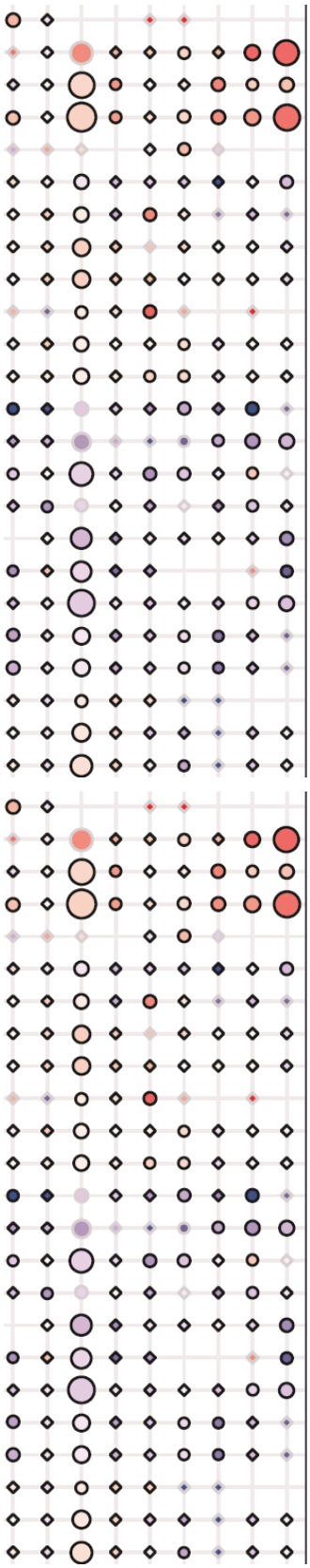


Identification of genetic risk factors within the human leukocyte antigen region for ulcerative colitis

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

Frauke Degenhardt
Kiel, 2020

Identification of genetic risk factors within the human leukocyte antigen region for ulcerative colitis

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

Vorgelegt von

Frauke Degenhardt

Kiel, 2020

Erster Gutachter:	Prof. Dr. Andre Franke
Zweiter Gutachter:	Prof. Dr. Tal Dagan
Tag der mündlichen Prüfung:	16.06.2020
Dekan:	Prof. Dr. Frank Kempken

TABLE OF CONTENTS

I List of main publications i
II List of further publications ii
III List of Figures vii
IV List of Tables viii
V List of Figures and Tables in Papers ix
VI Abbreviations xi
1. Zusammenfassung 1
2. Summary 2
3. Overview of main publications 3
4. Introduction 4
4.1. Inflammatory bowel disease 5
4.1.1. Clinical characteristics of IBD 5
4.1.2. Treatment of IBD 9
4.1.3. IBD demographics 10
4.1.4. PAPER A – IBD genetics 14
4.1.5. Other research areas in IBD 28
4.2. The human leukocyte antigen region (HLA) 32
4.2.1. HLA genes and their function 32
4.2.2. HLA nomenclature 35
4.2.3. Fine mapping of the HLA in IBD 37
4.3. Methodological considerations 40
4.3.1. Linkage Disequilibrium 40
4.3.2. Genotyping and the Illumina ImmunoChip 42
4.3.3. Concepts behind SNP association analyses 44
4.3.4. Methods for HLA typing 49
4.3.5. HLA imputation 52
4.3.6. HLA-peptide binding 55
4.3.7. Useful resources 56
5. Results 60
5.1. PAPER B 60
5.2. PAPER C 77
5.3. Additional results 111
5.3.1. HLA haplotype maps 111

5.3.2. HLA imputation pipeline 114
6. Discussion 116
6.1. Potential roles of arginine and lysine in UC 117
6.2. HLA-*DRB1* expression 119
6.3. DRB1*15 in Crohn's disease and other immune diseases 121
6.4. Limitations 122
6.4.1. Limitations of statistical power to detect specific HLA alleles 122
6.4.2. Limitations of HLA imputation 124
6.4.3. Limitations of peptide prediction 125
6.5. Concluding remarks & Outlook 126
7. References 129
8. Appendix 139
8.1. Appendix A 140
8.2. Appendix B 154
9. Declaration 193
10. Acknowledgements 194

I List of main publications

Paper A

Degenhardt F, Franke A. Genetik des Morbus Crohn und der Colitis ulcerosa: Aktueller Stand 15 Jahre nach Entdeckung von NOD2. Der Gastroenterologe. 2017. doi: 10.1007/s11377-016-0127-z.

Paper B

Degenhardt F*, Wendorff M*, Wittig M, Ellinghaus E, Datta LW, Schembri J, Ng SC, Rosati E, Hübenthal M, Ellinghaus D, Jung ES, Lieb W, Abedian S, Malekzadeh R, Cheon JH, Ellul P, Sood A, Midha V, Thelma BK, Wong SH, Schreiber S, Yamazaki K, Kubo M, Boucher G, Rioux JD, Lenz TL, Brant SR, Franke A. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum Mol Genet. 2019 Jun 15;28(12):2078-2092. doi: 10.1093/hmg/ddy443. PMID: 30590525; PMCID: PMC6548229.

Paper C

Degenhardt F, Mayr G, Wendorff M, Boucher G, Ellinghaus E, Ellinghaus D, ElAbd H, Rosati E, Hübenthal M, Juzenas S, Abedian S, Alizadeh B, BK T, Yang S-K, Duk Ye B, Cheon JH, Datta LW, Daryani NE, Ellul P, Esaki M, Fuyuno Y, McGovern DPB, Haritunians T, Hong M, Juyal G, Jung ES, Kubo M, Kugathasan S, Lenz TL, Leslie S, Malekzadeh R, Midha V, Motyer A, Ng SC, Okou DT, Raychaudhuri S, Schembri J, Schreiber S, Song K, Sood A, Takahashi A, Torres EA, Umeno J, Vahedi H, Weersma RK, Wong SH, Yamazaki K, Karlsen TH, Rioux JD, Brant SR for the MAAIS Recruitment Center, Franke A for the International IBD Genetics Consortium. Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals common disease signatures. Currently under review.

II List of further publications

Manuscripts under review

Koller AC, Greve C, Kaurani L, Klockmeier K, Sivalingam S, **Degenhardt F**, Heilmann-Heimbach S, Freudenberg F, Bittner R, Kittel-Schneider S, Winkler LE, Neukirch F, Breuer D, Herrera-Rivero M, Andlauer T, Ludwig KU, Thiele H, Petersen B, Forstner AJ, Maaser A, Motameny S, Kawalia A, Herms S, Hoffmann P, Lieb W, Franke A, Streit F, Sananbenesi F, Hänig C, Blanc E, Närnberg P, Fischer A, Wanker EE., Rietschel M, Nöthen MM, Reif A, Degenhardt F. Exome sequencing in multiply affected families identifies genes coding for the epigenetic machinery as candidates for schizophrenia. Under review in the Journal of Translational Psychiatry.

Rosati E, **Degenhardt F**, Pogorelyy MV, Minervina AA, Fazio A, Sabet S, Dowds M, Jaekel C, Wendorff M, Hübenthal M, Schreiber S, Mamedov I, Lieb W, Bokemeyer B, Hendricks A, Schafmayer C, Egberts J-H, Bacher P, Franke A. Profiling of T-cell receptor repertoires identifies distinctive disease features and candidate disease-biomarkers in inflammatory bowel diseases. Under review in Gut.

Wendorff M, Garcia Alvarez HM, Østerbye T, ElAbd H, Rosati E, **Degenhardt F**, Buus S, Franke A, Nielsen. Unbiased characterisation of HLA-peptide class II interactions based on large-scale peptide microarrays: assessment of the impact on HLA class II ligand and epitope prediction. Under review in Frontiers of Immunology.

Published manuscripts

1: González-Serna D, Ochoa E, López-Isac E, Julià A, **Degenhardt F**, Ortego-Centeno, N, Radstake TRDJ, Franke A, Marsal S, Mayes MD, Martín J, Márquez A; Scleroderma Genetic Consortium. A cross-disease meta-GWAS identifies four new susceptibility loci shared between systemic sclerosis and Crohn's disease. Sci Rep. 2020 Feb 5;10(1):1862. doi: 10.1038/s41598-020-58741-w. PubMed PMID: 32024964; PubMed Central PMCID: PMC7002703.

2: Mucha S, Baurecht H, Novak N, Rodríguez E, Bej S, Mayr G, Emmert H, Stölzl D, Gerdes S, Jung ES, **Degenhardt F**, Hübenthal M, Ellinghaus E, Kässens JC, Wienbrandt L, Lieb W, Müller-Nurasyid M, Hotze M, Dand N, Grosche S, Marenholz I, Arnold A, Homuth G, Schmidt CO, Wehkamp U, Nöthen MM, Hoffmann P, Paternoster L, Standl M; Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, Bønnelykke K, Ahluwalia TS, Bisgaard H, Peters A, Gieger C, Waldenberger M, Schulz H, Strauch K, Werfel T, Lee YA, Wolfien M, Rosenstiel P, Wolkenhauer O, Schreiber S, Franke A, Weidinger S, Ellinghaus D. Protein-coding variants contribute to the risk of atopic dermatitis and skin-specific gene expression. J Allergy Clin Immunol. 2019 Nov 9. pii: S0091-6749(19)31480-0. doi: 10.1016/j.jaci.2019.10.030. [Epub ahead of print] PubMed PMID: 31707051.

3: Teumer A, Li Y, Ghasemi S, Prins BP, Wuttke M, Hermle T, Giri A, Sieber KB, Qiu C, Kirsten H, Tin A, Chu AY, Bansal N, Feitosa MF, Wang L, Chai JF, Cocca M, Fuchsberger C, Gorski M, Hoppmann A, Horn K, Li M, Marten J, Noce D, Nutile T, Sedaghat S, Sveinbjornsson G, Tayo BO, van der Most PJ, Xu Y, Yu Z, Gerstner L, Ärnlöv J, Bakker SJL, Baptista D, Biggs ML, Boerwinkle E, Brenner H, Burkhardt R, Carroll RJ, Chee ML, Chee ML, Chen M, Cheng CY, Cook JP, Coresh J, Corre T, Danesh J, de Borst MH, De Grandi A, de Mutsert R, de Vries APJ, **Degenhardt F**, Dittrich K, Divers J, Eckardt KU, Ehret G, Endlich K, Felix JF, Franco OH, Franke A, Freedman BI, Freitag-Wolf S, Gansevoort RT, Giedraitis V, Gögele M, Grundner-Culemann F, Gudbjartsson DF, Gudnason V, Hamet P, Harris TB, Hicks AA, Holm H, Foo VHX, Hwang SJ, Ikram MA, Ingelsson E, Jaddoe VWV, Jakobsdottir J, Josyula NS, Jung B, Kähönen M, Khor CC, Kiess W, Koenig W, Körner A, Kovacs P, Kramer H, Krämer BK, Kronenberg F, Lange LA, Langefeld CD, Lee JJ, Lehtimäki T, Lieb W, Lim SC, Lind L, Lindgren CM, Liu J, Loeffler

M, Lyytikäinen LP, Mahajan A, Maranville JC, Mascalzoni D, McMullen B, Meisinger C, Meitinger T, Miliku K, Mook-Kanamori DO, Müller-Nurasyid M, Mychaleckyj JC, Nauck M, Nikus K, Ning B, Noordam R, Connell JO, Olafsson I, Palmer ND, Peters A, Podgornaia AI, Ponte B, Poulain T, Pramstaller PP, Rabelink TJ, Raffield LM, Reilly DF, Rettig R, Rheinberger M, Rice KM, Rivadeneira F, Runz H, Ryan KA, Sabanayagam C, Saum KU, Schöttker B, Shaffer CM, Shi Y, Smith AV, Strauch K, Stumvoll M, Sun BB, Szymczak S, Tai ES, Tan NYQ, Taylor KD, Teren A, Tham YC, Thiery J, Thio CHL, Thomsen H, Thorsteinsdottir U, Tönjes A, Tremblay J, Uitterlinden AG, van der Harst P, Verweij N, Vogelezang S, Völker U, Waldenberger M, Wang C, Wilson OD, Wong C, Wong TY, Yang Q, Yasuda M, Akilesh S, Bochud M, Böger CA, Devuyst O, Edwards TL, Ho K, Morris AP, Parsa A, Pendergrass SA, Psaty BM, Rotter JI, Stefansson K, Wilson JG, Susztak K, Snieder H, Heid IM, Scholz M, Butterworth AS, Hung AM, Pattaro C, Köttgen A. Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. Nat Commun. 2019 Sep 11;10(1):4130. doi: 10.1038/s41467-019-11576-0. PubMed PMID: 31511532; PubMed Central PMCID: PMC6739370.

4: Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, Tin A, Wang L, Chu AY, Hoppmann A, Kirsten H, Giri A, Chai JF, Sveinbjornsson G, Tayo BO, Nutile T, Fuchsberger C, Marten J, Cocca M, Ghasemi S, Xu Y, Horn K, Noce D, van der Most PJ, Sedaghat S, Yu Z, Akiyama M, Afaq S, Ahluwalia TS, Almgren P, Amin N, Ärnlöv J, Bakker SJL, Bansal N, Baptista D, Bergmann S, Biggs ML, Biino G, Boehnke M, Boerwinkle E, Boissel M, Bottinger EP, Boutin TS, Brenner H, Brumat M, Burkhardt R, Butterworth AS, Campana E, Campbell A, Campbell H, Canouil M, Carroll RJ, Catamo E, Chambers JC, Chee ML, Chee ML, Chen X, Cheng CY, Cheng Y, Christensen K, Cifkova R, Ciullo M, Concas MP, Cook JP, Coresh J, Corre T, Sala CF, Cusi D, Danesh J, Daw EW, de Borst MH, De Grandi A, de Mutsert R, de Vries APJ, **Degenhardt F**, Delgado G, Demirkan A, Di Angelantonio E, Dittrich K, Divers J, Dorajoo R, Eckardt KU, Ehret G, Elliott P, Endlich K, Evans MK, Felix JF, Foo VHX, Franco OH, Franke A, Freedman BI, Freitag-Wolf S, Friedlander Y, Froguel P, Gansevoort RT, Gao H, Gasparini P, Gaziano JM, Giedraitis V, Gieger C, Girotto G, Giulianini F, Gögele M, Gordon SD, Gudbjartsson DF, Gudnason V, Haller T, Hamet P, Harris TB, Hartman CA, Hayward C, Hellwege JN, Heng CK, Hicks AA, Hofer E, Huang W, Hutri-Kähönen N, Hwang SJ, Ikram MA, Indridason OS, Ingelsson E, Ising M, Jaddoe VWV, Jakobsdottir J, Jonas JB, Joshi PK, Josyula NS, Jung B, Kähönen M, Kamatani Y, Kammerer CM, Kanai M, Kastarinen M, Kerr SM, Khor CC, Kiess W, Kleber ME, Koenig W, Kooner JS, Körner A, Kovacs P, Kraja AT, Krajcoviechova A, Kramer H, Krämer BK, Kronenberg F, Kubo M, Kühnel B, Kuokkanen M, Kuusisto J, La Bianca M, Laakso M, Lange LA, Langefeld CD, Lee JJ, Lehne B, Lehtimäki T, Lieb W; Lifelines Cohort Study, Lim SC, Lind L, Lindgren CM, Liu J, Liu J, Loeffler M, Loos RJF, Lucae S, Lukas MA, Lyytikäinen LP, Mägi R, Magnusson PKE, Mahajan A, Martin NG, Martins J, März W, Mascalzoni D, Matsuda K, Meisinger C, Meitinger T, Melander O, Metspalu A, Mikaelsdottir EK, Milaneschi Y, Miliku K, Mishra PP; V. A. Million Veteran Program, Mohlke KL, Mononen N, Montgomery GW, Mook-Kanamori DO, Mychaleckyj JC, Nadkarni GN, Nalls MA, Nauck M, Nikus K, Ning B, Nolte IM, Noordam R, O'Connell J, O'Donoghue ML, Olafsson I, Oldehinkel AJ, Orho-Melander M, Ouwehand WH, Padmanabhan S, Palmer ND, Palsson R, Penninx BWJH, Perls T, Perola M, Pirastu M, Pirastu N, Pistis G, Podgornaia AI, Polasek O, Ponte B, Porteous DJ, Poulain T, Pramstaller PP, Preuss MH, Prins BP, Province MA, Rabelink TJ, Raffield LM, Raitakari OT, Reilly DF, Rettig R, Rheinberger M, Rice KM, Ridker PM, Rivadeneira F, Rizzi F, Roberts DJ, Robino A, Rossing P, Rudan I, Rueedi R, Ruggiero D, Ryan KA, Saba Y, Sabanayagam C, Salomaa V, Salvi E, Saum KU, Schmidt H, Schmidt R, Schöttker B, Schulz CA, Schupf N, Shaffer CM, Shi Y, Smith AV, Smith BH, Soranzo N, Spracklen CN, Strauch K, Stringham HM, Stumvoll M, Svensson PO, Szymczak S, Tai ES, Tajuddin SM, Tan NYQ, Taylor KD, Teren A, Tham YC, Thiery J, Thio CHL, Thomsen H, Thorleifsson G, Toniolo D, Tönjes A, Tremblay J, Tzoulaki I, Uitterlinden AG, Vaccargiu S, van Dam RM, van der Harst P, van Duijn CM, Velez Edward DR, Verweij N, Vogelezang S, Völker U, Vollenweider P, Waeber G, Waldenberger M, Wallentin L, Wang YX, Wang C, Waterworth DM, Bin Wei W, White H, Whitfield JB, Wild SH, Wilson JF, Wojczynski MK, Wong C, Wong TY, Xu L, Yang Q, Yasuda M, Yerges-Armstrong LM, Zhang W, Zonderman AB, Rotter JI, Bochud M, Psaty BM, Vitart V, Wilson JG, Dehghan A, Parsa A, Chasman DI, Ho K, Morris AP, Devuyst O, Akilesh S, Pendergrass SA, Sim X, Böger CA, Okada Y, Edwards TL, Snieder H, Stefansson K, Hung AM, Heid IM, Scholz M, Teumer A, Köttgen A, Pattaro C. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet. 2019 Jun;51(6):957-972. doi: 10.1038/s41588-019-0407-x. Epub 2019 May 31.

PubMed PMID: 31152163; PubMed Central PMCID: PMC6698888.

5: Westphal S, Stoppe C, Gruenewald M, Bein B, Renner J, Cremer J, Coburn M, Schaelte G, Boening A, Niemann B, Kletzin F, Roesner J, Strouhal U, Reyher C, Laufenberg-Feldmann R, Ferner M, Brandes IF, Bauer M, Kortgen A, Stehr SN, Wittmann M, Baumgarten G, Struck R, Meyer-Treschan T, Kienbaum P, Heringlake M, Schoen J, Sander M, Treskatsch S, Smul T, Wolwender E, Schilling T, **Degenhardt F**, Franke A, Mucha S, Tittmann L, Kohlhaas M, Fuernau G, Brosteanu O, Hasenclever D, Zacharowski K, Meybohm P; RIPHeart-Study Collaborators. Genome-wide association study of myocardial infarction, atrial fibrillation, acute stroke, acute kidney injury and delirium after cardiac surgery – a sub-analysis of the RIPHeart-Study. BMC Cardiovasc Disord. 2019 Jan 24;19(1):26. doi: 10.1186/s12872-019-1002-x. PubMed PMID: 30678657; PubMed Central PMCID: PMC6345037.

6: Tsoi LC, Rodriguez E, **Degenhardt F**, Baurecht H, Wehkamp U, Volks N, Szymczak S, Swindell WR, Sarkar MK, Raja K, Shao S, Patrick M, Gao Y, Uppala R, Perez White BE, Getsios S, Harms PW, Maverakis E, Elder JT, Franke A, Gudjonsson JE, Weidinger S. Atopic Dermatitis Is an IL-13-Dominant Disease with Greater Molecular Heterogeneity Compared to Psoriasis. J Invest Dermatol. 2019 Jul;139(7):1480-1489. doi: 10.1016/j.jid.2018.12.018. Epub 2019 Jan 11. PubMed PMID: 30641038; PubMed Central PMCID: PMC6711380.

7: Gassner C, **Degenhardt F**, Meyer S, Vollmert C, Trost N, Neuenschwander K, Merki Y, Portmann C, Sigurdardottir S, Zorbas A, Engström C, Gottschalk J, Amar El Dusouqui S, Waldvogel-Abramovski S, Rigal E, Tissot JD, Tinguely C, Mauvais SM, Sarraj A, Bessero D, Stalder M, Infanti L, Buser A, Sigle J, Weingand T, Castelli D, Braisch MC, Thierbach J, Heer S, Schulzki T, Krawczak M, Franke A, Frey BM. Low-Frequency Blood Group Antigens in Switzerland. Transfus Med Hemother. 2018 Jul;45(4):239-250. doi: 10.1159/000490714. Epub 2018 Jul 10. PubMed PMID: 30283273; PubMed Central PMCID: PMC6158591.

8: Kamm F, Strauch U, **Degenhardt F**, Lopez R, Kunst C, Rogler G, Franke A, Klebl F, Rieder F. Correction: Serum anti-glycan-antibodies in relatives of patients with inflammatory bowel disease. PLoS One. 2018 Sep 4;13(9):e0203709. doi: 10.1371/journal.pone.0203709. eCollection 2018. PubMed PMID: 30180207; PubMed Central PMCID: PMC6122814.

9: Kamm F, Strauch U, **Degenhardt F**, Lopez R, Kunst C, Rogler G, Franke A, Klebl F, Rieders F. Serum anti-glycan-antibodies in relatives of patients with inflammatory bowel disease. PLoS One. 2018 Mar 29;13(3):e0194222. doi: 10.1371/journal.pone.0194222. eCollection 2018. Erratum in: PLoS One. 2018 Sep 4;13(9):e0203709. PubMed PMID: 29596443; PubMed Central PMCID: PMC5875751.

10: **Degenhardt F**, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics datasets. Brief Bioinform. 2019 Mar 22;20(2):492-503. doi: 10.1093/bib/bbx124. PubMed PMID: 29045534; PubMed Central PMCID: PMC6433899.

11: Rühlemann MC, **Degenhardt F**, Thingholm LB, Wang J, Skiecevičienė J, Rausch P, Hov JR, Lieb W, Karlsen TH, Laudes M, Baines JF, Heinsen FA, Franke A. Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. Gut Microbes. 2018 Jan 2;9(1):68-75. doi: 10.1080/19490976.2017.1356979. Epub 2017 Aug 28. PubMed PMID: 28816579; PubMed Central PMCID: PMC5939986.

12: Kreutzer C, Peters S, Schulte DM, Fangmann D, Türk K, Wolff S, van Eimeren T, Ahrens M, Beckmann J, Schafmayer C, Becker T, Kerby T, Rohr A, Riedel C, Heinsen FA, **Degenhardt F**, Franke A, Rosenstiel P, Zubek N, Henning C, Freitag-Wolf S, Dempfle A, Psilopanagioti A, Petrou-Papadaki H, Lenk L, Jansen O, Schreiber S, Laudes M. Hypothalamic Inflammation in Human Obesity Is Mediated by Environmental and Genetic Factors. Diabetes. 2017 Sep;66(9):2407-2415. doi: 10.2337/db17-0067. Epub 2017 Jun 2. PubMed PMID: 28576837.

13: Yadav P, Ellinghaus D, Rémy G, Freitag-Wolf S, Cesaro A, **Degenhardt F**, Boucher G, Delacre M;

International IBD Genetics Consortium, Peyrin-Biroulet L, Pichavant M, Rioux JD, Gosset P, Franke A, Schumm LP, Krawczak M, Chamaillard M, Dempfle A, Andersen V. Genetic Factors Interact With Tobacco Smoke to Modify Risk for Inflammatory Bowel Disease in Humans and Mice. Gastroenterology. 2017 Aug;153(2):550-565. doi: 10.1053/j.gastro.2017.05.010. Epub 2017 May 12. PubMed PMID: 28506689; PubMed Central PMCID: PMC5526723.

14: Nowak-Göttl U, Limperger V, Kenet G, **Degenhardt F**, Arlt R, Domschikowski J, Clausnizer H, Liebsch J, Junker R, Steppat D. Developmental hemostasis: A lifespan from neonates and pregnancy to the young and elderly adult in a European white population. Blood Cells Mol Dis. 2017 Sep;67:2-13. doi: 10.1016/j.bcmd.2016.11.012. Epub 2016 Dec 5. PubMed PMID: 28017497.

15: Brüwer G, Limperger V, Kenet G, Klostermeier UC, Shneyder M, **Degenhardt F**, Finckh U, Heller C, Holzhauer S, Trappe R, Kentouche K, Knoefler R, Kurnik K, Krümpel A, Lauten M, Manner D, Mesters R, Junker R, Nowak-Göttl U. Impact of high risk thrombophilia status on recurrence among children and adults with VTE: An observational multicenter cohort study. Blood Cells Mol Dis. 2016 Nov;62:24-31. doi: 10.1016/j.bcmd.2016.10.024. Epub 2016 Nov 5. PubMed PMID: 27838551.

16: Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, **Degenhardt F**, Heinsen FA, Rühlemann MC, Szymczak S, Holm K, Esko T, Sun J, Pricop-Jeckstadt M, Al-Dury S, Bohov P, Bethune J, Sommer F, Ellinghaus D, Berge RK, Hübenthal M, Koch M, Schwarz K, Rimbach G, Hübbe P, Pan WH, Sheibani-Tezerji R, Häsler R, Rosenstiel P, D'Amato M, Cloppenborg-Schmidt K, Künzel S, Laudes M, Marschall HU, Lieb W, Nöthlings U, Karlsen TH, Baines JF, Franke A. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet. 2016 Nov;48(11):1396-1406. doi: 10.1038/ng.3695. Epub 2016 Oct 10. PubMed PMID: 27723756; PubMed Central PMCID: PMC5626933.

17: **Degenhardt F**, Dirmeier A, Lopez R, Lang S, Kunst C, Roggenbuck D, Reinhold D, Szymczak S, Rogler G, Klebl F, Franke A, Rieder F. Serologic Anti-GP2 Antibodies Are Associated with Genetic Polymorphisms, Fibrostenosis, and Need for Surgical Resection in Crohn's Disease. Inflamm Bowel Dis. 2016 Nov;22(11):2648-2657. PubMed PMID: 27753692; PubMed Central PMCID: PMC5082182.

18: Rivas MA, Graham D, Sulem P, Stevens C, Desch AN, Goyette P, Gudbjartsson D, Jonsdottir I, Thorsteinsdottir U, **Degenhardt F**, Mucha S, Kurki MI, Li D, D'Amato M, Annese V, Vermeire S, Weersma RK, Halfvarson J, Paavola-Sakki P, Lappalainen M, Lek M, Cummings B, Tukiainen T, Haritunians T, Halme L, Koskinen LL, Ananthakrishnan AN, Luo Y, Heap GA, Visschedijk MC; UK IBD Genetics Consortium; NIDDK IBD Genetics Consortium, MacArthur DG, Neale BM, Ahmad T, Anderson CA, Brant SR, Duerr RH, Silverberg MS, Cho JH, Palotie A, Saavalainen P, Kontula K, Färkkilä M, McGovern DP, Franke A, Stefansson K, Rioux JD, Xavier RJ, Daly MJ. Erratum: A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nat Commun. 2016 Sep 13;7:12869. doi: 10.1038/ncomms12869. PubMed PMID: 27619887; PubMed Central PMCID: PMC5027274.

19: Rivas MA, Graham D, Sulem P, Stevens C, Desch AN, Goyette P, Gudbjartsson D, Jonsdottir I, Thorsteinsdottir U, **Degenhardt F**, Mucha S, Kurki MI, Li D, D'Amato M, Annese V, Vermeire S, Weersma RK, Halfvarson J, Paavola-Sakki P, Lappalainen M, Lek M, Cummings B, Tukiainen T, Haritunians T, Halme L, Koskinen LL, Ananthakrishnan AN, Luo Y, Heap GA, Visschedijk MC; UK IBD Genetics Consortium; NIDDK IBD Genetics Consortium, MacArthur DG, Neale BM, Ahmad T, Anderson CA, Brant SR, Duerr RH, Silverberg MS, Cho JH, Palotie A, Saavalainen P, Kontula K, Färkkilä M, McGovern DP, Franke A, Stefansson K, Rioux JD, Xavier RJ, Daly MJ, Barrett J, de Lane K, Edwards C, Hart A, Hawkey C, Jostins L, Kennedy N, Lamb C, Lee J, Lees C, Mansfield J, Mathew C, Mowatt C, Newman B, Nimmo E, Parkes M, Pollard M, Prescott N, Randall J, Rice D, Satsangi J, Simmons A, Tremelling M, Uhlig H, Wilson D, Abraham C, Achkar JP, Bitton A, Boucher G, Croitoru K, Fleshner P, Glas J, Kugathasan S, Limbergen JV, Milgrom R, Proctor D, Regueiro M, Schumm PL, Sharma Y, Stempak JM, Targan SR, Wang MH. A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nat Commun. 2016 Aug 9;7:12342. doi: 10.1038/ncomms12342.

Erratum in: Nat Commun. 2016 Sep 13;7:12869. PubMed PMID: 27503255; PubMed Central PMCID: PMC4980482.

20: **Degenhardt F**, Niklowitz P, Szymczak S, Jacobs G, Lieb W, Menke T, Laudes M, Esko T, Weidinger S, Franke A, Döring F, Onur S. Genome-wide association study of serum coenzyme Q10 levels identifies susceptibility loci linked to neuronal diseases. Hum Mol Genet. 2016 Jul 1;25(13):2881-2891. Epub 2016 May 5. PubMed PMID: 27149984.

21: Hübenthal M, Hemmrich-Stanisak G, **Degenhardt F**, Szymczak S, Du Z, Elsharawy A, Keller A, Schreiber S, Franke A. Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease. PLoS One. 2015 Oct 14;10(10):e0140155. doi: 10.1371/journal.pone.0140155. eCollection 2015. PubMed PMID: 26466382; PubMed Central PMCID: PMC4605644.

22: Schmitt J, Schwarz K, Baurecht H, Hotze M, Fölster-Holst R, Rodríguez E, Lee YAE, Franke A, **Degenhardt F**, Lieb W, Gieger C, Kabesch M, Nöthen MM, Irvine AD, McLean WHI, Deckert S, Stephan V, Schwarz P, Aringer M, Novak N, Weidinger S. Atopic dermatitis is associated with an increased risk for rheumatoid arthritis and inflammatory bowel disease, and a decreased risk for type 1 diabetes. J Allergy Clin Immunol. 2016 Jan;137(1):130-136. doi: 10.1016/j.jaci.2015.06.029. Epub 2015 Aug 4. PubMed PMID: 26253344.

List of Figures

Figure 1 – Anatomical & endoscopic differences between Crohn's disease and ulcerative colitis. 5
Figure 2 – Differences in location of UC location across different age groups. 8
Figure 3 – Treatment of inflammatory bowel disease. 10
Figure 4 – Global prevalence of IBD. 11
Figure 5 – Global prevalence rates of IBD and trends across time. 12
Figure 6 – Manhattan plot of IBD, UC and CD in different ethnicities. 26
Figure 7 – Association analysis in African American IBD, CD and UC. 27
Figure 8 – Overview of gut microbiome composition across IBD. 29
Figure 9 – Fungal microbiome in IBD. 30
Figure 10 – IBD as a mucosal barrier defect. 31
Figure 11 – Gene map of the human leukocyte antigen (HLA) region. 33
Figure 12 – Molecular mechanisms of antigen presentation by the HLA. 34
Figure 13 – Current HLA nomenclature. 35
Figure 14 – Fine mapping study of the HLA in UC. 38
Figure 15 – Fine mapping study of the HLA in CD. 39
Figure 16 – Linkage disequilibrium explained. 40
Figure 17 – Experimental workflow for SNP genotyping. 42
Figure 18 – RNA typing protocol used for the analysis in this study. 50
Figure 19 – GUI of the HLAssign software for a chosen locus – HLA-*C*. 51
Figure 20 – Overview of the HIBAG prediction algorithm. 53
Figure 21 – Workflow of imputation using HIBAG. 54
Figure 22 – Crystal structure of the HLA class II binding groove. 55
Figure 23 – Number of HLA sequences available at the IPD-IMGT/HLA up to release 3.39.0. 57
Figure 24 – Populations for which information is available in the AFND. 58
Figure 25 – Example haplotype plot for DRB1*15:01. 112
Figure 26 – Example haplotype plot for DRB1*15:02. 113
Figure 27 – Example haplotype plot for DRB1*15:03. 113
Figure 28 – HLAPipe workflow. 114
Figure 29 – Epitope vs. non epitope properties. 118
Figure 30 – Weighted Kullback-Leibler motifs of DRB1*15:01 and DRB5*01:01. 121
Figure 31 – Power to detect a true effect in our study. 123
Figure 32 – P-values vs. OR in the Indian population for different allele frequencies. 124
Figure 33 – Motif plots of DRB1 *15:02 from experimental data. 126

IV List of Tables

Table 1 – Summary of Paper A, Paper B and Paper C. 3
Table 2 – GWAS in non-Caucasian populations. 25
Table 3 – Number HLA class I and HLA class II alleles in the IMGT/HLA database. 35
Table 4 – Timeline of the description of HLA loci in the HLA nomenclature report. 36
Table 5 – Example table to explain the concept of the odds ratio (OR). 45
Table 6 – Explanation of Equation 5 with an example. 111

V List of Figures and Tables in Papers

Paper A

Figures

Abbildung 1 – Genidentifikationen führen zu zahlreichen weiterführenden Forschungsarbeiten zum entsprechenden Kandidatengen, dessen Protein und Stoffwechselweg. 16
Abbildung 2 – Studien über chronisch-entzündliche Darmerkrankungen (CED) in nichteuropäisch-stämmigen Ethnizitäten sind nach wie vor unterrepräsentiert. 19
Abbildung 3 – Das Spektrum der bekannten genetischen Risikovarianten chronisch entzündlicher Erkrankungen (CED). 20

Tables

Tabelle 1 – Wichtige genomweite Assoziationsstudien (GWAS) und deren Ergebnisse. 18

Paper B

Figures

Figure 1 – Flowchart of steps taken in preparation and benchmark of our multi-ethnic reference panel. 64
Figure 2 – MDS analysis of HLA typed allele data. 65
Figure 3 – Imputation accuracies employing the multi-ethnic reference panel. 66
Figure 4 – Known architecture of HLA-*DRB3/4/5*. 72

Tables

Table 1 – Frequencies of HLA-*DRB3/4/5* in our multi-ethnic reference panel. 67
Table 2 – Imputation accuracies for 1000 Genomes populations. 67
Table 3 – Imputation accuracies of the imputation with the multi-ethnic reference panel. 68
Table 4 – Previously reported imputation accuracies. 69

Paper C

Figures

Figure 1 – HLA regional association plots. 105
Figure 2 – HLA single allele association analysis results at 2- and 4-digit resolution for MHC class II loci *-DRB3/4/5, -DRB1, -DQA1-DQB1*. 106
Figure 3 – Haplotypes for associated HLA alleles. 107
Figure 4 – Clustering of DRB1 proteins according to preferential peptide binding and combined peptide binding motifs. 108
Figure 5 – Cluster according to chosen physico-chemical properties of amino acids within the peptide binding pockets. 109
Figure 6 – Frequency of DRB1*01:03 across populations available in the allele frequency net database. 110

VI Abbreviations

Abbreviation	Explanation
AF	Allele frequency
AFND	Allele frequency network database
APC	Antigen presenting cell
ARG1	Arginine synthetase 1
ARG2	Arginine synthetase 2
ASA	5-Aminosacylates
ASCA	Anti-*Saccharomyces cerevisiae* antibody
ATG16L1	Autophagy related 16 like 1
B1-B3	Montreal classification of disease behaviour in CD
B1-B3p	Montreal classification of disease behaviour in CD (with modifier p for perianal disease)
BCR	B-cell receptor
BMI	Body mass index
bp	Base pairs
CD	Crohn's disease
$CD4^+$	CD4 positive; carrying co-receptor CD4
$CD8^+$	CD8 positive; carrying co-receptor CD8
CDAI	Crohn's disease activity index
CED	Chronisch entzündliche Darmerkrankungen, *engl.* inflammatory bowel disease (IBD)
CI	Confidence interval
CLIP	Class II invariant chain peptide
DNA	Deoxyribonucleic acid
DSS	Dextran sulfate sodium
E1 - E3	Montreal classification of disease extent in UC
EIM	Extraintestinal manifestation
eNOS	Endothelial nitric oxide synthetase
ER	Endoplasmic reticulum
GBD	Global burden of disease
GPU	Graphic processing unit
GUI	Graphical user interface
GWA	Genome wide association
GWAS	Genome wide association study
HapMap	Human haplotype project
HLA	Human leukocyte antigen
HLA-*A*	HLA gene locus *A*
HLA-*B*	HLA gene locus *B*
HLA-*C*	HLA gene locus *C*
HLA-*DPA1*	HLA gene locus *DPA1*
HLA-*DPA1-DPB1*	HLA haplotype *DPA1-DPB1*
HLA-*DPB1*	HLA gene locus *DPB1*
HLA-DQ	Antigen specificity DQ
HLA-*DQA1*	HLA gene locus *DQA1*
HLA-*DQA1-DQB1*	HLA haplotype *DQA1-DQB1*
HLA-*DQB1*	HLA gene locus *DQB1*
HLA-DR	Antigen specificity DR
HLA-*DRB3/4/5*	HLA gene loci *DRB3*, *DRB4* and *DRB5*

