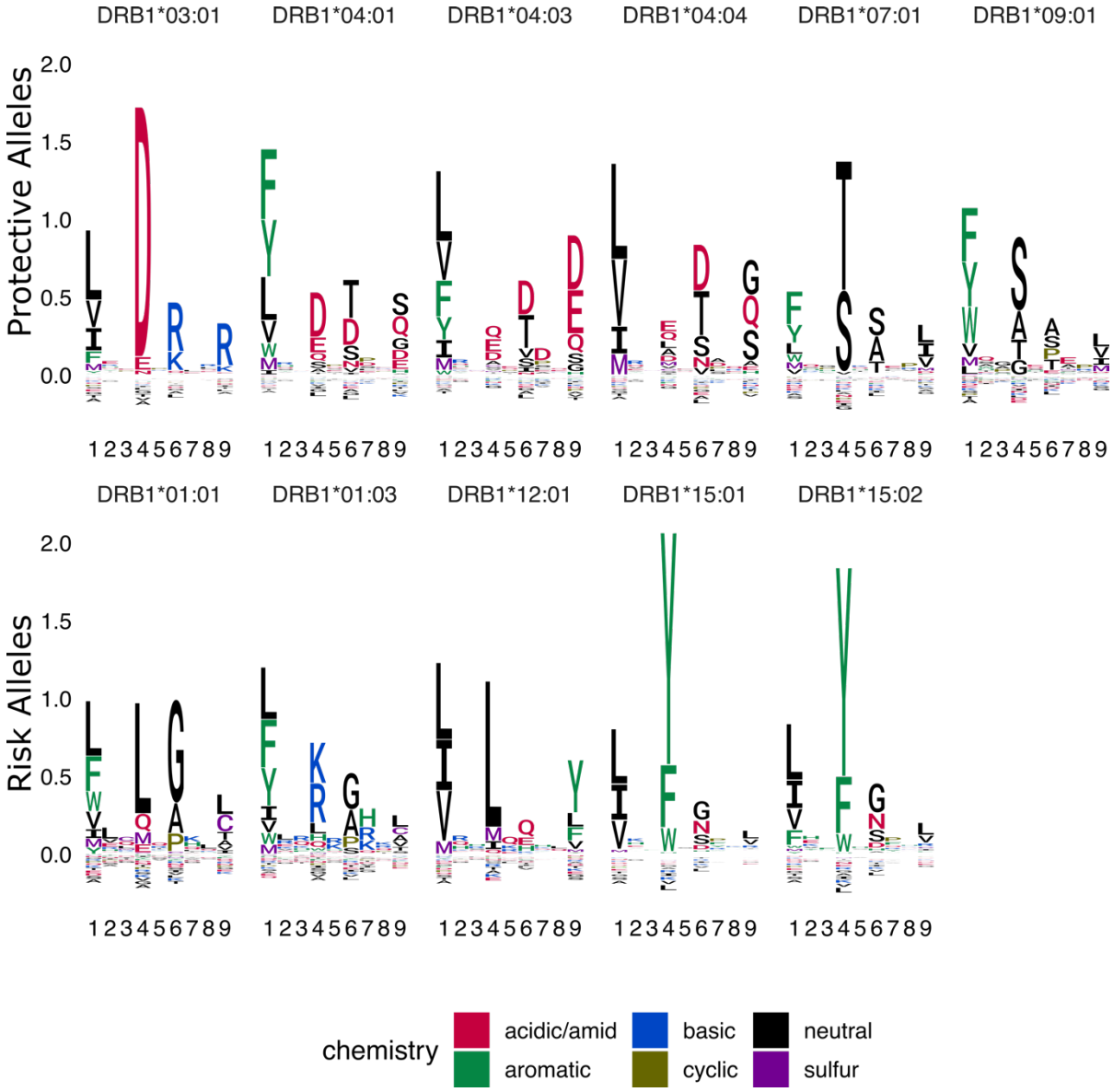
Supplementary Figure 30: *Power analysis based on the GWAS catalog data. The corrected power is calculated based on the median odds ratios and standard errors as listed in the GWAS catalog, when removing the strongest effect if more than one association is given. The power is calculated for the nominal significance level 1×10^{-5} and the frequencies given in our data.*



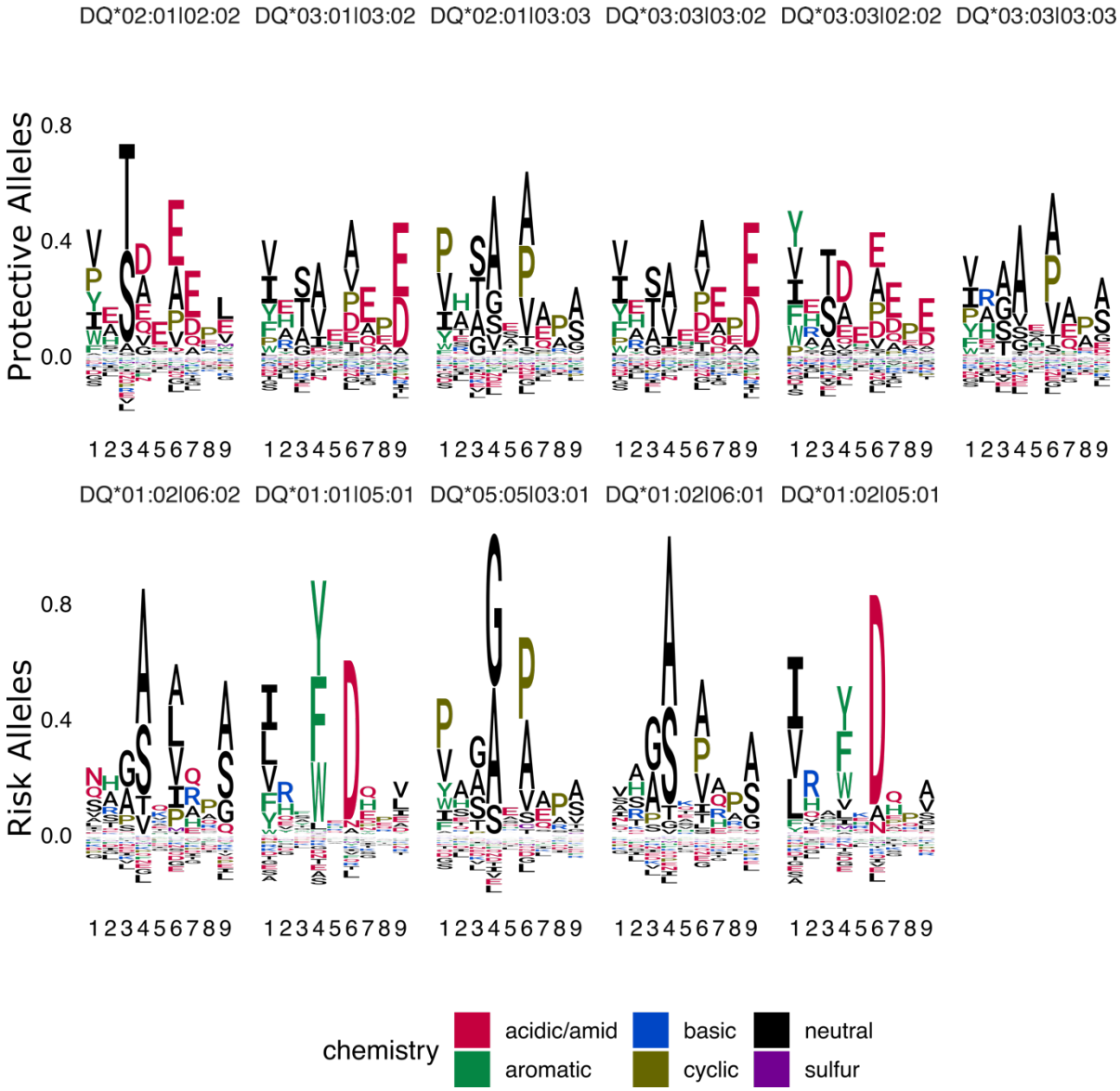
Supplementary Figure 31: *Finemapping of the HLA region including the imputed alleles (orange), amino acids (purple) and nucleotides (dark blue) generated from the HIBAG imputation. The light blue dots in the background are the topmed imputed variants.*