HLA-*DRB1*	HLA gene locus *DRB*1
HLA-*DRB3*	HLA gene locus *DRB3*
HLA-*DRB4*	HLA gene locus *DRB4*
HLA-*DRB5*	HLA gene locus *DRB5*
IBD	Inflammatory bowel disease
IBD	Identity by descent
IBD-U	Unclassified IBD
IEC	Intestinal epithelial cells
IL23	Interleukin receptor 23
IMGT	Immunogenetics project
IND	Indian population
InDel	Insertion/deletion variation
iNOS	Inducible nitric oxide synthetase
IRGM	Immunity-related GTPase family M protein
IRN	Iranian population
JAK	Januskinase
JPN	Japanese population
kb	Kilo base pairs
K	Amino acid lysine
KKK	Lysine tripeptide
KIR	Killer-cell immunoglobulin-like receptor
KOR	Korean population
L1 - L4	Montreal classification of disease location in CD
LD	Linkage disequilibrium
LMM	Linear mixed model
MAF	Minor allele frequency
MAX	Maximum
Mb	Mega base pairs
MHC	Major histocompatibility complex; termed HLA in humans
MIC	Human MHC class I chain related
MIN	Minimum
MLE	Maximum likelihood estimator
MLT	Maltese population
MMP	Matrix metalloprotease
MTS1	Macrophage-stimulating protein 1
MUC	Mucin
N	Amino acid asparagine
NHGRI	National human genome research institute
NGS	Next generation sequencing
nNOS	Neuronal nitric oxide synthetase
NO	Nitric oxide
NOD2	Nucleotide oligomerization domain 2
NOS1	Nitric oxide synthetase 1
NOS2	Nitric oxide synthetase 2
NOS3	Nitric oxide synthetase 3
OR	Odds ratio
P1-P9	Pockets 1 to 9 of the HLA protein
PAH	Pulmonary arterial heart syndrome
pANCA	Perinuclear antineutrophil cytoplasmic antibodies
PCA	Principal component analysis
PCR	Polymerase chain reaction
PPAR	Peroxisome proliferator-activated receptor

PRI	Puerto Rican population
PSC	Primary sclerosing cholangitis
QC	Quality control
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
R	Amino acid arginine
RRR	Arginine tripeptide
SAS	South Asian population of the 1000 Genomes/HapMap population (https://www.internationalgenome.org/category/population/)
SBT	Sequence based typing
SE	Standard error
SLC7A2	Solute carrier family member 2
SNP	Single nucleotide polymorphism (MAF >=1%)
SNV	Single nucleotide variation (MAF <1%)
SNX20	Sorting nexin 20
SSOP	Sequence specific oligonucleotide probe
SSP	Sequence specific primer
STAT	Signal transducers and activators of transcription
TCR	T-cell receptor
TNF	Tumor necrosis factor
TNFSF15	TNF superfamily member 15
UC	Ulcerative colitis
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World health organization
xHLA	Extended HLA region from 25-34 Mb
YRI	Yoruba in Ibadan, Nigeria population of the 1000 Genomes/HapMap population (https://www.internationalgenome.org/category/population/)

1. Zusammenfassung

Das Humane Leukozyten Antigen System (HLA) ist eine hochvariable Genkassette auf Chromosom *6p21,* das in vielen Funktionen des Immunsystems eine Rolle spielt. Das HLA ist der am stärksten mit chronisch entzündlichen Darmerkrankungen (CED) und dessen Subformen Morbus Crohn und Colitis ulcerosa assoziierte Genlokus. Dieser Suszeptibilitätslokus ist genetisch hoch-komplex und kennzeichnet sich durch enge Strukturen des Kopplungsungleichgewichts (Linkage Disequilibriums; LD), die die Zuordnung kausaler Varianten zum Assoziationssignal erschweren. In vorangegangenen Publikationen wurde das HLA Allel DRB1*01:03 als wichtiger krankheitsrelevanter Faktor für CED in der kaukasischen Population genannt[1]. Für den großen Anteil des Assoziationsignals in der HLA Region ist eine solche Zuordnung bisher jedoch nicht möglich. Ziel dieser Arbeit war es daher, das HLA Signal in CED weiter zu charakterisieren und zu analysieren, ob das starke LD zwischen den HLA Genorten HLA-*DQ* und -*DR* mithilfe eines ethnienübergreifenden Ansatzes aufgelöst werden könnte. Dabei war Fokus dieser Arbeit die Colitis ulcerosa. In Kollaboration mit dem Internationalen inflammatory bowel disease (IBD, zu Deutsch CED) Genetik Konsortium, trugen wir einen einzigartigen Datensatz bestehend aus Individuen 8 verschiedener Ethnizitäten zusammen, die auf Illuminas Immunochip genotypisiert wurden. Da die NGS-basierte Typisierung des HLA für eine große Anzahl an Individuen teuer ist, entschieden wir uns für einen Ansatz, der die Herleitung von HLA- Allelen (Imputation) ermöglichte. Zunächst erstellten wir eine optimal für unsere Anwendung zugeschnittene Imputationsreferenz. Danach imputierten wir die HLA Allele und führten eine Feinkartierung der HLA Region über die verschiedenen Populationen hinweg durch. Wir identifizierten genetische Risikofaktoren, die ethnienübergreifend geteilt werden und sowohl in der Effektstärke als auch -richtung korrelieren. Wir ermittelten ganze Allelgruppen, die mit der Colitis ulcerosa assoziiert sind, von denen wir die HLA-DRB1*15 Gruppe hervorheben wollen. Durch die Bestimmung von HLA-Haplotypen konnten wir die starke Korrelation zwischen verschiedenen Allelen der HLA-*DQ* und -*DR* Genorte erklären, und sogar gesamte HLA-*DQ*-*DR* Haplotypen, die über die Ethnizitäten hinweg geteilt werden, identifizieren. Wir zeigten, dass das DRB1*01:03 Allel ein populationsspezifisches Allel ist, welches hauptsächlich in Individuen westeuropäischer Herkunft und selten in nicht-kaukasischen Individuen vorkommt. Durch die Analyse der physikalisch-chemischen Eigenschaften der HLA Proteine und der Analyse von Peptiden, die von diesen HLA Allelen vorrangig gebunden werden, konnten wir außerdem die Komplexität des HLA Signals weiter reduzieren und zeigten, dass Risikoallele und protektive Allele bezüglich ihrer Eigenschaften und Peptidbindungspräferenzen zusammengefasst werden können. Die Ergebnisse dieser Studie dürften auch für andere Forschungsbereiche, im Speziellen der Genetik und Peptidomics, sowie auch in Verbindung mit der Analyse anderer Immunerkrankungen, wie zum Beispiel Multiple Sklerose und Lepra, die ebenso mit Allelen der HLA-DRB1*15 Gruppe assoziiert sind, von Interesse sein.

2. Summary

The human leukocyte antigen (HLA) region is a highly variable gene cassette located on chromosome *6p21* that is involved in diverse functions of the host-immune system. The HLA is the strongest but also the most complex susceptibility locus for inflammatory bowel disease (IBD) and its subforms Crohn's disease (CD) and ulcerative colitis (UC). Tight structures of linkage disequilibrium within this highly complex gene locus complicate the identification of causal variants to an association signal. In previous publications, the DRB1*01:03 allele was identified as a determining factor for IBD in the Caucasian population[1]. However, for most of the association signal such a delineation was not possible. The aim of this thesis was therefore to further characterise the HLA signal and to analyse whether tightly linked signals between the HLA-*DQ* and -*DR* loci could be resolved using a trans-ethnic approach. The focus of this study was UC. To this end, we assembled a unique dataset of individuals from 8 different ethnicities in collaboration with partners of the International IBD Genetics Consortium. The individuals were typed on Illumina's ImmunoChip. With NGS-based typing of HLA alleles being expensive for a larger number of samples, we opted to inferring HLA alleles using an imputation approach. First, we compiled an HLA imputation reference panel specially tailored to our cohorts of interests. Next, we imputed HLA alleles and performed HLA fine mapping across the different populations. For the first time, we identified genetic factors shared across different ethnicities both in effect direction and magnitude. We found whole allele groups that are associated with UC, highlighting alleles of the HLA-DRB1*15 group. Using haplotype phasing of HLA alleles, we were able to explain the tightly linked signals of the HLA-*DQ* and -*DR* loci and showed that there are entire HLA-*DQ*-*DR* haplotypes shared across different ethnicities. In addition, we identified the previously reported DRB1*01:03 as a population-specific signal that is mostly present in individuals of Western European descent and hardly present in non-Caucasian individuals. Furthermore, by analysis of physico-chemical properties of the respective HLA proteins and analysis of peptides that are preferentially bound by them, we were able to further reduce the complexity of the signal, showing that risk and protective proteins, respectively, share common features both in regard to their chemical properties as well as to the peptides they preferentially bind. Our findings should be of broad interest to the research community, both to genetics and peptidomics experts also in the context of the analysis of other immune-related diseases, for instance multiple sclerosis or leprosy, which are also associated with the HLA-DRB1*15 group.

3. Overview of main publications

This thesis has resulted in the publication of one review article and two original research articles which have been incorporated into this thesis as **Paper A**, **Paper B** and **Paper C** (for paper references see Chapter I). The background and aims, results and conclusions of these papers are described briefly below. A more detailed summary of the papers can be found in Chapter 4.1.4 (**Paper A**), Chapter 5.1 (**Paper B**) and Chapter 5.2 (**Paper C**).

Table 1 – Summary of Paper A, Paper B and Paper C.

	Paper A	Paper B	Paper C
Background and aims	This paper is a review paper that was written for clinicians to give an update on the knowledge obtained from genetic association studies of inflammatory bowel disease (IBD). It describes the most relevant and most consistent findings in IBD in the Caucasian population and shortly also elucidates to knowledge obtained in non-Caucasian populations.	HLA reference imputation panels are mostly available for Caucasian populations. More diverse panels are needed in trans-ancestry approaches. The aim of this study was therefore to construct such a panel using 1,360 individuals from different populations that could be used for accurate HLA imputation in **Paper C**.	Most of what is known on HLA association with IBD has been studied in Caucasian populations. Here, causal variants could not be identified except for the highly associated DRB1*01:03 due to the tight linkage disequilibrium (LD) across the HLA region. The aim of this study was therefore to investigate whether differences in in LD across populations of different ancestries could be leveraged to identify causal variants, and the characterisation of the HLA signal across different ancestries in general. The focus of the study was ulcerative colitis (UC).
Results		A benchmark was conducted using a 5x cross validation approach on a subset of samples from each of the cohorts analysed in **Paper C** and the independent 1000 Genomes population. We achieved highly accurate results across all analysed HLA loci.	The HLA-DRB1*15 group is associated with UC across different ancestries with different alleles of the group present in different populations. Using a newly developed phasing approach, we were able to explain the LD across the HLA further but could not identify causal variants. We found that the HLA-DRB1*01:03 allele is specific to countries linked to the British Empire and that peptides binding to the risk HLA alleles are rich in lysine and arginine.
Conclusions		The reference panel can be used in diverse populations to predict HLA genotypes with high accuracy.	We developed a new approach for the analysis of HLA data in the context of trans-ancestry fine mapping. For the first time, we identified how potential antigens in the disease may look like. We could not identify causal variants due to tight LD structures observed within the HLA region even across all our analysed populations of divergent ancestries.

4. Introduction

In this following chapter, inflammatory bowel disease (IBD) and its related demographic genetic and epidemiological aspects are introduced (Chapter 4.1). As the results of this thesis detail analyses performed across populations of different ancestries, differences in these aspects regarding ancestry are discussed. The introduction continues with an overview of the human leukocyte antigen (HLA), its nomenclature and function. Some historical aspects of HLA typing are described (Chapter 4.2). Subsequently, databases and tools useful in the analysis of the HLA are introduced. In reference to **Paper A**, studies performed on the HLA in the context of IBD are discussed. This includes the most important findings of the largest HLA fine mapping study performed using HLA imputation in the Caucasian population (Goyette *et al.*[1]). The introduction ends with a description of the most important methodologies used in this thesis (Chapter 4.3).

For the description of clinical features of the disease and its treatment, references were taken from the book chapters of the 2nd edition of “Molecular Genetics of Inflammatory Bowel Disease”[2], publications on Crohn’s disease (Crohn’s disease – your guide) and ulcerative colitis (Ulcerative colitis – your guide) by the Crohn’s & Colitis UK[3,4] and Shi *et al.*[5], which focuses especially on differences across populations of different ancestries. Sources used in the other chapters are detailed in the respective chapter.

4.1. Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a chronic inflammatory disease of the gut. The disease aetiology of IBD is still not completely understood, though research strongly indicates an interplay of genetic predisposition and environmental factors, especially related to the gut microbiome (referring to the bacterial, fungal and archaeal cells inhabiting the human gut)[2,6].

4.1.1. Clinical characteristics of IBD

IBD can be subdivided into Crohn's disease (CD) and ulcerative colitis (UC). While manifestations of ulcerative colitis are mainly localized to the colon and rectum, Crohn's disease can extend throughout the whole gastrointestinal tract. CD and UC present with different endoscopic findings as detailed in Figure 1.

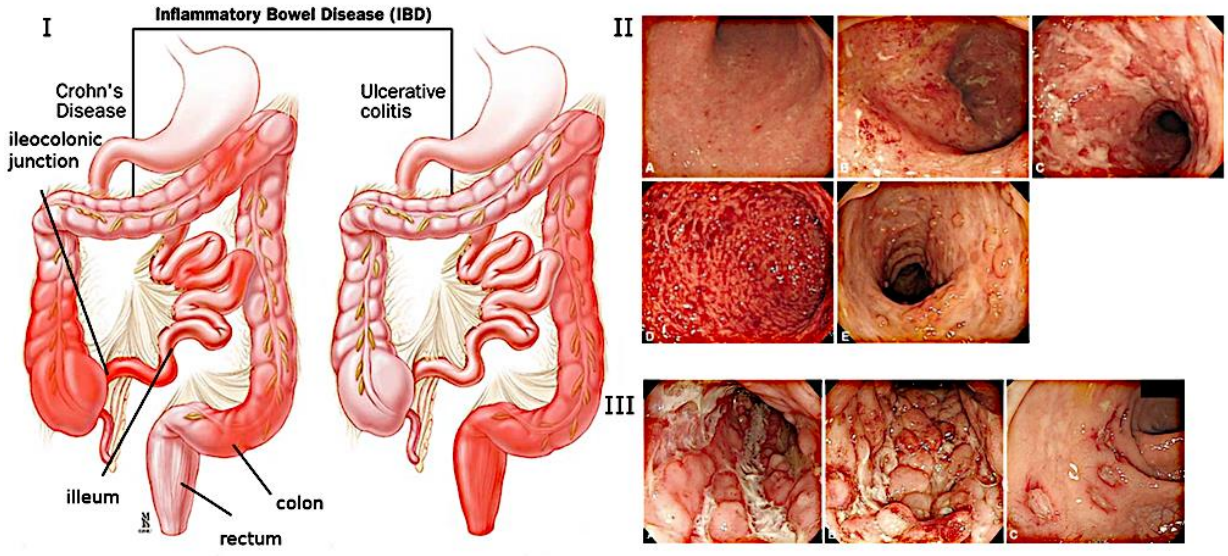


Figure 1 – Anatomical & endoscopic differences between Crohn's disease and ulcerative colitis. (I) Crohn's disease (CD) can affect the whole gastrointestinal tract while ulcerative colitis (UC) is restricted to the colon and rectum. Graphic figures are property of the Johns Hopkins university and were downloaded from www.hopkinsmedicine.org/inflammatory_bowel_disease_center/about_ibd. Figure was modified to include anatomical descriptions. Endoscopic images of UC (II) and CD (III). Figures taken from Lee *et al.*[7]. Original descriptions: (II) Figure 1: Typical endoscopic features of ulcerative colitis. (A) Mild: mucosal erythema [reddening of the mucosal walls], fine granularity, decreased vascular making. (B) Moderate marked erythema, loss of vascular marking, erosions. (C) severe ulcers. (D) Severe spontaneous bleeding. (E) Luminal narrowing with pseudopolyps. (III) Figure 2: Typical endoscopic features of Crohn's disease. (A) Longitudinal ulcers, (B) cobblestone appearance, (C) aphthous ulcers showing longitudinal array.

Symptoms of IBD

Clinically, the inflammatory bowel diseases present with several symptoms. These symptoms include abdominal pain or cramping, diarrhoea, rectal bleeding (more prominent in UC), fever and general fatigue. Common disease symptoms outside the gut (extraintestinal manifestations (EIMs)) include joint problems, liver inflammation in primary sclerosing cholangitis (PSC), skin problems like erythema nodosum, inflammation of the eyes and malnutrition causing weaker bone structure and anemia[2–4]. Approximately 6% to 47% of all Caucasian IBD patients suffer from EIMs of the disease[8,9]. The frequency of EIMs in non-Caucasian, e.g. Asian and African American populations, is not well studied, though lower prevalence rates for some of EIMs in IBD patients have been reported[10,11].

IBD risk factors

Risk factors for IBD include smoking (increased risk for CD and decreased risk for UC), geographical location (i.e. living in Western or Westernized countries: influences amongst others from (childhood) hygiene, dietary factors play a role here), age, family history of IBD and ancestral background[2].

Diagnosis of IBD

The diagnosis of IBD is predominantly made at ages 20-30, while mean and median ages for the diagnosis of CD are typically 5-10 years less[12]. A second larger peak at the sixth and seventh decade has been reported in smaller studies, though this has failed to replicate in multiple studies and is also not observed in non-Caucasian individuals[12,13]. Up to 25% of all IBD patients are paediatric patients which often suffer from a more severe form of IBD than adult patients[14]. IBD is diagnosed based on a scale of histological and endoscopic findings (i.e. disease location as detailed below), physical examination, and markers of inflammation present in the blood or stool. Markers of inflammation analysed in IBD include C reactive protein (blood: acute phase protein produced in the liver) and calprotectin or lactoferrin (stool: antimicrobial proteins). Additional serological markers include anti-*Saccharomyces cerevisiae* (ASCA) and perinuclear antineutrophil cytoplasmic antibodies (pANCA). These markers guide the differential diagnosis of CD and UC. While pANCA blood levels are higher in in UC patients, ASCA levels are higher in CD patients. Combined information on pANCA and ASCA enable differential diagnosis of CD against UC with a sensitivity (frequency of positive test result when patient suffers from a disease) of 46-64% and a specificity (frequency of negative test result, when patient is not affected) of 92-99% in adult-onset IBD[15,16]. Another important criterium for the diagnosis of IBD is **family history** for IBD (i.e. do/did family members suffer from IBD), which puts an individual at a larger risk to develop the disease. Approximately 12% [MIN=11%, MAX=13%] of all first-degree relatives of IBD patients are also affected by IBD in Caucasians. In Middle East Asians about 13% [10%, 16%] of the relatives are affected, while family history seems to be of lesser importance in the Black (7% [4%,10%]) and East or South Asian populations (3% [2%, 4%], 4% [2%, 7%])[5]. Regarding

association of IBD with **gender**, Shi *et al.*[5] reported, that UC was observed about equally in males and females in the Caucasian, Asian and Hispanic populations, while a preponderance in females was seen in the Black population. For Crohn's disease a male preponderance was observed in East Asian, South Asian, and Hispanic individuals, while in Middle East Asian and Caucasian populations the ratio was about equal.

Classification of IBD

Both CD and UC can be classified using the so-called Montreal classification system. This system was designed by a working party at the Montreal World Congress of Gastroenterology in 2006 as a revision of the preceding Vienna classification of IBD, for more accurate diagnostics of IBD[17].

Classification of Crohn's disease

Crohn's disease can be classified using the Montreal classification system according to age of onset, disease location (L) and disease behaviour (B). The latter is discussed in more detail in the next paragraph. Regarding disease location, CD is classified as ileal (affecting the small bowl, L1), ileocolonic (pertaining to the ileum and colon, L2) or colonic (L3) (Figure 1). CD that only effects the upper gastrointestinal tract is classified as L4. In Caucasians, 24-30% of all patients suffer from ileal, 32-37% from ileocolonic and 35-43% from colonic CD, while 3-5% only show symptoms in the upper GI tract[5]. Especially in East Asian, Middle East Asian and Black populations, CD with L3 is reported to be most common, while ileal disease is least common in the Black population with a prevalence of 6-22%[5]. Cleynen *et al.* showed that, genetically, colonic CD places in between UC and ileal CD, thus suggesting the necessity to subdivide CD into colonic and ileocolonic CD[18]. Regarding age groups, CD patients are categorized into 3 age groups, i.e. younger than 17, aged between 17 and 40 and older than 40. Typically, a more severe form of the disease is seen in paediatric patients.

Complications observed in CD patients

Common complications of CD are stricturing (narrowing of the gut lumen) and development of fistulae (formation of abnormal connections between skin/other organs and the bowel). A rarer complication is the perforation of the bowel (formation of holes in the mucosal linings of the bowel)[3]. Using the Montreal classification system for disease behaviour (B) of CD, inflammatory (B1), stricturing (B2) and penetrating (B3) disease are distinguished with an additional modifier for perianal disease (p). Complications are observed more often in non-Caucasian populations than in Caucasian populations[5]. With increasing age of the individuals affected by CD, the development of a complication becomes more likely with 74% of all patients suffering from B2 or B3 within 30 years of the diagnosis, while at the point of diagnosis this number is at 30%[19]. This leads to many patients needing surgery. For Caucasians Cleynen *et al*. report a value of more than 50% of all CD patients needing surgery within approximately 10 years after first diagnosis[19]. Shi *et al.* report that approximately 33% [27%, 39%] of

patients require surgery within 5 years of the diagnosis. Numbers observed in the Black population are higher (40% [33%, 48%]). Lower numbers were observed in the Asian population (17% [8%, 26%]). The severity of the disease is assessed using the Crohn's disease activity index (CDAI) which includes several clinical parameters like disease behaviour, stool frequency and pain. It is further described in Best *et al.*[20].

Classification of UC

UC is classified using the Montreal classification system by disease extent and age of onset. Regarding disease extent, it is classified as either proctitis (E1), where inflammation is localized to the terminal colon, left-sided (E2), where inflammation is localized to the descending colon and terminal colon and extensive (E3), where inflammation occurs across the entire colon. In Caucasians approximately 28% [24%, 32%] of patients suffer from inflammation localized to the terminal colon, while 36% [33%, 38%] each suffer from left-sided or extensive UC[21]. These values vary across different ethnicities and age groups. A more severe form of the disease is seen in paediatric UC patients (age <17, Figure 2).

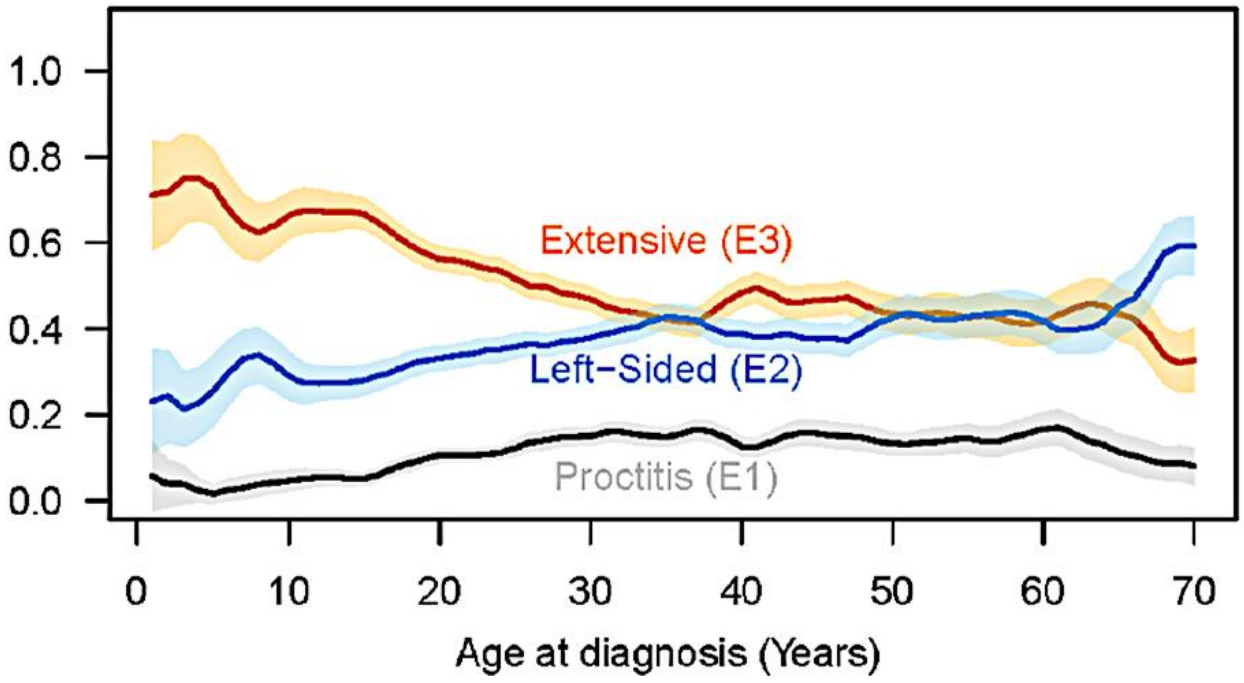


Figure 2 – Differences in location of UC location across different age groups.
Figure was taken from the Supplement of Cleynen et al.19. Original and shortened description: Supplementary Figure 2: UC disease extent versus age at diagnosis. Observed distribution of disease extent versus age at diagnosis is shown as smoothed proportions, with 95% confidence bands. Extensive, left-sided and proctitis are shown respectively in orange, blue and grey. Confidence bands are not corrected for disease duration, which is correlated to age at diagnosis.

Black, Hispanic and South Asian individuals are more likely to suffer from left-sided or extensive disease (approximately 40% each). In East Asian individuals each localization is approximately equally

likely and in Middle East Asian individuals extensive disease is most prominent with a prevalence of 44% [40%, 47%][5]. Severity of ulcerative colitis is most commonly assessed using the Mayo score first introduced by Kenneth *et al.* in 1987 (original publication list affiliation as from the Division of Gastroenterology and Internal Medicine, Mayo Clinic – hence the name)[22]. This score is based on a combination of stool frequency, occurrence of rectal bleeding, physician assessment, endoscopy and a physician's global assessment and is described further in the original publication.

Complications observed in UC patients

Complications of UC observed in individuals suffering from the disease, include the development of a toxic megacolon, colon cancer and extraintestinal manifestations like PSC as described in the section detailing general symptoms of IBD. A toxic megacolon manifests as the widening of the colon which most likely results from damage to nerve cells innervating the smooth muscles cells within the outer mucosal layer of the colon. It is accompanied by a fulminant acute inflammation of the colon. This leads to retention of faeces in the colon, which in turn poses a risk for rupture[4]. Overall, about 22% of UC patients have been reported to need surgery within 10 years of the diagnosis in the Caucasian population[19].

Unclassified IBD

In up to 10% of IBD patients, a differential diagnosis of CD or UC cannot be made and they are subsequently diagnosed with indeterminate IBD[23], also termed unclassified IBD or IBD-U. In up to 14%-18% of patients classified with UC or CD the diagnosis changes over time[24,25].

4.1.2. Treatment of IBD

The inflammatory bowel diseases are treated according to a regimen that is commonly aimed at reducing inflammation and avoiding surgery (Figure 3). The first line of medication includes 5-Aminosacylates (5-ASA: derivatives of salicylic acid bearing an amino-group; mesalazine, sulphalasalzine, olsalazine) as well as antibiotics, to treat bacterial infections (ciprofloxacin, metronidazole). Corticosteroids (prednisone, budenoside, hydrocortisone) are administered to reduce inflammation. Other medications aimed at reducing inflammation include immunosuppressants (azathioprine, mercaptopurine, methotrexate), and biological drugs (produced biotechnologically and are similar to substances produced in the human body). Biological drugs include monoclonal antibodies as anti-integrin antibodies (natalizumab, vedolizumab), anti-tumour necrosis factor alpha (TNFα) antibodies (infliximab, adalimumab, golimumab), anti-interleukin 12/23 antibodies (ustekinumab) and inhibitors of the JAK-STAT pathway (tofatinicinib; JAK: Januskinase, STAT: signal transducers and activators of transcription). Anti-integrin antibodies target integrins, which play a role in leukocyte homing to the gut. Anti-TNFα antibodies target TNFα, which is a cytokine that can induce several other

inflammatory cytokines via the NFκB-pathway.

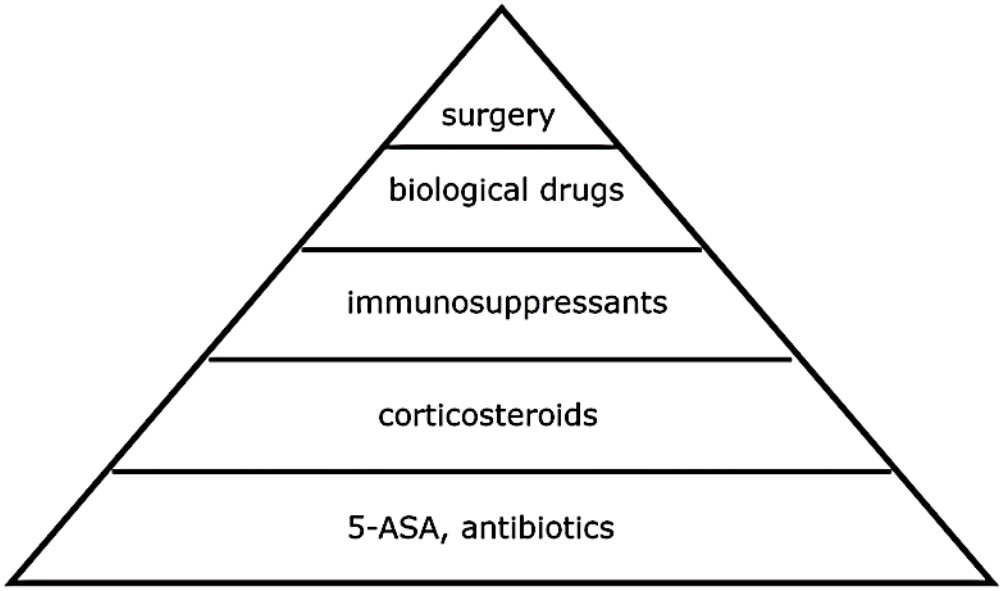


Figure 3 – Treatment of inflammatory bowel disease.
At the bottom of the treatment regimen 5-ASA and antibiotics are used with surgery at the top of the pyramid as a last resort.

The JAK-STAT pathway is a signalling pathway that induces production of certain cytokines[3,4,26]. Other treatment options include surgery of the gut, in which the most affected parts of the gut are removed. However, this option is only considered at the end of the treatment spectrum and used only in severe cases of IBD. Additional treatments may be used to alleviate symptoms outside the gut. This includes medication that resolves cramps, pain medication, diuretics (for diarrhoea), laxatives (for constipation) or a diet tailored to reducing symptoms of the disease (e.g. decreased fibre intake, avoiding food that causes bloating).

4.1.3. IBD demographics

Regarding the overall global demographical burden of IBD, the disease is most prevalent in Western and Westernized countries. It has evolved into a global disease with 6.8 million cases and a global prevalence of 84.3 [MIN=79.2, MAX=89.9] per 100,000 individuals[27]. Figure 4 shows age-standardized prevalence rates of IBD for 2017 across the globe, as well as the prevalence change from 1990-2017 taken from Atalab *et al.*[27]. Here, age-standardized prevalence rates were calculated as the mean of the IBD prevalence observed in chosen age groups weighted by the number of individuals in the respective age groups.

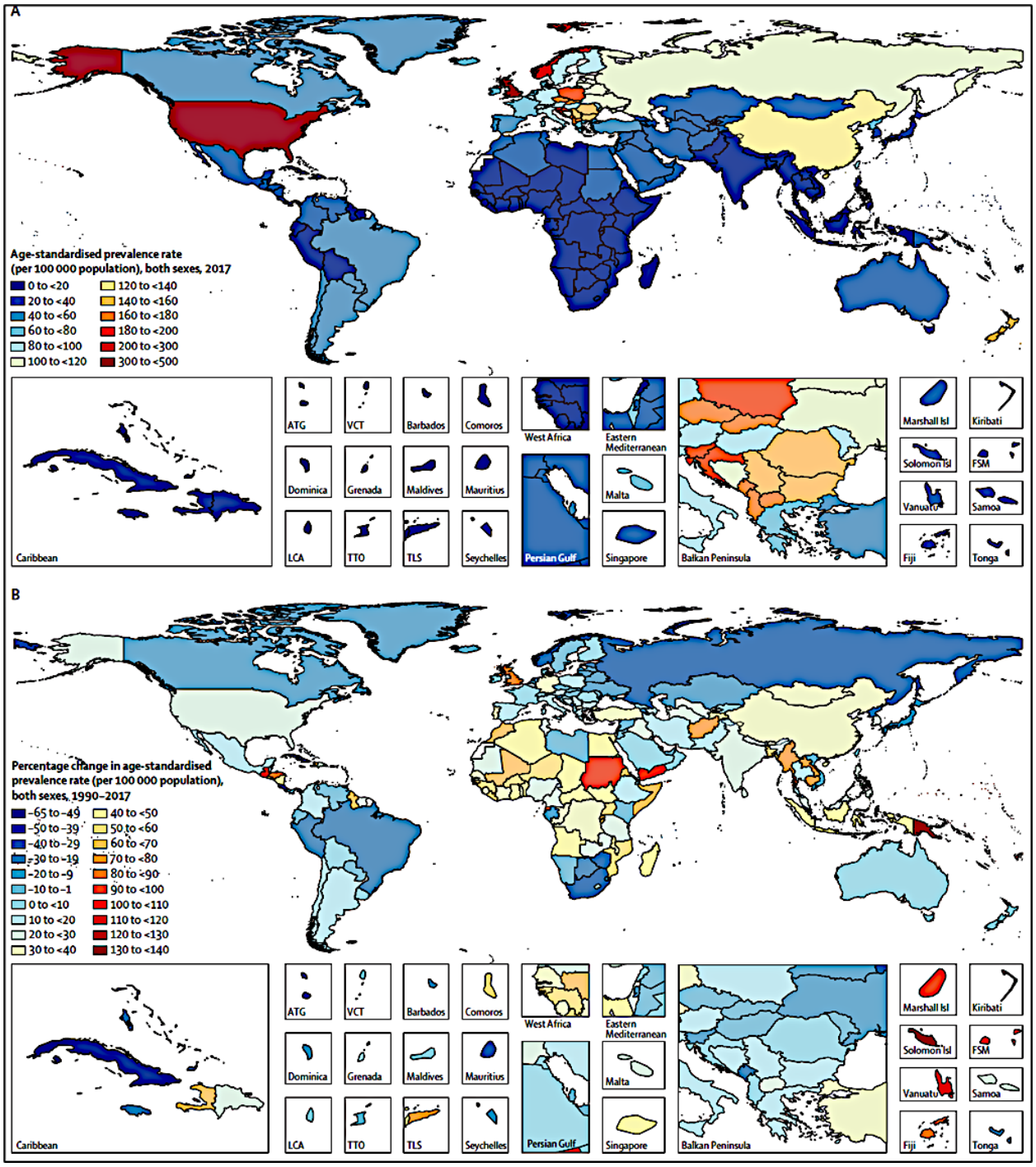


Figure 4 – Global prevalence of IBD.
Figure taken from Atalab *et al.*[27]. Original and shortened description: Figure 1: (A) Age-standardized prevalence rate (per 100,000 population) of IBD for 195 countries and territories, 2017. (B) Percentage change in age-standardized prevalence rate (per 100,000 population) of IBD. 1990-2017; ATG=Antigua and Barbuda. VCT=Saint Vincent and the Grenadines. LCA=Saint Lucia. TTO=Trinidad and Tobago.

Figure 4 nicely shows that prevalence rates have especially increased in Africa and East Asia, while they have not changed to a large degree in the Western and Westernized countries (North America and Europe). The increase of prevalence rates may be attributable to the increasing adaptation of a Western lifestyle – such as diet changing from traditional to Western diet and changes in hygiene practices. A total of $2.2 billion dollar are spent on medical care of IBD patients per year in the US alone[19]. Overall, regions with a high socio-demographic-index (SDI), a measure that includes demographical factors such as income, education and children born, had higher age-standardized prevalence rates than low SDI regions (Figure 5). Among the high SDI regions, North America had the highest prevalence (422.0 [398.7, 446.1])[27].

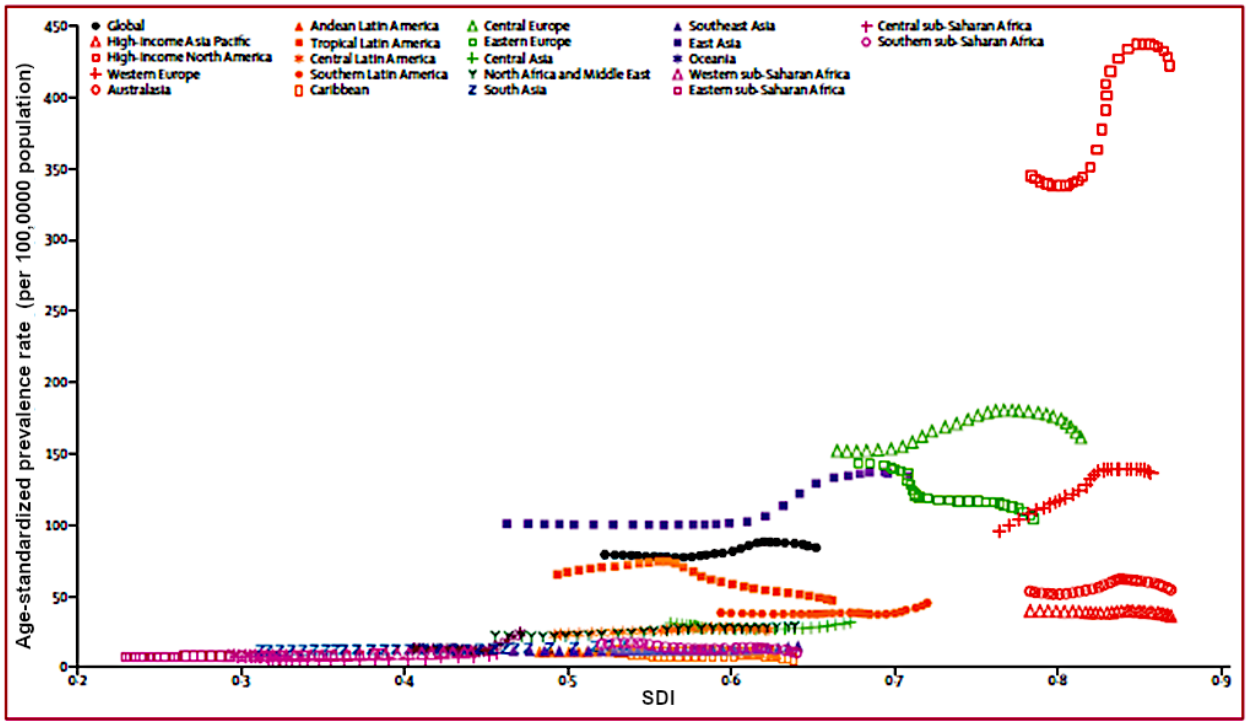


Figure 5 – Global prevalence rates of IBD and trends across time.
Figure taken and modified from Atalab *et al.*[27]. Publication by the Global Burden of Disease (GBD) 2017 Inflammatory Bowel Disease Collaborators. Original description: Figure 7 Age-standardized prevalence rate of IBD globally and for 21 GBD regions by SDI, 1990-2017.

In the German population, age-standardized values were reported to be 88.3 cases per 100,000 individuals (Supplement of Atalab *et al.*). Unstandardized incidence and prevalence rates for CD and UC separately can be found in Ng *et al.*[21]. However, the reported measures vary drastically even within the same country group. Maximal prevalence rates of UC in Asian populations (including East Asian, South Asian, South East Asian and Western Asian populations) were reported to be 106.2 per 100,000 individuals, while for CD prevalence rates were reported to be 53.1 (both values measured in Western Asia, values in other Asian countries are substantially lower by a factor of about 10). In Africa, a maximum prevalence of 10.57 per 100,000 individuals for UC and 19.02 for CD were reported, while

in South America UC was measured at a maximum prevalence of 44.3 and CD had a maximum prevalence of 44.1. The Supplementary Material of Ng *et al.*[21] also provides information on the UC to CD ratio. In general, UC has been reported to be more prevalent in the Asian population (UC to CD ratio of 2.75 (South Korea) to 4.28 (Taiwan)), while for the Caucasian population prevalence of UC over prevalence of CD was reported to be only slightly higher (UC to CD ratio of 1.03 to 1.62 in the USA). Some exceptions, as for instance parts of Southern Europe (UC to CD ratio 8.27 in Sardinia) and Finland (UC to CD ratio of 2.56-7.42), exist. Overall, incidence rates of IBD are levelling off in Caucasian populations and on the incline in newly Westernized countries (Figure 5), while prevalence rates are continuing to rise across the globe with an increase in overall life-expectancy.

4.1.4. PAPER A – IBD genetics

Paper A

Degenhardt F, Franke A. Genetik des Morbus Crohn und der Colitis ulcerosa: Aktueller Stand 15 Jahre nach Entdeckung von NOD2. Der Gastroenterologe. 2017. doi: 10.1007/s11377-016-0127-z.

This is a concise review of the current understanding of the genetics of Inflammatory bowel disease (IBD) and its two subforms Crohn's disease (CD) and ulcerative colitis (UC). It is directed at the broad medical community in Germany. Here, we describe genes that are relevant for IBD pathogenesis and that were identified to be associated with IBD by linkage analysis and in genome-wide association studies (GWAS). We highlight that the earliest identified markers are also among the most robust signals in many GWAS. Most of the discovered loci are shared between UC and CD, showing a strong commonality between the two inflammatory diseases with effect sizes that are correlated across different ethnicities. Genetic polymorphisms within the HLA region are highly associated with UC and explain a significant proportion of the heritability of this disease. We conclude that since the heritability explained by the markers identified until now is not exhausting its limits, environmental factors as well as genetic factors and their interplay must play a vital role in the disease aetiology of IBD.

Author contributions

This review was conceived and written by Andre Franke with contributions by Frauke Degenhardt.

Schwerpunkt

Gastroenterologe 2017 · 12:38–48
DOI 10.1007/s11377-016-0127-z
Online publiziert: 3. Januar 2017

F. Degenhardt · A. Franke
Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität zu Kiel, Kiel, Deutschland

Genetik des Morbus Crohn und der Colitis ulcerosa

Aktueller Stand 15 Jahre nach Entdeckung von *NOD2*

Die beiden häufigsten chronisch-entzündlichen Darmerkrankungen (CED) sind Morbus Crohn und Colitis ulcerosa. Die meisten CED-Patienten besitzen eine genetische Veranlagung für ihre Erkrankung und sind Träger von mehreren genetischen Risikovarianten. Was seit dem Jahr 2001, in dem das für Morbus Crohn bedeutsamste Krankheitsgen *NOD2* entdeckt wurde, in der genetischen Forschung passiert ist und welche klinische Bedeutung den neuen Erkenntnissen beigemessen werden kann, ist im Folgenden zusammengefasst. Der Schwerpunkt liegt dabei vor allem auf der prägnanten Zusammenfassung großer multizentrischer Studien in europäischstämmigen Patienten mit entsprechender statistischer Signifikanz.

Epidemiologie

Über eine Million US-Amerikaner und 2,5 Millionen Europäer leiden unter CED, mit steigender Inzidenz auch außerhalb des europäischen Kontinents (z. B. Asien und dem mittleren Osten) sowie Südamerika. Damit wird die Erkrankung zunehmend zu einem globalen Problem und einer immer größeren Last für das Gesundheitssystem [18]. Trotz intensiver Forschung gilt CED noch immer als eine idiopathische Erkrankung, was bedeutet, dass man zwar viele Risikofaktoren kennt, aber die Pathologie noch nicht vollständig erklären kann. Die CED sind multifaktorielle polygene Erkrankungen und lassen sich entsprechend nicht auf eine einzelne Ursache zurückführen: Es handelt sich um eine Verbindung von genetischer Veranlagung, noch näher zu bestimmenden Risiko- und Umweltfaktoren (etwa Rauchen, Ernährungsgewohnheiten, Hygiene) und einer Störung der Barrierefunktion der Darmschleimhaut, sodass Darmbakterien sich auf der Darmwand ansiedeln und eine starke Immunreaktion auslösen können. Auch die Zusammensetzung des Darmmikrobioms (die Gesamtheit aller Bakterien, Viren, Pilze, Protozoen und Archaeen im Darm) spielt eine wichtige Bedeutung bei der Krankheitsentstehung. Ein krankheitsauslösendes Pathogen konnte bis heute nicht eindeutig identifiziert werden, besonders das *Mycobacterium avium subsp. paratuberculosis* (MAP) wurde in der Vergangenheit (zu Unrecht) als Krankheitsauslöser verdächtigt [9].

Psychische Ursachen werden derzeit als Krankheitsauslöser ausgeschlossen, anerkannt ist aber, dass sich psychische Faktoren auf den Verlauf der Erkrankung auswirken können. Immer mehr aktuelle Studien zeigen auch eine Verbindung zwischen psychischen Faktoren und der Zusammensetzung des Darmmikrobioms („gut–brain axis").

» Epidemiologische Studien implizieren die Existenz genetischer Suszeptibilitätsfaktoren

Epidemiologische Studien haben eine familiäre Häufung von CED-Patienten gezeigt: 2–14 % der Patienten mit Morbus Crohn haben mindestens einen Verwandten, der ebenfalls an Morbus Crohn erkrankt ist [13]. Sie implizieren damit die Existenz genetischer Suszeptibilitätsfaktoren. Eine Metaanalyse von 6 Zwillingsstudien lieferte im Jahr 2011 Konkordanzraten für Morbus Crohn von 30,3 % für eineiige (d. h. nahezu 100 % genetisch identisch) und 3,6 % für zweieiige Zwillinge [10]. Zwar überschätzen Zwillingsstudien gegebenenfalls die Vererbbarkeit (Heritabilität) – eine ganz aktuelle Studie zeigt, dass „familiäre Umweltfaktoren" dringend berücksichtig werden müssen [24] – jedoch fehlt vielen Zwillingsstudien auch ein mehrjähriges Follow-up, d. h. man unterschätzt die Zahl der neuerkrankten Zwillingspartner aufgrund der Momentaufnahme der einmal erhobenen Daten. Zusammenfassend lässt sich aber mit großer Sicherheit sagen, dass die CED-Pathogenese eine signifikante genetische Komponente hat, jedoch CED (abgesehen von den *seltenen* frühkindlichen Formen) keine Erbkrankheiten im klassischen Sinne sind.

Entdeckung von *NOD2*

Im Jahr 1996 lieferte eine Kopplungsanalyse den ersten Hinweis auf ein CED-Krankheitsgen auf Chromosom 16 [15]. Im Jahr 2001 wurden dann 3 Risikovarianten für Morbus Crohn in *NOD2* auf Chromosom 16 identifiziert (R702W, G908R und L1007 fs; [16]). Das Odds-Ratio für die Varianten wurde mit 2–4 für heterozygote Träger und 20–40 für homozygote Träger beschrieben. Mindestens eine dieser Varianten liegt bei

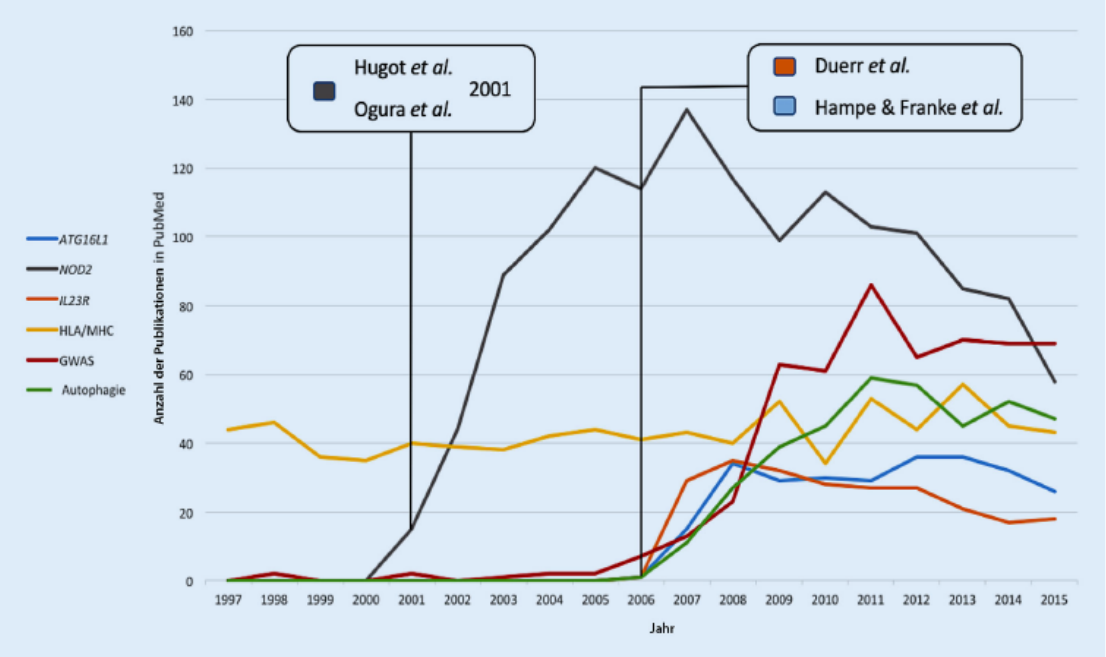


Abb. 1 ▲ Genidentifikationen führen zu zahlreichen weiterführenden Forschungsarbeiten zum entsprechenden Kandidatengen, dessen Protein und Stoffwechselweg. Die Y-Achse zeigt die Anzahl der in PubMed gelisteten Artikel für chronisch-entzündliche Darmerkrankungen (CED; oder Morbus Crohn oder Colitis ulcerosa) und die entsprechenden bekannten Kandidatengene. Vor der Identifikation von *ATG16L1* als Risikogen für Morbus Crohn wurde z. B. der Stoffwechselweg Autophagie im Kontext der CED noch nicht untersucht, danach ist ein nahezu exponentieller Anstieg zu erkennen (*grüne Linie*). *GWAS* genomweite Assoziationsstudie

30–40 % der Patienten mit Morbus Crohn vor, verglichen mit 6–7 % bei gesunden Europäern [22]. In den Jahren 1996–2004 wurden insgesamt 11 Kopplungsanalysen für Morbus Crohn publiziert (Review in [22], Metaanalyse: [31]). Zusammenfassend waren die Ergebnisse der Kopplungsanalysen eher enttäuschend, nur wenige robuste Signale wurden neben *NOD2* identifiziert.

> » *NOD2* ist das am besten validierte Krankheitsgen für Morbus Crohn

Schon im Jahr 1996 zeigten Risch et al. [27], dass 18.000 betroffene Geschwisterpaare in einer Kopplungsanalyse benötigt würden, um Odds-Ratios größer als 1,5 für Varianten mit einer Allelfrequenz von 50 % zu identifizieren. Es gilt: Je höher die Allelfrequenz und je größer das Odds-Ratio, d. h. die Effektstärke, desto größer ist die statistische Power, eine Risikovariante zu identifizieren. Nach wie vor ist *NOD2* das am besten untersuchte und am besten validierte Krankheitsgen für Morbus Crohn: Für die Ätiologie der Colitis ulcerosa sind *NOD2*-Risikovarianten sehr wahrscheinlich nicht relevant.

Genomweite Assoziationsstudien

Umfangreiche „genetische Landkarten", die über eine Million „single nucleotide polymorphisms" (Einzelnukleotidpolymorphismen, SNPs) im menschlichen Genom katalogisierten, und die Etablierung der Chiptechnologie in der Biomedizin machten schließlich genomweite Assoziationsstudien (GWAS) mit kommerziellen SNP-Arrays und über 500.000 zu testenden SNPs pro Probe möglich. Typischerweise werden dabei in einem GWAS-Ansatz DNA-Proben von mindestens 2000 nicht verwandten Patienten und 3000 Kontrollpersonen untersuch. Einige der wichtigsten Studien und Befunde sind in ◘ **Tab. 1** zusammengefasst. Die nunmehr 10 Jahre andauernde sogenannte „GWAS-Ära" lässt sich als äußerst erfolgreiches Kapitel der CED-Genetikforschung zusammenfassen. War bis zum Jahr 2001 noch kein Krankheitsgen für die CED bekannt, wurden mittlerweile dank GWAS über 200 Krankheitsloci beschrieben. Die Beschreibung der assoziierten Stoffwechselwege hat das Krankheitsverständnis der CED enorm verbessert. ◘ **Abb. 1** zeigt exemplarisch, dass die Identifikation eines einzelnen Krankheitsgens – und im konkreten Fall eines neuen Stoffwechselwegs – eine regelrechte „Forschungslawine" lostreten kann.

Gezielte Studien, z. B. zur Autophagie im Mausmodell, wurden erst nach Identifikation des Kandidatengens *ATG16L1* möglich und lieferten zusätzliche mechanistische Hinweise zur Relevanz dieses Stoffwechselwegs.

Zusammenfassung · Abstract

Gastroenterologe 2017 · 12:38–48 DOI 10.1007/s11377-016-0127-z

F. Degenhardt · A. Franke

Genetik des Morbus Crohn und der Colitis ulcerosa. Aktueller Stand 15 Jahre nach Entdeckung von *NOD2*

Zusammenfassung

Modernste Technologien der genetischen Forschung bieten der Medizin einen ganz neuen Zugangsweg zur Entdeckung von genetischen Krankheitsursachen. Innerhalb der letzten 15 Jahre wurden die genetischen Untersuchungstechniken in einem solchen Umfang weiterentwickelt, dass heute auch die Untersuchung „komplexer" Erkrankungen und die Entschlüsselung kompletter Patientengenome innerhalb kürzester Zeit möglich sind.

Für die chronisch-entzündlichen Darmerkrankungen (CED) Morbus Crohn und Colitis ulcerosa sind mittlerweile über 200 assoziierte Genloci bekannt. Die Mehrheit der identifizierten Loci überschneidet sich nicht nur zwischen Morbus Crohn und Colitis ulcerosa, was deren klinische Ähnlichkeit widerspiegelt, sondern auch mit anderen chronisch-entzündlichen Erkrankungen, vor allem mit Psoriasis, Morbus Bechterew und primären Immundefizienzen. Die beiden wichtigsten und am besten validierten CED-Krankheitsloci sind nach wie vor *NOD2* (für Morbus Crohn) und die HLA-Region (für Colitis ulcerosa). Genetische Analysen in anderen Ethnizitäten (z. B. Asiaten) zeigen nahezu die gleichen assoziierten Risikoloci wie in europäischstämmigen Patienten. Interessanterweise ist *NOD2* hier eine Ausnahme.

Die jüngsten großen genetischen Studien bestätigen die Assoziation von Risikovarianten in *NOD2* mit Dünndarmbefall bei Morbus Crohn und suggerieren die Existenz von eher drei als zwei Subformen der CED: 1) Colitis ulcerosa, 2) Morbus Crohn mit Dickdarmbefall, 3) Morbus Crohn mit Dünndarmbefall.

Durch hochmoderne Sequenzieranalysen konnten in den letzten Jahren für frühkindliche Formen der CED einzelne Mutationen vor allem in bekannten Immundefizienzgenen (neben den bekannten Mutationen in *IL10RA/B*), identifiziert und die Erkrankungen damit aufgeklärt werden (monogener Defekt).

Schlüsselwörter

Morbus Crohn · Colitis Ulcerosa · Chronisch-entzündliche Darmerkrankungen · Genomweite Assoziationsstudie · Genetische Risikofaktoren · Humanes Leukozytenantigen · Genetische Pleiotropie

Genetics of Crohn's disease and ulcerative colitis. Current status 15 years after discovery of *NOD2*

Abstract

Modern technologies in genetic research hold the promise of identifying yet unknown causes of diseases, revealing the relevant pathways, and prioritizing therapeutic targets. Technological developments of the last 15 years tremendously contributed to novel genetic findings in complex disease research. Entire patient genomes can now be deciphered within a few days at reasonable and continuously decreasing costs.

More than 200 genetic susceptibility loci have been identified for the inflammatory bowel (IBD) diseases, Crohn's disease (CD), and ulcerative colitis (UC). Reflecting their clinical similarity, most of the risk loci are shared between CD and UC outside the major histocompatibility complex (MHC). The genetic risk map of IBD is also highly similar to other chronic immune-mediated diseases, especially psoriasis, ankylosing spondylitis, and primary immunedeficiencies. The two best validated IBD disease loci are still *NOD2* for CD and the HLA/MHC region on chromosome 6p21 for UC. Genetic investigations in non-European-ancestry cohorts (e. g., from Eastern Asia) also suggest that risk loci are shared across different ancestries. Interestingly, this is not true for *NOD2*.

A recent large-scale and multicenter study validated the association of susceptibility variants within the *NOD2* gene and ileal CD. These genetic analyses also support the theory that rather three than two subtypes of IBD exist: (1) UC, (2) colonic CD, and (3) ileal CD.

Modern high-throughput sequencing studies revealed several monogenic forms of early onset and very early onset IBD, implicating often known immunodeficiency disease loci, including the previously implicated interleukin-10 receptor (*IL10RA/B*) gene. Thus, identifying a single causative genetic variant in these young patients proves their exact disease cause.

Keywords

Morbus Crohn · Colitis Ulcerosa · Inflammatory bowel diseases · Genome-wide association study · Genetic susceptibility · Major Histocompatibilty Complex · Genetic pleiotropy

Die durchschnittliche durch GWAS identifizierte CED-Risikovariante ist in der Normalbevölkerung häufig (mediane Allelfrequenz etwa 30 %) und besitzt ein niedriges Odds-Ratio (mediane Effektstärke von 1,1). Trotz der großen statistischen Signifikanz der Varianten bedeuten die hohen Allelfrequenzen und niedrigen Effektstärken einen geringen prädiktiven Wert („area under receiver operator curve", AUC, von 0,60 bei Unterscheidung Morbus Crohn vs. Colitis ulcersosa; [5]). Und die Erfahrung belegt, was die Statistiker prognostiziert haben: Je mehr Fälle und Kontrollen (idealerweise zwei bis dreimal soviele Kontrollen wie Patienten) in eine GWAS eingeschlossen werden, desto größer ist die statistische Power und desto mehr signifikante Loci können genomweit identifiziert werden [32]. Die genomweite Signifikanz, ein p-Wert der kleiner als 5×10^{-8} ist, gilt in der Genetik heutzutage als etabliertes Signifikanzniveau.

Im Jahr 2016 wurde die größte internationale genetische Assoziationsstudie für CED-Subphänotypen veröffentlicht [5]. Genetische Daten und Patientendaten aus 49 Zentren und 16 Ländern (Europa, Nordamerika und Australasien) wurden bezüglich Genotyp-Phänotyp-Assoziationen untersucht. Insgesamt wurden 16.902 Patienten mit Morbus Crohn und

Tab. 1 Wichtige genomweite Assoziationsstudien (GWAS) und deren Ergebnisse[a]

Jahr	Studie	Ergebnisse (Beispiele)
2005	Yamazaki et al. [34]	Erstidentifikation von *TNFSF15*, spezifisches Krankheitsgen für Asiaten
2006	Duerr et al. [6]	Erstidentifikation von *IL23R*
2007	Hampe & Franke et al. [14]	Erstidentifikation von *ATG16L1* und Implikation des Autophagiestoffwechselwegs bei CED
2007	WTCCC [33]	11 Loci für Morbus Crohn, davon 4 neu (z. B. Autophagiegen *IRGM*)
2008	Barrett et al. [2]	Metaanalyse für Morbus Crohn, Identifikation von 32 Loci (davon 21 neu)
2008	Franke et al. [10]	Erste GWAS für Colitis ulcerosa, Identifikation von 4 neuen Loci inklusive *IL10*
2010	McGovern et al. [23]	Metaanalyse für Morbus Crohn, Identifikation von *FUT2*
2010	Franke et al. [11]	Metaanalyse für Morbus Crohn, Identifikation von 71 Loci (davon 30 neu)
2011	Anderson et al. [1]	Metaanalyse für Colitis ulcerosa, Identifikation von 47 Loci (davon 29 neu)
2012	Jostins et al. [17]	Genomweite Kandidatengenanalyse (Immunochip), Identifikation von 163 Loci für CED (davon 71 neu)
2015	Liu et al. [21]	Metaanalyse (Immunochip) zusammen mit knapp 10.000 Patienten aus Ostasien, Indien und dem Iran; Identifikation von 38 neuen Loci
2016	Ellinghaus et al. [7]	Metaanalyse/phänotypübergreifende Analyse (Immunochip) von 5 verschiedenen Erkrankungen: Morbus Crohn, Colitis ulcerosa, Morbus Bechterew, Psoriasis und primär sklerosierende Cholangitis (PSC); 244 Loci identifiziert, davon 27 neu (und davon wiederum 6 neu für Morbus Crohn)

[a]Eine detailliertere und vollständigere Aufstellung ist der Tab. 1 in Ellinghaus et al. [7] zu entnehmen.
CED chronisch-entzündliche Darmerkrankungen

12.597 Patienten mit Colitis ulcerosa in die Analysen eingeschlossen. Es konnten *keine* signifikanten genetischen Assoziationen mit dem Krankheitsverlauf oder Komplikationen gefunden werden. Eine starke Assoziation mit einem frühen Ersterkrankungsalter zeigten 3 Loci: 3p21 (*MST1*), *NOD2* und HLA. *NOD2* zeigte außerdem eine starke Assoziation mit „ilealem Befall" (aber *nicht* mit Befall des Dickdarms *und* höherem Schweregrad/Komplikationen wie vorher angenommen). Die wichtigste klinische Erkenntnis aus den genetischen Analysen der Studie war, dass die CED nicht mehr binär in Morbus Crohn und Colitis ulcerosa unterteilt werden sollten, sondern ein kontinuierliches Spektrum von Erkrankungen mit Colitis ulcerosa und ilealem Morbus Crohn an den entgegengesetzten Polen darstellen.

Morbus Crohn mit Dickdarmbefall ist dabei der Colitis ulcerosa auf molekularer Ebene ähnlicher, wie ileokolischer Morbus Crohn dem ilealen Morbus Crohn näher ist. Im Hinblick auf Therapien und klinische Studien ist dieses Ergebnis äußerst interessant, die exakten klinischen Implikationen werden sich aber erst aus weiterführenden Studien ergeben. Mit Sicht durch die Forscherbrille sollte ein besonderes Augenmerk nun auf die verschiedenen Darmabschnitte bei CED-Patienten gelegt und deren Unterschiede in Physiologie, Mikrobiom und molekularen Profilen hochaufgelöst analysiert werden. Patienten mit Morbus Crohn sollten zukünftig also in 3 Gruppen aufgeteilt („ileal" vs. „colonic" vs. „ileocolonic") und Daten innerhalb dieser Gruppen analysiert werden.

Bedeutung der HLA-/MHC-Region

Der Haupthistokompatibilitätskomplex (*engl.* MHC) auf Chromosom 6p21 (25–34 Megabasen, d. h. etwa 9 Millionen Basen groß) ist beim Menschen auch als HLA-Region bekannt (HLA: humanes Leukozytenantigen). Die HLA-Region spielt bei fast allen chronisch-entzündlichen Erkrankungen die wichtigste Rolle, hier liegen meist die signifikantesten Risikovarianten mit den größten Effektstärken. Bei den CED sind vor allem die klassischen HLA-Gene am stärksten assoziiert, wobei die Assoziation für die Colitis ulcerosa deutlich stärker ist als für den Morbus Crohn (hier Klasse I stärker assoziiert als bei Colitis ulcerosa). Für beide Erkrankungen ist das Klasse-II-Gen *DRB1* der am stärksten assoziierte Genort [12]. Allerdings ist die Assoziation am HLA-Lokus nach wie vor äußerst komplex und lässt sich aufgrund der großen Dichte an genetischen Varianten/Kandidatengenen und der starken Kopplung (große Chromosomenabschnitte werden während der Meiose zwischen väterlichen und mütterlichen Chromosomen *nicht* „gemischt") genetisch nur schwer auflösen.

» *HLA-DRB1**01:03 ist das am stärksten assoziierte HLA-Allel für Colitis ulcerosa

Das am stärksten assoziierte HLA-Allel für Colitis ulcerosa ist *HLA-DRB1**01:03 (Odds-Ratio = 3,59; $p = 3 \times 10^{-19}$), das Klasse-II-Antigen-präsentierende Zellen (dendritische Zellen, Makrophagen, B-Zellen) und damit CD4-positive T-Zellen pathophysiologisch impliziert. Große funktionelle genomische Studien, die Genexpressions- und Epigenetikdaten von verschiedenen Zelltypen mit genetischen Informationen integriert analysierten, haben parallel CD4-positive Zellen (v. a. T_H0) als wichtigsten Zelltyp bei der Colitis ulcerosa identifiziert [8]. Eine offene Frage ist, ob ein exogenes Antigen bei der Colitis ulcerosa eine zentrale Rolle spielt oder welches das genaue körpereigene Antigen ist.

Krankheitsübergreifende Studien

Wie im Ausblick nachfolgend geschildert gibt es eine signifikante genetische Schnittmenge zwischen den CED und der klassischen Infektionserkrankung Lepra. Über die Hälfte der bekann-

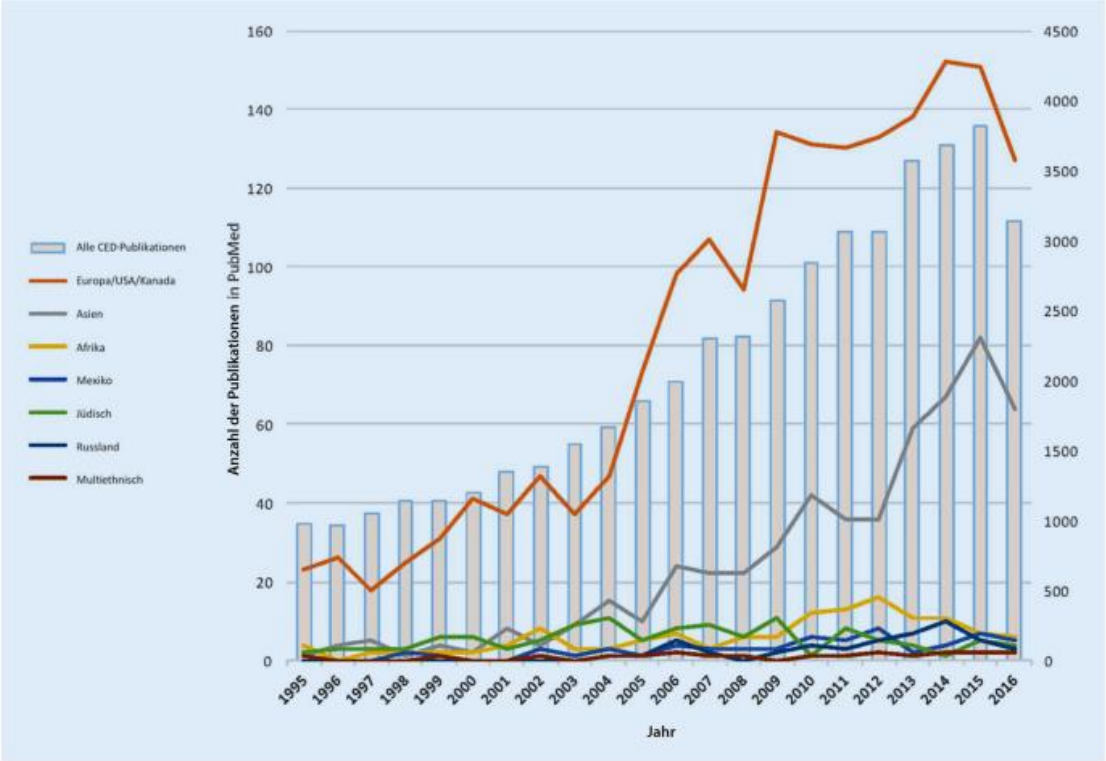


Abb. 2 ▲ Studien über chronisch-entzündliche Darmerkrankungen (CED) in nichteuropäischstämmigen Ethnizitäten sind nach wie vor unterrepräsentiert. Klar zu erkennen ist ein steiler Anstieg der CED-Veröffentlichungen seit dem Jahr 2003, vor allem für europäischstämmige Patienten (*orange Linie*). Seit etwa 2005 ist ein steiler Anstieg für Untersuchungen in asiatischen Kohorten (*graue Linie*) zu verzeichnen. Die linke Y-Achse zeigt die Anzahl der CED-Publikationen pro Jahr für die genannte Ethnie, die rechte Y-Achse zeigt die Anzahl der gesamten CED-Publikationen in PubMed pro Jahr (*Balkendiagramm im Hintergrund*)

ten CED-Risikoloci überlappen mit bekannten Risikoloci für andere chronisch-entzündliche Erkrankungen [17]. Ob es letztlich auch die gleichen Gene oder vielleicht sogar die gleichen Varianten mit gleichen Effektrichtungen sind, lässt sich nur durch aufwändige Feinkartierungsstudien und systematische krankheitsübergreifende Analysen herausfinden. Ellinghaus et al. (siehe **Tab. 1**) haben solch eine systematische Analyse im Jahr 2016 veröffentlicht. Durch die kombinierte Analyse von 5 verschiedenen Erkrankungen, inklusive Morbus Crohn und Colitis ulcerosa, konnten Ellinghaus et al. zeigen, dass in der Tat eine große genetische Ähnlichkeit zwischen den genannten 5 Erkrankungen besteht. Allerdings sind die Beziehungen komplexer als zuvor angenommen und teilweise unterscheiden sich die Effektrichtungen und -stärken der einzelnen Varianten stark. Oft sind es auch verschiedene Varianten am gleichen Locus, die bei der anderen Erkrankung assoziiert sind.

Die Studie kommt zu dem Schluss, dass genetische Pleiotropie – und nicht die klinische Heterogenität der Erkrankungen (z. B. exzessive Komorbidität) – die primäre Erklärung für die molekulare Ähnlichkeit verschiedener Erkrankungen ist. Genetische Pleiotropie bedeutet, dass ein einzelnes Gen (oder eine einzelne Variante) mehrere phänotypische Merkmale ausprägen kann. Eine Analyse sämtlicher elektronischer Patientenakten in Dänemark aus den letzten Jahren bestätigte, dass die 5 untersuchten Erkrankungen hochsignifikant miteinander assoziiert sind und häufig im gleichen Patienten auftreten ($p < 1{,}21 \times 10^{-9}$ für 10 der 12 möglichen Krankheitskombinationen). Für weitere Informationen zum Thema werden die Übersichtsartikel von Lees et al. [20] beziehungsweise Parkes et al. [26] empfohlen.

Genetische Studien in anderen Ethnien

Abb. 2 veranschaulicht, dass der überwiegende Teil an CED-Studien aus Europa oder Nordamerika kommt. Jährlich werden nur wenige CED-Studien zu ethnischen Minderheiten in Nordamerika (Asiaten, Mexikaner, Afroamerikaner) oder aus den entsprechenden Heimatländern publiziert. Erfreulich ist generell insgesamt der große Anstieg an CED-Fachartikeln pro Jahr seit 2003 und die immer stärker werdende CED-Forschung in Asien (linearer

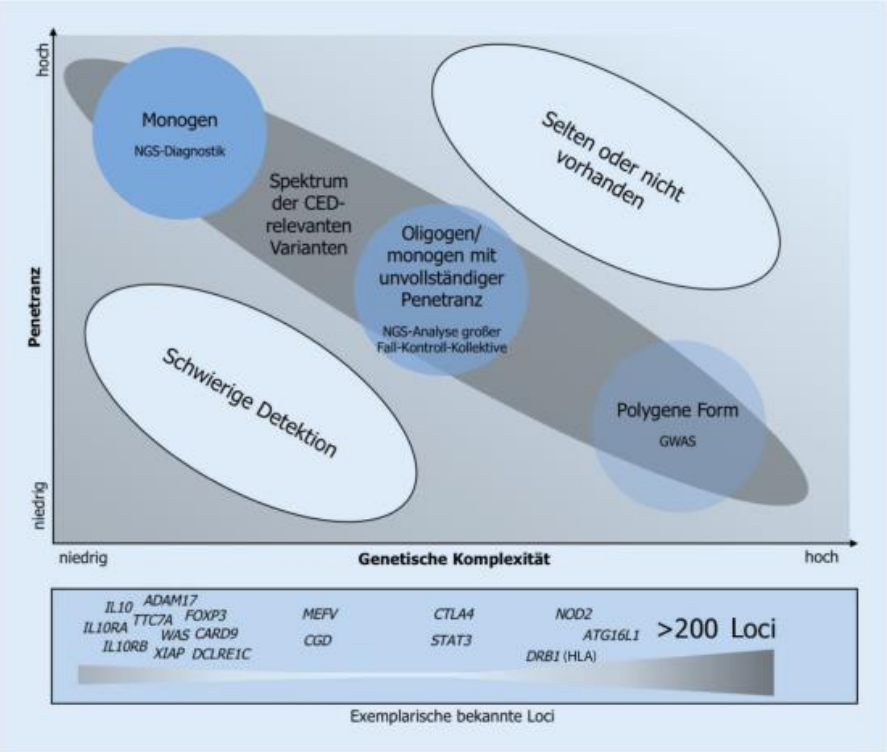


Abb. 3 ◀ Das Spektrum der bekannten genetischen Risikovarianten chronisch-entzündlicher Erkrankungen (CED) reicht von seltenen bis häufigen Varianten (in Bezug auf deren Frequenz in der Normalbevölkerung) und von niedrig penetranten Varianten (polygene Form der CED, d. h. hohe Komplexität oder „disease modifier" bei frühkindlichen Formen) zu hochpenetranten Mutationen. Im oberen Achsensystem sind verschiedene „Variantenklassen" und die Methoden, die für deren Detektion üblicherweise genutzt werden, skizziert. Das untere Panel listet exemplarische bekannte Krankheitsgene. *GWAS* genomweite Assoziationsstudie, *NGS* Next Generation Sequencing

Anstieg seit 2005). Dieser Anstieg lässt sich wahrscheinlich auch durch die steigende CED-Inzidenz und -Prävalenz in Asien und die damit verbundene Bedeutung für das Gesundheitssystem und das öffentliche Leben erklären. Eine der größten genetischen Studien zu CED für Ostasien (Japan, Korea, Hong-Kong-Chinesen), Indien und den Iran (siehe ◘ **Tab. 1**, [21]) zeigte, dass die Effektstärken und -richtungen für die bekannten CED-Risikoloci nahezu identisch mit den Daten von europäischstämmigen Patienten sind. Nur wenige Loci zeigten signifikante Unterschiede in den Ostasiaten (*NOD2* → in Ostasien nicht relevant, HLA → signifikante Unterschiede bei den assoziierten Allelen/Genen, *TNFSF15* → in Ostasiaten deutlich relevanter). Unterschiede in Umwelteinflüssen, Ernährungsverhalten, genetischer Abstammung (und in entsprechend anderer evolutionärer Historie) und menschlicher Physiologie erklären wahrscheinlich die wenigen, aber signifikanten Unterschiede zwischen Ostasiaten und europäischstämmigen Patienten.