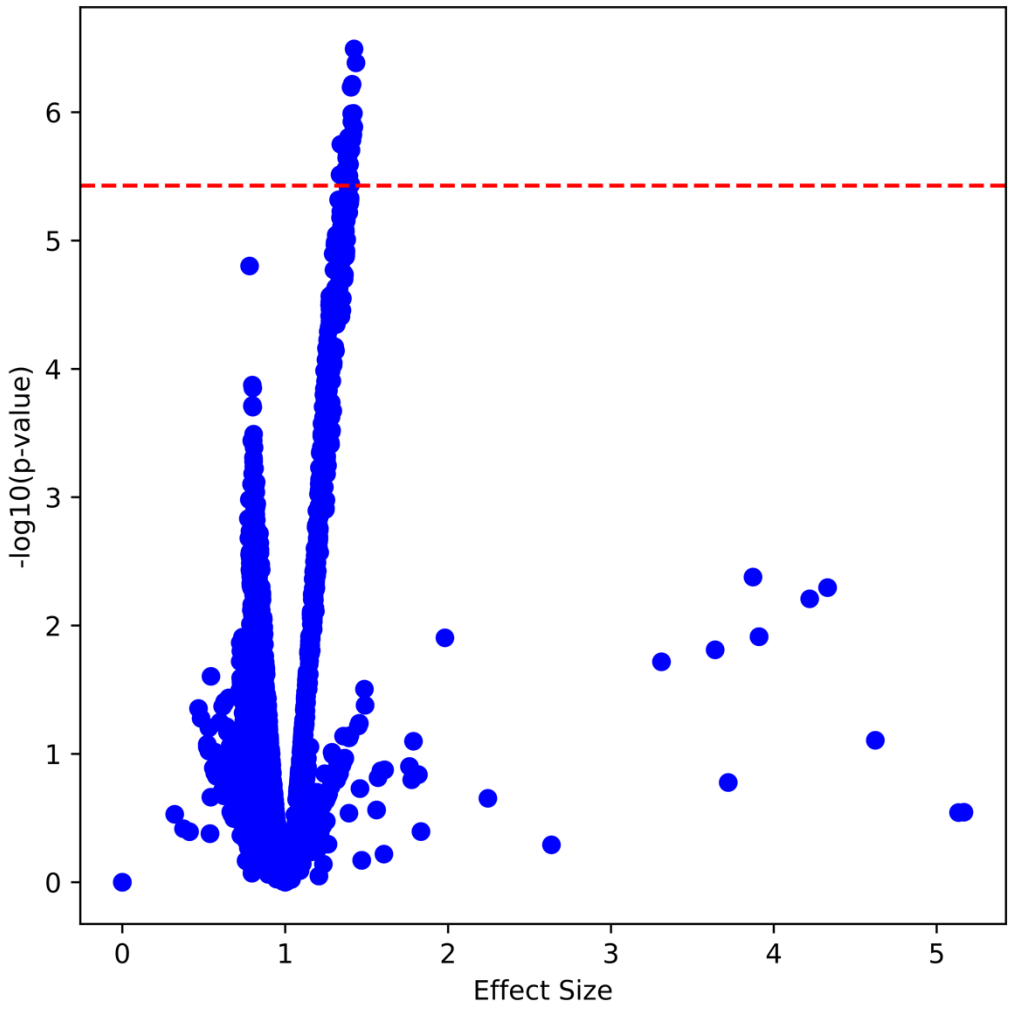
Supplementary Figure 32: *Dendrogram of HLA-DRB1 alleles. The distances within the dendrogram are based on the Pearsons' correlation coefficient between the predicted eluted ligand mass spectrometry score from NetMHCIIpan-4.0. The main clusters of risk and protective alleles are highlighted. The colored box below shows the effect direction shown in our data as well as in the two most important publications on HLA in UC (blue: protective, red: risk, grey: no information, shading based on the p-value: grayish red/blue: no significance, mid red/blue: p<0.05, pure red/blue: p<5 ×10^{-8}). All straight lines were also stable on other prediction subsets of the human proteome, while dotted lines represent lines that differ dependent on the proteome.*



Supplementary Figure 33: *Binding logo plot of associated HLA-DR alleles in differentiation to alleles with the other direction of effect. The upper row represents the protective associated alleles, the bottom line the logos of the risk alleles. The motifs are based on the NetMHCIIpan-4.0 predictions of the binding cores of all peptides at least annotated as weak binders, excluding peptides binding against one of the alleles of different direction of effect. The single letters represent the one letter amino acid code colored by the chemical properties of the amino acids.*



Supplementary Figure 34: *Binding logo plot of associated HLA-DQ alleles. The allele names are shortened, e.g., DQA1*02:01-DQB1*02:02 is written as DQ*02:01|02:02. The motifs are based on the NetMHCIIpan-4.0 predictions of the binding cores of all peptides at least annotated as weak binders. The single letters represent the one letter amino acid code colored by the chemical properties of the amino acids.*



Supplementary Figure 35: *Vulcano plot of the PepWAS analysis. Each dot represents the PepWAS results for one peptide. The red dotted line represents the Bonferroni corrected P-value* (P-value $<3.73 \times 10^{-6}$ based on 13,411 peptides).

Supplementary Tables

Supplementary Table 1: *Sample number before, during, and after QC. Numbers in white lines represent sample numbers to be removed.*

	UC	Control	All traits
Metadata	950	4681	20563
GSA+Exome (GSA/Exome)			19769 (20554/19772)
GSA input	950	4680	20554
-non German samples (Metadata)	0	0	1312
-only Exome data	0	0	3
-only GSA data	31	171	785
-duplicate samples	1	46	256
-unique blacklist	32	217	2065
GSA QC input			**18492**
-Missingness outlier	12	77	392
-Heterozygosity outlier	2	5	41
-PCA outliers			506
-duplicates	0	0	2
-unique QC removed	53	121	903
GSA after QC	865	4342	17589
-Relatives	2	157	1672
-relatives not already removed			1634
GSA final Qced	863	4185	**15955**
Exome input	119	4509	19772
-het/hom	3	11	43
-TiTv	0	0	1
-singletons	6	11	65
-missingness	0	0	7
-sex	0	0	13
-unique QC removed	6	16	84
Exome final Qced	913	4493	**19688**
GSA and Exome Qced (including relatives)			17138
Association (no relatives, only UC and Controls)	863	4185	5048

Supplementary Table 2: *Genes and transcripts used to generate the proteome. Given are next to the genetic hg38 location and the ensemble-ids also the biotype of the transcript, the hgnc symbol and the uniprot gene ids.*

[See separate xlsx-file]

Supplementary Table 3: *At least nominal significantly associated lead variants identified in the "imputed genotyping" dataset or the "Exome" dataset. For each locus an identifier is given in the "NR" column and the band information, further in case of a locus top hit it is annotated whether the locus is LD supported. For comparison additional entries from external sources were added. Those sources are the GWAS catalog (dataset named by the first author of the publication; namely: Liu JZ[3], de Lange K[4], Silverberg MS[5], McGovern DP[6], Jostins L[7], Barrett JC[8], Anderson CA[9], Asano K[10], Okamoto D[11], Ellinghaus D[12]) as well as information for the exact same variants or variants in high-LD (R^2>0.9) from the publicly available RICOPILI summary statistics (IBD_UC_1KG_oct13). The specific variant is characterized by its rs-id, the chromosomal position in*

GRCh38 (hg38) and GRCh37 (hg19) as well as the given alleles. From the association analysis the P-value (p.value) the OR with its 95% confidence interval (CI_L95 and CI_U95) as well as the beta and standard error (SE) are given. Further, the MAPPED_TRAIT is shown, which is always ulcerative colitis for our own data and the dataset extracted from RICOPILI but varies for data from the GWAS catalog as also variants including IBD are listed in case no association with UC could be identified for this locus in the database but another IBD related trait. Further the allele frequencies separated by patients (AF.Cases) and controls (AF.Controls) is given. For the RICOPILI data the frequencies are as given by Hapmap. For imputed variants, the imputation info is given. The R2 is related to imputed genotyping lead variant, the corresponding dataset is also given in the dataset_ld column. The mapped gene and the related impact on the protein structure (change) are listed as well as gtex associated genes associated with the variant.

[See separate xlsx-file]

Supplementary Table 4: *The association results with the HLA imputed data. In the "type" column the type of the analyzed variant is given as one of the following: 1-field HLA, 2-field HLA, nuc, or prot. The HLA gene name (locus) and the exact description of the variant (name) together with the chromosome (CHR) and position in the human reference genome GRCh38 (hg38) and GRCh37 (hg19) and the position within the protein sequence in case of nucleotides (nuc) and amino acids (prot) specify the exact variant. The alleles A1 and A2 are either the one letter nucleotide or amino acid code or in the case of polymorphic variants and HLA alleles noted as present-absent (P and A). From the association analysis the P-value (p.value), the OR with its 95% confidence interval (CI_L95 and CI_U95), as well as the standard error (SE) are given. Further the minor allele frequency (MAF) as well as the allele frequency separated by cases (AF.Cases) and controls (AF.Controls) is included. Additionally, for the HLA alleles the posterior probability is listed as a reliability score of the imputation.*

[See separate xlsx-file]

Supplementary Table 5: *The significant associated peptides from the PepWAS analysis (**sheet: peptides**) and the information condensed on the transcript (sheet: transcripts) and gene level (sheet: genes). For each peptide the association statistics are given as effect size (effect_size) and p-value (p) further the corresponding transcripts (ENSTs) with the position in GRCh38, the ensemble gene ids (ENSG) and the hgnc symbol is given. It is noted whether the peptide is present in the reference proteome (mutations) and in how many samples (n). The sample numbers with a specific peptide are also listed separated by cases (n_nase) and controls (n_controls). Those values are also given as frequencies (f, f_case and f_control). Also, the OR of the frequencies (OR_freq) is included and the p-value of the fisher test on the frequencies. The mutations related to the single peptides are listed: First the mutations that are needed to generate the sequence (nucchanges and protchanges), further the different amino acid positions in different transcripts are listed and then all mutations, that would change the sequence of the peptide, are listed. The column HLA lists the HLA alleles predicted to bind the peptide. The number (n_immunopeptidome_blood) and sequences (immunopeptidome_blood) of identified peptides in the 25 immunopeptidomes published in ElAbd et al.*[13] *are presented. On the transcript level (**sheet: transcripts**) for each ensemble transcript (ENST) the uniprot gene id and the ensemble protein id together with the length of the amino acid sequence (lengthAA) are noted. The number of peptides per transcript are given (n_hits). As the peptides are generated by a sliding window approach and peptides binding HLA class II are longer than the binding pocket, often neighboring peptides are predicted as similar good peptides and a single missense mutation might have only a small impact on the binding affinity, therefore also the number of peptides where less than 9 amino acids are in the same order are given (n_no9AAoverlap). Further, the number of PepWAS hits is separated by those with a mutation (n_hits_mut) and those present in the reference (n_hits_ref). Additional numbers of related mutations (n_relevant_mut) are given as mutations necessary to form a PepWAS hit (necessary_mut), mutations that are changing the present peptide but both peptides are PepWAS hits (possible_mut), and mutations that modify a peptide in a way that it is not predicted as PepWAS hit anymore (forbidden_mut). On the gene level (**sheet:genes**) the different transcripts are summarized, with the majority of the previously described attributes, and whether all peptides annotated to one gene are expressed within one transcript (AllHitsInOneENST). Further information about the expression of the genes is given as reported in Taman et al.*[14] *and in Linggi et al*[15]*.. Further if the confidence set of associated variants includes any GTEx*[16] *variants the effect is given in comparison to the risk variants (risk_variant_lead_to_expression) as "decreased" or "increased" expression. The immunopeptidome data are given for the whole genes independent of the location of the PepWAS hits.*

[See separate xlsx-file]

References

1. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

2. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).

3. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

4. De Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

5. Silverberg, M. S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41**, 216–220 (2009).

6. McGovern, D. P. B. *et al.* Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332–337 (2010).

7. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H. & Dermot, P. Host-microbe interactions have shaped the genetic architecture of Inflammatory Bowel Disease. **491**, 119–124 (2012).

8. Barrett, J. C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330–1334 (2009).

9. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).

10. Asano, K. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.* **41**, 1325–1329 (2009).

11. Okamoto, D. *et al.* Genetic analysis of ulcerative colitis in Japanese individuals using population-specific SNP array. *Inflamm. Bowel Dis.* **26**, 1177–1187 (2020).

12. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).

13. Elabd, H. *et al.* Predicting Peptide HLA-II Presentation Using Immunopeptidomics , Transcriptomics and Deep Multimodal Learning. *bioRxiv* (2022) doi:https://doi.org/10.1101/2022.09.20.508681.

14. Taman, H. *et al.* Transcriptomic landscape of treatment-naïve ulcerative colitis. *J. Crohn's Colitis* **12**, 327–336 (2018).

15. Linggi, B. *et al.* Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis. *Sci. Rep.* **11**, 1–12 (2021).

16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

BIBLIOGRAPHY

[1] Abul K. Abbas, Andrew H. Lichtman, and Shiv Pillai. *Cellular and Molecular Immunology*. Vol. 8. Saunders, 2014, p. 535. ISBN: 978-0-323-31614-9.

[2] Jennifer G. Abelin et al. "Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction." In: *Immunity* 51.4 (2019), 766–779.e17. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2019.08.012. URL: https://doi.org/10.1016/j.immuni.2019.08.012.

[3] Alex T. Adams et al. "Two-stage genome-wide methylation profiling in childhood-onset Crohn's disease implicates epigenetic alterations at the VMP1/MIR21 and HLA loci." In: *Inflammatory Bowel Diseases* 20.10 (2014), pp. 1784–1793. ISSN: 1536-4844. DOI: 10.1097/MIB.0000000000000179.

[4] Sharon D Adams, Kathleen C Barracchini, Deborah Chen, FuMeei Robbins, Lu Wang, Paula Larsen, Robert Luhm, and David F Stroncek. "Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification." In: *Journal of translational medicine* 2.1 (Sept. 2004), p. 30. ISSN: 1479-5876. DOI: 10.1186/1479-5876-2-30. URL: http://www.ncbi.nlm.nih.gov/pubmed/15363110.

[5] Tariq Ahmad, Sara E. Marshall, and Derek Jewell. "Genetics of inflammatory bowel disease: The role of the HLA complex." In: *World Journal of Gastroenterology* 12.23 (2006), pp. 3628–3635. ISSN: 1007-9327. DOI: 10.3748/wjg.v12.i23.3628.

[6] David M. Altshuler et al. "An integrated map of genetic variation from 1,092 human genomes." In: *Nature* 491.7422 (2012), pp. 56–65. ISSN: 1476-4687. DOI: 10.1038/nature11632.

[7] Miguel Álvaro-Benito, Eliot Morrison, Esam T. Abualrous, Benno Kuropka, and Christian Freund. "Quantification of HLA-DM-dependent major histocompatibility complex of class II immunopeptidomes by the peptide landscape antigenic epitope alignment utility." In: *Frontiers in Immunology* 9.MAY (2018). ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.00872.

[8] Ashwin N. Ananthakrishnan. "Environmental Risk Factors for Inflammatory Bowel Diseases: A Review." In: *Digestive Diseases and Sciences* 60.2 (2015), pp. 290–298. ISSN: 15732568. DOI: 10.1007/s10620-014-3350-9.

[9] Carl A. Anderson, Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. "Data quality control in genetic case-control association studies." In: *Nature Protocols* 5.9 (2010), pp. 1564–1573. ISSN: 1750-2799. DOI: 10.1038/nprot.2010.116.

[10] Massimo Andreatta, Ole Lund, and Morten Nielsen. "Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach." In: *Bioinformatics* 29.1 (2013), pp. 8–14. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts621.

[11] Vito Annese. "Genetics and epigenetics of IBD." In: *Pharmacological Research* 159.May (2020), p. 104892. ISSN: 1096-1186. DOI: 10.1016/j.phrs.2020.104892. URL: https://doi.org/10.1016/j.phrs.2020.104892.

[12] Anthony Nolan Research Institute. *Nomenclature for Factors of the HLA System*. 2019. URL: http://hla.alleles.org/nomenclature/naming.html.

[13] James J Ashton, Katy Latham, Robert Mark Beattie, and Sarah Ennis. "Review article: the genetics of the human leucocyte antigen region in inflammatory bowel disease." In: *Alimentary Pharmacology & Therapeutics* 50.8 (Oct. 2019), pp. 885–900. ISSN: 0269-2813. DOI: 10.1111/apt.15485. URL: http://doi.wiley.com/10.1111/apt.15485.

[14] Raja Atreya, Markus F. Neurath, and Britta Siegmund. "Personalizing Treatment in IBD: Hype or Reality in 2020? Can We Predict Response to Anti-TNF?" In: *Frontiers in Medicine* 7.September (2020), pp. 1–14. ISSN: 2296858X. DOI: 10.3389/fmed.2020.00517.

[15] A Auton et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (2015), pp. 68–74. DOI: https://doi.org/10.1038/nature15393. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=26432245.

[16] Shawn Babiuk, Benjamin Horseman, Chenhong Zhang, Mik Bickis, Anthony Kusalik, Lawrence B. Schook, Mitchell S. Abrahamsen, and Reno Pontarollo. "BoLA class I allele diversity and polymorphism in a herd of cattle." In: *Immunogenetics* 59.2 (2007), pp. 167–176. ISSN: 0093-7711. DOI: 10.1007/s00251-006-0173-7.

[17] Rodrigo Barquera, Evelyn Collen, Da Di, Stéphane Buhler, João Teixeira, Bastien Llamas, José M. Nunes, and Alicia Sanchez-Mazas. "Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide." In: *Hla* 96.3 (2020), pp. 277–298. ISSN: 2059-2310. DOI: 10.1111/tan.13956.

[18] Michal Bassani-Sternberg and David Gfeller. "Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions." In: *The Journal of Immunology* 197.6 (2016), pp. 2492–2499. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1600808. URL: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1600808.

[19] Valia Bravo-Egana, Holly Sanders, and Nilesh Chitnis. "New challenges, new opportunities: Next generation sequencing and its place in the advancement of HLA typing." In: *Human Immunology* 82.7 (2021), pp. 478–487. ISSN: 1879-1166. DOI: 10.1016/j.humimm.2

021.01.010. URL: https://doi.org/10.1016/j.humimm.2021.01.010.

[20] Brian L. Browning, Xiaowen Tian, Ying Zhou, and Sharon R. Browning. "Fast two-stage phasing of large-scale sequence data." In: *American Journal of Human Genetics* 108.10 (2021), pp. 1880–1890. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2021.08.005. URL: https://doi.org/10.1016/j.ajhg.2021.08.005.

[21] Brian L. Browning, Ying Zhou, and Sharon R. Browning. "A One-Penny Imputed Genome from Next-Generation Reference Panels." In: *American Journal of Human Genetics* 103.3 (2018), pp. 338–348. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2018.07.015. URL: https://doi.org/10.1016/j.ajhg.2018.07.015.

[22] V Brusic, G Rudy, G Honeyman, J Hammer, and L Harrison. "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network." In: *Bioinformatics (Oxford, England)* 14.2 (1998), pp. 121–130. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.2.121.

[23] Huynh-Hoa Hh Bui et al. "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications." In: *Immunogenetics* 57.5 (2005), pp. 304–14. ISSN: 0093-7711. DOI: 10.1007/s00251-005-0798-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/15868141.

[24] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." In: *Nucleic Acids Research* 47.D1 (2019), pp. D1005–D1012. ISSN: 1362-4962. DOI: 10.1093/nar/gky1120.

[25] Marta Byrska-Bishop et al. "High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios." In: *SSRN Electronic Journal* (2021). DOI: 10.2139/ssrn.3967671.

[26] Marta Byrska-Bishop et al. "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios." In: *Cell* 185.18 (2022), 3426–3440.e19. ISSN: 10974172. DOI: 10.1016/j.cell.2022.08.004.

[27] Stewart T. Chang, Debashis Ghosh, Denise E. Kirschner, and Jennifer J. Linderman. "Peptide length-based prediction of peptide-MHC class II binding." In: *Bioinformatics (Oxford, England)* 22.22 (2006), pp. 2761–2767. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl479.

[28] Vylyny Chat, Robert Ferguson, Leah Morales, and Tomas Kirchhoff. "Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies." In: *Frontiers in Genetics* 12.February (2022), pp. 1–9. ISSN: 1664-8021. DOI: 10.3389/fgene.2021.790445.

[29] Binbin Chen et al. "Predicting HLA class II antigen presentation through integrated deep learning." In: *Nature biotechnology* (2019). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0280-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/31611695.

[30] Geng Chen, Baitang Ning, and Tieliu Shi. "Single-cell RNA-seq technologies and related computational data analysis." In: *Frontiers in Genetics* 10.APR (2019), pp. 1–13. ISSN: 16648021. DOI: 10.3389/fgene.2019.00317.

[31] Rui Chen, Kelly M. Fulton, Susan M. Twine, and Jianjun Li. "Identification of Mhc Peptides Using Mass Spectrometry for Neoantigen Discovery and Cancer Vaccine Development." In: *Mass Spectrometry Reviews* (2019). ISSN: 1098-2787. DOI: 10.1002/mas.21616. URL: http://dx.doi.org/10.1002/mas.21616.

[32] Jun Cheng, Kaïdre Bendjama, Karola Rittner, and Brandon Malone. "BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning." In: *Bioinformatics* 37.June (2021), pp. 4172–4179. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab422.

[33] Suzanne Clancy. "RNA Splicing: Introns, Exons and Spliceosome." In: *Nature Education* (2008). URL: http://www.nature.com/scitable/topicpage/RNA-Splicing-Introns-Exons-and-Spliceosome-12375.

[34] Isabelle Cleynen et al. "Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study." In: *Lancet (London, England)* 387.10014 (Jan. 2016), pp. 156–67. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(15)00465-1. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4714968.