Frühkindliche Formen chronisch-entzündlicher Darmerkrankungen

Die meisten CED-Patienten haben eine komplexe polygene Form der Erkrankung, d. h. mehrere genetische Risikovarianten im Zusammenspiel mit pathophysiologisch deutlich relevanteren (unbekannten) Umweltfaktoren führen zur Erkrankung (siehe ◘ **Abb. 3**). In den letzten Jahren – und seit dem ersten Artikel von Glocker et al. zum Thema im *The New England Journal of Medicine* im Jahr 2009 – sind zunehmend Berichte über Patienten mit monogenen beziehungsweise oligogenen Erkrankungsformen der CED erschienen. Hierbei handelt es sich überwiegend um frühkindliche äußerst schwere Formen der Erkrankung mit Erstmanifestation in den ersten 8 Jahren, jedoch auch in einzelne Fälle bis zum 16. Lebensjahre. Bahnbrechende Entwicklungen in der Sequenziertechnologie – ein menschliches Genom kann heute innerhalb kürzester Zeit in einem Labor vollständig entschlüsselt werden – ebneten den Weg für die einfachere Aufklärung von monogenen Mendel-Erkrankungen. Die genetischen Defekte dieser CED-ähnlichen monogenen Erkrankungen beeinträchtigen meist die intestinale Barrierefunktion oder sind ursächlich für einen Immundefekt (d. h. gestörte Granulozyten- und Phagozytenaktivität, hyper- und autoinflammatorische Störungen oder gestörte B- sowie T-Lymphozyten-Funktionen). Einen detaillierten und aktuellen Überblick liefert der Übersichtsartikel von Uhlig et al. [30]. Die Identifikation einzelner Gene und Mutationen hat das Verständnis der Pathogenese von CED enorm verbessert und erlaubt in einigen Fällen auch die exakte Klärung der Krankheitsursachen. Seit dem 1.7.2016 ist eine Genpaneldiagnostik mit Next Generation Sequencing (NGS) auch gemäß einheitlichem Bewertungsmaßstab (EBM) abrechenbar

Infobox 1 Weiterführende internetbasierte Informationen

- International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). https://www.ibdgenetics.org. Zugegriffen: 6. Dezember 2016
- Kassenärztliche Bundesvereinigung. Praxisnachrichten. Weiterentwicklung der Gebührenordnungspositionen zum 1. Juli. http://www.kbv.de/html/1150_21752.php. Zugegriffen: 6. Dezember 2016

(Infobox 1) und sollte bei Verdacht auf eine frühkindliche schwere CED-ähnliche Erkrankung entsprechend (über ein Institut für Humangenetik und Speziallabore) beauftragt werden. Die Identifikation eines entsprechenden Immundefekts kann neue therapeutische Optionen in therapierefraktären Patienten eröffnen. Entscheidend ist das Ersterkrankungsalter; ältere Patienten mit frühem Ersterkrankungsalter können natürlich bei entsprechendem Verdacht auch einem diagnostischen NGS-Test unterzogen werden.

Ausblick

Moderne genetische Analysen sind äußerst exakt, quantitativ (Allelzahlen können sehr präzise gemessen werden) und liefern extrem hochaufgelöste Daten des menschlichen Genoms. Die Identifikation von über 200 Krankheitsvarianten für die CED und die Aufklärung diverser frühkindlicher Formen der CED haben das Verständnis dieser Erkrankung enorm verbessert. Über 1444 Kandidatengene liegen in 163 der 200 CED-assoziierten genomischen Regionen. Von diesen konnten Jostins et al. [17] 300 Gene als wahrscheinlich krankheitsrelevant mithilfe bioinformatischer Verfahren priorisieren. Abgesehen von *NOD2, ATG16L1, IL23R, IL10* und anderen wenigen Ausnahmen ist bis jetzt also unklar, welches die eigentlichen Risikogene in den assoziierten genomischen Regionen und letztlich welche genomischen Varianten relevant sind: Sind es häufige oder seltene kausative Varianten? Welche Art von Varianten, d. h. SNPs, „singlenucleotidevariants" (SNVs), Insertionen/Deletionen oder Kopienzahlvariationen, liegen vor?

Weiterführende Untersuchungen bis hin zu Studien im Tiermodell sind nötig, um die genaue Rolle der Kandidatengene zu untersuchen. Systematische Sequenzierexperimente in mehreren 10.000 Patienten sind außerdem nötwendig, um auch die selteneren genetischen Varianten (Allelfrequenz <5 % und v. a. <1 % in der Normalbevölkerung) auf Assoziation mit CED zu testen. Alle genetischen Befunde für die polygene Form der CED erklären bisher nur etwa 13,6 % der phänotypischen Varianz für Morbus Crohn und 7,5 % für Colitis ulcerosa [17]. Einen großen Anteil hat dabei die MHC-Region. Erklärt für Morbus Crohn der MHC-Index-SNP (dies ist der am stärksten assoziierte SNP in der Region) rs9264942 noch 0,3 % der Varianz, sind es für alle HLA-Allele zusammen 3,1 %. Für Colitis ulcerosa ergeben sich 2,3 % der Varianz durch den Index-SNP rs6927022 und 6,2 % durch alle HLA-Allele [12]. Diese Zahlen zeigen, dass abgesehen vom Rauchen, das ein etablierter Risikofaktor für Morbus Crohn ist, vor allem bislang noch nicht identifizierte Umweltfaktoren einen großen Anteil an der CED-Ätiologie haben müssen.

» Bislang nicht identifizierte Umweltfaktoren müssen einen großen Anteil an der CED-Ätiologie haben

Die bisher identifizierten genetischen Risikofaktoren können natürlich Hinweise auf entsprechende Umweltfaktoren liefern, wie z. B. *NOD2* das angeborene Immunsystem im Darmepithel (und die Rolle der Bakterien im Darm) auf das Tableau gebracht hat.

Noch größere Anstrengungen werden nötig sein, um nichtgenetische vererbbare molekulare Faktoren zu untersuchen. Jüngste Analysen konnten z. B. zeigen, dass eine chronische Darmentzündung im Mausmodell [29] auch Auswirkungen auf die nachfolgenden Generationen hat (F_1- und F_2-Generation) und Störungen der Genexpression im Darmepithel hervorruft. Epigenetische Veränderungen der Keimbahn-DNA, z. B. DNA-Methylierung, sind demnach als nichtgenetische Erbfaktoren zu berücksichtigen.

Nach Einschätzung der Autoren liegt das größte Potenzial, zumindest für die Colitis ulcerosa, in der Erforschung der HLA-Region auf Chromosom 6p21 und in gezielten immunogenetischen Analysen. Die HLA-Assoziation der Colitis ulcerosa kann mit der HLA-Assoziation bei Zöliakie verglichen werden (MHC-Klasse-II-Signal, hochsignifikante Assoziation; [19]). Bei der Zöliakie kennt man das Antigen „Gluten" als wesentlichen krankheitsverursachenden Umweltfaktor. Es könnte sich also lohnen, ein entsprechendes Antigen bei der Colitis ulcerosa zu suchen.

Proteine in der „modernen" Ernährung sowie des Darmmikrobioms erscheinen hier als plausible Testkandidaten. Das Zusammenspiel zwischen Ernährungsfaktoren, Darmmikrobiom und Wirt (hier: Gesamtheit der genetischen Varianten eines Menschen) ist Gegenstand zahlreicher laufender Untersuchungen und wird in den nächsten Jahren mit großer Wahrscheinlichkeit viele neue Erkenntnisse zur CED-Ätiologie liefern. In diesem Zusammenhang ist auch die Tatsache interessant, dass Varianten am NOD2-Genort, allerdings andere als die CED-Risikovarianten, die Hauptrisikovarianten für Lepra, eine klassische Infektionserkrankung (*Mycobacterium leprae*), sind. Dieses konnte im Jahr 2010 in einer chinesischen GWAS für Lepra gezeigt werden [35]. Die Liste der Risikoloci für Lepra und CED überschneidet sich nahezu perfekt. Neben *NOD2* wurden die bekannten CED-Loci *C13orf31, RIPK2, TNFSF15* und die HLA-Region als Risikofaktoren für Lepra identifiziert. Damit ist die Mykobakterieninfektionshypothese für die CED wieder aktueller denn je. Wahrscheinlich ist jedoch, dass nicht nur Mykobakterien, sondern auch andere Bakterien des Mikrobioms eine entscheidende Rolle bei der Krankheitsentstehung spielen. Zusammenfassend könnte die Pathogenese der CED wie folgt verlaufen: Eine beeinträchtigte Darmbarriere (gestörtes Mikrobiom mit geringerer Artenvielfalt und mehr „schlechten" als „guten" Bakterien, Stress, Antibiotika, für die

Mukosa toxische Ernährungsbestandteile) erleichtert Bakterien eine Ansiedelung auf dem Darmepithel. Die Barriere ist aufgrund genetischer Suszeptibilität durchlässiger (z. B. Störung der „tight junctions") und das angeborene Immunsystem geschwächt (z. B. NOD2-Defekt, Autophagiedefekt). Somit kann der Darminhalt leichter mit Zellen des adaptiven Immunsystems (z. B. mit dendritischen Zellen) in Berührung kommen. Genetische Suszeptibilität (die meisten der CED-relevanten Risikogene sind in Immunzellen exprimiert), mangelndes „immunologisches Training" im frühen Kindesalter (Hygienehypothese) und die große „Dosis" an Bakterien und sonstigem Darminhalt hinter der Barriere führen zu einer übersteigerten Immunantwort. Diese Immunreaktion kann in genetisch suszeptiblen Personen auch nur langsam wieder herunterreguliert werden (z. B. Defekt im antiinflammatorischen Interleukin-10-Stoffwechselweg). Wahrscheinlich ist auch, dass das entsprechende Antigen „chronisch" vorhanden ist, d. h. Bestandteil des dysregulierten Mikrobioms des Patienten ist. Zwar ist es durchaus zu begrüßen, dass mittlerweile eine kleine Auswahl an Biologikatherapien zur Behandlung der CED zur Verfügung steht. Jedoch sollte das in keiner Weise die Suche nach den eigentlichen Krankheitsursachen verlangsamen, zumal die Genetik der Lösung des Rätsels schon ein Stück weit entgegengekommen ist.

Fazit für die Praxis

- **Chronisch-entzündliche Darmerkrankungen umfassen ein Spektrum von Entzündungsformen, in dem der Morbus Crohn mit reinem Dünndarmbefall und die Colitis ulcerosa die beiden Extreme eines fließenden Übergangs darstellen. Morbus Crohn mit reinem Befall des Dickdarms sollte als distinkte 3. Entität zwischen den beiden Extremen betrachtet werden.**
- **Die häufigen genetischen Suszeptibilitätsfaktoren für die CED wurden für europäischstämmige Patienten zwischenzeitlich sämtlich identifiziert, Patienten mit anderem ethnischen Hintergrund unterscheiden sich nur gering in ihrem genetischen Risikoprofil.**
- **Die Aufklärung der ursächlichen genetischen Prinzipien inklusive der damit verbundenen primären pathophysiologischen Ereignisse hat die Entwicklung neuer Therapieverfahren beeinflusst.**
- **Die genetische Ätiologie chronisch-entzündlicher Darmerkrankungen ist überraschend polygen. Zudem sind viele der Krankheitsgene auch für chronische Erkrankungen relevant, die vordergründig nicht mit dem Darm assoziiert sind.**
- **Patienten mit einem CED-Ersterkrankungsalter <16 Jahre sollten eine genetische Diagnostik (seit 1.7.2016 abrechenbar) durchlaufen. Es gilt: Je jünger der Patient bei Ersterkrankung, desto wahrscheinlicher ist das Vorliegen einer monogenen Form (nur *eine* kausative Mutation) der CED.**

Korrespondenzadresse

Prof. Dr. A. Franke
Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität zu Kiel
Rosalind-Franklin-Str. 12, 24105 Kiel, Deutschland
a.franke@mucosa.de

Einhaltung ethischer Richtlinien

Interessenkonflikt. F. Degenhardt und A. Franke geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine von den Autoren durchgeführten Studien an Menschen oder Tieren.

Literatur

1. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, Lagacé C, Scott R, Amininejad L, Bumpstead S, Baidoo L, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Colombel JF, Denson LA, De Vos M, Dubinsky M, Edwards C, Ellinghaus D, Fehrmann RS, Floyd JA, Florin T, Franchimont D, Franke L, Georges M, Glas J, Glazer NL, Guthery SL, Haritunians T, Hayward NK, Hugot JP, Jobin G, Laukens D, Lawrance I, Lémann M, Levine A, Libioulle C, Louis E, McGovern DP, Milla M, Montgomery GW, Morley KI, Mowat C, Ng A, Newman W, Ophoff RA, Papi L, Palmieri O, Peyrin-Biroulet L, Panés J, Phillips A, Prescott NJ, Proctor DD, Roberts R, Russell R, Rutgeerts P, Sanderson J, Sans M, Schumm P, Seibold F, Sharma Y, Simms LA, Seielstad M, Steinhart AH, Targan SR, van den Berg LH, Vatn M, Verspaget H, Walters T, Wijmenga C, Wilson DC, Westra HJ, Xavier RJ, Zhao ZZ, Ponsioen CY, Andersen V, Torkvist L, Gazouli M, Anagnou NP, Karlsen TH, Kupcinskas L, Sventoraityte J, Mansfield JC, Kugathasan S, Silverberg MS, Halfvarson J, Rotter JI, Mathew CG, Griffiths AM, Gearry R, Ahmad T, Brant SR, Chamaillard M, Satsangi J, Cho JH, Schreiber S, Daly MJ, Barrett JC, Parkes M, Annese V, Hakonarson H, Radford-Smith G, Duerr RH, Vermeire S, Weersma RK, Rioux JD (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43(3):246–252. doi:10.1038/ng.764
2. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Wellcome Trust Case Control Consortium, Belgian-French IBD Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40(8):955–962. doi:10.1038/ng.175
3. Beaudoin M, Goyette P, Boucher G, Lo KS, Rivas MA, Stevens C, Alikashani A, Ladouceur M, Ellinghaus D, Törkvist L, Goel G, Lagacé C, Annese V, Bitton A, Begun J, Brant SR, Bresso F, Cho JH, Duerr RH, Halfvarson J, McGovern DP, Radford-Smith G, Schreiber S, Schumm PL, Sharma Y, Silverberg MS, Weersma RK, Quebec IBD Genetics Consortium, NIDDK IBD Genetics Consortium, International IBD Genetics Consortium, D'Amato M, Vermeire S, Franke A, Lettre G, Xavier RJ, Daly MJ, Rioux JD (2013) Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLOS Genet 9(9):e1003723. doi:10.1371/journal.pgen.1003723
4. Brant SR (2011) Update on the heritability of inflammatory bowel disease: the importance of twin studies. Inflamm Bowel Dis 17(1):1–5. doi:10.1002/ibd.21385
5. Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, Andersen V, Andrews JM, Annese V, Brand S, Brant SR, Cho JH, Daly MJ, Dubinsky M, Duerr RH, Ferguson LR, Franke A, Gearry RB, Goyette P, Hakonarson H, Halfvarson J, Hov JR, Huang H, Kennedy NA, Kupcinskas L, Lawrance

IC, Lee JC, Satsangi J, Schreiber S, Théâtre E, van der Meulen-de Jong AE, Weersma RK, Wilson DC, International Inflammatory Bowel Disease Genetics Consortium, Parkes M, Vermeire S, Rioux JD, Mansfield J, Silverberg MS, Radford-Smith G, McGovern DP, Barrett JC, Lees CW (2016) Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. Lancet 387(10014):156–167. doi:10.1016/S0140-6736(15)00465-1

6. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314(5804):1461–1463
7. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, Park YR, Raychaudhuri S, Pouget JG, Hübenthal M, Folseraas T, Wang Y, Esko T, Metspalu A, Westra HJ, Franke L, Pers TH, Weersma RK, Collij V, D'Amato M, Halfvarson J, Jensen AB, Lieb W, Degenhardt F, Forstner AJ, Hofmann A, International IBD Genetics Consortium (IIBDGC), International Genetics of Ankylosing Spondylitis Consortium (IGAS), International PSC Study Group (IPSCSG), Genetic Analysis of Psoriasis Consortium (GAPC), Psoriasis Association Genetics Extension (PAGE), Schreiber S, Mrowietz U, Juran BD, Lazaridis KN, Brunak S, Dale AM, Trembath RC, Weidinger S, Weichenthal M, Ellinghaus E, Elder JT, Barker JN, Andreassen OA, McGovern DP, Karlsen TH, Barrett JC, Parkes M, Brown MA, Franke A (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 48(5):510–518. doi:10.1038/ng.3528
8. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518(7539):337–343. doi:10.1038/nature13835
9. Feller M, Huwiler K, Stephan R, Altpeter E, Shang A, Furrer H, Pfyffer GE, Jemmi T, Baumgartner A, Egger M (2007) Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis. Lancet Infect Dis 7(9):607–613
10. Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, Domingues FS, Albrecht M, Nothnagel M, Ellinghaus D, Sina C, Onnie CM, Weersma RK, Stokkers PC, Wijmenga C, Gazouli M, Strachan D, McArdle WL, Vermeire S, Rutgeerts P, Rosenstiel P, Krawczak M, Vatn MH, BSEN study group, Mathew CG, Schreiber S (2008) Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat Genet 40(11):1319–1323. doi:10.1038/ng.221
11. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42(12):1118–1125. doi:10.1038/ng.717
12. Goyette P, Boucher G, Mallon D, Ellinghaus E, Jostins L, Huang H, Ripke S, Gusareva ES, Annese V, Hauser SL, Oksenberg JR, Thomsen I, Leslie S, International Inflammatory Bowel Disease Genetics Consortium, Australia and New Zealand IBDGC, Belgium IBD Genetics Consortium, Italian Group for IBD Genetic Consortium, NIDDK Inflammatory Bowel Disease Genetics Consortium, United Kingdom IBDGC, Wellcome Trust Case Control Consortium, Quebec IBD Genetics Consortium, Daly MJ, Van Steen K, Duerr RH, Barrett JC, McGovern DP, Schumm LP, Traherne JA, Carrington MN, Kosmoliaptsis V, Karlsen TH, Franke A, Rioux JD, Quebec IBD Genetics Consortium, Daly MJ, Van Steen K, Duerr RH, Barrett JC, McGovern DP, Schumm LP, Traherne JA, Carrington MN, Kosmoliaptsis V, Karlsen TH, Franke A, Rioux JD (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(2):172–179. doi:10.1038/ng.3176
13. Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K (2006) Family and twin studies in inflammatory bowel disease. World J Gastroenterol 12(23):3668–3672
14. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, Günther S, Prescott NJ, Onnie CM, Häsler R, Sipos B, Fölsch UR, Lengauer T, Platzer M, Mathew CG, Krawczak M, Schreiber S (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 39(2):207–211
15. Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, Bonaiti-Pellie C, Weissenbach J, Mathew CG, Lennard-Jones JE, Cortot A, Colombel JF, Thomas G (1996) Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature 379(6568):821–823
16. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411(6837):599–603
17. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Büning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, International IBD Genetics Consortium (IIBDGC), Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491(7422):119–124. doi:10.1038/nature11582
18. Kaplan GG (2015) The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 12(12):720–727. doi:10.1038/nrgastro.2015.150
19. Karlsen TH, Chung BK (2015) Genetic risk and the development of autoimmune liver disease. Dig Dis 33(Suppl 2):13–24. doi:10.1159/000440706
20. Lees CW, Barrett JC, Parkes M, Satsangi J (2011) New IBD genetics: common pathways with other diseases. Gut 60(12):1739–1753. doi:10.1136/gut.2009.199679
21. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, Abedian S, Cheon JH, Cho J, Daryani NE, Franke L, Fuyuno Y, Hart A, Juyal RC, Juyal G, Kim WH, Morris AP, Poustchi H, Newman WG, Midha V, Orchard TR, Vahedi H, Sood A, Sung JJ, Malekzadeh R, Westra HJ, Yamazaki K, Yang SK, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Barrett JC, Franke A, Alizadeh BZ, Parkes M, Daly MJ, Kubo M, Anderson CA, Weersma RK (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47(9):979–986. doi:10.1038/ng.3359
22. Mathew CG, Lewis CM (2004) Genetics of inflammatory bowel disease: progress and prospects. Hum Mol Genet 13(1):R161–R168
23. McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, Ippoliti A, Vasiliauskas E, Berel D, Derkowski C, Dutridge D, Fleshner P, Shih DQ, Melmed G, Mengesha E, King L, Pressman S, Haritunians T, Guo X, Targan SR, Rotter JI, International IBD Genetics Consortium (2010) Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet 19(17):3468–3476. doi:10.1093/hmg/ddq248
24. Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A (2016) Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. Nat Genet 48(9):980–983. doi:10.1038/ng.3618
25. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nuñez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411(6837):603–606
26. Parkes M, Cortes A, van Heel DA, Brown MA (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 14(9):661–673. doi:10.1038/nrg3502

27. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281):1516–1517
28. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagacé C, Neale B, Lo KS, Schumm P, Törkvist L, National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, Lettre G, Franke A, D'Amato M, McGovern DP, Cho JH, Rioux JD, Xavier RJ, Daly MJ (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43(11):1066–1073. doi:10.1038/ng.952
29. Tschurtschenthaler M, Kachroo P, Heinsen FA, Adolph TE, Rühlemann MC, Klughammer J, Offner FA, Ammerpohl O, Krueger F, Smallwood S, Szymczak S, Kaser A, Franke A (2016) Paternal chronic colitis causes epigenetic inheritance of susceptibility to colitis. Sci Rep 6:31640. doi:10.1038/srep31640
30. Uhlig HH, Schwerd T, Koletzko S, Shah N, Kammermeier J, Elkadri A, Ouahed J, Wilson DC, Travis SP, Turner D, Klein C, Snapper SB, Muise AM, COLORS in IBD Study Group and NEOPICS (2014) The diagnostic approach to monogenic very early onset inflammatory bowel disease. Gastroenterology 147(5):990–1007.e3. doi:10.1053/j.gastro.2014.07.023
31. van Heel DA, Fisher SA, Kirby A, Daly MJ, Rioux JD, Lewis CM, Genome Scan Meta-Analysis Group of the IBD International Genetics Consortium (2004) Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. Hum Mol Genet 13(7):763–770
32. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7–24. doi:10.1016/j.ajhg.2011.11.029
33. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678
34. Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, Cardon L, Takazoe M, Tanaka T, Ichimori T, Saito S, Sekine A, Iida A, Takahashi A, Tsunoda T, Lathrop M, Nakamura Y (2005) Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet 14(22):3499–3506 (Nov)
35. Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y, Cui Y, Yan XX, Yang HT, Yang RD, Chu TS, Zhang C, Zhang L, Han JW, Yu GQ, Quan C, Yu YX, Zhang Z, Shi BQ, Zhang LH, Cheng H, Wang CY, Lin Y, Zheng HF, Fu XA, Zuo XB, Wang Q, Long H, Sun YP, Cheng YL, Tian HQ, Zhou FS, Liu HX, Lu WS, He SM, Du WL, Shen M, Jin QY, Wang Y, Low HQ, Erwin T, Yang NH, Li JY, Zhao X, Jiao YL, Mao LG, Yin G, Jiang ZX, Wang XD, Yu JP, Hu ZH, Gong CH, Liu YQ, Liu RY, Wang DM, Wei D, Liu JX, Cao WK, Cao HZ, Li YP, Yan WG, Wei SY, Wang KJ, Hibberd ML, Yang S, Zhang XJ, Liu JJ (2009) Genomewide association study of leprosy. N Engl J Med 361(27):2609–2618. doi:10.1056/NEJMoa0903753

Fachnachrichten

Mitteldeutsche Gesellschaft für Gastroenterologie
Hessen — Sachsen — Sachsen-Anhalt — Thüringen

Ausschreibung des Förderpreises der Mitteldeutschen Gesellschaft für Gastroenterologie e.V.

Die Mitteldeutsche Gesellschaft für Gastroenterologie e.V. (MGG) vergibt auf ihrem 26. Jahreskongress in Frankfurt a.M. 2017 wieder einen Förderpreis für junge Kliniker und Wissenschaftler.

Der Preis wird für Forschungsarbeiten von Klinikern und Wissenschaftlern, die in den Mitgliedsländern der MGG (Hessen, Thüringen, Sachsen, Sachsen-Anhalt) tätig und unter 45 Jahre alt sind, vergeben und ist mit 3.000 € dotiert.

Die eingereichten Arbeiten müssen ein Forschungsthema aus dem Gebiet der klinischen Gastroenterologie behandeln und dürfen nicht länger als ein Jahr vor Ablauf der Ausschreibung fertig gestellt worden sein. Sie sollen außerdem erstmals und nicht bereits zu anderen Wettbewerben eingereicht worden sein. Die Arbeiten müssen auf eigenen wissenschaftlichen Untersuchungen beruhen, die in der Hauptsache in den Mitgliedsländern der MGG durchgeführt wurden.

Alle an der Durchführung der Untersuchungen beteiligten Mitarbeiter sind als Co-Autoren der Arbeit namentlich zu benennen und sollen ihr Einverständnis zur Teilnahme am Wettbewerb schriftlich erklären. Damit erkennen sie auch an, dass die Arbeit nicht von einem der Co-Autoren an anderer Stelle eingereicht wird. Ein wissenschaftliches Curriculum vitae sowie ein Votum informativum des Leiters der Einrichtung (Direktor/Chefarzt) sind den Bewerbungsunterlagen beizufügen.

Die Arbeiten sind in deutscher oder englischer Sprache zu **5 Exemplaren** bis zum **24. März 2017** an folgende Adresse einzureichen:

Prof. Dr. med. Joachim Glaser
Schriftführer der MGG
Vitalisklinik Bad Hersfeld GmbH
Am Weinberg 3
36251 Bad Hersfeld

Additional information on IBD genetics

In **Paper A**, we summarized the most important genetic studies conducted in the last decades of the IBD research. These studies were largely conducted in Caucasian populations. Since the publication of **Paper A**, two other large studies, also performed in the Caucasian population, have been published (Huang *et al.*[28] and de Lange *et al.*[29]). For the majority of the markers identified to be associated with IBD, only few have been "conclusively resolved to specific functional variants"[28]. Huang *et al.* therefore performed a fine mapping analysis of IBD associated SNV markers. Out of 94 inflammatory bowel disease loci, Huang *et al.* assigned 45 associations to a single causal variant with a certainty of greater than 50% (18 with a certainty of >95%). Most of these markers were enriched for protein-coding changes and tissue-specific epigenetic markers in immune cells and the gut mucosa. De Lange *et al.* analysed genetic association with IBD in 25,305 previously unpublished individuals combined by a meta-analysis with published summary statistics of other studies. They added 25 loci to the list of known IBD susceptibility loci. Three of these genes were integrin genes. Integrins are involved in T cell homing to the tissue and hence have an important function in the adaptive immune response.

To complement studies focusing on Caucasian populations described in **Paper A**, the largest GWAS analyses in non-Caucasian populations are listed in Table 2.

Table 2 – GWAS in non-Caucasian populations.
This table is as an extension to Table 1 in **Paper A.** Only the largest studies within each population are shown. Studies are listed by geographical location of populations and year.

Year	Study	Ancestry	Results (examples)
2017	Brant *et al.*[30]	African American	Largest analysis for IBD
2015	Juyal *et al.*[31]	North Indian	Largest study for CD
2016	Fuyuno *et al.*[32]	Japanese	Largest study for IBD
2016	Yang *et al.*[33]	Korean	Largest study for IBD
2014	Yang *et al.*[34]	Korean	Largest study for CD
2013	Yang *et al.*[35]	Korean	Largest study for UC

Prior to the studies listed in Table 2, other smaller scale studies or non-GWA studies (only targeting a limited number of SNPs) exist in these cohorts. They are not shown here. Apart from the African American population, the studies listed in Table 2 are also (partly) included in the large cross-ancestry GWAS published by Liu *et al.*[36].

In Figure 6 and Figure 7, which were taken from Liu *et al.*[36] and Brant *et al.*[30] and the Manhattan plots of the genome-wide association results obtained in the respective studies are shown. Most strikingly, a clear and consistent association of IBD with variants at the HLA locus can be observed in these plots.

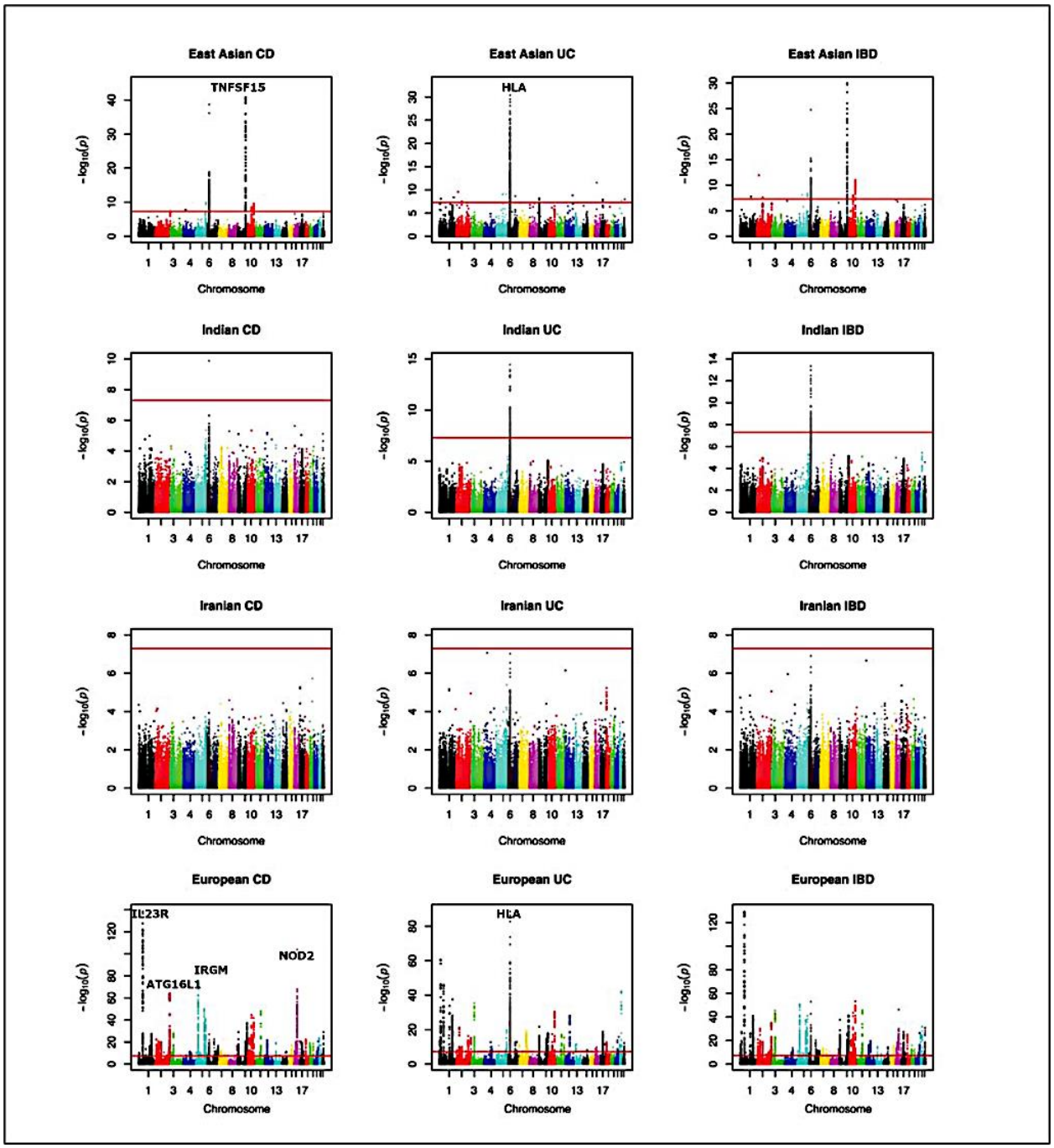


Figure 6 – Manhattan plot of IBD, UC and CD in different ethnicities.
Figure taken from the Supplement of Liu *et al.*[36]. Original and shortened description: Supplementary Figure 6. Manhattan plot for each separate ancestral cohort. The x-axis of each plot indicates the position of all tested SNPs per chromosome. The y-axis shows the strength of association (-log10 P-value). In rows from top to bottom: East Asian, Indian, Iranian and European. In columns from left to right: Crohn's disease (CD), ulcerative colitis (UC) and combined inflammatory bowel disease (IBD). Added description on: ATG16L1: Autophagy related 16 like 1, IL23: Interleukin receptor 23, NOD2: Nucleotide oligomerization domain, IRGM: Immunity-related GTPase family M protein, TNFSF15: TNF superfamily member 15.

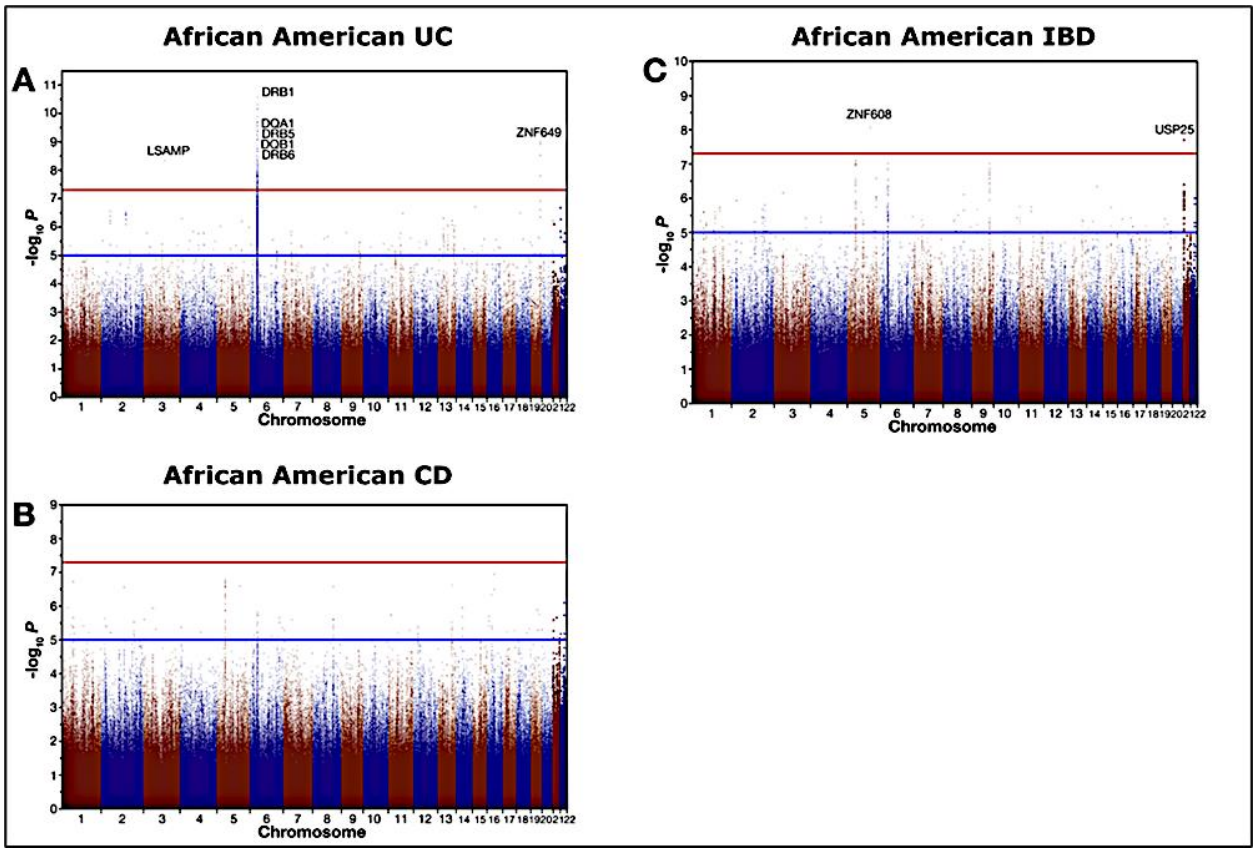


Figure 7 – Association analysis in African American IBD, CD and UC.
Figure taken from Brant *et al*.[30]. Original and shortened description: Figure 2. (A, B, C) Meta-analysis Manhattan plots for UC (A), CD (B), and IBD (C) phenotypes, respectively. All SNPs are plotted according to their position on each chromosome on x-axis, against their association on y-axis. The red and blue lines indicate the genome-wide significance ($P=5x10^{-8}$) and the suggestive significance threshold ($P=1x10^{-5}$), respectively. GWS signals are labeled with corresponding gene names.