[35] Seungho Cook, Wanson Choi, Hyunjoon Lim, Yang Luo, Kunhee Kim, Xiaoming Jia, Soumya Raychaudhuri, and Buhm Han. "Accurate imputation of human leukocyte antigens with CookHLA." In: *Nature Communications* 12.1 (2021), pp. 1–11. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21541-5. URL: http://dx.doi.org/10.1038/s41467-021-21541-5.

[36] Daniele Corridoni et al. "Single-cell atlas of colonic CD8+ T cells in ulcerative colitis." In: *Nature Medicine* 26.9 (Sept. 2020), pp. 1480–1490. ISSN: 1078-8956. DOI: 10.1038/s41591-020-1003-4. URL: https://www.nature.com/articles/s41591-020-1003-4.

[37] Adrian Cortes and Matthew A Brown. "Promise and pitfalls of the Immunochip." In: (2011), pp. 2010–2012.

[38] Eduardo P. Costa, Gerben Menschaert, Walter Luyten, Kurt De Grave, and Jan Ramon. "PIUS: Peptide identification by unbiased search." In: *Bioinformatics* 29.15 (2013), pp. 1913–1914. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt298.

[39] Sayantan Das et al. "Next-generation genotype imputation service and methods." In: *Nature Genetics* 48.10 (2016), pp. 1284–1287. ISSN: 1546-1718. DOI: 10.1038/ng.3656.

[40] Katrina M. De Lange et al. "Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease." In: *Nature Genetics* 49.2 (2017), pp. 256–261. ISSN: 1546-1718. DOI: 10.1038/ng.3760.

[41] Séverine Dédier, Stefan Reinelt, Séverine Rion, Gerd Folkers, and Didier Rognan. "Use of fluorescence polarization to monitor MHC-peptide interactions in solution." In: *Journal of Immunological Methods* 255.1-2 (2001), pp. 57–66. ISSN: 0022-1759. DOI: 10.1016/S0022-1759(01)00423-9.

[42] F. Degenhardt and A. Franke. "Genetik des Morbus Crohn und der Colitis ulcerosa: Aktueller Stand 15 Jahre nach Entdeckung von NOD2." In: *Gastroenterologe* 12.1 (2017), pp. 38–48. ISSN: 1861-969X. DOI: 10.1007/s11377-016-0127-z.

[43] Frauke Degenhardt et al. "Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles." In: *Human molecular genetics* 28.12 (2019), pp. 2078–2092. ISSN: 1460-2083. DOI: 10.1093/hmg/ddy443.

[44] Frauke Degenhardt et al. "Transethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals not only shared but also ethnicity-specific disease associations." In: *Human molecular genetics* 30.5 (Apr. 2021), pp. 356–369. ISSN: 1460-2083. DOI: 10.1093/hmg/ddab017. URL: http://www.ncbi.nlm.nih.gov/pubmed/33555323.

[45] Olivier Delaneau, Jean François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. "Accurate, scalable and integrative haplotype estimation." In: *Nature Communications* 10.1 (2019), pp. 24–29. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13225-y. URL: http://dx.doi.org/10.1038/s41467-019-13225-y.

[46] Alexander T. Dilthey, Loukas Moutsianas, Stephen Leslie, and Gil McVean. "HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes." In: *Bioinformatics* 27.7 (2011), pp. 968–972. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr061.

[47] P. Dönnes and O. Kohlbacher. "SVMHC: a server for prediction of MHC-binding peptides." In: *Nucleic Acids Research* 34.Web Server (July 2006), W194–W197. ISSN: 0305-1048. DOI: 10.1093/nar/gkl284. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl284.

[48] J. L. Duke et al. "Determining performance characteristics of an NGS-based HLA typing method for clinical applications." In: *Hla* 87.3 (2016), pp. 141–152. ISSN: 2059-2310. DOI: 10.1111/tan.12736.

[49] Hesham ElAbd et al. "Predicting Peptide HLA-II Presentation Using Immunopeptidomics , Transcriptomics and Deep Multimodal Learning." In: *bioRxiv* (2022). DOI: https://doi.org/10.1101/2022.09.20.508681.

[50] David Ellinghaus, Jörn Bethune, Britt-Sabina Petersen, and Andre Franke. *The genetics of Crohn's disease and ulcerative colitis – status quo and beyond*. Tech. rep. 1. 2015, pp. 13–23. DOI: 10.3109/00365521.2014.990507. URL: http://informahealthcare.com/doi/abs/10.3109/00365521.2014.990507.

[51] EMBL-EBI. *IMGT/HLA*. 2022. URL: https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics/.

[52] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L. Tress. "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes." In: *Human Molecular Genetics* 23.22 (2014), pp. 5866–5878. ISSN: 1460-2083. DOI: 10.1093/hmg/ddu309.

[53] P. Faridi, M. Dorvash, and A. W. Purcell. "Spliced HLA-bound peptides: a Black Swan event in immunology." In: *Clinical and Experimental Immunology* 204.2 (2021), pp. 179–188. ISSN: 1365-2249. DOI: 10.1111/cei.13589.

[54] Andre Franke et al. "Systematic Association Mapping Identifies NELL1 as a Novel IBD Disease Gene." In: *PLoS ONE* 2.8 (Aug. 2007). Ed. by Greg Gibson, e691. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000691. URL: https://dx.plos.org/10.1371/journal.pone.0000691.

[55] Kelly A. Frazer et al. "A second generation human haplotype map of over 3.1 million SNPs." In: *Nature* 449.7164 (2007), pp. 851–861. ISSN: 0028-0836. DOI: 10.1038/nature06258.

[56] GenDx. *NGSengine® - Analysis software for NGS-based typing*.

[57] Genetic Alliance. "Understanding Genetics: A New York, Mid- Atlantic Guide for Patients and Health Professionals." In: *Genetic Alliance, The New York - Mid-Atlantic Consortium for Genetic and Newborn Screening Services* Appendix E Inheritance Patterns (2008), p. 105.

[58] GenScript. *Advancing genomics, medicine and health together – by semiconductor DNA synthesis technology*. URL: https://www.genscript.com/advancing-genomics-medicine-and-health-together-by-semiconductor-dna-synthesis-technology-summary.html.

[59] Walter Gilbert. "Why genes in pieces?" In: *Nature* 271.5645 (Feb. 1978), pp. 501–501. ISSN: 0028-0836. DOI: 10.1038/271501a0. URL: http://www.nature.com/articles/271501a0.

[60] J. Glas, K. Martin, G. Brünnler, R. Kopp, C. Folwaczny, E. H. Weiss, and E. D. Albert. "MICA, MICB and C1_4_1 polymorphism in Crohn's disease and ulcerative colitis." In: *Tissue Antigens* 58.4 (2001), pp. 243–249. ISSN: 00012815. DOI: 10.1034/j.1399-0039.2001.580404.x.

[61] M. H. Gleeson, J. S. Walker, J. Wentzel, J. A. Chapman, and R. Harris. "Human leucocyte antigens in Crohn's disease and ulcerative colitis." In: *Gut* 13.6 (1972), pp. 438–440. ISSN: 0017-5749. DOI: 10.1136/gut.13.6.438.

[62] Hannah Gordon, Frederik Trier Moller, Vibeke Andersen, and Marcus Harbord. "Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies." In: *Inflammatory bowel diseases* 21.6 (June 2015), pp. 1428–34. ISSN: 1536-4844. DOI: 10.1097/MIB.0000000000000393. URL: http://www.ncbi.nlm.nih.gov/pubmed/25895112.

[63] Philippe Goyette et al. "High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis." In: *Nature Genetics* 47.2 (2015), pp. 172–179. ISSN: 1061-4036. DOI: 10.1038/ng.3176. URL: http://www.nature.com/doifinder/10.1038/ng.3176.

[64] Daniel B. Graham and Ramnik J. Xavier. "Pathway paradigms revealed from the genetics of inflammatory bowel disease." In: *Nature* 578.7796 (Feb. 2020), pp. 527–539. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2025-2. URL: http://www.nature.com/articles/s41586-020-2025-2.

[65] Jochen Graw. *Genetik*. Springer Lehrbuch, 2010. ISBN: 978-3-64-204998-9. DOI: 10.1007/978-3-642-04999-6.

[66] GRC. *Human Genome Assembly GRCh37*. 2009. URL: https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37.

[67] GRC. *Human Genome Assembly GRCh38*. 2013. URL: https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38.

[68] GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues." In: *Science (New York, N.Y.)* 369.6509 (2020), pp. 1318–1330. ISSN: 1095-9203. DOI: 10.1126/science.aaz1776. URL: http://www.ncbi.nlm.nih.gov/pubmed/32913098.

[69] Colleen T. Harrington, Elaine I. Lin, Matthew T. Olson, and James R. Eshleman. "Fundamentals of pyrosequencing." In: *Archives of Pathology and Laboratory Medicine* 137.9 (2013), pp. 1296–1303. ISSN: 0003-9985. DOI: 10.5858/arpa.2012-0463-RA.

[70] Robert Häsler et al. "Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease." In: *Gut* 66.12 (2017), pp. 2087–2097. ISSN: 1468-3288. DOI: 10.1136/gutjnl-2016-311651. URL: http://134.245.63.235/ikmb-tools/mucosaRNA/.

[71] Graham A. Heap et al. "Clinical Features and HLA Association of 5-Aminosalicylate (5-ASA)-induced Nephrotoxicity in Inflammatory Bowel Disease." In: *Journal of Crohn's and Colitis* 10.2 (Feb. 2016), pp. 149–158. ISSN: 1873-9946. DOI: 10.1093/ecco-jcc/jjv219. URL: http://www.ncbi.nlm.nih.gov/pubmed/26619893.

[72] James M. Heather and Benjamin Chain. "The sequence of sequencers: The history of sequencing DNA." In: *Genomics* 107.1 (2016), pp. 1–8. ISSN: 1089-8646. DOI: 10.1016/j.ygeno.2015.11.003. URL: http://dx.doi.org/10.1016/j.ygeno.2015.11.003.

[73] A. S. Hinrichs. "The UCSC Genome Browser Database: update 2006." In: *Nucleic Acids Research* 34.90001 (2006), pp. D590–D598. ISSN: 0305-1048. DOI: 10.1093/nar/gkj144.

[74] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. "Next-generation sequencing technologies: An overview." In: *Human Immunology* 82.11 (Nov. 2021), pp. 801–811. ISSN: 0198-8859. DOI: 10.1016/j.humimm.2021.02.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0198885921000628.

[75] Carolyn Katovich Hurley. "Naming HLA diversity: A review of HLA nomenclature." In: *Human Immunology* 82.7 (2021), pp. 457–465. ISSN: 1879-1166. DOI: 10.1016/j.humimm.2020.03.005.

[76] Illumina. *Infinium™ Global Screening Array-24 v3.0 BeadChip*. 2016. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-global-screening-array-data-sheet-370-2016-016.pdf.

[77] Illumina. *Read length recommendations*. 2022. URL: https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html.

[78] Illumina. *Trusted bead-based technology - A fundamentally different approach to high-density arrays*. URL: https://www.illumina.com/science/technology/microarray.html.

[79] Kamilla Kjaergaard Jensen, Massimo Andreatta, Paolo Marcatili, Søren Buus, Jason A. Greenbaum, Zhen Yan, Alessandro Sette, Bjoern Peters, and Morten Nielsen. "Improved methods for predicting peptide binding affinity to MHC class II molecules." In: *Immunology* 154.3 (July 2018), pp. 394–406. ISSN: 0019-2805. DOI: 10.1111/imm.12889. URL: http://doi.wiley.com/10.1111/imm.12889.

[80] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei Min Chen, Patrick J. Concannon, Stephen S. Rich, Soumya Raychaudhuri, and Paul I W de Bakker. "Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens." In: *PLoS ONE* 8.6 (2013). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0064683.

[81] Tiira Johansson, Dawit A. Yohannes, Satu Koskela, Jukka Partanen, and Päivi Saavalainen. "HLA RNAseq reveals high allele-specific variability in mRNA expression." In: *bioRxiv* (2018), p. 413534. DOI: 10.1101/413534. URL: https://www.biorxiv.org/content/early/2018/09/10/413534.

[82] Seulgi Jung et al. "Identification of three novel susceptibility loci for inflammatory bowel disease in Koreans in an extended genome-wide association study." In: *Journal of Crohn's & colitis* 54 (Apr. 2021), pp. 1–54. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjab060. URL: http://www.ncbi.nlm.nih.gov/pubmed/33853113.

[83] Mollie M Jurewicz, Richard A Willis, Vasanthi Ramachandiran, John D Altman, and Lawrence J Stern. "MHC-I peptide binding activity assessed by exchange after cleavage of peptide covalently linked to β2-microglobulin." In: *Analytical Biochemistry* 584 (Nov. 2019), p. 113328. ISSN: 0003-2697. DOI: 10.1016/j.ab.2019.05.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0003269719301800.

[84] Assaf Kacen et al. "Uncovering the modified immunopeptidome reveals insights into principles of PTM-driven antigenicity." In: *bioRxiv* (2021), p. 2021.04.10.438991. URL: https://doi.org/10.1101/2021.04.10.438991.

[85] Yoichi Kakuta et al. "Repertoire analysis of memory T-cell receptors in Japanese patients with inflammatory bowel disease." In: *JGH Open* 4.4 (2020), pp. 624–631. ISSN: 2397-9070. DOI: 10.1002/jgh3.12305.

[86] Masahiro Kanai, Toshihiro Tanaka, and Yukinori Okada. "Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set." In: *Journal of Human Genetics* 61.10 (2016), pp. 861–866. ISSN: 1435-232X. DOI: 10.1038/jhg.2016.72.

[87] Edita Karosiene, Michael Rasmussen, Thomas Blicher, Ole Lund, Søren Buus, and Morten Nielsen. "NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ." In: *Immunogenetics* 65.10 (Oct. 2013), pp. 711–24. ISSN: 1432-1211. DOI: 10.1007/s00251-013-0720-y. URL: http://link.springer.com/10.1007/s00251-013-0720-y.

[88] Arthur Kaser, Sebastian Zeissig, and Richard S. Blumberg. "Inflammatory Bowel Disease." In: *Annual Review of Immunology* 28.1 (Mar. 2010), pp. 573–621. ISSN: 0732-0582. DOI: 10.1146/annurev-immunol-030409-101225. URL: http://www.annualreviews.org/doi/10.1146/annurev-immunol-030409-101225.

[89] Jan Christian Kässens, Lars Wienbrandt, and David Ellinghaus. "BIGwas: Single-command quality control and association testing for multi-cohort and biobank-scale GWAS/PheWAS data." In: *GigaScience* 10.6 (2021), pp. 1–12. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab047.

[90] Stefan H. E. Kaufmann. *Basiswissen Immunologie*. Springer-Lehrbuch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, p. 133. ISBN: 978-3-642-40324-8. DOI: 10.1007/978-3-642-40325-5. URL: http://link.springer.com/10.1007/978-3-642-40325-5.

[91] Jan H. Kessler, Willemien E. Benckhuijsen, Tuna Mutis, Cornelis J.M. Melief, Sjoerd H. van der Burg, and Jan W. Drijfhout. "Competition-based cellular peptide binding assay for HLA class I." In: *Current protocols in immunology* Chapter 18 (2004). ISSN: 1934-368X. DOI: 10.1002/0471142735.im1812s61.

[92] Emily A. King, J. Wade Davis, and Jacob F. Degner. "Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval." In: *PLoS Genetics* 15.12 (2019), pp. 1–20. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008489. URL: http://dx.doi.org/10.1371/journal.pgen.1008489.

[93] Robert J. Klein et al. "Complement Factor H Polymorphism in Age-Related Macular Degeneration." In: *Science* 308.5720 (Apr. 2005), pp. 385–389. ISSN: 0036-8075. DOI: 10.1126/science.1109557. URL: https://www.science.org/doi/10.1126/science.1109557.

[94] Jerzy K. Kulski, Takashi Shiina, Tatsuya Anzai, Sakae Kohara, and Hidetoshi Inoko. "Comparative genomic analysis of the MHC: The evolution of class I duplication blocks, diversity and complexity from shark to man." In: *Immunological Reviews* 190.1 (2002), pp. 95–122. ISSN: 0105-2896. DOI: 10.1034/j.1600-065X.2002.19008.x.

[95] Thomas LaFramboise. "Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances." In: *Nucleic Acids Research* 37.13 (2009), pp. 4181–4193. ISSN: 0305-1048. DOI: 10.1093/nar/gkp552.

[96] E. S. Lander and D. Botstein. "Mapping mendelian factors underlying quantitative traits using RFLP linkage maps." In: *Genetics* 121.1 (Jan. 1989), pp. 185–99. ISSN: 0016-6731. DOI: 10.1093/genetics/121.1.185. URL: https://academic.oup.com/genetics/article/121/1/185/5997927.

[97] Tuuli Lappalainen, Alexandra J. Scott, Margot Brandt, and Ira M. Hall. "Genomic Analysis in the Age of Human Genome Sequencing." In: *Cell* 177.1 (2019), pp. 70–84. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.02.032. URL: https://doi.org/10.1016/j.cell.2019.02.032.

[98] Sneh Lata, Manoj Bhasin, and Gajendra P.S. Raghava. *Immunoinformatics*. Ed. by Darren R. Flower. Vol. 409. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2007, pp. 201–215. ISBN: 978-1-60327-118-9. DOI: 10.1007/978-1-60327-118-9. URL: http://link.springer.com/10.1007/978-1-60327-118-9.

[99] Aonghus Lavelle and Harry Sokol. "The Gut Microbiome in Inflammatory Bowel Disease." In: *Molecular Genetics of Inflammatory Bowel Disease*. Cham: Springer International Publishing, 2019, pp. 347–377. DOI: 10.1007/978-3-030-28703-0_16. URL: http://link.springer.com/10.1007/978-3-030-28703-0%7B%5C_%7D16.

[100] Mark N. Lee and Matthew Meyerson. "Antigen identification for HLA class I– and HLA class II–restricted T cell receptors using cytokine-capturing antigen-presenting cells." In: *Science Immunology* 6.55 (Jan. 2021), pp. 139–148. ISSN: 2470-9468. DOI: 10.1126/sciimmunol.abf4001. URL: https://www.science.org/doi/10.1126/sciimmunol.abf4001.

[101] Seunggeung Lee, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. "Rare-Variant Association Analysis : Study Designs and Statistical Tests." In: (2014), pp. 5–23. DOI: 10.1016/j.ajhg.2014.06.009.

[102] Evangelia Legaki. "Influence of environmental factors in the development of inflammatory bowel diseases." In: *World Journal of Gastrointestinal Pharmacology and Therapeutics* 7.1 (2016), p. 112. ISSN: 2150-5349. DOI: 10.4292/wjgpt.v7.i1.112. URL: http://www.wjgnet.com/2150-5349/full/v7/i1/112.htm.

[103] Florence Lichou and Gosia Trynka. "Functional studies of GWAS variants are gaining momentum." In: *Nature Communications* 11.1 (2020), pp. 2–5. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20188-y. URL: http://dx.doi.org/10.1038/s41467-020-20188-y.

[104] Bryan Linggi et al. "Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis." In: *Scientific Reports* 11.1 (2021), pp. 1–12. ISSN: 2045-2322. DOI: 10.1038/s41598-021-97366-5. URL: https://doi.org/10.1038/s41598-021-97366-5.

[105] Jimmy Z Liu et al. "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations." In: *Nature Genetics* 47.October 2014 (Sept. 2015), pp. 979–986. ISSN: 1061-4036. DOI: 10.1038/ng.3359. URL: http://www.nature.com/articles/ng.3359.

[106] Zhonghao Liu, Jing Jin, Yuxin Cui, Zheng Xiong, Alireza Nasiri, Yong Zhao, and Jianjun Hu. "DeepSeqPanII: an interpretable recurrent neural network model with attention mechanism for peptide-HLA class II binding prediction." In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021), pp. 1–9. ISSN: 1557-9964. DOI: 10.1109/TCBB.2021.3074927.