Associated variants in the nucleotide oligomerization domain 2 (*NOD2*) gene, located on chromosome 16, are absent in the Asian population. The NOD2 protein is a pattern recognition receptor that can recognize components of the bacterial cell wall (peptidoglycans). In the Asian population, variants within the TNF superfamily member 15 (*TNFSF15*) gene, located on chromosome 9, are most strongly associated with CD. TNFSF15 is a cytokine of the TNF family and is involved in cell signalling. In the African American population, the strongest associated SNPs were reported to be located in the HLA region, with other associations hypothesized to having arisen from European admixture (*NOD2*; association by LD with another gene, the Sorting nexin 20 (*SNX20*) gene). Overall, Liu *et al*. [36] reported that genetic association of associated markers were widely consistent in their effect directions and magnitudes across different populations. They also showed that some heterogeneity (different allele frequencies and different effect sizes) does exist for IBD markers (e.g. variants in the Interleukin receptor 23 (*IL23R*), Autophagy related 16 like 1 (*ATG16L1*) and *NOD2* genes). A large part of the disease variability in European and East Asian UC can be explained by the HLA region (Chapter 4.2.3). With sample sizes varying across the datasets and UC having a stronger association signal for the

HLA than CD (hence variants within the HLA had a higher power (Chapter 6.4.1) to be detected in smaller studies), we chose to focus on UC in the context of this thesis. This decision was also made considering the complexity of the analyses presented in **Paper C**.

Genetic association with IBD subphenotypes

Genetic association with IBD subphenotypes has so far only been conducted in the Caucasian population[19]. Strongest genetic associations were observed with CD disease location (ileal vs. colonic) and extent of the disease for UC. Here, variants within the *NOD2* gene and the HLA region were reported to be associated with ileal and colonic CD disease location respectively. For colonic CD, the HLA allele HLA DRB1*01:03 specifically was associated with the disease. A less severe extent of UC was associated with a SNP located in HLA-B*08. In **Paper C**, we showed that alleles of this group are located on the same haplotype as the DRB1*03:01 allele. Associations with variants in *NOD2*, the HLA and Macrophage-stimulating protein 1 (*MTS1*) were also observed with age of onset of IBD.

With genetic research into IBD reaching its limits for the analysis of common variants (minor allele frequency (MAF) >1%; i.e. not many more common variants are likely to be discovered), still, only about 13.1% – 13.6% percent of the heritability (phenotypic variance explained by genetics) of CD and 7.5 – 8.2% of the heritability of UC can be explained[6,36]. Larger efforts like the analysis of whole **exomes** and **rare variants** measured on genotyping chips may explain some of the remaining "missing heritability". These efforts require even larger datasets (>100K) than included in already published studies to give valuable insights into association with rare variants. Exome analysis has also especially proven to be of diagnostic value in parent-child trios or family analyses in paediatric IBD cases, where *de novo* mutations are more commonly found[37–39].

4.1.5. Other research areas in IBD

Beside genetic predisposition, other factors such as environmental factors like a Western diet are attributed to contributing to the disease aetiology. With genetic associations including genes resulting in immune regulatory proteins of the innate *(NOD2*) and adaptive (HLA) immune system, the human microbiome seems to be a prominent target for the identification of disease related factors (i.e. antigens derived from bacteria and fungi inhabiting the gut may be recognized by the immune system as is the case for NOD2 and bacterial peptidoglycans). These antigens could stem from the microbiome itself (i.e. bacterial cell wall components) or microbiome-related dietary factors. Therefore, knowledge obtained from this active research area, which is currently also a "hot topic" for many other diseases, is described briefly below. Larger scale analysis of the microbiome is facilitated by the development and decreasing costs of high-throughput next generation sequencing based analysis of 16S bacterial and 18S fungal ribonucleic acid (RNA). IBD has been associated with a shift in the **microbial diversity**

(i.e. number of observed species) of the gut microbiome for both the bacterial and fungal microbiomes[40]. If this is cause or result of the disease is still unclear[41]. Overall, IBD is associated with a decreased bacterial diversity, especially in the most prominently present bacterial phyla (with phylogeny described from top to bottom across Kingdom – Domain – Life – Phylum – Order – Class – Family – Genus – Species) Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria (Figure 8).

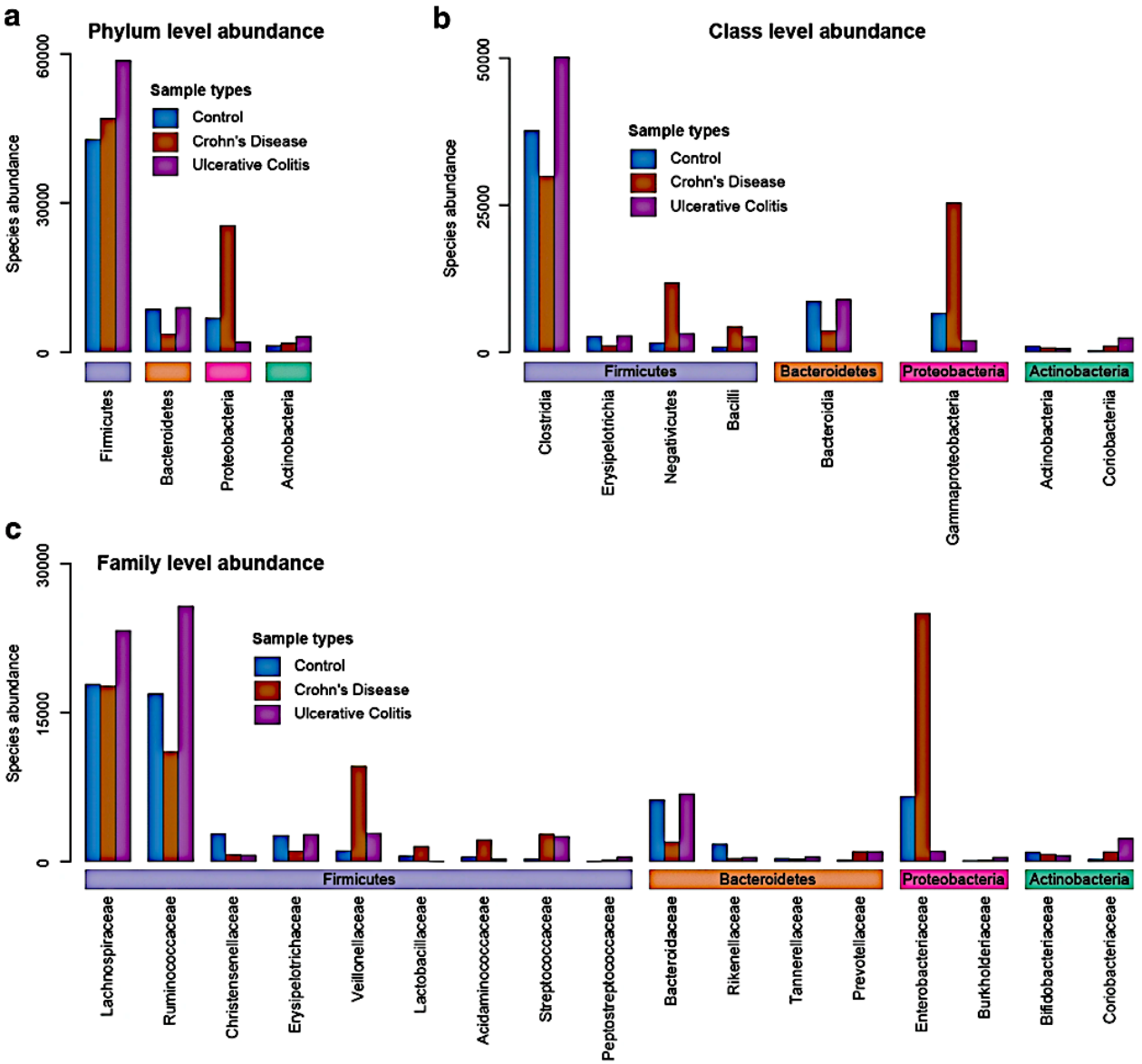


Figure 8 – Overview of gut microbiome composition across IBD.
In total 9 CD patients, 11 UC patients and 10 healthy volunteers were analysed. Figure taken from Alam *et al.*[42]. Original description: The gut microbial abundance. a Phylum, b Class and c Family level abundance in different conditions. Classes and families belonging to the four most abundant phylum across conditions are grouped according to phylum.

Together these phyla constitute >98% of the gut microbiome[42,43]. This decrease in microbial diversity is more significant in CD patients than in UC patients[42,43].

A decrease in the diversity is also observed in the fungal microbiome, with Ascomycota and Basidiomycota being most abundant (Figure 9). In CD especially, a shift in the Basidiomycota/Ascomycota ratio is observed along with an increase in *Candida albicans* and a decrease in Sac*charomyces cerevisiae* species[40].

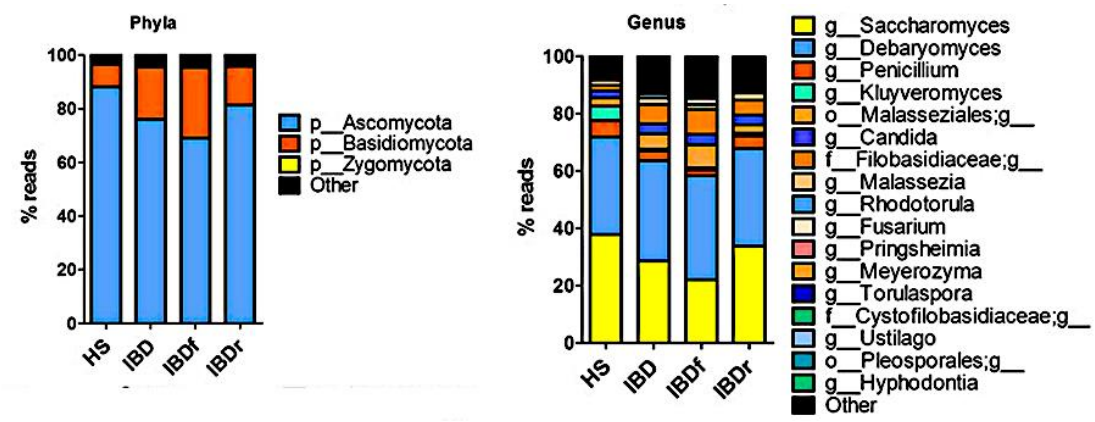


Figure 9 – Fungal microbiome in IBD.
Figure taken from Sokol *et al.*[40]. Original and shortened description: Figure 2 – Altered fungal microbiota biodiversity and composition in IBD. Global composition of fungal microbiota at the phyla and genus levels. Healthy subjects (HS) and patient subgroups are labelled on the x-axis and expressed as relative OTUs abundance for each group. OTU: Operational Taxonomical Unit, HS: Healthy subjects, IBDf: IBD flare, IBDr: IBD in remission.

Microbes in the gut are separated from the underlying tissue by intestinal epithelial cells (IECs) lining the gut lumen. These are constantly replenished from resident stem cells every 4-5 days. Thus IECs form a **physical and chemical barrier** against microbes[44,45]. To keep microbes in the gut at bay, the epithelial layer secretes mucus (Goblet cells) and antimicrobial peptides (Paneth cells). IECs are also involved in nutrient uptake from the gut. A current hypothesis states that upon disease, the mucosal barrier is disrupted due to a reduction in mucosal thickness. Consequently, increasing amounts of microbes are given access to the underlying host tissue, thereby causing an excessive immune response (Figure 10). Which processes lead to loss of the mucosal integrity, is not completely understood. Some genes suggested to play a role in epithelial cell function (i.e. regulators of epithelial cell homeostasis and physiology including mucin (*MUC*) and matrix metalloprotease (*MMP*) genes) have been implicated in genome wide association studies and are listed in detail in “Molecular Genetics of Inflammatory Bowel Disease”, written by IBD experts across the globe[2].

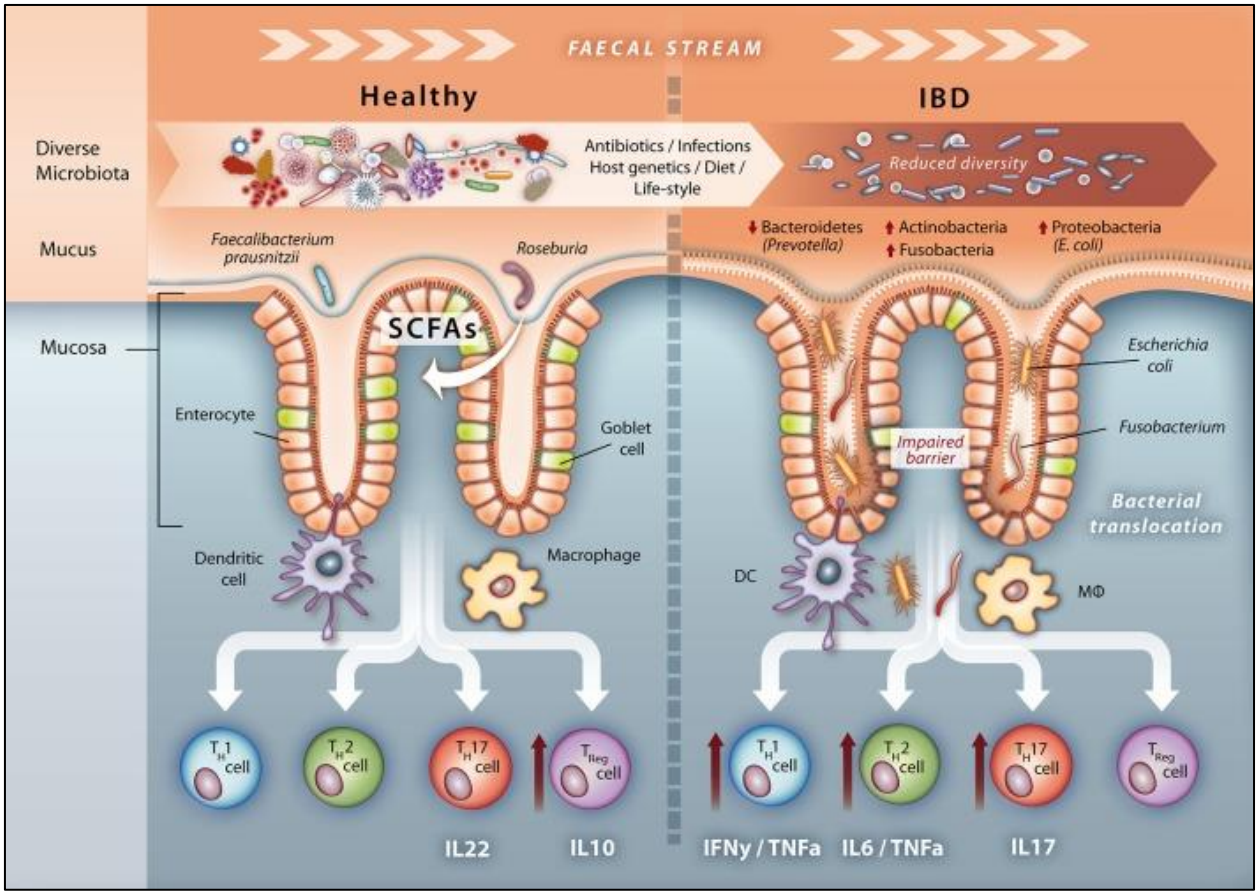


Figure 10 – IBD as a mucosal barrier defect.
Figure taken from Sommer *et al.*[43]. Original and shortened description: Figure 1 – Microbial signatures of a healthy gut and IBD. Under healthy homeostasis the microbiota is diverse. Goblet cells produce a thick colonic mucus layer, which creates a physical barrier against the microbiota, but also harbours a specific mucus-resident microbiota enriched in, for example, short-chain fatty acid producing bacteria *Roseburia* and *Faecalibacterium prausnitzii.* However, the composition of the microbiota is less diverse in patients with IBD with fewer Bacteroidetes mainly attributed to loss of *Prevotella* species and expansion of Actinobacteria, Proteobacteria such as adherent invasive *Escherichia coli* and *Fusobacteria*. Mucosal function is also altered, for example, lipid metabolism, illustrating a cometabolism of the metaorganism. DC, dendritic cells; IFNγ, interferon gamma; IL, interleukin; MΦ, macrophages; TNFα, tumor necrosis factor alpha.

A relatively young research area in IBD involves sequencing-based analysis of **T-cell** and **B-cell receptors** (BCRs and TCRs). Here, associations of TCRs and BCRs with IBD are investigated, to identify if there are specific receptors found in the BCR/TCR repertoire (sum of all receptors observed in an individual in the respective cell type) of IBD patients[46,47]. As I describe further below, T/B cells are activated after they have been presented with a specific HLA-peptide combination, which they recognize via the respective TCR/BCR receptors. Another important factor within this setting is the identification of culprit antigens (i.e. peptides that in the combination with specific HLA proteins and TCRs/BCRs elicit an immune response). Peptidomics analyses (i.e. analysis of peptides) are just emerging in the research field of IBD and are discussed further in the outlook (Chapter 6.5)

Overall, IBD is a complex genetic disease with many factors involved in the disease aetiology and disease progression. A lot of research spanning a wide spectrum of different scientific disciplines has been performed. A selection of active research areas, with a focus on the more bioinformatic-heavy research has been introduced above. However, in the past decades, little of the research has entered the clinical setting, where new developments in treatment (i.e. medication) are still the most dominant additions. This is in part because of variability of results seen across studies, especially for microbiome studies. Also, conclusions from current research regarding the disease aetiology are still challenging. A deeper understanding of mechanisms behind what is observed warrants further research and validation (especially functional validation of IBD susceptibility genes). However, microbiome and immune system related peptidomics research has the potential to identify possible culprit antigens in IBD.

4.2. The human leukocyte antigen region (HLA)

Information on the function of the HLA was mainly taken from “Janeway’s Immunobiology” by Kenneth Murphy and Casey Weaver[48].

4.2.1. HLA genes and their function

The HLA region (termed major histocompatibility complex (MHC) in general, and HLA in humans) is a genetic region that spans approximately 5 (Mega base pairs) Mb on chromosome *6p21* with genomic positions ranging from 29 to 34 Mb. Genes expressed from this region result in proteins that are involved in many complex functions of the adaptive and innate immune system. These include presentation of foreign peptides to the host immune system and factors of innate immunity like proteins of the complement system. The genes of the HLA can be classified into three different groups according to structure and function: the HLA class I, HLA class II and HLA class III genes (Figure 11). HLA class I and HLA class II genes code for proteins that are expressed at the cell surface or have regulatory functions, while HLA class III genes code for proteins of the complement system (e.g. complement component 2 and 4; *C2*, *C4*), cytokines (e.g. *TNF*) and heat shock proteins. HLA class I and HLA class II genes can be divided further into classical (HLA class I: HLA-*A*, -*B*, -*C*; HLA class II: HLA-*DRA1*, -*DRB1*, -*DQA1*, -*DQB1*, -*DPA1,- DPB1*) and non-classical genes (e.g. HLA class I: HLA-*E*, -*G*, -*F*; HLA class II: HLA-*DM*, -*DO*). While non-classical HLA class I genes have been shown to be expressed at the cell surface[49], non-classical HLA class II genes HLA-*DM* and -*DO* have regulatory functions in the process of HLA class II peptide loading[50]. The non-classical HLA class I genes are structurally related to the classical HLA class I genes and are to be expressed on specialized tissues (e.g. *HLA-G* on placenta and testis). HLA-*G* expression has previously also been linked to IBD[51,52].

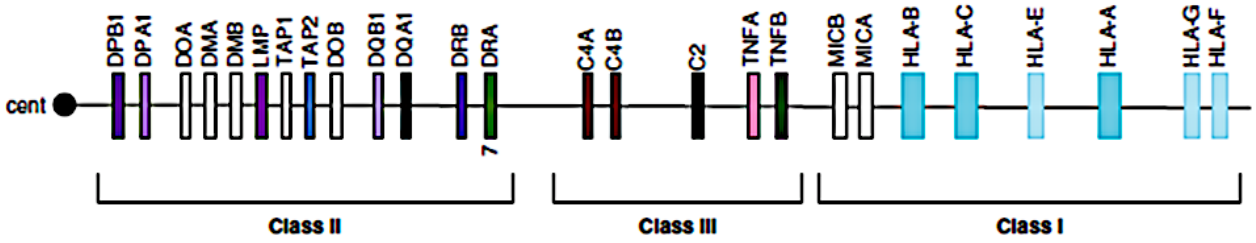


Figure 11 – Gene map of the human leukocyte antigen (HLA) region.
Figure taken and modified from Zhang *et al.*[53]. Modified description: Fig.1.1 Genetic map of the human MHC. Schematic map of human MHC on chromosome 6. Illustrations are not drawn to scale. The centromere (circle) and major HLA are indicated in order. Abbreviations not described in the main text: MICA A/B: MHC class I polypeptide-related sequence A/B, TAP1/2: Transporter associated with antigen processing, low molecular weight polypeptide (LMP), part of the human immunoproteasome.

The most investigated HLA class I genes HLA-*A*, *-B* and *-C* code for proteins that are expressed on all nucleated cells and present peptides to cytotoxic T cells that express co-receptor CD8 (also termed CD8 positive T cells, short $CD8^+$). Peptides presented by HLA class I proteins are produced by degradation of proteins expressed by or derived from intracellular pathogens like viruses in the cytosol. After degradation by the proteasome, a large multicatalytic protease complex present in the cytosol, antigenic peptides translocate to the endoplasmic reticulum (ER) where they are loaded onto the HLA class I complex and from there exported to the cell surface (Figure 12). HLA I class proteins are transmembrane proteins made of three α subunits (α_1, α_2, α_3) and a fourth subunit stemming from chromosome 15, the ß-microglobulin. The HLA-peptide complex is recognized by $CD8^+$ T cells, upon which an immune response is triggered that leads to the destruction of the affected cell by cytotoxic degradation by $CD8^+$ T cells. HLA class II proteins are expressed, with some exceptions (e.g. epithelium of the gut), on antigen presenting cells (APCs) like dendritic cells and macrophages to T cells that express co-receptor CD4 (also termed CD4 positive T cells, short $CD4^+$). HLA class II proteins are heterodimers composed of an α- and a β-chain that are encoded by *HLA-DRA*, *-DQA*, *-DPA* and *HLA-DRB*, *-DQB*, *-DPB*, respectively. While the α-chains of the DQ and the DP molecules are considerably polymorphic, there are only 29 known variations of the *HLA-DRA1* gene that are translated into 2 different DRA1 proteins (hla.alleles.org/nomenclature/stats.html on 05032020). Each α- and β-chain is made of two subdomains of which subdomains α_1 and β_1 come into contact with the peptide and the TCR receptor (Figure 12). Peptides presented by HLA class II molecules are produced from proteins derived from extracellular pathogens and taken up by macrophages and dendritic cells through phagocytes and micropinocytosis or by B cells by receptor-mediated endocytosis through surface immunoglobulins[50,54]. After their synthesis in the ER, HLA class II molecules subsequently enter the endocytic pathway by translocation to the Golgi apparatus.

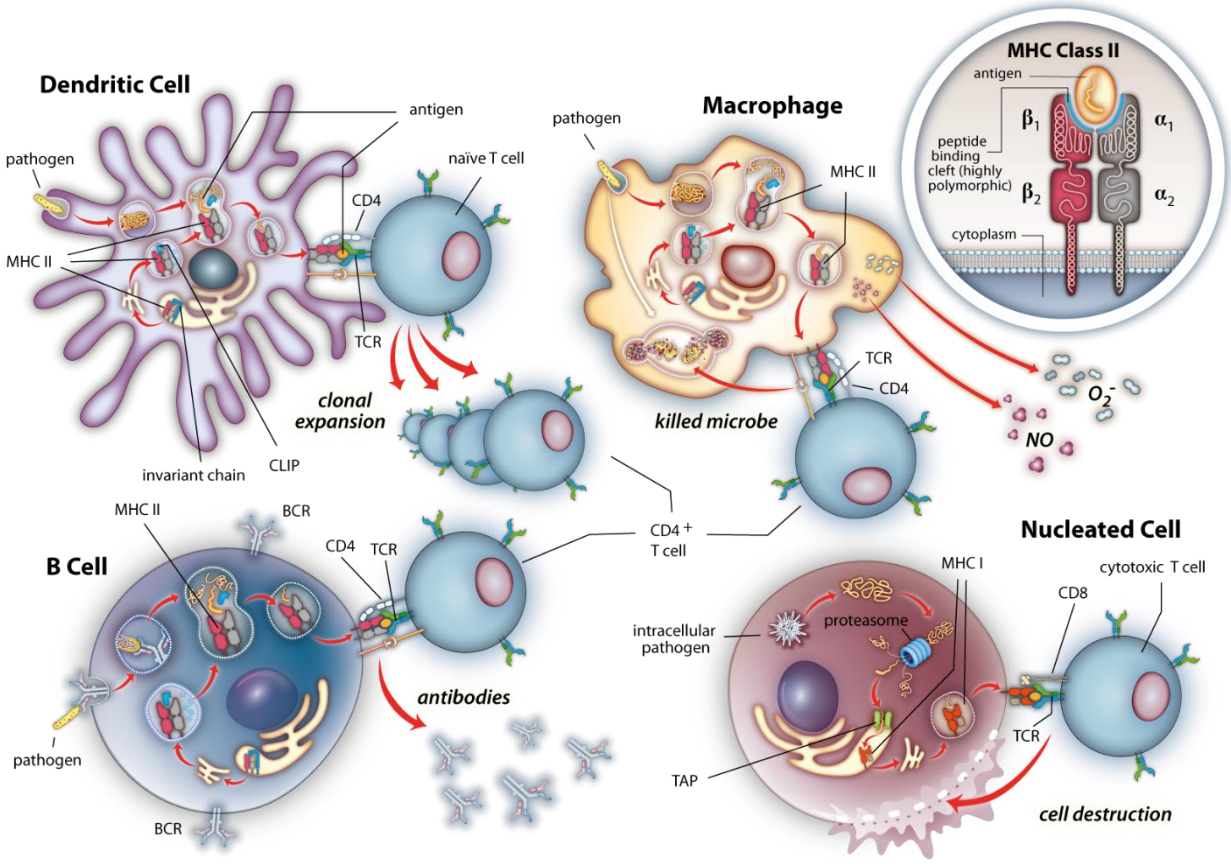


Figure 12 – Molecular mechanisms of antigen presentation by the HLA.
With special thanks to Mareike Wendorff and Renate Nikolaus, who jointly designed this figure.
BCR: B-cell receptor, CLIP: class II invariant chain peptide, NO: Nitric oxide, TAP: Transporter associated with antigen processing, TCR: T-cell receptor.

They are retained inside the endosomal vesicle by binding to the invariant protein chain. In an acidified endosome, the invariant chain is cleaved by acidic protease to produce a short peptide fragment, the class II invariant chain peptide (CLIP). Proteins that have entered the cell through endocytosis are delivered to endosomes in which they are degraded by proteases. Upon fusion of endosomes containing these peptides and the endosome containing the HLA class II molecules, the CLIP protein is exchanged for an antigenic peptide. This process is catalysed by the non-classical HLA molecule DM. DM is inhibited by the non-classical HLA class II molecule DO, that is also present in the acidified endosome. Constant interplay between these two molecules influences the selection of peptides bound to the HLA class II molecule. The HLA class II–peptide complex translocates to the surface where it is recognized by $CD4^+$ T cells and an immune response is mounted. This immune response leads to the destruction of pathogens and an increased humoral (involving antibodies) response by activation of B cells. Besides pathogenic peptides, both HLA class I and HLA class II molecules also present self-peptides to the immune system.

4.2.2. HLA nomenclature

The HLA region is a genetically highly polymorphic region with thousands of different SNP haplotypes existing for some of the known HLA genes. The genetic diversity of this region may allow distinct profiles of immune responses in individuals carrying distinct HLA polymorphisms. By March 2020, the Immunogenetics project (IMGT)/HLA database (release 3.39.0), a repository for all known HLA variation[55], listed a total of 19,031 HLA class I and 7,183 class II alleles. The numbers for the HLA loci present in this database are listed in Table 3.

Table 3 – Number HLA class I and HLA class II alleles in the IMGT/HLA database.
The number of different proteins that result from these alleles and the number of non-functional "null" proteins translated from the alleles for database version 3.39.0 are shown.

Gene	*A*	*B*	*C*	*DRA*	*DRB1*	*DQA1*	*DQB1*	*DPA1*	*DPB1*	*DRB3*	*DRB4*	*DRB5*
Alleles	5,907	7,126	5,709	29	2,690	229	1,795	168	1,537	340	165	123
Proteins	3,720	4,604	3,470	2	1,889	98	1,194	65	1,006	254	109	95
Nulls	308	244	243	0	89	6	77	3	80	16	19	17

The repository itself is available at https://www.ebi.ac.uk/ipd/imgt/hla/. In order to keep track of the different HLA alleles, a specific nomenclature was developed by the "World Health Organization (WHO) Nomenclature Committee for Factors of the HLA System". Genes are categorized according to four different criteria resulting in an "eight-digit" or "four-field" nomenclature system (Figure 13).

Figure 13 – Current HLA nomenclature.
HLA alleles are named in a 4-field notation. Each field is separated by the separator ":" and signify different properties of the HLA allele.

The first field names the allele group, which often corresponds to the serological antigen of an allele, determined by cross-reactivity of lymphocytic components in the blood. The next field shows a nonsynonymous nucleotide exchange within the exons of an HLA gene. The third field classifies synonymous nucleotide exchanges within the exons of an HLA gene and the last field shows changes in the intron sequence. For instance, HLA-A*02:01:01:01 and HLA-A*02:01:01:02 only differ in their intron sequence while HLA-A*02:01:01 and HLA-A*02:01:02 have the same amino acid sequence but different nucleotides within the coding regions. HLA-A*02:01 and HLA-A*02:02 differ in amino acid and

nucleotide sequence but belong to the same HLA-*A* group. A suffix added to the allele sequence may indicate differences in expression, i.e. the letter N denotes an allele that is not expressed, and Q denotes alleles whose expression is questionable.

The nomenclature of the HLA has changed and expanded since the discovery of the first antigens on human leukocytes by serological typing with the latest major update introduced in 2010. Responsibilities for introducing and updating HLA nomenclature lie with the WHO Nomenclature Committee, formed in 1967 during the second international Histocompatibility Workshop. Since then several HLA nomenclature reports have been published. To the largest part, the following information is taken from the webpage hla.alllees.org and the nomenclature reports listed in Table 4.

Table 4 – Timeline of the description of HLA loci in the HLA nomenclature report.

1968	Publication the first HLA nomenclature report
1975	First description of *HLA-A, -B, -C* and *-D*, HLA nomenclature report[56]
1977	First description of HLA-*DR*, HLA nomenclature report[57]
1984	First description of HLA-*DQ*, *-DP* and split of DR5 into DR11/DR12, split of DR6 into DR13/14, HLA nomenclature report[58]
1987	Introduction of eight-digit system and first description of HLA-*DRB3*, *-DRB4*. Split of DR2 into DR15/16, HLA nomenclature report[59]
1989	First description of HLA-*DRB5*, HLA nomenclature report[60]

First, human leukocyte antigens were grouped into a single serological antigen group based on cross-reaction of lymphocyte components in the blood. Without knowing that products of different genes were evaluated, the identified antigens were named HL-A and a sequential number, i.e. HL-A1 and HL-A2. These serological antigen groups were then split into HLA-*A* and HLA-*B* and later HLA-*C*, *-DR*, *-DQ* and *-DP* when evidence accumulated that these antigens must originate from different genes (Table 4). A two-field based system was introduced in the 1987 nomenclature report which enabled the distinction of HLA alleles on the sequence level i.e. B*52:01 and B*52:02 could now be distinguished, whereas before the serotype group B*52 accommodated both alleles. For most of the more common HLA alleles, HLA groups correlate with their serological specificities, i.e. DRB1*01:01, DRB1*01:02 and DRB1*01:03 all have the serological specificity DR1, whereas some alleles of the DRB1*03 are assigned serological specificities of DR17 and DR18. In the subsequent years more digits were added until a four-field based system, like the one described above, was established with each field allowing for two numbers only. This meant however, that only up to 99 alleles of the same group could be accommodated. Roll-over groups were defined when this number was reached. The first groups for which problems arose were the A*02 and B*15 groups. When the 100th allele for the A*02 group was identified, it was subsequently named A*92:01, similarly the 100th allele of the B*15 group was named B*95:01. When it became clear that this system was not able to deal with the fast-growing numbers of alleles being typed and maybe several hundred alleles could be identified for a single HLA gene group, the new system described above was introduced. First, separators were added to distinguish the HLA

group fields, i.e. A*01010101 became A*01:01:01:01. All alleles with roll-over groups were renamed to their original groups, i.e. A*92 and B*95 were renamed A*02 and B*15 respectively. For alleles for which more than 100 alleles had already been reached the number 100 would not be assigned, i.e. there would not be A*02:100 and B*15:100. Allele groups yet to reach this number would be assigned the number 100, i.e. A*24:99 and A*24:100 would exist. With this nomenclature change new allele groups were defined based on exon 1 and exon 2 identities for HLA class I molecules and exon 2 identities in HLA class II molecules, thus making custom group assignments, made by researchers within their own specific research projects obsolete and increasing comparability between studies (**Paper B**). Alleles with amino acid sequence identity in exons 1 and 2 (HLA class I) or exon 2 only were grouped into the same so-called (protein) “P-group” and alleles with nucleotide sequence identity across the respective exons were grouped in the so called “G-group”. In most cases the assigned P- or G-group of an allele corresponds to the most frequent allele in this group.

4.2.3. Fine mapping of the HLA in IBD

Even before the HLA was implicated with convincing statistical evidence as a susceptibility locus in IBD by means of linkage-analysis in large IBD families[61], it was investigated using smaller scale serological and genetic studies with some of these studies also focusing on the association of the HLA with IBD subphenotypes. Though reproducibility across the studies was limited, many of their results still hold true today. Comprehensive reviews of the findings are presented in Stokkers *et al.*[62] and Ahmad *et al.*[63]. Stokkers *et al.* performed a meta-analysis across 29 HLA associations studies related to IBD that were published between the years of 1966 to 1998. 15 studies included CD and 18 studies included UC, some included both. These studies were mainly conducted in Caucasian populations. For UC, an overall positive association of DR2, comprised of the DR15 and DR16 serotypes, pinpointed later to an association with DRB1*15:02, as well as DRB1*01:03 was identified. Negative association with UC was reported for DR4. For CD, a positive association was found for DR7, DRB3*03:01 and DQ4 and a negative association was found for DR3 and DR2. Ahmad *et al.* report on findings on the HLA until 2006 and review the major findings of subphenotype analysis. HLA DRB1*07 and DRB1*01:03 and DRB1*04 were implicated in the location of Crohn’s disease. While DRB1*07 and DRB1*04 were positively associated with ileal Crohn’s disease, DRB1*01:03 was associated with colonic CD. For UC extensive disease was associated with DRB1*01:03 and DRB1*04:01 was associated with protection against UC. Multiple studies conducted at this time in the Japanese, Korean and Caucasian individuals implicated DRB1*15:02, B*52:01 and C*12:02 in UC. Since then additional analyses have been conducted in the Chinese and Iranian population. UC in Iranians (from the city of Kerman) has been associated with HLA-DRB1*04. HLA-DRB1*13 was significantly associated with disease severity in the same study[18]. An analysis of Chinese Han and Uyghur UC patients found that HLA-DRB1*04/08/13/14 may contribute to the clinical heterogeneity of UC in these ethnicities. Han *et*

al. analysed amino acid polymorphisms in the Korean and Japanese population[64,65]. To this date, the largest fine mapping study within the Caucasian population based on the imputation of the HLA has been performed by Goyette *et al.*[1]. This study analysed 18,405 individuals with CD, 14,308 individuals with UC and 34,241 control individuals, typed on the Illumina ImmunoChip. They generated HLA types by HLA imputation using the imputation tools SNP2HLA[66] and HLA*IMP2[67]. This study demonstrated how HLA haplotype-like correlations could be observed from single association statistics using conditional reciprocal logistic regression and drew comprehensive maps of susceptibility alleles associated with IBD, for alleles for which such an assignment could be made (Figure 14, Figure 15).

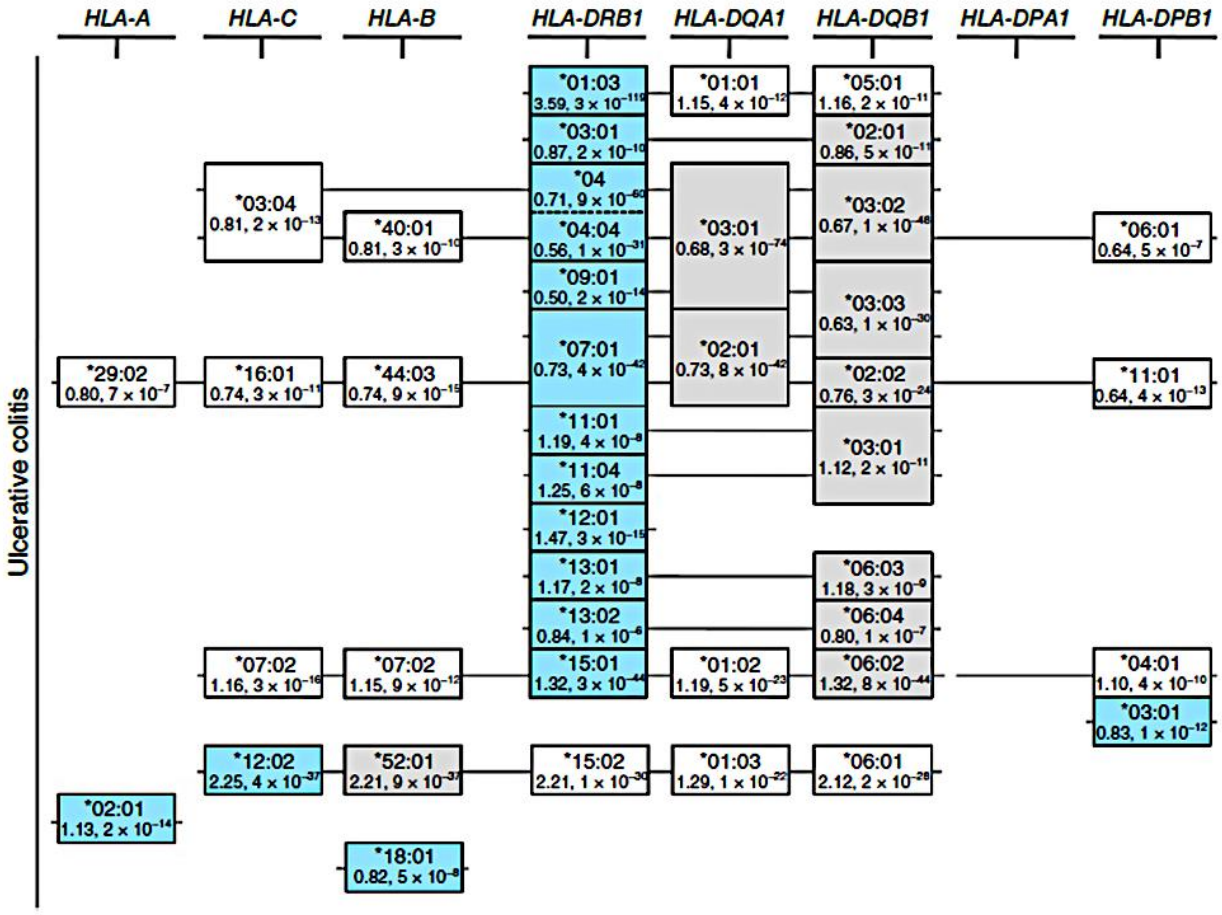


Figure 14 – Fine mapping study of the HLA in UC.
Figure taken from Goyette *et al.*[1]. Original and shortened description: The structures illustrated are not classically defined haplotype structures but were identified entirely on the basis of the correlation of signal defined through pairwise reciprocal conditional logistic regression analyses. Alleles identified as primary tags for independent association signals in our HLA-DRB1–focused models are shown in light-blue boxes, and alternate alleles with equivalent effects are shown in gray boxes. Alleles in white boxes show study-wide significant secondary effects that can be explained entirely by the selected HLA alleles.

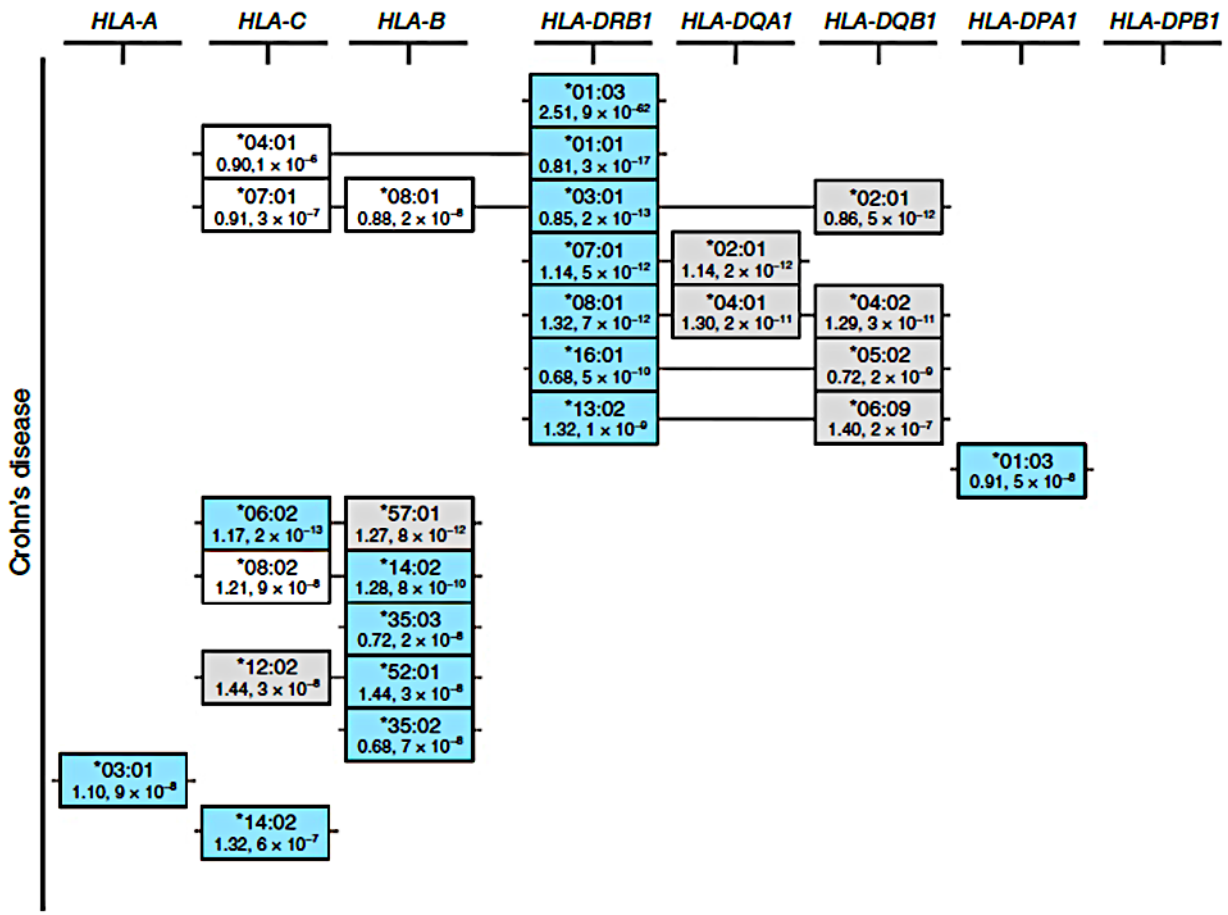


Figure 15 – Fine mapping study of the HLA in CD.
Figure taken from Goyette *et al.*[1]. Original and shortened description: The structures illustrated are not classically defined haplotype structures but were identified entirely on the basis of the correlation of signal defined through pairwise reciprocal conditional logistic regression analyses. Alleles identified as primary tags for independent association signals in our HLA-DRB1–focused models are shown in light-blue boxes, and alternate alleles with equivalent effects are shown in gray boxes. Alleles in white boxes show study-wide significant secondary effects that can be explained entirely by the selected HLA alleles.

Overall, this study not only confirmed DRB1*01:03 as one of the most important risk alleles across both CD and UC, but also demonstrated that HLA alleles are differentially associated with the two diseases. In this study, UC showed a stronger association with the HLA than UC. The authors also showed that a large part of the heritability, especially in UC, can be explained by HLA alleles as opposed to single SNPs within the HLA region. When considering HLA alleles as opposed to the most associated SNP (rs6927022, located in the HLA-*DQA1* gene locus), the variance explained in UC could be increased by a factor of 3 from 2.3% to 6.2%. In CD, the variance explained was increased by a factor of 10 (from 0.3% to 3.1%) when considering the HLA alleles over the top associated SNP rs9264942 (located in the HLA-*B* gene locus). The authors however stated that the measure used for this analysis could not be compared to traditional heritability estimates.

4.3. Methodological considerations

The study populations used in **Paper B** and **Paper C** where typed on the Illumina ImmunoChip, which is a custom SNP genotyping array that was specially designed for GWAS of immune-related diseases. It is described in more detail below. Additionally, the concepts of HLA typing and imputation (**Paper B** and **Paper C**) are introduced and genetic association analyses in the context of the HLA are discussed. This chapter ends with the description of useful resources that were also used to validate findings in **Paper C**.

4.3.1. Linkage Disequilibrium

Linkage disequilibrium (LD) is a term used to describe the non-random association of alleles at different genomic locations: As a very basic description, LD results from the fact that neighbouring genomic regions are inherited together (on a haplotype) across generations[68]. This concept is shown in Figure 16.

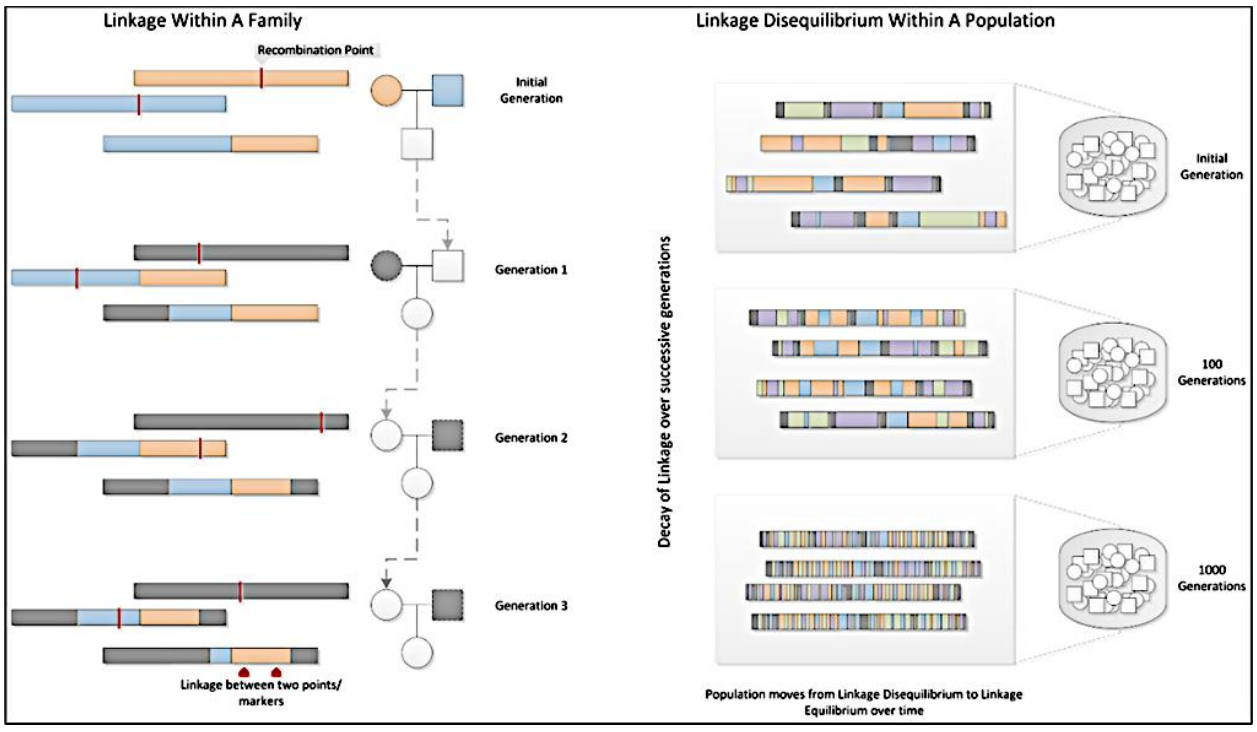


Figure 16 – Linkage disequilibrium explained.
Figure taken and modified from Bush and Moore[68]. Original description: Figure 2. Linkage and Linkage Disequilibrium. Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis, shown as red lines. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers or points on a chromosome in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome.

The term haplotype is used to describe the co-occurrence of alleles located on the same genomic strand. In turn, linkage equilibrium describes the situation in which alleles at different genomic locations are inherited completely independently of each other. Several other factors influence the patterns of linkage disequilibrium besides recombination events in gametes. These are described in detail in Ardlie *et al.*[69] and only shortly introduced here. LD values are influenced by recombination rates across the genome – i.e. recombination hot spots exist, where recombination events occur more frequently. Factors influencing the LD at the population level include inbreeding within a population (generally higher values of LD are observed here) and migration/admixture between populations of different ancestries (lower LD values observed in the first generations). Another factor at the population level includes genetic drift, which refers to the fact, that only part of the genetic material of a generation is given to the next generation, and processes of natural selection in which of whole haplotype stretches or single markers may be kept in or deleted from the pool of genomes. Finally, mutation events and gene conversion events (also termed homologous recombination, in which gene segments are replaced from homologous sequences of the opposite chromosome[70]) may influence the LD. Several measures to calculate the LD exist[68]. The most frequently used measure of LD in genetic association studies is the r^2-measure. To calculate r^2, the observed haplotype frequency of two alleles at different genetic positions/loci is compared to the expected haplotype of these alleles under the assumption of independence as:

Equation 1

$$\mathrm{r}^2 = \frac{(\mathrm{p_{AB}} - \mathrm{p_A} * \mathrm{p_B})^2}{\mathrm{p_A}(1-\mathrm{p_A}) * \mathrm{p_B}(1-\mathrm{p_B})}$$

Here, p_{AB} denotes the frequency of the haplotype with allele A at the first marker and allele B at the second marker. $p_A * p_B$ denotes the haplotype frequency of the same haplotype alleles at markers A and B under the assumption of statistical independence, which is calculated as the product of the frequencies of allele A (p_A) and allele B (p_B). The denominator $\sqrt{p_A(1-p_A) * p_B(1-p_B)}$ is used to standardize the measure. The r^2 can range between 0 and 1. An $r^2 = 1$ describes a situation of complete LD (i.e. markers are highly correlated and typically located closely within a genome) and $r^2 = 0$ describes incomplete LD (i.e. markers are not correlated and typically located far away from each other within a genome)[68]. The confidence of LD estimates can be calculated by bootstrapping of the study population and determination of the 2.5% and 97.5% percentiles of the resulting distribution.

When the haplotype frequency of two alleles is unknown, i.e. data were not phased before calculation of LD, it can be approximated by appropriate methods as described for instance by Gaunt *et al.*[71]. For the analysis of multiallelic markers, classical LD measures are inaccurate and other measures, that take the contribution of more than two alleles to observed SNP haplotypes into account, need to be considered[72]. This was observed especially during the computation of classical LD measures within

the HLA for the preparation of **Paper C**, since one HLA locus can have many different alleles. Here, I calculated the percentage an HLA allele of interest occurred with another allele at a different locus. This accounted for the fact that infrequent alleles may be located on the same haplotype as very frequent alleles with overall small LD, due to the high number of diverse haplotypes of the frequent allele. As a result of this thesis this method and the production of haplotype maps using this method, is described further in Chapter 5.

4.3.2. Genotyping and the Illumina ImmunoChip

Genotyping of single nucleotide variants (SNVs; minor allele frequency (MAF) <1%) and single nucleotide polymorphisms (SNP; MAF >=1%) was carried out in this study using a genotyping array produced by the company Illumina – the ImmunoChip. Genotyping arrays are designed for the high-throughput analysis of SNPs, using fluorescently labelled oligonucleotide sequence specific probes. SNP genotypes are inferred from the intensity of the fluorescence emitted when probes (each carrying the sequence of the respective SNP allele) labelled with a red or green fluorophore bind complementarily to the DNA. To infer the SNP genotypes computationally, calling algorithms (i.e. Opticall[73], GenomeStudio automatic clustering: software provided by Illumina) that are based on clustering of the measured SNP intensities (Figure 17).

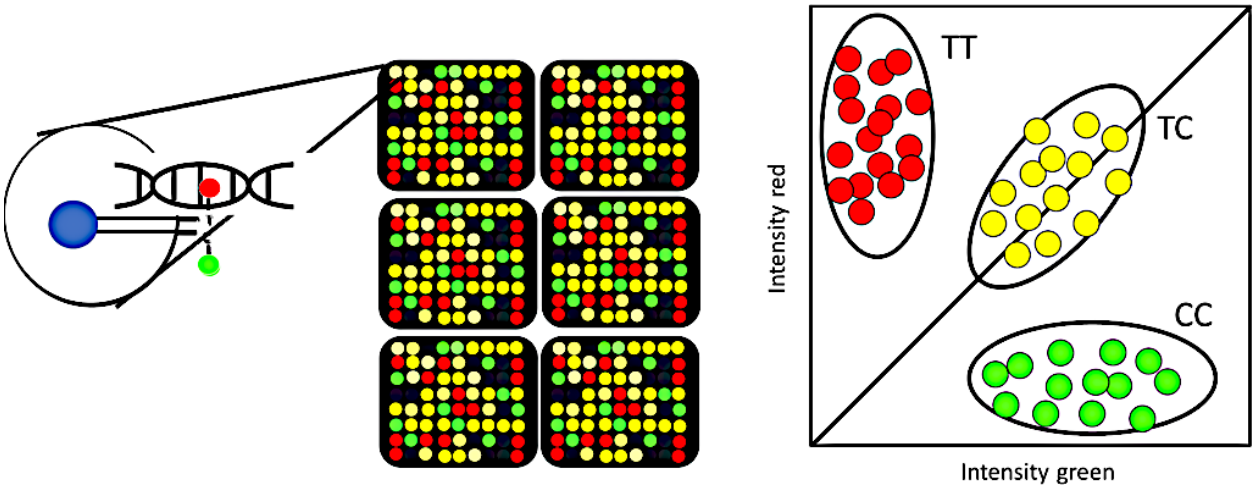


Figure 17 – Experimental workflow for SNP genotyping.
For each individual, genotypes are measured on a genotyping array using fluorescently labelled probes that are complementary to stretches of the genomic DNA. Upon binding, fluorescence is emitted and registered. Genotypes are called using clustering algorithms across all individuals measured on the array based on the emitted light intensities.

The design of SNP genotyping arrays was aided by large genome sequencing efforts including the Human Genome Project in which polymorphic nucleotides in the human genome were identified, as well as the Human Haplotype Project (HapMap), in which the LD structure was used to establish SNP

haplotype blocks. Harnessing LD structure, the variation of the genome could now be explained by single “tag-SNPs”, reducing the number of SNPs that needed to be typed in order to analyse a large portion of the genome.

The Illumina ImmunoChip is a custom genotyping array specially designed for the analysis of immune-related diseases. It was used in two different versions in this study – the HumanImmuno v1 Beadchip and the Illumina ImmunoArray-24 v2. The larger part of the data was genotyped on HumanImmuno version 1 Beadchip, while the Maltese dataset in **Paper C** was genotyped on the Illumina ImmunoArray-24 version 2. This is because at the time the Maltese dataset was measured, the version 1 chip had gone out of production. The HumanImmuno version 1 Beadchip covers 195,524 variants of which 718 are small insertion/deletion polymorphisms and 194,806 are SNPs. Insertion/deletion polymorphisms – short InDels – refer to deletions or insertions of nucleotide sequences rather than variation of single nucleotides. It contents are discussed in great detail in the doctoral thesis of Jostins[74] and described also by Cortes *et al.*[75]. The ImmunoChip was initiated by the Wellcome Trust Case-Control Consortium and designed by the Illumina ImmunoChip consortium consisting of leading investigators in the research fields of immune-mediated diseases with the goal of deep replication of established susceptibility loci for known immune-mediated diseases and fine mapping thereof[74,75]. In total 50,000 SNPs from 186 loci implicated in genome wide association studies (GWAS) of 17 major autoimmune diseases including, type 1 diabetes, celiac disease, inflammatory bowel disease, multiple sclerosis, ankylosing spondylitis, rheumatoid arthritis, vitiligo, and systemic lupus erythematosus were included. In addition to these 186 loci, the HumanImmuno version 1 Beadchip contains 100 SNPs from the Killer-cell immunoglobulin-like receptor (KIR) region, 6378 SNPs across the HLA and 848 SNPs from the National Human Genome Research Institute (NHGRI) GWAS catalogue. Additionally, SNPs from gene-candidate and rare SNPs from sequencing projects were added.[74]

The ImmunoArray-24 version 2 Beadchip was designed by Illumina with new content “produced by the research community” based on unspecified research publications. It contains 86,234 new SNP markers and approximately 164,000 markers retained from the HumanImmuno version 1 Beadchip. It contains additional loci from the NHGRI GWAS catalogue and published ancestry markers. In total, again 17 different immune-related are represented on this chip. More information on the chip could not be retrieved even upon request to Illumina. In this study, pre-called data were used for the larger part of the data (data from Liu *et al.*[36]), i.e. data for which the genotypes had already been inferred using the genotype caller Opticall. For newly added data, the idat-files (containing intensities) were received. The dataset from Malta was measured in house. Intensities for the newly added data were extracted using the Immuno_Beadchip_111419691_B manifest file and called using Opticall.

4.3.3. Concepts behind SNP association analyses

Association analyses of SNPs, SNVs and HLA variants are used for the study of complex, non-Mendelian diseases. Here the association of genomic variants with a trait – typically the binary trait case-control status – is analysed with single tests for each genomic variant measured in a study. Metric traits (e.g. blood measurement) can also be analysed, typically in a control population. This analysis is however less commonly conducted. When SNV variants within the HLA and HLA alleles are analysed, this study is referred to as an HLA fine mapping study in the literature. Association analysis of SNPs/SNVs across the whole genome are called genome-wide association studies (GWAS). In **Paper C**, we performed an HLA fine mapping study, analysing SNP and HLA allele variation within the HLA region only. The concepts needed to understand the outcome of such an analysis are explained below. In case-control type association studies, the association between the binary disease outcome and genomic variant is analysed by statistically assessing whether a certain allele occurs significantly more often in one group than in the other.

Minor and major allele

Association analysis are typically conducted on the SNP, SNV or HLA variant that is less common within a population – referred to as the minor allele. The allele frequency of this allele is calculated as the minor allele frequency – short MAF. The major allele refers to the more frequent allele in the study. In **Paper C**, the association of HLA alleles was also conducted on the major allele in some instances (explained below). Their frequency is therefore referred to by AF.

Coding of biallelic SNVs and InDels

SNVs are described by the observed nucleotides (A, C, T, G). InDels are either described as I or D (for insertion or deletion) or the nucleotide sequences of the InDel (e.g. A, ATTTTC). In the most used multiplicative model, that assumes that logarithmic odds (see below) of the disease are linearly dependent, i.e. additive on a log scale, genotypes are coded as 0,1 or 2 based on the minor allele. Let A be the minor allele and B be the major allele, then the genotype BB is coded as 0, the genotype AB is coded as 1 and the genotype AA is coded as 2. SNVs and InDels can also be multiallelic (i.e. have more than 2 alleles), in which case each variation is considered separately (i.e. the most frequent allele against the others). Other genotype models, which describe a recessive or dominant effect of an allele, are coded differently and described for instance by Clarke *et al*.[76].

Coding of multiallelic HLA alleles

HLA alleles are per definition multiallelic. In an HLA fine mapping study, each HLA variation is considered separately and coded as either absent (A) or present (P). The minor allele for most HLA alleles is its presence “P”. For some alleles at less diverse genetic loci (HLA-*DPA1*, HLA-*DRB3/4/5*)

the minor allele is its absence “A”. To maintain a consistent coding of these HLA alleles compared to the alleles for which P was observed as the minor allele, statistics were calculated on their major allele “P”. HLA genotypes were coded as 0=AA, 1=PA, 2= PP.

Confounders

In a SNP/SNV/HLA association analysis, confounders are variables that, besides the genotype, also have an influence on the trait of interest These include for instance the population diversity within a study measured in the form of principal components (PCs) or other variables that have clinically been shown to have an influence on a disease (i.e. age, gender, body mass index (BMI)). PCs are calculated as a step in a typical GWAS quality control[77] to detect outliers (e.g. outliers based on ethnicity, individuals that do not map to the population of interest). They are included in a GWAS to control for *population stratification* (i.e. differences in allele frequency merely based on differences in ancestry), even after exclusion of outliers. Typically, 5-10 PCs are included in a GWAS and chosen based on which PCs reduce the so-called genomic inflation factor. The genomic inflation factor is an indicator for population stratification and is calculated based on selected variants using the χ^2- distribution.

Contingency tables & the odds ratio (OR)

Statistical tests that do not consider cofounders for the analysis of binary traits include the χ^2 – and fisher exact test (the latter is especially used for the analysis of low-frequency variants). These tests compute statistics on so called contingency tables. In a GWAS context, contingency tables, summarize the number of individuals that are affected or unaffected by a disease and carry/or don’t carry a specific nucleotide or HLA variant as described in Table 5.

Table 5 – Example table to explain the concept of the odds ratio (OR).
Allele A is the less frequent allele that calculations are based on. The table shows observed frequencies of each allele in the two groups in a hypothetical sample.

	variant A (minor allele)	**variant B (major allele)**	
affected	5 (a)	10 (b)	15
unaffected	7 (c)	10 (d)	17
	12	20	Σ =32

The P-value of association is then calculated based on these tables as described in more detail elsewhere[78]. Since we conducted association analysis *with* confounders in **Paper C**, the concept of contingency tables is only used to introduce the concept of the odds ratio in the following. The effect of an allele (whether it is a protective or risk allele, and the extent of the effect) is calculated as an odds ratio (OR), usually reported in combination with the 95% confidence interval (CI) calculated from the standard error (SE) of the OR and the OR.

Equation 2

$$\mathbf{OR} = \frac{\mathbf{a}}{\mathbf{b}} * \frac{\mathbf{c}}{\mathbf{d}}, \qquad \mathbf{SE} = \sqrt{\frac{\mathbf{1}}{\mathbf{a}} + \frac{\mathbf{1}}{\mathbf{b}} + \frac{\mathbf{1}}{\mathbf{c}} + \frac{\mathbf{1}}{\mathbf{d}}}, \qquad \mathbf{95\%CI} = \mathbf{exp}(\mathbf{log}(\mathbf{OR}) \pm \mathbf{1.96} * \mathbf{SE})$$

In the context of association analysis carried out for a SNP or HLA alleles, the OR is the quotient of the odds of getting a disease when carrying a specific allele, and the odds of getting the disease when not carrying this allele. An OR equal to 1 indicates that the allele of reference (i.e. minor allele in most cases) does not have an effect, an OR<1 indicates that an allele is protective for a disease and an OR>1 indicates, that the allele of reference confers risk. The question of "how big is a big odds ratio" is addressed by Chen *et al.*[79]. At a disease frequency of 1%, they state that odds ratios indicative of small, medium and large effects should have values of 1.68, 3.47 and 6.71 (this is based on the more easily interpretable Cohen's d (unrelated to d in Table 5) – the standardized effect measure in a classical t-test, in which d = 0.2 ~ OR=1.68, ad=0.5~OR=3.47 and d=0.6~6.71)[79]. In GWAS of complex diseases, the ORs are typically small with OR<1.5. These allele-based tests assume equilibrium according to Hardy-Weinberg.

Logistic regression analysis

Regression analyses are used to analyse a trait under the assumption that there are additional confounders that influence the trait. In a case-control study, logistic regression is typically used. The general logistic regression (also termed Logit-model) estimates

Equation 3

$$\ln\left(\frac{\mathrm{P}(\mathrm{Y_i}=1)}{\mathrm{P}(\mathrm{Y_i}=0)}\right) = \mathrm{x_{i1}}\beta_1 + \mathrm{x_{i2}}\beta_2 + \cdots .. + \mathrm{x_{in}}\beta_\mathrm{n} + \epsilon$$

for i=1,….,N individuals, with error ϵ, regression coefficients vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ and independent variables (i.e. genotype + confounders) $\mathrm{x_i}$. In logistic regression the effect estimates cannot be calculated analytically, and iterative algorithms are used. For better interpretation of the effects estimate, it is transformed into the odds ratio.

Equation 4

$$\mathrm{OR} = \exp(\beta)\,, 95\%\mathrm{CI} = \exp\left(\beta \pm 1.96 * \mathrm{SE}\right)$$

The P-value of association is calculated from the beta coefficient of the SNP using appropriate statistics like the Wald-statistic.

Multiple testing problem

In a typical GWAS several hundred thousand to more than a million markers are analysed in parallel. In general, the errors made in statistical inferences increase with the number of tests performed (false positives) if the nominal significance level of 5% is used. This needs to be considered in the interpretation of the association P-values. Several different methods to adjust for multiple testing exist. In the context of GWAS, the correction by the Bonferroni method is used. Here the P-value of an association is simply multiplied by the number of tests. In a typical GWAS, two thresholds are used to define significantly associated variants, instead of a direct correction for multiplicity. These thresholds are static thresholds at which the number of possible tests needed to be performed in a GWAS was estimated at 10^6. Specifically, the number of common independent variants with a MAF > 5% was estimated to be 150 per 500 kilo base pairs by the HapMap consortium. With an extrapolation to 3.3 giga base pairs this amount corresponds to about 10^6 independent variants[80–82]. The P-value cutoff for significance was then adjusted according to Bonferroni-Holm as the quotient of 0.05 and 10^6, which is $5x10^{-8}$. Nominally significantly associated P-values are smaller than 10^{-5}. In the recent past and considering the growing number of variants that can be tested in a GWAS, whole exome analyses (WES) or whole genome analyses (WGS), these thresholds have been reanalysed by various groups. They generally confirmed the threshold of $5x10^{-8}$ for variants with a frequency of >5% and suggest more stringent thresholds for the analysis of lower frequency variants in the European population (based on LD analyses)[80–82]. In **Paper C**, we focused more on the consistency of effects measures regarding effects direction (risk or protective) and magnitude, simply because the sample size of some the analysed cohorts – and hence the power (discussed in Chapter 6.4.1) to detect an effect – is small.

Metaanalysis concepts

Information in this paragraph is mainly based on Borenstein *et al.*, Morris *et al.* and Lee *et al.*[83–85].

Results from multiple association studies can be combined using meta-analysis approaches, which increase statistical power for the detection of association but also allow for the analysis of the robustness of an effect (i.e. are effect sizes consistent across multiple studies and is the variation of an effect random or not)[83]. Meta-analysis methods combine effect sizes – and their variance – across different studies under model dependent assumptions. In the analysis of binary traits, the meta-analysis is conducted on the OR and its standard error. The standard error is a measure of the precision of the estimated effect primarily driven by the number of individuals that were used (i.e. it is larger for small studies) and study design (i.e. matched case-control studies, in which clinically relevant outcomes other than the outcome of interest are used to pair a case with a control, usually result in higher precision). Standard meta-analysis approaches differentiate between fixed effects and random effects.

Primer on fixed-effects meta-analysis

In a fixed effects meta-analysis, it is assumed that all factors that influence the estimate are the same in each analysed study and the true effect size is the same for each study. The true effect refers to the effect that one would measure in a population of infinite size (rather than a subsample of this population). The estimate of an effect is assumed to differ from the true effect only by a random sampling error. This error may be larger in small populations. In the fixed-effects meta-analysis the number of samples is therefore used to weight the effect estimates derived from the different studies.

Primer on random-effects meta-analysis

The random effects meta-analysis assumes that factors that influence the estimate are different between each study (e.g. patients in one study are leaner or younger than in the other). Therefore, in random effect studies the between-study variation is considered additionally to the random sampling error.[83]

Primer on trans-ethnic meta-analysis

A trans-ancestry meta-analyses, as apparent from the name, is a meta-analysis across studies of diverse ancestries. Neither fixed-effects models – that cannot adequately account for heterogeneity across studies, nor random effects models – that cannot account for similarity of allele frequencies in closely related studies while simultaneously considering differences in allele frequencies in more divergent populations, are applicable. In **Paper C**, we used a method called RE2C to analyse heterogeneity of effects across the populations of interest. This method was developed by Lee *et al.* and extends the random effects model RE proposed by Han and Eskin[85,86]. Using simulated and real data, Han and Eskin showed, that the classical random effects model is underpowered to detect an effect when large inter-study heterogeneity is observed. They proposed the new random effects model RE. RE is based on a likelihood-ratio test, while assuming no heterogeneity under the null hypothesis. In traditional random effects models the assumption is that heterogeneity exists under the null hypothesis. RE2C can also cope with correlated statistics from studies with overlapping samples. The latter is extremely helpful in meta-analyses today, with samples – most often unknowingly – being shared across biobanks and genotyping initiatives such as *23AndMe* and *UK Biobank*. In our study, no samples overlapped, such that we used RE2C only for the analysis of heterogeneous effect sizes across the different analysed populations. In RE2C, a P-value of associations across the studies is produced, while no common effects estimate is computed. RE2C also introduces a new method to analyse genetic variation across studies where heterogeneity of effect sizes is present. To facilitate the interpretation of results across different ethnicities and HLA alleles at the first glance, I developed a special meta-analysis plot (Figure 2 of **Paper C**).

4.3.4. Methods for HLA typing

In the mid-20th century and from thereon for a long period of time, typing of the HLA was mainly performed by serotyping using amongst others micro toxicity assays. In these assays the serum of an individual is incubated with different antibodies against known HLA antigens and factors of the complement system contained in rabbit serum are added. Upon antibody binding, cells are lysed by activation of the complement system and a fluorochrome dye can enter the cell. This dye is subsequently detected by microscopy using a phase-contrast microscope[87]. Individuals are assigned the HLA type for which HLA antigens stimulated a response in the analysed serum. This evolved to sequence-based typing (SBT) by the sequence specific oligonucleotide probe (SSOP) and sequence specific primer (SSP) as well as Sanger sequencing of exons 2 and 3 (HLA class I) and exon 2 (HLA class II), which are the most variable sections of the HLA region[87,88]. The SSOP method is based on fluorescently labelled probes that complementarily bind to the polymorphic sequence of the HLA. To type an HLA allele, sequence specific oligonucleotide probes complimentary to different HLA sequences are added to immobilized Polymerase Chain Reaction (PCR)-amplified locus-specific DNA. A light signal is emitted upon probe binding that is detected by chemiluminescence[87]. SSP is based on sequence specific primers that only bind to a target HLA sequence. This sequence is then amplified and can be detected as a fragment on an agarose gel[87]. Depending on the number of probes or primer pairs used, these methods can distinguish HLA alleles with varying resolution. Typically, they follow an iterative protocol in which the allelic group is determined first, followed by a more accurate typing of allelic diversity within this group[87,88]. With decreasing costs and more accurate sequence alignment protocols, NGS-based HLA typing has become an increasingly popular method for the typing of HLA alleles. In this thesis a protocol developed by Wittig, Juzenas *et al.*[88] was used to type the individuals that were used for the compilation of our high-resolution, manually curated HLA imputation reference (**Paper B**) and is described in more detail in “HLA Typing – Methods and Protocols” published by Springer[89] and also shown in Figure 18. Using nucleotide sequences of HLA alleles documented in the IMGT/HLA database, Wittig *et al.* designed 16,351 biotinylated RNA baits covering a length of 215.5 kilo base pairs (kb) for target enrichment. Here, sequences derived from the HLA region are targeted from fragmented whole genome DNA and extracted. To determine an individual’s HLA profile using this method, their DNA is fragmented, and a standard DNA library is prepared. Biotinylated RNA baits that are complimentary to HLA variations from the IMGT/HLA database bind to the fragmented DNA, which is captured using streptavidin coated magnetic beads that attach themselves to the biotinylated RNA baits. The captured fragments are then amplified using PCR and sequenced on a second-generation sequencer resulting in paired-end sequences of length 100-150 bp.