[107] Po Ru Loh et al. "Reference-based phasing using the Haplotype Reference Consortium panel." In: *Nature Genetics* 48.11 (2016), pp. 1443–1448. ISSN: 1546-1718. DOI: 10.1038/ng.3679.

[108] C. Lundegaard, O. Lund, C. Kesmir, S. Brunak, and M. Nielsen. "Modeling the adaptive immune system: predictions and simulations." In: *Bioinformatics* 23.24 (2007), pp. 3265–3275. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm471. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm471.

[109] Yang Luo et al. "A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response." In: *Nature Genetics* 53.10 (2021), pp. 1504–1516. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00935-7.

[110] Ana Marcu et al. "HLA Ligand Atlas: A benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy." In: *Journal for ImmunoTherapy of Cancer* 9.4 (2021), pp. 16–18. ISSN: 2051-1426. DOI: 10.1136/jitc-2020-002071.

[111] Karen Maresso and Ulrich Broeckel. "5 Genotyping Platforms for Mass-Throughput Genotyping with SNPs , Including Human Genome-Wide Scans." In: vol. 60. 07. 2008. DOI: 10.1016/S0065-2660(07)00405-1.

[112] Marcel Margulies et al. "Genome sequencing in microfabricated high-density picolitre reactors." In: *Nature* 437.7057 (2005), pp. 376–380. ISSN: 0028-0836. DOI: 10.1038/nature03959.

[113] Shane McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation." In: *Nature Genetics* 48.10 (2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[114] Shutao Mei, Rochelle Ayala, Sri H. Ramarathinam, Patricia T. Illing, Pouya Faridi, Jiangning Song, Anthony W. Purcell, and Nathan P. Croft. "Immunopeptidomic analysis reveals that deamidated HLA-bound peptides arise predominantly from deglycosylated precursors." In: *Molecular and Cellular Proteomics* 19.7 (2020), pp. 1236–1247. ISSN: 1535-9484. DOI: 10.1074/mcp.RA119.001846. URL: http://dx.doi.org/10.1074/mcp.RA119.001846.

[115] Janine Meienberg et al. "New insights into the performance of human whole-exome capture platforms." In: *Nucleic Acids Research* 43.11 (2015). ISSN: 1362-4962. DOI: 10.1093/nar/gkv216.

[116] "Method of the year." In: *Nature Methods* 5.1 (2008), p. 1. ISSN: 1548-7091. DOI: 10.1038/nmeth1153.

[117] *Michigan Imputation Server - Free Next-Generation Imputation Service*. URL: https://imputationserver.sph.umich.edu/.

[118] *Minimac4*. URL: https://genome.sph.umich.edu/wiki/Minimac4.

[119] Vanessa Mitsialis et al. "Single-Cell Analyses of Colon and Blood Reveal Distinct Immune Cell Signatures of Ulcerative Colitis and Crohn's Disease." In: *Gastroenterology* 159.2 (2020), 591–608.e10. ISSN: 15280012. DOI: 10.1053/j.gastro.2020.04.074. URL: http://dx.doi.org/10.1053/j.gastro.2020.04.074.

[120] Manuel Muro, Ruth López-Hernández, and Anna Mrowiec. "Immunogenetic biomarkers in inflammatory bowel diseases: Role of the IBD3 region." In: *World Journal of Gastroenterology* 20.41 (2014), pp. 15037–15048. ISSN: 2219-2840. DOI: 10.3748/wjg.v20.i41.15037.

[121] National Heart, Lung and Blood Institute. *TOPMed imputation Server*. URL: https://imputation.biodatacatalyst.nhlbi.nih.gov.

[122] National Human Genome Research Instutute. *Allele*. 2022. URL: https://www.genome.gov/genetics-glossary/Allele.

[123] National Human Genome Research Instutute. *Deoxyribonucleic Acid (DNA)*. 2022. URL: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid.

[124] National Institutes of Health in the Department of Health an Human Services. *Immune Epitope Database (IEDB)*. 2022. URL: https://www.iedb.org.

[125] NCBI. *GRCh37*. 2009. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF%7B%5C_%7D000001405.13/%7B%5C#%7D/def%7B%5C_%7Dasm%7B%5C_%7DALT%7B%5C_%7DREF%7B%5C_%7DLOCI%7B%5C_%7D9.

[126] NCBI. *GRCh38*. 2013. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF%7B%5C_%7D000001405.26/.

[127] Jacques Neefjes, Marlieke L. M. Jongsma, Petra Paul, and Oddmund Bakke. "Towards a systems understanding of MHC class I and MHC class II antigen presentation." In: *Nature Reviews Immunology* 11.12 (Dec. 2011), pp. 823–836. ISSN: 1474-1733. DOI: 10.1038/nri3084. arXiv: arXiv:1011.1669v3. URL: http://www.nature.com/articles/nri3084.

[128] Matthew R. Nelson et al. "The support of human genetic evidence for approved drug indications." In: *Nature Genetics* 47.8 (2015), pp. 856–860. ISSN: 1546-1718. DOI: 10.1038/ng.3314.

[129] Sarah C Nelson, Kimberly F Doheny, Cathy C Laurie, and Daniel B Mirel. "Is 'forward' the same as 'plus'?. . . and other adventures in SNP allele nomenclature." In: *Trends in genetics : TIG* 28.8 (Aug. 2012), pp. 361–3. ISSN: 0168-9525. DOI: 10.1016/j.tig.2012.05.002. URL: http://www.ncbi.nlm.nih.gov/pubmed/22658725.

[130] Siew C. Ng et al. "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies." In: *Lancet (London, England)* 390.10114 (Dec. 2017), pp. 2769–2778. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(17)32448-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/29050646.

[131] Morten Nielsen. *NetMHCIIpan-4.1*. URL: https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-4.1.

[132] Morten Nielsen and Massimo Andreatta. "NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions." In: *Nucleic Acids Research* 45.April (2017), pp. 2–7. ISSN: 0305-1048. DOI: 10.1093/nar/gkx276. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx276.

[133] Morten Nielsen and Ole Lund. "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction." In: *BMC Bioinformatics* 10.1 (Dec. 2009), p. 296. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-296. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-296.

[134] Morten Nielsen, Ole Lund, Søren Buus, and Claus Lundegaard. "MHC Class II epitope predictive algorithms." In: *Immunology* 130.3 (July 2010), pp. 319–328. ISSN: 00192805. DOI: 10.1111/j.1365-2567.2010.03268.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2567.2010.03268.x.

[135] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Bjoern Peters, Alessandro Sette, Sune Justesen, Søren Buus, and Ole Lund. "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan." In: *PLoS Computational Biology* 4.7 (2008), pp. 1–10. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1000107.

[136] Morten Nielsen, Claus Lundegaard, and Ole Lund. "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method." In: *BMC Bioinformatics* 8.1 (2007), p. 238. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-238. URL: http://www.biomedcentral.com/1471-2105/8/238.

[137] Nurulamin M. Noor, Bram Verstockt, Miles Parkes, and James C. Lee. "Personalised medicine in Crohn's disease." In: *The lancet. Gastroenterology & hepatology* 5.1 (2020), pp. 80–92. ISSN: 2468-1253. DOI: 10.1016/S2468-1253(19)30340-1. URL: http://dx.doi.org/10.1016/S2468-1253(19)30340-1.

[138] Eduardo Nunes et al. "Definitions of histocompatibility typing terms." In: *Blood* 118.23 (2011), pp. 180–183. ISSN: 1528-0020. DOI: 10.1182/blood-2011-05-353490.

[139] Omixon Inc. *HLA TWIN\texttrademark*. URL: https://www.omixon.com/products/hla-twin/.

[140] Ingrid Ordás, Lars Eckmann, Mark Talamini, Daniel C. Baumgart, and William J. Sandborn. "Ulcerative colitis." In: *Lancet (London, England)* 380.9853 (Nov. 2012), pp. 1606–19. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(12)60150-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/22914296.

[141] Kouichi Ozaki et al. "Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction." In: *Nature Genetics* 32.4 (2002), pp. 650–654. ISSN: 1061-4036. DOI: 10.1038/ng1047.

[142] Britt Sabina Petersen, Broder Fredrich, Marc P. Hoeppner, David Ellinghaus, and Andre Franke. "Opportunities and challenges of whole-genome and -exome sequencing." In: *BMC Genetics* 18.1 (2017), pp. 1–13. ISSN: 1471-2156. DOI: 10.1186/s12863-017-0479-5.

[143] Poomarin Phloyphisut, Natapol Pornputtapong, Sira Sriswasdi, and Ekapol Chuangsuwanich. "MHCSeqNet: A deep neural network model for universal MHC binding prediction." In: *BMC Bioinformatics* 20.1 (2019), pp. 1–10. ISSN: 1471-2105. DOI: 10.1186/s12859-019-2892-4.

[144] Anthony W. Purcell, Sri H. Ramarathinam, and Nicola Ternette. "Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics." In: *Nature Protocols* 14.6 (June 2019), pp. 1687–1707. ISSN: 1754-2189. DOI: 10.1038/s41596-019-0133-y. URL: http://www.nature.com/articles/s41596-019-0133-y.

[145] Julien Racle et al. "Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes." In: *Nature Biotechnology* 37.11 (2019), pp. 1283–1286. ISSN: 15461696. DOI: 10.1038/s41587-019-0289-6. URL: http://dx.doi.org/10.1038/s41587-019-0289-6.

[146] H. -G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović. "SYFPEITHI: database for MHC ligands and peptide motifs." In: *Immunogenetics* 50.3-4 (Nov. 1999), pp. 213–219. ISSN: 0093-7711. DOI: 10.1007/s002510050595. URL: http://link.springer.com/10.1007/s002510050595.

[147] Hans-Georg Rammensee, T. Friede, and S. Stevanoviíc. "MHC ligands and peptide motifs: first listing." In: *Immunogenetics* 41.4 (1995), pp. 178–228. ISSN: 0093-7711. DOI: 10.1007/BF00172063.

[148] Greeshma Ray and Michelle S. Longworth. "Epigenetics, DNA Organization, and Inflammatory Bowel Disease." In: *Inflammatory Bowel Diseases* 25.2 (2019), pp. 235–247. ISSN: 1536-4844. DOI: 10.1093/ibd/izy330.

[149] Pedro A. Reche, Hong Zhang, John Paul Glutting, and Ellis L. Reinherz. "EPIMHC: A curated database of MHC-binding peptides for customized computational vaccinology." In: *Bioinformatics* 21.9 (2005), pp. 2140–2141. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti269.

[150] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. "NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data." In: *Nucleic Acids Research* 48.W1 (July 2020), W449–W454. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa379.

[151] J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. E. Marsh. "The IPD and IMGT/HLA database: allele variant databases." In: *Nucleic Acids Research* 43.D1 (2015), pp. D423–D431. ISSN: 0305-1048. DOI: 10.1093/nar/gku1161. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1161.

[152] James Robinson, Dominic J. Barker, Xenia Georgiou, Michael A. Cooper, Paul Flicek, and Steven G.E. Marsh. "IPD-IMGT/HLA Database." In: *Nucleic Acids Research* 48.D1 (2020), pp. D948–D955. ISSN: 1362-4962. DOI: 10.1093/nar/gkz950.

[153] Kenneth L. Rock, Eric Reits, and Jacques Neefjes. "Present Yourself! By MHC Class I and MHC Class II Molecules." In: *Trends in Immunology* 37.11 (2016), pp. 724–737. ISSN: 1471-4981. DOI: 10.1016/j.it.2016.08.010. URL: http://dx.doi.org/10.1016/j.it.2016.08.010.

[154] Elisa Rosati et al. "A novel unconventional T cell population enriched in Crohn's disease." In: *Gut* Published (2022), gutjnl–2021–325373. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2021-325373.

[155] Alan B. Rose. "Introns as gene regulators: A brick on the accelerator." In: *Frontiers in Genetics* 10.FEB (2019), pp. 1–6. ISSN: 1664-8021. DOI: 10.3389/fgene.2018.00672.

[156] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. "Genotype imputation using the Positional Burrows Wheeler Transform." In: *PLoS Genetics* 16.11 (2020), pp. 1–19. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1009049. URL: http://dx.doi.org/10.1371/journal.pgen.1009049.

[157] Milagros D Samaniego, Yolanda T Becker, and Joshua David Lindsey. "14 - Replacement Therapy for Kidney Failure: Transplantation." In: ed. by A Vishnu B T - Pathophysiology of Kidney Disease Moorthy and Hypertension. W.B. Saunders, 2009, pp. 169–177. ISBN: 978-1-4160-4391-1. DOI: https://doi.org/10.1016/B978-1-4160-4391-1.50020-5. URL: https://www.sciencedirect.com/science/article/pii/B9781416043911500205.

[158] Jose L. Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A. Reche. "Fundamentals and Methods for T- and B-Cell Epitope Prediction." In: *Journal of Immunology Research* 2017 (2017). ISSN: 2314-7156. DOI: 10.1155/2017/2680160.

[159] Melanie Schirmer, Ashley Garner, Hera Vlamakis, and Ramnik J. Xavier. "Microbial genes and pathways in inflammatory bowel disease." In: *Nature Reviews Microbiology* 17.August (2019), pp. 497–511. ISSN: 1740-1534. DOI: 10.1038/s41579-019-0213-6. URL: http://dx.doi.org/10.1038/s41579-019-0213-6.

[160] Detlef Schuppan et al. "A Randomized Trial of a Transglutaminase 2 Inhibitor for Celiac Disease." In: *New England Journal of Medicine* 385.1 (2021), pp. 35–45. ISSN: 0028-4793. DOI: 10.1056/nejmoa2032441.

[161] Robert J. Seward, Elise E. Drouin, Allen C. Steere, and Catherine E. Costello. "Peptides Presented by HLA-DR Molecules in Synovia of Patients with Rheumatoid Arthritis or Antibiotic-Refractory Lyme Arthritis." In: *Molecular & Cellular Proteomics* 10.3 (2011), p. M110.002477. ISSN: 1535-9476. DOI: 10.1074/mcp.M110.002477. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.M110.002477.

[162] Xiaoshan M Shao et al. "High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets." In: *Cancer Immunology Research* 8.3 (Mar. 2020), pp. 396–408. ISSN: 2326-6066. DOI: 10.1158/2326-6066.CIR-19-0464. URL: http://cancerimmunolres.aacrjournals.org/lookup/doi/10.1158/2326-6066.CIR-19-0464.

[163] Jay Shendure et al. "Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome." In: *Science* 309.5741 (2005), pp. 1728–1732. ISSN: 0036-8075. DOI: 10.1126/science.1117389.

[164] Takashi Shiina, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K. Kulski. "The HLA genomic loci map: Expression, interaction, diversity and disease." In: *Journal of Human Genetics* 54.1 (2009), pp. 15–39. ISSN: 1434-5161. DOI: 10.1038/jhg.2008.5.

[165] H Singh and G P Raghava. "ProPred: prediction of HLA-DR binding sites." In: *Bioinformatics (Oxford, England)* 17.12 (2001), pp. 1236–1237. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.12.1236.

[166] Montgomery Slatkin. "Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future." In: *Nature Reviews Genetics* 9.6 (2008), pp. 477–485. ISSN: 1471-0056. DOI: 10.1038/nrg2361.

[167] Christopher S. Smillie et al. "Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis." In: *Cell* 178.3 (2019), 714–730.e22. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.06.029.

[168] Wade P. Smith, Quyen Vu, Shuying Sue Li, John A. Hansen, Lue Ping Zhao, and Daniel E. Geraghty. "Toward understanding MHC disease associations: Partial resequencing of 46 distinct HLA haplotypes." In: *Genomics* 87.5 (2006), pp. 561–571. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2005.11.020.

[169] Ludvig M Sollid and Bana Jabri. "Triggers and drivers of autoimmunity: lessons from coeliac disease." In: *Nature reviews. Immunology* 13.4 (2013), pp. 294–302. ISSN: 1474-1741. DOI: 10.1038/nri3407. arXiv: NIHMS150003. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3818716%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[170] Felix Sommer et al. "Microbiomarkers in inflammatory bowel diseases: Caveats come with caviar." In: *Gut* 66.10 (2017), pp. 1734–1738. ISSN: 1468-3288. DOI: 10.1136/gutjnl-2016-313678.

[171] Katharina Stahl, Damian Gola, and Inke R. König. "Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data." In: *Frontiers in Genetics* 12.September (2021), pp. 1–12. ISSN: 1664-8021. DOI: 10.3389/fgene.2021.724037.

[172] Curt Stern. "The Hardy-Weinberg Law." In: *Science* 97.2510 (Feb. 1943), pp. 137–138. ISSN: 0036-8075. DOI: 10.1126/science.97.2510.137. URL: https://www.science.org/doi/10.1126/science.97.2510.137.

[173] P. C.F. Stokkers, P. H. Reitsma, G. N.J. Tytgat, and S. J.H. Van Deventer. "HLA-DR and -DQ phenotypes in inflammatory bowel disease: A meta-analysis." In: *Gut* 45.3 (1999), pp. 395–401. ISSN: 0017-5749. DOI: 10.1136/gut.45.3.395.

[174] T Sturniolo et al. "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices." In: *Nature biotechnology* 17.6 (1999), pp. 555–561. ISSN: 1087-0156. DOI: 10.1038/9858.

[175] Christopher Szeto et al. "Impact of HLA-DR antigen binding cleft rigidity on T cell recognition." In: *International Journal of Molecular Sciences* 21.19 (2020), pp. 1–20. ISSN: 1422-0067. DOI: 10.3390/ijms21197081.

[176] Sasha Taleban et al. "Ocular manifestations in inflammatory bowel disease are associated with other extra-intestinal manifestations, gender, and genes implicated in other immune-related traits." In: *Journal of Crohn's and Colitis* 10.1 (2016), pp. 43–49. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjv178.

[177] Daniel Taliun et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." In: *Nature* 590.7845 (2021), pp. 290–299. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03205-y.

[178] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. "Benefits and limitations of genome-wide association studies." In: *Nature Reviews Genetics* 20.August (2019). ISSN: 1471-0064. DOI: 10.1038/s41576-019-0127-1. URL: http://dx.doi.org/10.1038/s41576-019-0127-1.

[179] Hagar Taman, Christopher G. Fenton, Inga V. Hensel, Endre Anderssen, Jon Florholmen, and Ruth H. Paulssen. "Genome-wide DNA methylation in treatment-naïve ulcerative colitis." In: *Journal of Crohn's and Colitis* 12.11 (2018), pp. 1338–1347. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjy117.

[180] Hagar Taman, Christopher G. Fenton, Inga V. Hensel, Endre Anderssen, Jon Florholmen, and Ruth H. Paulssen. "Transcriptomic landscape of treatment-naïve ulcerative colitis." In: *Journal of Crohn's and Colitis* 12.3 (2018), pp. 327–336. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjx139.

[181] Nitima Tatiya-Aphiradee, Waranya Chatuphonprasert, and Kanokwan Jarukamjorn. "Immune response and inflammatory pathway of ulcerative colitis." In: *Journal of Basic and Clinical Physiology and Pharmacology* 30.1 (2019), pp. 1–10. ISSN: 2191-0286. DOI: 10.1515/jbcpp-2018-0036.

[182] Martin Christen Frolund Thomsen and Morten Nielsen. "Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion." In: *Nucleic Acids Research* 40.W1 (2012), pp. 281–287. ISSN: 0305-1048. DOI: 10.1093/nar/gks469.

[183] Stefka Tyanova, Tikira Temu, and Juergen Cox. "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics." In: *Nature Protocols* 11.12 (2016), pp. 2301–2319. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.136. URL: http://dx.doi.org/10.1038/nprot.2016.136.

[184] Ryan Ungaro, Saurabh Mehandru, Patrick B Allen, Laurent Peyrin-Biroulet, and Jean-Frédéric Colombel. "Ulcerative colitis." In: *Lancet (London, England)* 389.10080 (Apr. 2017), pp. 1756–1770. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(16)32126-2. URL: https

://linkinghub.elsevier.com/retrieve/pii/S0140673616321262.