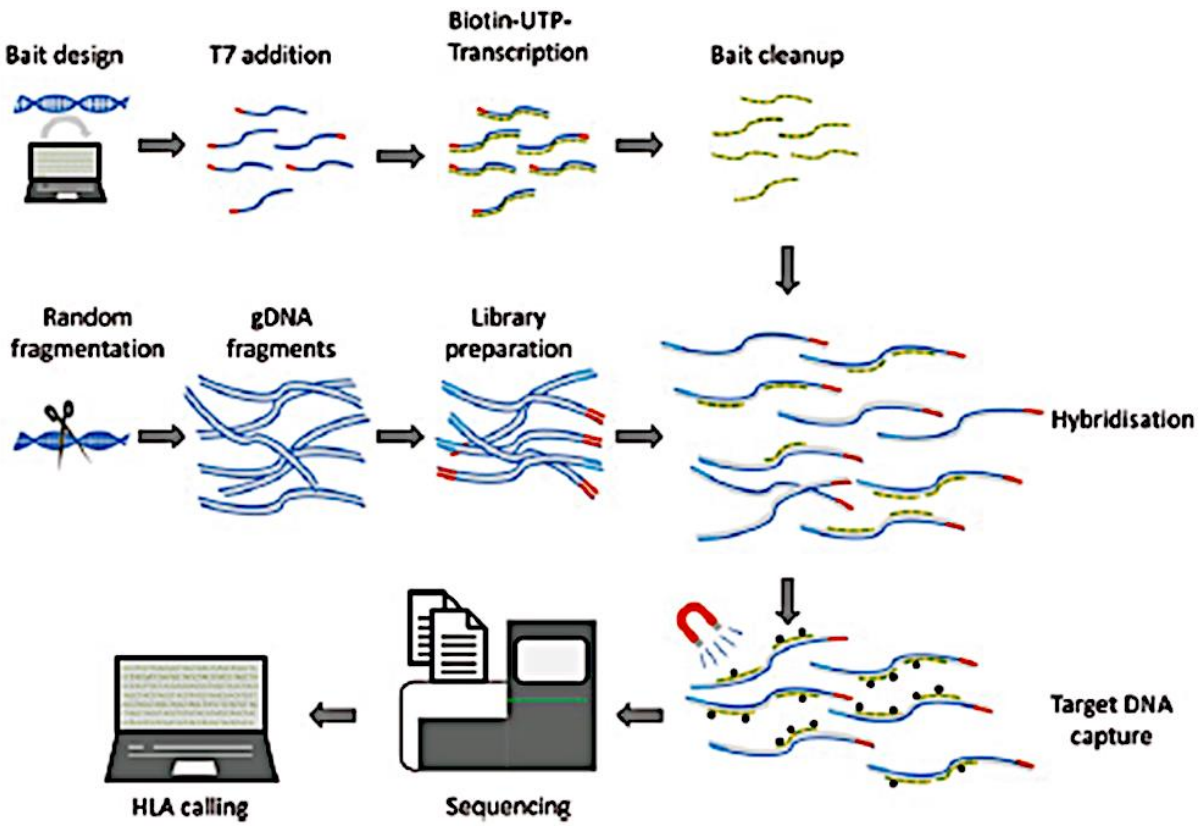


Figure 18 – RNA typing protocol used for the analysis in this study
Figure was taken from "HLA Typing – Methods and Protocols" published by Springer[89]. Original and shortened description: Fig. 1 Workflow. The workflow starts with two independent parts, the bait synthesis and the library preparation. For the bait synthesis a tiling of the target sequence is performed, which means all target sequence is fragmented into 120 bp parts. This can be performed in overlapping or nonoverlapping manner. For the HLA design nonoverlapping tiling was performed as the template sequence was a collection of all known HLA allele sequences, which is already a very redundant template. For the gDNA perform random fragmentation followed by a library preparation of the fragmented DNA. The library preparation adds sequences to the fragment flanks, which are needed during the sequencing reaction. These sequences consist of primer-binding sites, inserts, and indices (molecular barcodes) and the so manipulated gDNA is ready for sequencing. Perform a hybridization of the RNA baits and ready-to-sequence DNA to bind the target DNA to the biotinylated RNA baits. Using Streptavidin coated magnetic beads it is possible to capture the RNA-DNA-hybrids and to get rid of the majority of unspecific DNA. The captured DNA goes on a 2nd generation sequencer which generates a pair of fastq files (sequence data plus quality data). These fastq files can be analysed with appropriate HLA calling software.

Fastq files, produced by the sequencer, are subsequently analysed using a computational method developed by Wittig *et al.* on a graphical user interface (GUI) (Figure 19)[88]. The software can be downloaded from https://hlassign.org/.

Figure 19 – GUI of the HLAssign software for a chosen locus – HLA-*C*.
Figure taken from hlassign.org/release-notes/. Original description: New HLAssign user interface design. This screenshot shows the HLA typing results in the graphical quality control mode of HLAssign. Additionally, the evenness of the read distribution (i.e. larger gaps between two homologous parts of other loci (this may for instance be the case for HLA-*A* and HLA-*H* and HLA-*DRB3/4/5* and HLA-*DRB1*)) are considered.

Reads obtained from the sequencing are mapped to the cDNA collection of the IMGT/HLA database. Matching of the sequenced reads to the reference is evaluated as follows: Only perfect matches between the reference and the sequenced reads are allowed, and only single-start point are retained (i.e. copies of the same read are reduced to 1). cDNAs with a central read coverage, calculated as a ratio of central reads)/(noncentral reads), <0.2 are discarded as overlapping reads.

For each cDNA pair (i.e. pair of alleles) the coverage calculated "as the area under the curve (auc), the number of reads that map exclusively to one of the two alleles (allele-specific mappings, asm), the number of mapped pairs per read (mppr), the length of the mappable sequence (msl)" and the read equality (req) are calculated[88]. "The read equality is calculated as the ratio of min(asm)/max(asm). After these values are calculated for all allele combinations, they are scaled between 0 and 1. The most likely genotype is the one that had the highest harmonic mean for all these measures"[88]. During manual selection of the alleles, allele pairs per locus are ranked according to the harmonic mean of the measures described above. For the construction of the HLA imputation reference in **Paper B**, Michael Wittig manually called the main part of the HLA data. I called HLA-*DRB3/4/5* in all populations and the Japanese cohort.

4.3.5. HLA imputation

GWAS typically analyse genotypic information from several thousand individuals to increase the statistical power of detecting an effect. Large disease consortia like the International IBD Genetics Consortium (IIBDGC), of which our group is a member, combine studies across the globe to achieve this goal. Statistical power of detection especially needs to be considered in the analysis of the highly polymorphic HLA region, where not only two to four but thousands of different variations may exist, and even larger study cohorts are needed. This soon becomes a problem of financial feasibility. Typing of the classical HLA alleles cost > 50€/per sample using NGS and several hundred € using conventional SBT methods. With 20,000 and more individuals analysed in larger association analyses, the endeavour of HLA typing becomes cost- and highly labour-intensive and other solutions are needed. HLA imputation refers to the inference of HLA types for individuals with an unknown HLA profile but known SNP genotypes from a reference dataset containing known HLA allele and HLA SNP genotype information. It was introduced by Leslie *et al.* in 2008[90] as the LDHLA algorithm which was further developed by Dilthey *et al.*[91,67] as HLA*IMP (2011) and HLA*IMP:02 (2013). HLA imputation makes use of genotypes measured using SNP array technology. This is data that is most commonly already available for individuals included in studies of GWA. To create a reference dataset, HLA allele information is measured using suitable typing methods and used together with the genotype information to create a reference dataset. The accuracy of HLA imputation highly relies on the size of and allelic diversity covered in a reference dataset as well as ancestry background of both input and reference data (**Paper B**) and does not reach diagnostic accuracy. Approaches for the inference of HLA alleles including HLA*IMP/HLA*IMP:02, SNP2HLA[66] and HLA Genotype Imputation with Attribute Bagging (HIBAG)[92] and reviews thereof[93–95] have been published in in the last decade. For our own analysis we considered both SNP2HLA and HIBAG, which were easily available, straight-forward in their application and had been shown to be accurate[92,96]. However, the libraries of HLA alleles considered in SNP2HLA were out-of-date and a larger effort would have been required to update the tool. HIBAG, on the other hand, is a tool that is independent of predefined HLA libraries, with the drawback, that at the time we constructed the model, it was computationally time and resource intensive. Since then faster algorithms for computation on the Graphic Processing Unit (GPU) have been developed. **HIBAG** is an HLA imputation approach based on supervised machine learning that uses attribute bagging and is available as an R-package (https://github.com/zhengxwen/HIBAG)[97]. It was published by Zheng *et al.* in 2014. The HIBAG model is shown in Figure 20. The general workflow of an HLA imputation is shown in Figure 21.

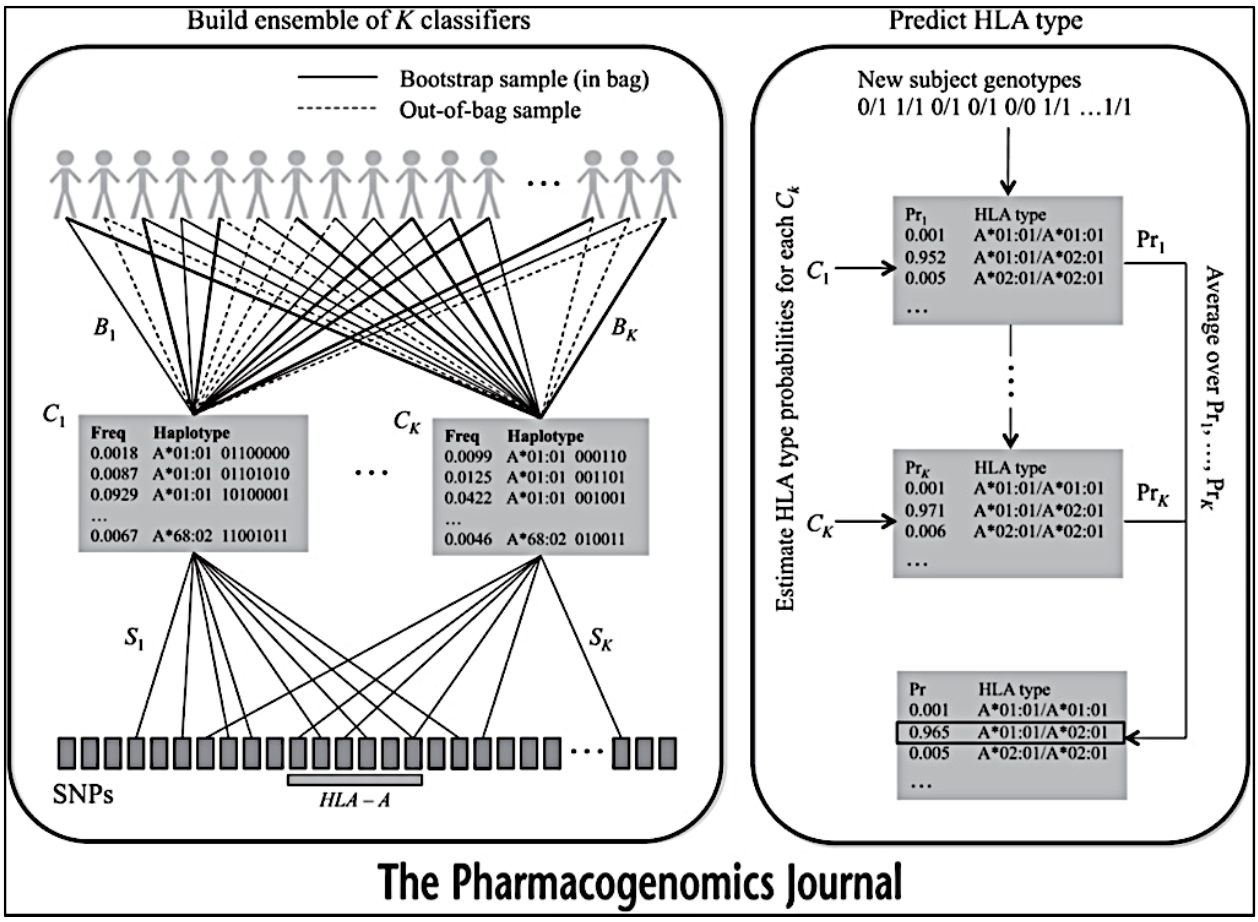


Figure 20 – Overview of the HIBAG prediction algorithm.
Figure was taken from Zheng *et al.*[92]. Original description: Overview of the HIBAG prediction algorithm. HIBAG is an ensemble classifier consisting of individual classifiers (C_k) with human leukocyte antigen (HLA) and single-nucleotide polymorphism (SNP) haplotype probabilities estimated from bootstrapped samples (B_k) and SNP subsets (S_k). The SNP subsets are determined by a variable selection algorithm with a random component. HLA-type predictions are averaged over the posterior probabilities from all classifiers.

To explain HLA imputation using HIBAG, we mainly cite from the original paper by Zheng *et al.* and a book chapter entitled “Imputation-Based HLA Typing with SNPs in GWAS Studies” from “HLA Typing – Methods and Protocols” published by Springer[89].

As introduced above, HIBAG is short for HLA Genotype Imputation with Attribute Bagging. Attribute bagging here refers to a repeated random sub sampling with replacement of individuals with known HLA and SNP genotype profiles (i.e. an individual can be present more than once in a subsample of the size of the original dataset; termed bootstrap aggregating or bagging), combined with random sub sampling with replacement of the feature space (here the features space is represented by SNPs) information at chosen positions. With each of these subsamples, a classifier is constructed that groups the SNP genotypes of individuals in this sub sample to the observed alleles. This grouping is performed for each locus separately based on the HLA alleles the individuals in the respective subsample carry. SNP haplotypes and their frequencies are estimated using an expectation maximization algorithm on

the joint probability of an HLA type and the SNP genotypes and is described in more detail in Zheng *et al.*[92]. Genomic positions that are most predictive for differentiating between all alleles of a locus are identified for each classifier using a stepwise procedure aimed at minimizing the error of classification (i.e. predicted HLA allele) calculated from the samples not used within a classifier (termed also out-of-bag samples). Highly predictive SNPs are likely to be used across multiple classifiers. The different classifiers are then combined to build an ensemble classifier.

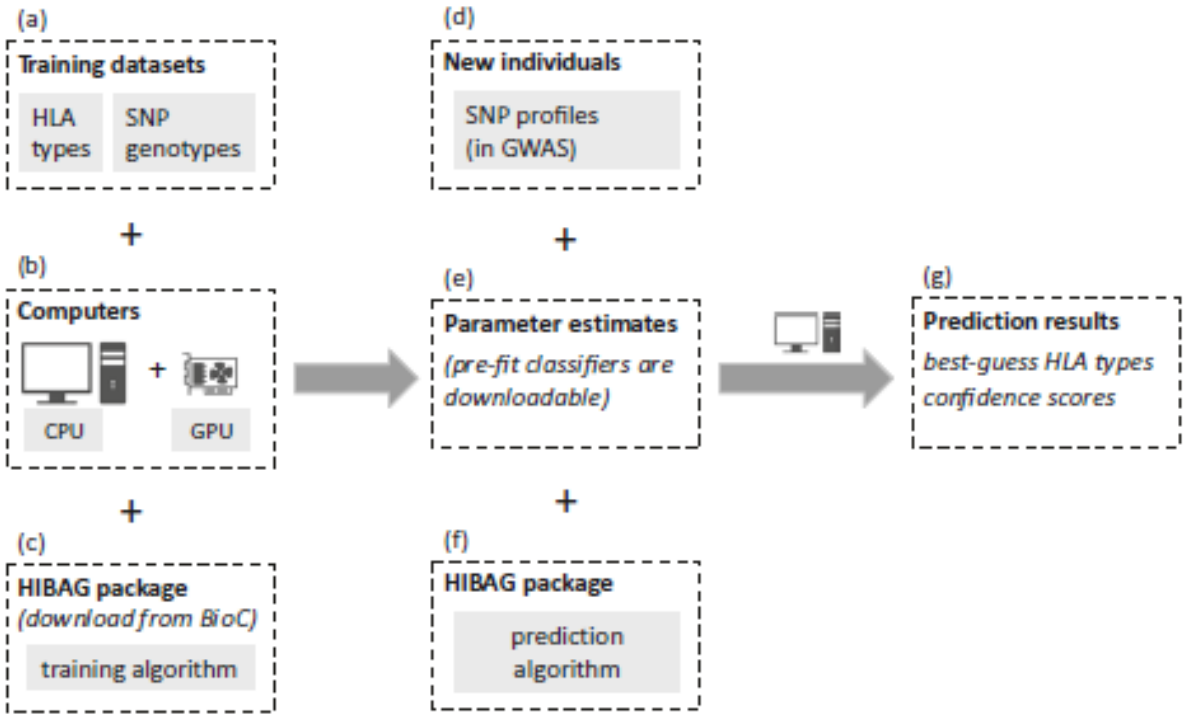


Figure 21 – Workflow of imputation using HIBAG.
Figure was taken from the book chapter entitled "Imputation-Based HLA Typing with SNPs in GWAS Studies" from "HLA Typing – Methods and Protocols" published by Springer[89]. Original description: Fig. 1 Computational workflow for HLA imputation using SNP data. (a) the training datasets consist of HLA and SNP genotypes; (b) computing equipment can be a desktop, a cluster of compute node, or a graphics processing unit (GPU); (c) the HIBAG R package can be downloaded from Bioconductor (BioC) and the training algorithm is freely available in the package; (d) SNP profiles are required for imputing HLA types of new individuals; (e) the parameter estimates are the output of the training algorithm and the pre-fit classifiers are published online; (f) the prediction algorithm is freely available in the HIBAG package; (g) the prediction results are best-guess HLA types and their confidence scores.

The HLA allele information for an individual with previously unknown HLA genotype is then inferred by comparing this individual's SNP genotype information to the haplotypes stored in the ensemble classifier and the HLA alleles with the highest probability across all classifiers are chosen. The alleles that best fit to the observed SNP genotype information across all classifiers is predicted. The accuracy of HLA imputation depends on the number of samples used in the HLA reference, the number of SNPs, the frequency of the HLA alleles in a population and the ancestry of the individuals used for construction

of the HLA reference panel – i.e. the reference population and study population should be of closely related ancestries. Both sample number and divergent populations within a reference panel, increase the allelic diversity in the HLA reference. HIBAG models were released with the software and are available on http://zhengxwen.github.io/HIBAG/hibag_index.html showcasing many different SNP arrays from Affymetrix and Illumina for African American, Asian, European and Hispanic ancestries. A notion of caution is advised, since these datasets were generated before a major HLA nomenclature update in 2010 and it is not quite clear whether they follow standard G- or P-group notations. The model produced in **Paper B** was added to the HIBAG repository and is available at http://zhengxwen.github.io/HIBAG/hibag_index.html.

4.3.6. HLA-peptide binding

HLA molecules bind peptides inside the peptide binding groove. While HLA class I proteins preferentially bind to peptides of length 9-15 amino acids, peptides that are bound by HLA class II proteins can be longer and typically have a length of 15-30 amino acids. This is due to the fact that the peptide binding groove of HLA class I molecules is closed at both ends, while the peptide binding groove of HLA class II molecules is open at both ends. The core recognition site of HLA class II molecules has a length of 9 amino acids and is displayed in Figure 22.

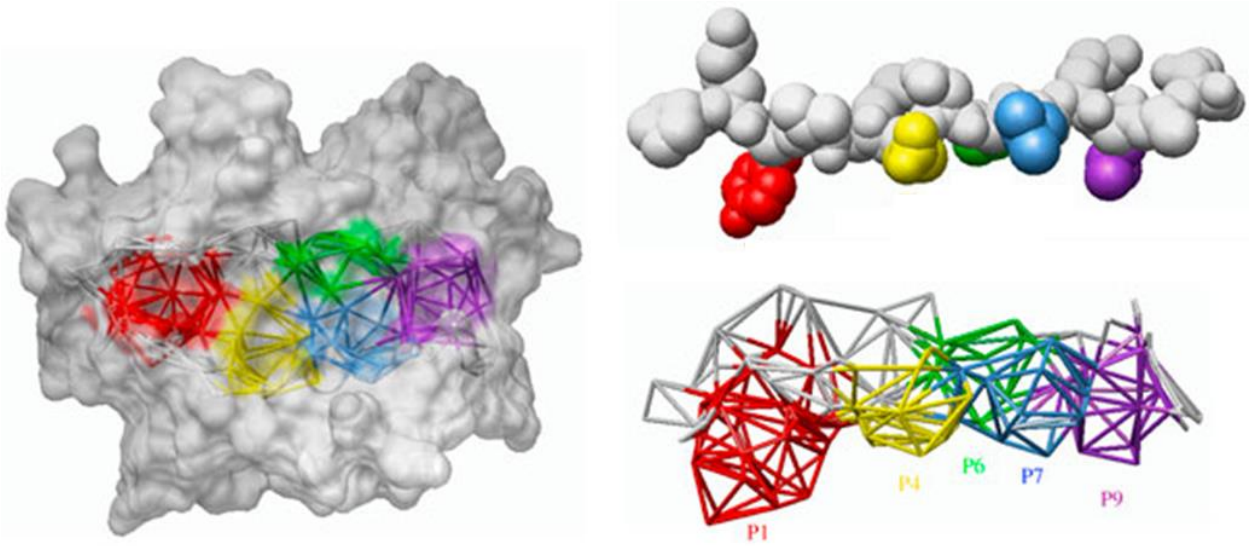


Figure 22 – Crystal structure of the HLA class II binding groove.
Figure was taken from Yeturu *et al.*[98]. Original and shortened description: Figure 1: Structure of an MHC class II binding groove. (A) Binding domain of HLA-DR1 [PDB:1DLH], with the five pockets in the binding groove highlighted (P1 – red; P4 – yellow; P6 – green; P7 – blue; P9 – purple).

For HLA class II proteins, the amino acids of the HLA that come into contact with the amino acids of the peptide are grouped into so called pockets (P). In total 9 pockets (P1 to P9) are defined for HLA class II proteins, with anchor residues, i.e. residues that are most important for HLA-peptide binding,

at pockets 1, 4, 6, 7 and 9. Flanking regions are either side of the core recognition site also affect HLA-peptide binding. Several machine learning algorithms to predict HLA-peptide binding exist. These algorithms are trained on experimental data mostly measured manually in the lab that are deposited in the Immuno Epitope Database or HLA peptides identified by Mass Spectrometry and peptide immunopurification studies in cell lines[99–101]. In this study we used the HLA-peptide binding tool NetMHCIIpan-3.2. NetMHCIIpan can predict the peptides that bind to any HLA allele by extrapolation across HLA alleles that it was trained on using pseudosequences generated for the alleles, while NetMHCII predicts HLA-peptide binding only for alleles that this tool was trained on.

4.3.7. Useful resources

These resources are also described in more detail in "HLA Typing – Methods and Protocols" published by Springer[89].

The IPD-IMGT/HLA database

The IMGT/HLA database is a central repository for human sequences of the HLA. It was first released in 1998 and was later incorporated into the Immuno Polymorphism Database (IPD) that was released 2002[55]. It currently contains 26,518 (IPD-IMGT/HLA version 3.39.0) allele sequences of 43 genes (6 HLA class I, 12 HLA class I pseudogenes, 21 HLA class II and 5 other non-HLA class I and class II alleles) and is updated every three months. Since its release there have been 90 releases (1.0-1.16, 2.0-2.28, 3.0.0-3.39; Figure 23). The IMGT/HLA database, amongst others, provides FASTA files for all known HLA alleles based on full gene, coding gene and amino acid information as well as alignment files for each locus. Additionally, updates in nomenclature as well as changes in allele status are recorded. The current releases all follow the nomenclature from 2010 (Chapter 4.2.2). Allele status here refers to whether an allele is retained in or deleted from a data release based on additional information gathered on an allele between releases, e.g. errors in the sequence or information on expression of an allele. Reasons why alleles are changed from release to release, are for instance the identification of errors in a submitted sequence. Most often alleles with errors in the sequence are found to be equal to another already submitted sequence and are deleted. Alleles are changed when longer read information is obtained or the expression status is evaluated. Apart from data on the human allele specifications, sequence information on an additional 80 organisms can be accessed from the IDP-MHC database (Release 3.4.0.0).

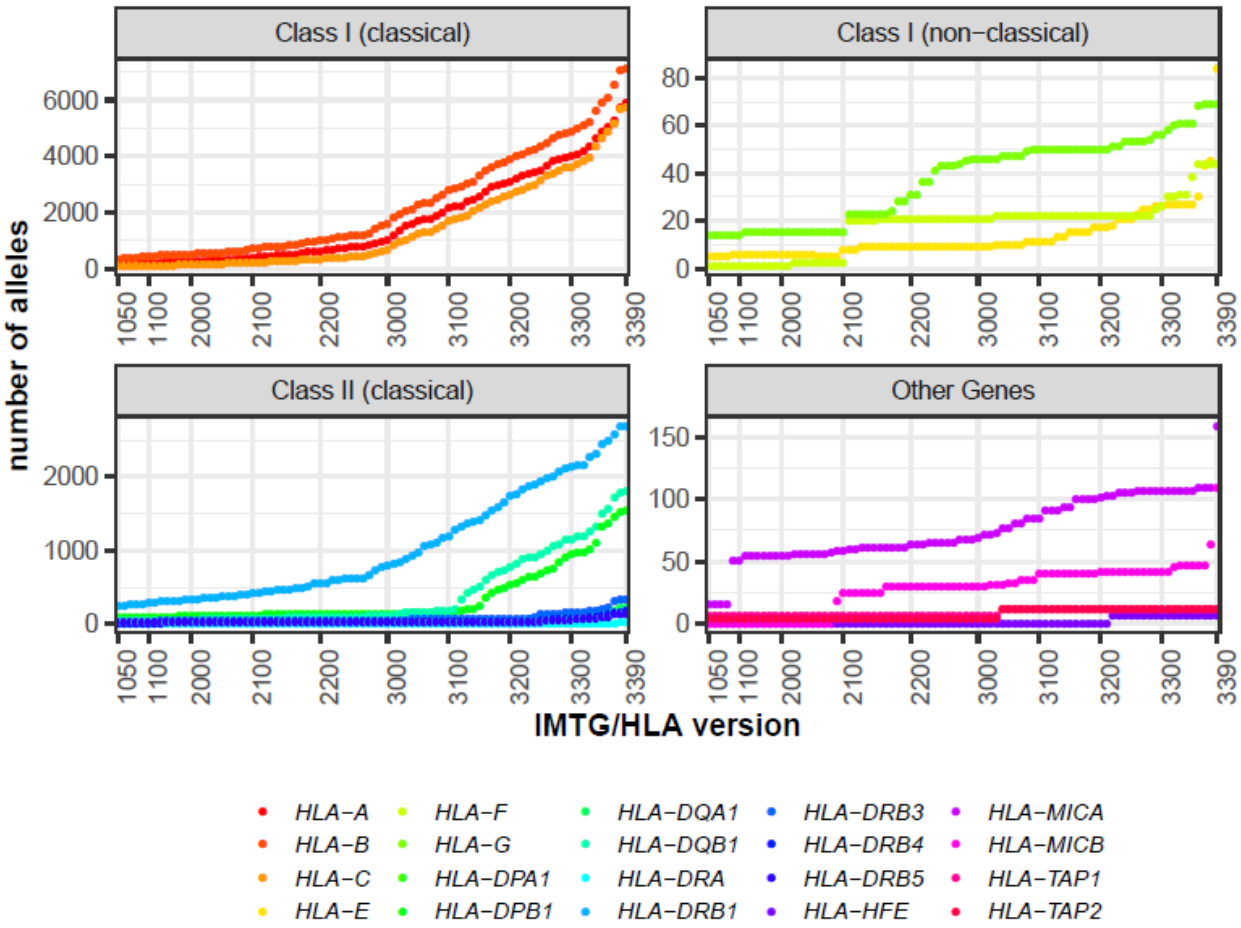


Figure 23 – Number of HLA sequences available at the IPD-IMGT/HLA up to release 3.39.0. This is a selection of genes present in the database. Figure was created from ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/Allelelist_history.txt. We noted the first and last version number of a base version (i.e. version 1,2,3), where 3390 corresponds to version 3.39.0.

The IMGT/HLA database is available at https://www.ebi.ac.uk/ipd/imgt/hla/. The web front provides additional web-based services for aligning and blasting of HLA allele sequences and queries for HLA alleles with specific submission status or ethnical origin. Historically, many of the alleles uploaded to the IMGT/HLA database have included only exons I and II for HLA class I molecules and exon II for HLA class II molecules, which are the regions that are functionally most variable. Currently the IMGT/HLA database still excepts partial sequences in addition to full length genomic sequences for submission into the database. With decreasing costs of NGS and the generation of an increasing number of genomic sequences, the number of sequences submitted to the IMGT/HLA database is expected to increase substantially. The impact of NGS-based HLA typing can also be seen in Figure 23. Still, the coverage of HLA class I alleles is at present significantly higher than the coverage of HLA class II alleles.

The HLA alleles website

The website hla.alleles.org presents a useful resource to learn about HLA nomenclature, its history and updates thereof. It is tightly linked to the IMGT/HLA database, listing alleles present in the databases and providing an overview of which loci have been typed. It presents information on HLA nomenclature reports – which contain changes in nomenclature across different timepoints – and basic knowledge on how alleles are named.

HLA frequency database

The allele frequency net database (AFND) is a web-based resource (http://allelefrequencies.net/) that stores HLA allele and HLA haplotype frequency information for worldwide populations[102]. In addition, information on other immune genes (Killer-cell immunoglobulin-like receptor: KIR, cytokines, Human MHC class I chain related genes: MIC) can be queried in various populations (Figure 24).

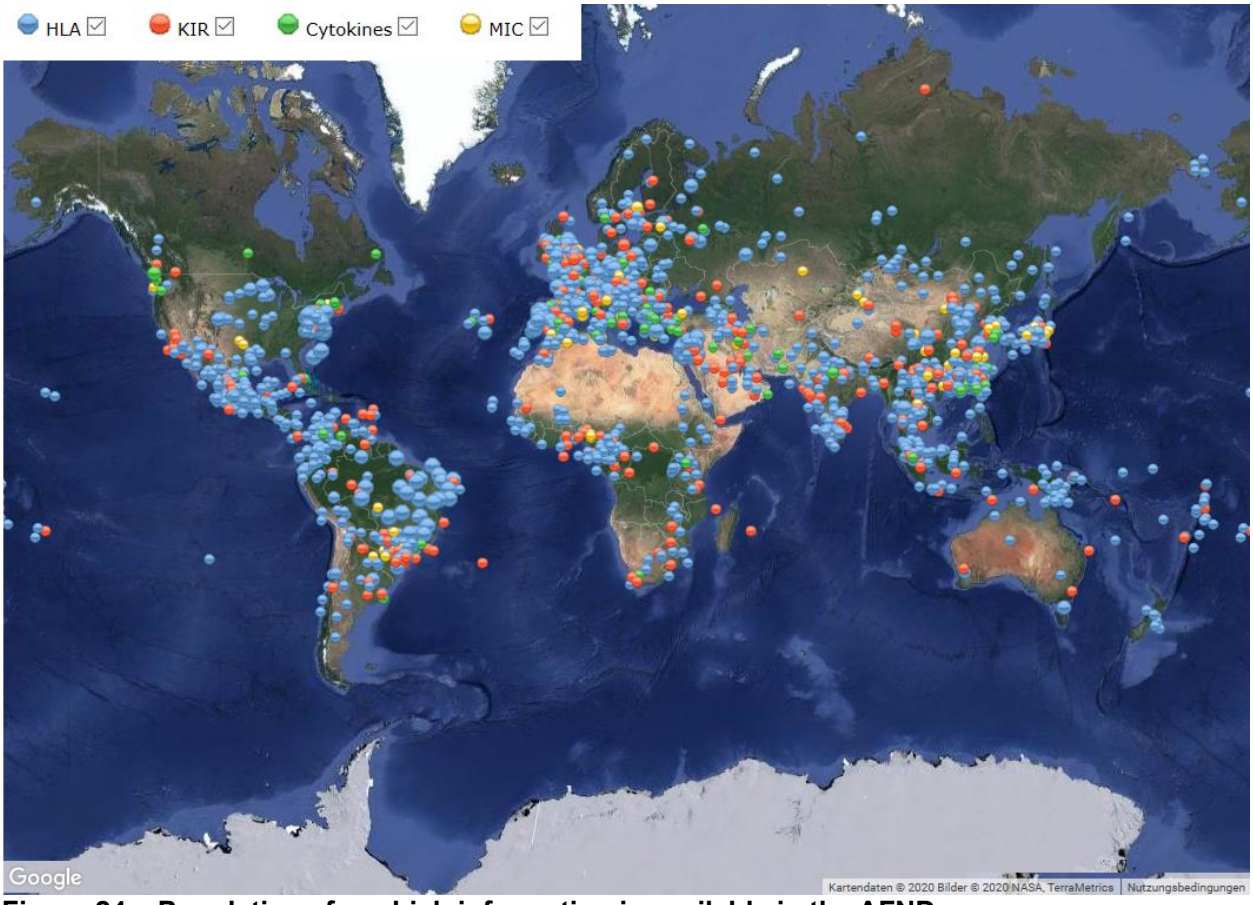


Figure 24 – Populations for which information is available in the AFND.
Figure was taken and modified from the AFND webpage. The map used here is based on the google maps and was retrieved on the 1st of April 2020. A satellite image was chosen. Original description: Map 1. Populations in the Allele Frequency Net Database (blue: HLA, red: KIR, green: cytokines, yellow: MIC).

In March 2020, frequency information from 10,688,428 individuals originating from 1,717 different populations were available on the AFND across all stored frequencies for the genes described in Figure 24. The AFND provides several additional resources, including an epitope frequency search currently available for *HLA-A, -B* and *-C* and information of HLA drug reverse reaction for HLA alleles as well as a frequency conversion tool to determine amino acid frequencies in a population. Allele frequencies and haplotypes are searchable for any allele present in the AFND, including amongst other filter options to select only population samples of specific country of origin, ethnic background, the number of individuals recorded in the population sample and the level of resolution (1-field to 4-field).

5. Results

5.1. PAPER B

Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, Ng SC, Rosati E, Hübenthal M, Ellinghaus D, Jung ES, Lieb W, Abedian S, Malekzadeh R, Cheon JH, Ellul P, Sood A, Midha V, Thelma BK, Wong SH, Schreiber S, Yamazaki K, Kubo M, Boucher G, Rioux JD, Lenz TL, Brant SR, Franke A. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum Mol Genet. 2019 Jun 15;28(12):2078-2092. doi: 10.1093/hmg/ddy443. PMID: 30590525; PMCID: PMC6548229.

Background & Aims: Next generation sequencing (NGS)-based HLA typing is cost-intensive and therefore limits the number of samples that can be analysed. To address this problem, appropriate HLA imputation references are needed. Published reference panels, mostly addressing allelic diversity only in the Asian and Caucasian populations, exist for variable number of HLA loci and allelic resolutions. Here we aimed to create a HLA imputation reference panel that covers a wide spectrum of different ethnicities using 2-field HLA information for the genes *HLA-A, -B, -C* (HLA class I) as well as *HLA-DRB1*, *-DRB3/4/5*, *-DQA1*, *-DQB1*, *-DPA1* and *-DPB1* (HLA class II).

Methods: Using NGS, we typed 1,360 individuals from 8 different populations of divergent ancestries at the stated loci and combined this with genotype information derived for all individuals from Illumina's ImmunoChip. This data was used to create an HLA imputation reference panel using the machine learning tool HIBAG[92]. We benchmarked our approach by performing a 5x cross-validation across our 8 populations and the independent 1000 Genomes population[103–105]. Furthermore, the *HLA-DRB3/4/5* genes were analysed in detail. Additionally, we calculated statistical measures such as sensitivity and specificity for all alleles to evaluate how accurately each single allele could be imputed.

Results: Our benchmark showed high accuracy across partitions of our 8 different populations and the independent 1000 Genomes population. We validated known HLA-*DRB1-DRB3/4/5* haplotypes and addressed why imputation is challenging in general for some HLA alleles across different HLA imputation panels, including our panel, with a focus on alleles of the HLA-*DRB1* locus.

Conclusion: In summary, we created a highly diverse imputation panel for the imputation of the HLA with a broad range of classical HLA alleles available for imputation.

Authors contributions

F.D. performed statistical and computational analysis. M.We. performed computational analysis with contributions from M.Hü. 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., 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.

Supplementary files

Supplementary files for **Paper B** are shown in Appendix A (Chapter 8). Supplementary tables are stored on the enclosed CD.

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

[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.