[185] Werna T.C. Uniken Venema, Michiel D. Voskuil, Gerard Dijkstra, Rinse K. Weersma, and Eleonora A.M. Festen. "The genetic background of inflammatory bowel disease: from correlation to causality." In: *Journal of Pathology* 241.2 (2017), pp. 146–158. ISSN: 1096-9896. DOI: 10.1002/path.4817.

[186] Bianca J.C. Van Den Bosch and Marieke J.H. Coenen. "Pharmacogenetics of inflammatory bowel disease." In: *Pharmacogenomics* 22.1 (2021), pp. 55–66. ISSN: 1744-8042. DOI: 10.2217/pgs-2020-0095.

[187] Stephan R. Vavricka, Alain Schoepfer, Michael Scharl, Peter L. Lakatos, Alexander Navarini, and Gerhard Rogler. "Extraintestinal manifestations of inflammatory bowel disease." In: *Inflammatory Bowel Diseases* 21.8 (2015), pp. 1982–1992. ISSN: 1536-4844. DOI: 10.1097/MIB.0000000000000392.

[188] Gopalakrishnan Venkatesh, Aayush Grover, G Srinivasaraghavan, and Shrisha Rao. "MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model." In: *Bioinformatics* 36.Supplement_1 (July 2020), pp. i399–i406. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa479. URL: http://www.ncbi.nlm.nih.gov/pubmed/32657386.

[189] Bram Verstockt, Kenneth Gc Smith, and James C. Lee. "Genome-wide association studies in Crohn's disease: Past, present and future." In: *Clinical and Translational Immunology* 7.1 (2018), pp. 1–16. ISSN: 2050-0068. DOI: 10.1002/cti2.1001.

[190] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. "10 Years of GWAS Discovery: Biology, Function, and Translation." In: *American Journal of Human Genetics* 101.1 (2017), pp. 5–22. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.06.005. URL: http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

[191] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. "The Immune Epitope Database (IEDB): 2018 update." In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D339–D343. ISSN: 0305-1048. DOI: 10.1093/nar/gky1006. URL: https://academic.oup.com/nar/article/47/D1/D339/5144151.

[192] Karl V. Voelkerding, Shale A. Dames, and Jacob D. Durtschi. "Next-generation sequencing:from basic research to diagnostics." In: *Clinical Chemistry* 55.4 (2009), pp. 641–658. ISSN: 0009-9147. DOI: 10.1373/clinchem.2008.112789.

[193] Ji Wan, Wen Liu, Qiqi Xu, Yongliang Ren, Darren R. Flower, and Tongbin Li. "SVRMHC prediction server for MHC-binding peptides." In: *BMC Bioinformatics* 7 (2006), pp. 1–5. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-463.

[194] Peng Wang, John Sidney, Courtney Dow, Bianca Mothé, Alessandro Sette, and Bjoern Peters. "A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach." In: *PLoS Computational Biology* 4.4 (2008), e1000048. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000048. URL: http://dx.plos.org/10.1371/journal.pcbi.1000048.

[195] Peng Wang, John Sidney, Yohan Kim, Alessandro Sette, Ole Lund, Morten Nielsen, and Bjoern Peters. "Peptide binding predictions for HLA DR, DP and DQ molecules." In: *BMC bioinformatics* 11.1 (2010), p. 568. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-568. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2998531%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[196] Ralf Waßmuth. *Einführung in das HLA-System*. 2. edition. 2005, p. 166. ISBN: 3-609-16332-1.

[197] Mareike Wendorff, Heli M. Garcia Alvarez, Thomas Østerbye, Hesham ElAbd, Elisa Rosati, Frauke Degenhardt, Søren Buus, Andre Franke, and Morten Nielsen. "Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction." In: *Frontiers in Immunology* 11.August (2020), pp. 1–8. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.01705.

[198] L. Werner et al. "Altered T cell receptor beta repertoire patterns in pediatric ulcerative colitis." In: *Clinical and Experimental Immunology* 196.1 (2019), pp. 1–11. ISSN: 1365-2249. DOI: 10.1111/cei.13247.

[199] Lars Wienbrandt, Jan Christian Kässens, Matthias Hübenthal, and David Ellinghaus. "1000× faster than PLINK: Combined FPGA and GPU accelerators for logistic regression-based detection of epistasis." In: *Journal of Computational Science* 30 (2019), pp. 183–193. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2018.12.013. URL: https://doi.org/10.1016/j.jocs.2018.12.013.

[200] Horst Will. *Molekularbiologie kurz und bündig*. 2014. ISBN: 978-3-64-255109-3. DOI: 10.1007/978-3-642-55110-9.

[201] M. Wittig et al. "Development of a high-resolution NGS-based HLA-typing and analysis pipeline." In: *Nucleic Acids Research* (2015), pp. 1–8. ISSN: 0305-1048. DOI: 10.1093/nar/gkv184. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv184.

[202] Andrew R. Wood et al. "Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation." In: *PLoS ONE* 8.5 (2013). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0064343.

[203] Bingbing Xu, Yijun Meng, and Yongfeng Jin. "RNA structures in alternative splicing and back-splicing." In: *Wiley Interdisciplinary Reviews: RNA* 12.1 (2021), pp. 1–39. ISSN: 1757-7012. DOI: 10.1002/wrna.1626.

[204] Lee Min Yap, Tariq Ahmad, and Derek P. Jewell. "The contribution of HLA genes to IBD susceptibility and phenotype." In: *Best Practice and Research: Clinical Gastroenterology* 18.3 (2004), pp. 577–596. ISSN: 1521-6918. DOI: 10.1016/j.bpg.2004.01.003.

[205] Karma Yeshi, Roland Ruscher, Luke Hunter, Norelle L. Daly, Alex Loukas, and Phurpa Wangchuk. "Revisiting inflammatory bowel disease: Pathology, treatments, challenges and emerging therapeutics including drug leads from natural products." In: *Journal of Clinical Medicine* 9.5 (2020), pp. 1–39. ISSN: 2077-0383. DOI: 10.3390/jcm9051273.

[206] Nour Younis, Rana Zarif, and Rami Mahfouz. "Inflammatory bowel disease: between genetics and microbiota." In: *Molecular Biology Reports* 47.4 (2020), pp. 3053–3063. ISSN: 1573-4978. DOI: 10.1007/s11033-020-05318-5. URL: https://doi.org/10.1007/s11033-020-05318-5.

[207] Fengchao Yu, Guo Ci Teo, Andy T. Kong, Sarah E. Haynes, Dmitry M. Avtonomov, Daniel J. Geiszler, and Alexey I. Nesvizhskii. "Identification of modified peptides using localization-aware open search." In: *Nature Communications* 11.1 (2020), pp. 1–9. ISSN: 20411723. DOI: 10.1038/s41467-020-17921-y.

[208] Hao Zhang, Ole Lund, and Morten Nielsen. "The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding." In: *Bioinformatics (Oxford, England)* 25.10 (2009), pp. 1293–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp137. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2732311%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[209] Lianming Zhang, Yiqing Chen, Hau-San Wong, Shuigeng Zhou, Hiroshi Mamitsuka, and Shanfeng Zhu. "TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules." In: *PloS one* 7.2 (2012), e30483. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0030483. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3285624%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[210] Shanrong Zhao. "Alternative splicing, RNA-seq and drug discovery." In: *Drug Discovery Today* 24.6 (2019), pp. 1258–1267. ISSN: 1878-5832. DOI: 10.1016/j.drudis.2019.03.030. URL: https://doi.org/10.1016/j.drudis.2019.03.030.

[211] X Zheng, J Shen, C Cox, J C Wakefield, M G Ehm, M R Nelson, and B S Weir. "HIBAG-HLA genotype imputation with attribute bagging." In: *The Pharmacogenomics Journal* 14.2 (2014), pp. 192–200. ISSN: 1473-1150. DOI: 10.1038/tpj.2013.18. URL: https://pubmed.ncbi.nlm.nih.gov/23712092/.

[212] Xiuwen Zheng. "Imputation-Based HLA Typing with SNPs in GWAS Studies." In: *HLA Typing,Methods in Molecular Biology (Transplantation)*. Ed. by Sebastian Boegel. Humana Press, 2018, pp. 163–176. DOI: 10.1007/978-1-4939-8546-3_11. URL: http://link.springer.com/10.1007/978-1-4939-8546-3%7B%5C_%7D11.

[213] Wei Zhou et al. "Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts." In: *Nature Genetics* 52.6 (2020), pp. 634–639. ISSN: 1546-1718. DOI: 10.1038/s41588-020-0621-6. URL: http://dx.doi.org/10.1038/s41588-020-0621-6.

[214] Tao Zuo and Siew C. Ng. "The Gut Microbiota in the Pathogenesis and Therapeutics of Inflammatory Bowel Disease." In: *Frontiers in Microbiology* 9 (Sept. 2018), p. 2247. ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.02247. URL: https://www.frontiersin.org/article/10.3389/fmicb.2018.02247/full.

DECLARATION

DECLARATION

I hereby declare,

1. - that this thesis is completely the result of my own work. Apart from the advice of my supervisors, all sources are listed in the bibliography.
2. - that the PAPER A has already been published in the Journal of *Human Molecular Genetics*. PAPER B has already been published in *Frontiers in Immunology*. PAPER C has currently not been submitted to any scientific journal but a submission is under preparation.
3. - that the thesis has been prepared according to the rules of the DFG.
4. - that I do not have any academic degree that was withdrawn.

ERKLÄRUNG

Hiermit erkläre ich,

1. - dass diese Arbeit vollständig das Ergebnis meiner eigenen Arbeit ist. Abgesehen von der Beratung durch meine Betreuer sind alle Quellen im Literaturverzeichnis gelistet.
2. - dass die erste Publikation (PAPER A) bereits in der Zeitschrift *Human Molecular Genetics* veröffentlicht wurde. Die zweite Publikation (PAPER B) wurde bereits im *Frontiers in Immunology* veröffentlicht. Die dritte Publikation (PAPER C) wurde bislang noch bei keinem Journal eingereicht, aber ein Einreichen des Manuskripts wird vorbereitet.
3. - dass diese Arbeit unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden ist.
4. - dass mir kein akademischer Grad entzogen worden ist.

Kiel, November 2022

Mareike Wendorff