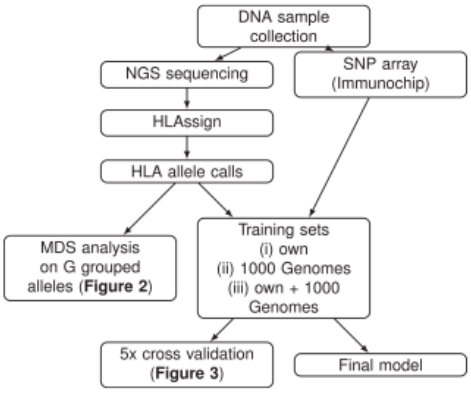


Figure 1. Flowchart of steps taken in preparation and benchmarking of our multi-ethnic reference panel. HLA allele calls were made based on NGS reads. Genotype information was measured using the Illumina Immunochip. These data were combined to train a HIBAG imputation model. Benchmarking was performed 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,

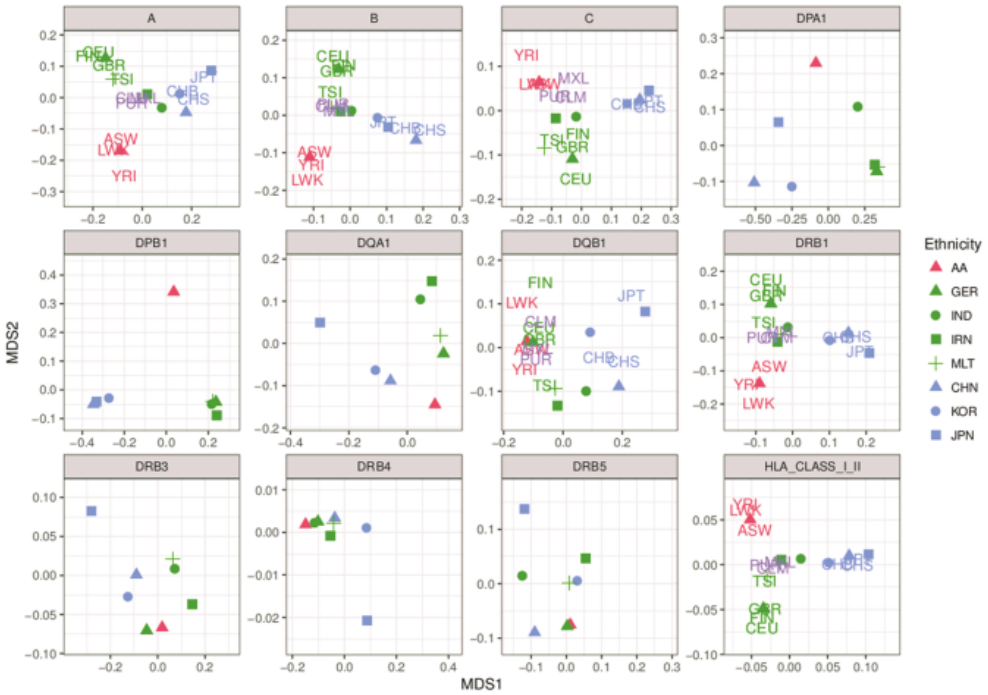


Figure 2. MDS analysis of HLA typed allele data: the MDS analysis was performed using a Euclidean distance measure. Alleles with a frequency <1% were excluded to produce a clustering that is not biased by similarity in low frequency variants. Colors show the origin of the cohort. Red: African American (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

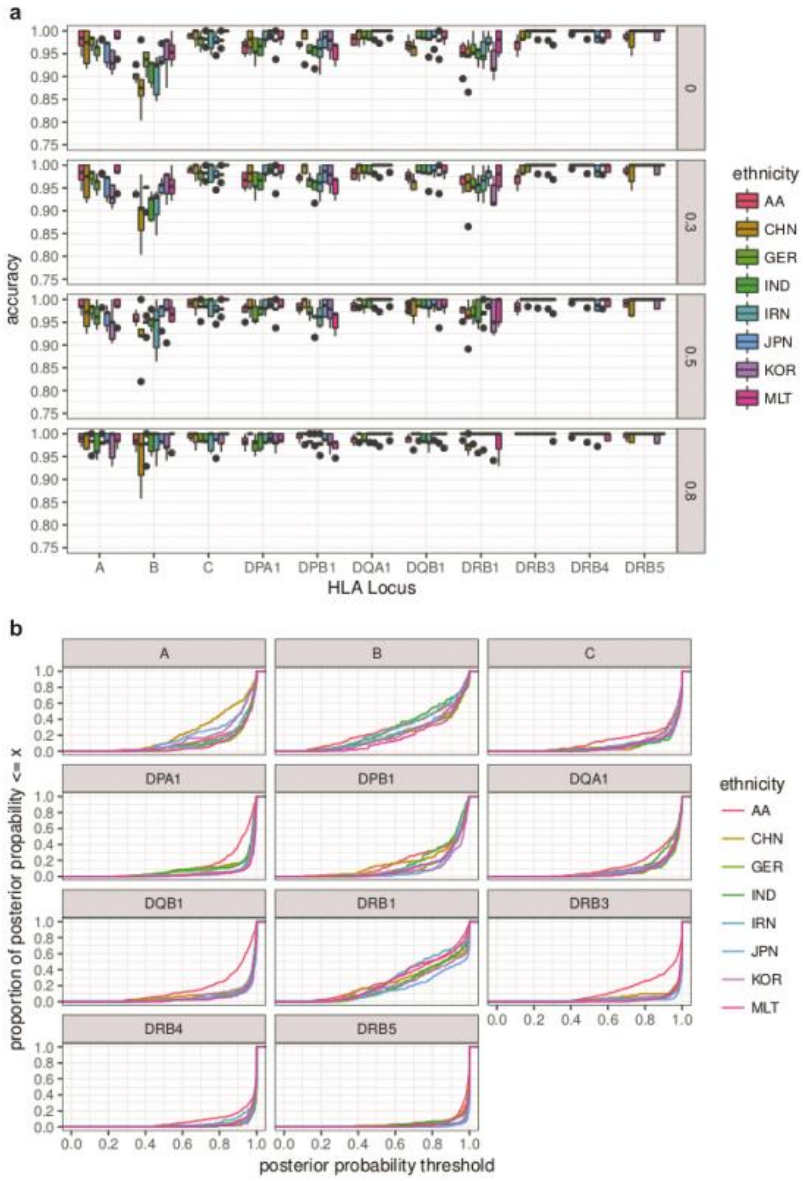


Figure 3. Imputation accuracies employing the multi-ethnic reference panel: accuracies and post-imputation probabilities of HLA imputation with HIBAG using a 5-fold cross-validation scheme and the multi-ethnic data set with full four-digit allele information. 20% of the data with a specific ethnic background 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).

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

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 MLT 0.967	GER 0.961 MLT 0.967	AA 0.937	CHN 0.924 JPN 0.958 KOR 0.947	CHN 0.924 JPN 0.958 KOR 0.947	CHN 0.924 JPN 0.958 KOR 0.947	CHN 0.924 JPN 0.958 KOR 0.947
	1000 Genomes						
	EUR 0.96	EUR 0.96	ASW 0.9 LWK 0.913 YRI 0.893	CHB 0.956 CHS 0.951 JPT 0.953	CHB 0.956 CHS 0.951 JPT 0.953	CHB 0.956 CHS 0.951 JPT 0.953	CHB 0.956 CHS 0.951 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	

(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	PUR	0.913
					YRI	0.893	JPT	0.953		

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

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-

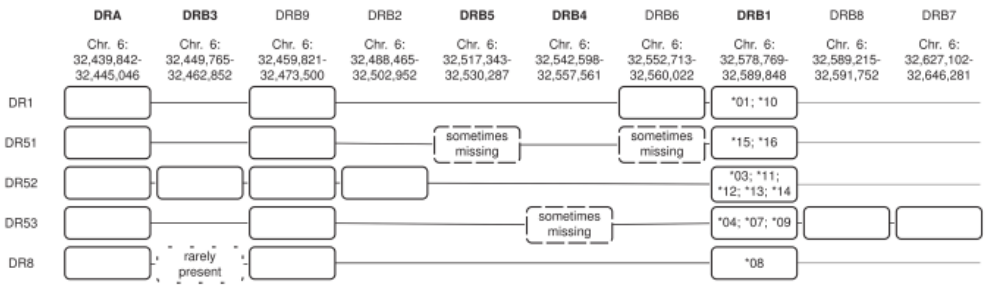


Figure 4. Known architecture of *HLA-DRB3/4/5*: HLA haplotypes that usually contain a specific *HLA-DRB1* allele (*HLA-DRB1* column) are shown. Two-digit alleles are denoted. All loci are depicted in order of their genomic location. *HLA-DRA*, *HLA-DRB1* and *HLA-DRB9* coincide with all haplotypes. The 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.

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.

5.2. PAPER C

Degenhardt F, Mayr G, Wendorff M, Boucher G, Ellinghaus E, Ellinghaus D, ElAbd H, Rosati E, Hübenthal M, Juzenas S, Abedian S, Alizadeh B, BK T, Yang S-K, Duk Ye B, Cheon JH, Datta LW, Daryani NE, Ellul P, Esaki M, Fuyuno Y, McGovern DPB, Haritunians T, Hong M, Juyal G, Jung ES, Kubo M, Kugathasan S, Lenz TL, Leslie S, Malekzadeh R, Midha V, Motyer A, Ng SC, Okou DT, Raychaudhuri S, Schembri J, Schreiber S, Song K, Sood A, Takahashi A, Torres EA, Umeno J, Vahedi H, Weersma RK, Wong SH, Yamazaki K, Karlsen TH, Rioux JD, Brant SR for the MAAIS Recruitment Center, Franke A for the International IBD Genetics Consortium. Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals common disease signatures. Currently under review.

Background and Aims: Inflammatory bowel disease (IBD) is a chronic inflammatory disease of the gut. Genetic association studies have identified the highly variable human leukocyte antigen (HLA) region as the strongest susceptibility locus for IBD, and specifically DRB1*01:03 as a determining factor for ulcerative colitis. However, for most of the association signal such a delineation could not be made due to tight structures of linkage disequilibrium within the HLA. The aim of this study was therefore to further characterise the HLA signal using a trans-ethnic approach.

Methods: We performed a comprehensive fine mapping of single HLA alleles in a cohort of 9,272 African American, East Asian, Puerto Rican, Indian and Iranian descent and 40,691 previously analysed Caucasians, additionally analysing whole HLA haplotypes. Using NetMHCIIpan-3.2, we performed peptide prediction for sets of 200,000 random unique human peptides for associated HLA alleles and additionally analysed their physico-chemical properties.

Results: We describe genetic factors and even entire *HLA-DQ-DR* haplotypes shared across different ethnicities, highlighting alleles of the HLA-DRB1*15 group. We identify the previously reported DRB1*01:03 as a population-specific signal that is mostly present in individuals of Western European descent and hardly present in non-Caucasian individuals. Peptides that preferentially bind to risk proteins are rich in positively charged amino acids such as arginine and lysine.

Conclusions: The HLA plays an important role for UC susceptibility across different ethnicities, though there are some population specific signals which should be considered dependent on the population of interest. This research implicates that there may be specific features of peptides that preferentially bind to risk and protective HLA proteins.

Author contributions are listed in the paper.

Authors contributions & Supplementary files

Author contributions are listed in the paper. Supplementary files for **Paper C** are shown in the Appendix B (Chapter 8). Supplementary tables and an additional supplementary analysis are stored on the enclosed CD.

Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals common disease signatures

Short title: Trans-ethnic analysis of the HLA in UC

Frauke Degenhardt[1], Gabriele Mayr[1], Mareike Wendorff[1], Gabrielle Boucher[2], Eva Ellinghaus[3], David Ellinghaus[1,4], Hesham ElAbd[1], Elisa Rosati[1], Matthias Hübenthal[1,5], Simonas Juzenas[1], Shifteh Abedian[6,7], Behrooz Alizadeh[6], Thelma BK[8], Suk-Kyun Yang[9], Byong Duk Ye[9], Jae Hee Cheon[10], Lisa Wu Datta[11], Naser Ebrahim Daryani[12], Pierre Ellul[13], Motohiro Esaki[14], Yuta Fuyuno[14,15], Dermot PB McGovern[16], Talin Haritunians[16], Myhunghee Hong[17], Garima Juyal[18], Eun Suk Jung[1,10], Michiaki Kubo[19], Subra Kugathasan[20,21], Tobias L. Lenz[22], Stephen Leslie[23], Reza Malekzadeh[7], Vandana Midha[24], Allan Motyer[23], Siew C Ng[25], David T Okou[26], Soumya Raychaudhuri[27,28,29,30,31], John Schembri[13], Stefan Schreiber[1,32], Kyuyoung Song[17], Ajit Sood[24], Atsushi Takahashi[33], Esther A Torres[34], Junji Umeno[14], Homayon Vahedi[7], Rinse K Weersma[35], Sunny H Wong[25], Keiko Yamazaki[15], Tom H Karlsen[4,36]* John D Rioux[2,*], Steven R Brant[11,37,*] for the MAAIS Recruitment Center, Andre Franke[1,*,#] for the International IBD Genetics Consortium

*joint senior authors

1 Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany.

2 Université de Montréal and the Montréal Heart Institute, Research Center, Montréal Heart Institute, Montréal, Québec, Canada.

3 K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway.

4 Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway.

5 Department of Dermatology, Venerology and Allergy, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.

6 Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands.

7 Digestive Disease Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran.

8 Department of Genetics, University of Delhi South Campus, New Delhi, India.

9 Department of Gastroenterology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea.

10 Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of Medicine, Seoul, Korea.

11 Harvey M. and Lyn P. Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, John Hopkins University School of Medicine, Baltimore, USA.

12 Department of Gastroenterology, Tehran University of Medical Sciences Emam Hospital, Tehran, Iran.

13 Department of Gastroenterology, Mater Dei Hospital, Msida, Malta.

14 Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan.

15 Laboratory for Genotyping Development, Center for Integrative Medical Sciences, Riken, Yokohama, Japan.

16 F.Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.

17 Department of Biochemistry and Molecular Biology, University of Ulsan College of Medicine, Seoul, Korea.

18 School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.

19 RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

2

20 Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA.

21 Pediatric Institute, Children's Healthcare of Atlanta, Atlanta, USA.

22 Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany.

23 Schools of Mathematics and Statistics and BioSciences and Melbourne Integrative Genomics, University of Melbourne, Australia.

24 Dayanand Medical College and Hospital, Ludhiana, India.

25 Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong.

26 Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Emory University School of Medicine, Atlanta, USA.

27 Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

28 Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

29 Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

30 Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.

31 Centre for Genetics and Genomics Versus Arthritis, Division of Muscolosceletal and Dermatological Sciences, School of Biological Sciences, University of Manchester, Manchester, UK.

32 Department of Medicine, Christian-Albrechts-University of Kiel, Kiel, Germany.

33 Laboratory for Statistical and Translational Genetics, Center for Integrative Medical Sciences, Riken, Yokohama, Japan.

34 Department of Medicine, University of Puerto Rico Center for IBD, University of Puerto Rico School of Medicine, Rio Piedras, Puerto Rico.

35 Department of Gastroenterology and Hepatology, University of Groningen and University Medical

Center Groningen, Groningen, The Netherlands.

36 Research Institute for Internal Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet and University of Oslo, Oslo, Norway.

37 Department of Medicine, Rutgers Robert Wood Johnson School of Medicine and Department of Genetics, Rutgers University Brunswick and Piscataway, New Jersey, USA.

4

ACKNOWLEDGEMENTS & GRANT SUPPORT

This project received infrastructure support from the DFG Excellence Cluster No. 306 "Inflammation at Interfaces". M.W. and H.E. are supported by the German Research Foundation (DFG) through the Research Training Group 1743, "Genes, Environment and Inflammation". E.E. received funding from the European Union Seventh Framework Program (FP7-PEOPLE-2013-COFUND; grant agreement No. 609020 (Scientia Fellows)). S.A. is supported by joint funding from the University Medical Center Groningen, Groningen, The Netherlands, and Institute for Digestive System Disease, Tehran University of Medical Sciences, Tehran, Iran. Funding for the Multicenter African American IBD Study (MAAIS) samples, for the GENESIS samples, and for the African Americans recruited by Cedars Sinai was provided by the U.S.A. National Institutes of Health (NIH) grants DK062431 (S.R.B.), DK 087694 (S.K.), and DK062413 (D.P.B.M), respectively. This work was supported by a grant from the 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. This research was supported by a Mid-career Researcher Program grant through the National Research Foundation of Korea to K.S. (2017R1A2A1A05001119), funded by the Ministry of Science, Information & Communication Technology and Future Planning, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (grant number: HI18C0094), Republic of Korea. Funding for the Indian samples was provided by the Centre of Excellence in Genome Sciences and Predictive Medicine (Grant # BT/01/COE/07/UDSC/2008) from the Department of Biotechnology, Government of India). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5

ABBREVIATIONS

(A) Alpha chain of an HLA protein

(B) Beta chain of an HLA protein

AA African American population of this study

AFR African American population of the 1000 Genomes/HapMap population see also https://www.internationalgenome.org/category/population/)

AMR Admixed American population of the 1000 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)

AF Allele Frequency

CEU Utah Residents (CEPH) with Northern and Western European Ancestry of the 1000 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)

CI Confidence Interval

EAS East Asian population of the 1000 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)

EUR Caucasian population of this population or (mentioned within the context of the 1000Genomes/HapMap population European data of the latter; see also https://www.internationalgenome.org/category/population/)

F1, F3 Atchley Factors 1 and 3, that contain information 54 on amino acid properties

HLA Human Leukocyte Antigen

HLA-*A* Human Leukocyte Antigen gene locus *A*

HLA-*B* Human Leukocyte Antigen gene locus *B*

HLA-*C* Human Leukocyte Antigen gene locus *C*

HLA-*DRA* Human Leukocyte Antigen gene locus *DRA*

HLA-*DRB1* Human Leukocyte Antigen gene locus *DRB1*

HLA-*DRB3* Human Leukocyte Antigen gene locus *DRB3*

HLA-*DRB4* Human Leukocyte Antigen gene locus *DRB4*

HLA-*DRB5* Human Leukocyte Antigen gene locus *DRB5*

HLA-*DQA1* Human Leukocyte Antigen gene locus *DQA1*

HLA-*DQB1* Human Leukocyte Antigen gene locus *DQB1*

HLA-*DPA1* Human Leukocyte Antigen gene locus *DPA1*

HLA-*DPB1* Human Leukozyten Antigen gene locus *DPB1*

IND Indian population

IRN Iranian population

JPN Japanese population

KOR Korean population

MAF Minor Allele Frequency

MLE Maximum Likelihood Estimator

MLT Maltese population

PRI Puerto Rican population

P1-P9 Pockets 1 to 9 of the HLA protein

QC Quality Control

SAS South Asian population of the 1000 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)

SNP Single Nucleotide Polymorphism (MAF >= 1%)

SNV Single Nucleotide Variation (MAF < 1%)

xHLA extended HLA region

YRI Yoruba in Ibadan, Nigeria population of the 1000 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)

7

Competing interests

The authors declare no competing financial interests.

Author information

Steve R Brant, Tom H Karlsen, John D Rioux and Andre Franke: These authors jointly supervised this work.

International IBD Genetics Consortium

A full list of members and affiliations appears in the **Supplementary Note.**

MAAIS Recruitment center

A full list of members and affiliations appears in the **Supplementary Note.**

Authors contributions

F.D. performed statistical and computational analysis, G.B. contributed to statistical analysis. M.W. and H.E. performed computational analysis with contributions from D.E., M. Hü, S.L., A.M., T.L. and S.R.. G.M. performed protein structure analysis and analysis of physico-chemical properties with contributions from F.D.. S.J. performed HLA typing in contribution to the HLA reference panel. F.D., G.M., E.E., E.R. wrote or revised this manuscript. S.A., B.A., T.B.K., S-K.Y., B.D.Y., J.H.C., L.W.D., N.E.D., P.E., M.E., Y.F., D.P.B.M., T.H., M.Ho., G.J., E.S.J., M.K., S.K., R.M., V.M., S.C.N., D.T.O, J.S., S.S., K.S., A.S., A.T., E.A.T, J.U., H.V., R.K:W.,S.H:W., K.Y. were involved in study subject recruitment, contributed genotype data and or/phenotype data. F.D., T.H.K., J.D.R., S.R.B. and A.F. conceived, designed and managed the study. All authors reviewed, edited and approved the final manuscript.

ABSTRACT

Background and Aims: Inflammatory bowel disease (IBD) is a chronic inflammatory disease of the gut. Genetic association studies have identified the highly variable human leukocyte antigen (HLA) region as the strongest susceptibility locus for IBD, and specifically DRB1*01:03 as a determining factor for ulcerative colitis. However, for most of the association signal such a delineation could not be made due to tight structures of linkage disequilibrium within the HLA. The aim of this study was therefore to further characterize the HLA signal using a trans-ethnic approach.

Methods: We performed a comprehensive fine mapping of single HLA alleles in a cohort of 9,272 African American, East Asian, Puerto Rican, Indian and Iranian descent and 40,691 previously analyzed Caucasians, additionally analyzing whole HLA haplotypes. Using NetMHCIIpan-3.2, we performed peptide prediction for sets of 200,000 random unique human peptides for associated HLA alleles and additionally analysed their physico-chemical properties.

Results: We describe genetic factors shared across different ethnicities, highlighting alleles of the HLA-DRB1*15 group and even entire HLA-*DQ*-*DR* haplotypes shared across different ethnicities. We identify the previously reported DRB1*01:03 as a population-specific signal that is mostly present in individuals of Western European descent and hardly present in non-Caucasian individuals. Peptides that preferentially bind to risk proteins are rich in positively charged amino acids such as arginine and lysine.

Conclusions: The HLA plays an important role for UC susceptibility across different ethnicities, though there are some population specific signals which should be considered dependent on the population of interest. This research implicates that there may be specific features of peptides that preferentially bind to risk and protective HLA proteins.

Keywords: HLA, trans-ethnic, ulcerative colitis (UC), inflammatory bowel diseases (IBD), fine mapping

INTRODUCTION

Ulcerative colitis (UC) is a chronic inflammatory disease of the gut. Like Crohn's disease (CD), the other main subphenotype of inflammatory bowel disease (IBD), it is most likely caused by an abnormal reaction of the immune system to microbial stimuli with environmental factors also playing a role. Currently more than 200 genetic susceptibility loci are known for IBD for Caucasian populations and many of them are shared between CD and UC (Jostins *et al.*, Liu *et al.*, de Lange *et al.*[1–3]). Strong genetic association signals with both diseases are seen in the human leukocyte antigen (HLA) region. The HLA is located on the long arm of chromosome 6 between 29 and 34 Mb and has many complex functions within the immune system. One of the major tasks of the HLA is the presentation of antigens to the host-immune system, which leads to increased immune response and the elimination of intracellular (HLA class I) and extracellular pathogens (HLA class II). In Caucasian IBD populations a large percentage of the phenotypic variation is explained by variants within the HLA class II, with DRB1*01:03 being the strongest risk allele for UC (P [P-value]=2.68×10^{-119}, OR [odds ratio]=3.59; 95% CI [confidence interval]=3.22-4.00[4]), specifically by alleles of the HLA-*DR* and -*DQ* loci, though tight structures of linkage disequilibrium have hindered the assignment of causality. Additionally, a systematic comparison across ethnicities for the HLA association in UC has not been performed, also due to the lack of HLA imputation panels that could accurately infer HLA alleles for trans-ethnic genetic data sets[4–13]. Recently, we created such a trans-ethnic HLA imputation reference including dense single nucleotide polymorphism (SNP) fine mapping data typed on Illumina's ImmunoChip, covering a large proportion of the HLA, within 8 populations of different ethnicities[14]. Here we report the first trans-ethnic fine mapping study of the HLA in UC and additionally analyze physico-chemical properties as well as how potential culprit antigens may look for IBD.

METHODS

Cohort description

A detailed description of the cohorts and recruitment sites can be found in the **Supplementary Methods** and **Supplementary Table 1**. In brief, a total of 52,550 individuals (18,142 ulcerative colitis cases (UC), 34,408 controls) were used in this study, of which 10,063 (3,517 UC cases and 6,546 controls were of non-Caucasian origin. The Caucasian, Iranian, Indian and Asian dataset (from which we extracted Japanese and Chinese individuals) are of part of the data freeze published in Liu *et al.*[2], while individuals of African American (Huang *et al.*[15]), Korean (Ye *et al.*[16]), Maltese, and Puerto Rican descent were added. The recruitment of study subjects was approved by the ethics committees or institutional review boards of all individual participating centers or countries. Written informed consent was obtained from all study participants.

Genotyping & Quality control

All individuals were typed on the Illumina HumanImmuno BeadChip v.1.0 or the Illumina Infimum ImmunoArray 24 v2.0 (Malta). Genotypes of the study subjects were quality controlled as described in the **Supplementary Methods**.

HLA Imputation

QC-ed genotype data for each cohort were imputed using Beagle version 4.1[17,18] based on genotypes observed in the respective cohort. We imputed HLA alleles at loci HLA-*A*, -*C*, -*B*, -*DRB3*, -*DRB5*, -*DRB4*, -*DRB1*, -*DQA1*, -*DQB1*, -*DPA1* and -*DPB1* at full context 4-digit level using the IKMB reference published in Degenhardt *et al.*[14] and the imputation tool HIBAG[19]. Imputation of the Caucasian panel was additionally performed with the HLARES panel published with HIBAG (ImmunoChip-European_HLARES-HLA4-hg19.RData). Alleles were not excluded by setting a posterior probability

threshold. However, we took the sensitivity and specificity measures we generated as previously reported (Degenhardt *et al.*[14]) into consideration during interpretation.

Phasing of single nucleotide variants

Using SHAPEIT2[20] version r727, we phased quality-controlled genotype data on chromosome 6, 25Mb to 34Mb of the respective cohorts using variants with MAF >1%. We excluded SNPs that did not match 1000 Genomes Phase III[21] (October 2014) alleles (published with the SNP imputation tool IMPUTE2[22,23]) and ATCG variants that did not match the AFR, EUR, SAS, EAS or AMR populations (+ strand assumed for both). AFR (used for comparison with our African American samples), EUR (Caucasian, Iranian, Maltese, SAS (Indian), EAS (Chinese, Korean, Japanese) and AMR (Puerto Rican). Using default values of SHAPEIT2 (--input-thr 0.9, --missing-code 0, --states 100, --window 2, --burn 7, --prune 8, --main 20 and –effective-size18,000, we first generated a haplotype graph and, as suggested by the authors of SHAPEIT2, calculated a value of phasing certainty based on 100 haplotypes generated from the haplotype graph for each population separately. Then, we excluded SNPs with a median phasing certainty<0.8 within each population separately.

Imputation of single nucleotide variants

SNP imputation was performed using IMPUTE2[22,23] and the 1000 Genomes Phase III reference[21](October 2014) using parameters: -Ne 20,000, -buffer 250, -burnin 10, -k 80, -iter 30, -k_hap 500, -outdp 3, -pgs_miss, -os 0 1 2 3, allowing additionally for the imputation of large regions (-allow_large_regions). Imputation of the Caucasian data set was performed in batches of 10,000 samples. Imputation quality control was performed post-imputation excluding variants with an IMPUTE2 info score <0.8. For the Caucasian data set we excluded variants with a median IMPUTE2 info score <0.8 and a minimum info score <0.3. Additionally, we imputed SNPs into the data set using imputed HLA allele information (i.e. translated imputed HLA information into real nucleotide information at each position of the allele) (**Supplementary Methods**).

Generation of HLA haplotypes

HLA haplotypes were generated by comparing SNP haplotypes generated by SHAPEIT2[20] for each individual and SNP haplotypes stored for the alleles within the classifiers of the HLA reference model[14] for the alleles that were imputed for each individual at a given locus. For 10 random classifiers, we calculated the minimal distance between the SNP haplotypes stored for the allele of interest in the HLA reference model and the SNP haplotypes generated by SHAPEIT2. We assigned alleles to a parental haplotype based on how often this allele had minimal difference to the haplotype. Phasing certainty was calculated as the percentage of times an allele was assigned to the chosen parental haplotype. In cases no decision could be made or both alleles were assigned to the same haplotype, phasing certainty was set to 0. If an individual was homozygous at a locus, phasing certainty was set to 1.

HLA haplotype benchmark

We tested this method using genotype information of trio samples (Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) and Yoruba in Ibadan, Nigeria (YRI)) extracted from the Hapmap Phase 3 project and HLA allele information published for these individuals in the 1000 Genomes HLA diversity panel[24] using the most common allele for ambiguous calls. In total, 27 CEU samples and 24 YRI samples and their parents were analyzed. Genotype data were downloaded from the HapMap Phase 3 Server (version 2015-05) and positions present on the Illumina ImmunoChip were extracted. We applied the procedure described above for phasing of HLA alleles. The results are shown in **Supplementary Table 2.**

Calculation of marginal probabilities for each allele

Since HIBAG stores a matrix of all posterior probability values of each allele combination per individual, we calculated the marginal sums of posterior probability for each allele per individual. The overall marginal probability of an allele was then calculated as the mean of the marginal sums of the posterior probability calculated for alleles predicted to carry this allele.

14

Association analysis

Associations of a marker with the disease outcome were calculated using genotype dosages based on the minor allele of imputed and genotyped variants. HLA alleles were coded as present (P) or absent (A) with genotype dosages (PP=2, AP=1 and AA=0) by simply counting the number of times an allele occurred for a specific individual. Additive association analyses for each marker were performed using

$$\log(\text{odds}_i) = \beta_0 + \beta_1 x_i + \beta_2 U_{1i} + \beta_3 U_{2i} + \beta_4 U_{3i} + \beta_5 U_{4i} + \beta_6 U_{5i} + (\beta_7 b_i)$$

for individual i=1,…..,N, genotype dose or call (x) and eigenvectors (U1-U5). For the analysis of the Puerto Rican and Indian cohort we additionally adjusted for batch (b) (**Supplementary Methods;** batches during QC).

Meta-analysis

Classical fixed-effects or random-effects meta-analyses are not optimal for the analysis across study estimates where underlying allele frequencies are different between cohorts or similar only for some of the analyzed cohorts (Morris *et al.*[25]) as in the case of trans-ethnic analyses. Using REC2 (Lee *et al.*[26]), a tool optimized for the analysis of heterogenous effects, we combined the association statistics for all 9 cohorts for SNPs and HLA alleles with MAF (SNPs) and AF (HLA alleles) >1% in the respective cohorts. to calculate a combined P-value, setting the correlation between studies to uniform.

Clustering according to preferential peptide binders

Using NetMHCIIpan-3.2[27], we predicted binding affinities for five sets of the 200,000 unique random peptides **(Supplementary Methods)** for all alleles that were significant in the meta-analysis and had a frequency of >1% in at least one of the 9 populations. We selected the top 2% (strong binders (SB)) preferential peptide binders as given by the NetMHCIIpan-3.2 software for each allele and calculated

the pairwise Pearson correlation between alleles based on complete observations for the respective allele combinations[27] using R (version 3.3.1), creating a matrix of correlations. Clustering was performed on this matrix using hclust of the R package stats. Correlation between the clusters was calculated using corrplot (version 0.84) and dendextend (version 1.12). Here, the correlation between cluster dendrograms (i.e. the concordance of the tree-structure) is calculated with a value of 0 signifying dissimilar tree-structures and 1 signifying highly similar tree-structures. Dendrograms were plotted using the ape (version 5.3) package, for DQ and DRB1.

Generation of combined peptide motifs

Based on the clusters generated above for the human peptides, we grouped the risk alleles and protective alleles into 2 clusters each **(Supplementary Methods).** For each of the 5 peptide sets, we concatenated the top 2% ranked binders (percentile rank of NetMHCIIpan-3.2) for alleles within each protective and risk group and excluded peptides that were among the 10% top ranked binders (percentile rank of NetMHCIIpan-3.2) in two or more of the groups. Based on this, we generated peptide binding motifs using Seq2Logo.[28] for each of the groups.

Clustering according to physico-chemical properties

Clustering of HLA proteins was performed using 5 different numerical scores: the Atchley scores F1 and F3[29], residue-volume[30] and self-defined parameters charge and hydrogen-acceptor capability **(Supplementary Table 3).** The amino acid sequence of each respective allele was extracted at positions noted in **Supplementary Table 4** for Pockets 1, 4, 6, 7 and 9 from HLA allele protein sequences that were retrieved from the IMGT/HLA database (version3.37.0)[31] and aligned using MUSCLE[32]. The alpha chain, of the HLA-*DR* locus is invariable and was not considered in the analysis of this locus. For the respective analysis, each amino acid was assigned its numerical score. Clustering was then performed on the scores using the hclust function of the R (version 3.3.1) package stats and Euclidian distances.

16

RESULTS

Here we imputed HLA alleles for a total of 9 cohorts **(Supplementary Figure 1, Supplementary Table 1)** within 3 HLA class I (HLA-*A*, -*C* and -*B*) and 8 class II loci (HLA- *DRB3*, -*DRB5*, -*DRB4*, -*DRB1*, -*DQA1*, -*DQB1*, -*DPA1* and -*DPB1*) at full context 4-digit level utilizing a median of 8,555 SNP genotypes (located within extended HLA between 25 and 34 Mb on chromosome 6p21) from Illumina's ImmunoChip. After QC, a total of 17,276 UC cases and 32,975 controls of which 13,927 cases and 26,764 controls were previously reported Caucasians[4] and 3,251 cases and 6,021 controls were non-Caucasian individuals (part of Liu *et al*.[2] study). To increase the density of single nucleotide variants (SNVs, including variants with minor allele frequency (MAF)<1%) within the HLA region, we used publicly available nucleotide sequences of HLA alleles and further imputed SNVs based on the HLA alleles imputed for each individual using IMPUTE2[18] with the 1000 Genomes Phase III[21] individuals as reference (a median of 88,087 SNVs with INFO score > 0.8 after imputation). Subsequently, we performed association analysis on single alleles and SNVs and a meta-analysis on SNP variants with a minor allele frequency (MAF) >1% and HLA alleles with an allele frequency (AF) >1% in respective cohorts only. The meta-analysis was performed using RE2C[26], a tool, that can analyze studies with heterogeneous effect sizes giving P-values that we hereafter call RE2Cp (optimized for highly heterogenious variants)and REC2p* We then obtained phased HLA haplotypes using HIBAG[19] and SHAPEIT2[20].

In line with our previous study in Caucasians[4], we observed strong, consistent association signals for SNPs and HLA alleles within the HLA class II region, featuring HLA-*DRB1*, HLA-*DQA1*, and HLA-*DQB1*, for all UC case-control panels except the small-sized Puerto Rican and Maltese cohorts **(Figure 1 and Supplementary Figure 2)**. The strongest association signal was seen for SNP rs28479879 (RE2Cp*=8.87×10^{-156}), located in the HLA-*DR* locus, including HLA-*DRB1* and HLA-*DRB3/4/5*. In the Japanese and Korean panels, we further observed a "roof-top"-like association signal spanning the HLA class I and II loci **(Figure 1)** that, as we subsequently demonstrated, was caused by strong linkage-disequilibrium between the most disease-associated class II alleles DRB1*15:02, DQA1*01:03, and DQB1*06:01, and the class I alleles B*52:01 and C*12:02. The "roof-top"-like signal disappeared when

conditioning on class I and class II alleles separately **(Supplementary Figure 3).** Likely due to lack of statistical power, e.g. for the Maltese data set, and/or diversity of the population, e.g. for the Puerto Ricans, association P-values for these populations did not achieve the genome-wide significance threshold ($P<5\times10^{-8}$).

The most strongly and consistently associated class II risk alleles within the meta-analysis were alleles of the DRB1*15 group ($RE2Cp^*=1.87\times10^{-116}$) **(Figure 2, Supplementary Table 5),** observed to be located on the same haplotype as DQA1*01:02/03 and DQB1*06:01/02 **(Figure 3, Supplementary Table 6)**. DRB1*15:02 was most frequent in the Asian populations (Japanese, Korean), while DRB1*15:03 was specific to the African American population and DRB1*15:01 had the stronger association and higher allele frequency in the Chinese and Caucasian population **(Figure 2, Supplementary Table 5)**, which is consistent to data published in the HLA allele frequency database[33]. Other associated class II alleles included DQA1*03 alleles ($RE2Cp^*=5.83\times10^{-81}$) that were observed to be located on a haplotype with DRB1*04 ($RE2Cp^*=2.36\times10^{-55}$), DRB1*07:01 ($REC2p^*=5.99\times10^{-35}$) or DRB1*09:01 ($RE2Cp^*=2.73\times10^{-12}$). DRB1*04/07/09 alleles are all located on the same haplotype as HLA-*DRB4* alleles[14], therefore absence of HLA-*DRB4*, hereafter named DRB4*00:00, was significantly associated with high risk ($RE2Cp^*=2.35\times10^{-127}$). Along the same line HLA-*DRB5* is located on the same haplotype as DRB1*15. Its absence was therefore observed to be protective. We identified DRB1*10:01 as a novel association signal ($RE2Cp^*=1.03\times10^{-6}$). It was observed to be most frequent in the Iranian (3.2% controls and 1.6% cases) and Indian (6.7% controls and 3.3% cases) populations and rare in other populations **(Supplementary Table 5)**, which is most likely why it has not been described before**.** Among population-specific signals, we also observed significant association of UC with DRB1*14:04 (P=0.004, OR=1.64 95%CI: 1.18-2.29) in the Indian population **(Figure 2)**. Overall, alleles of 11 of the 13 known HLA-*DRB1* 2-digit groups and all 5 known -*DQB1* groups were associated with UC across the different cohorts **(Figure 2, Supplementary Table 5, Supplementary Figure 4)**, with more DRB1 alleles conferring protection than risk. Effect sizes in the larger Caucasian and Japanese populations were observed to be moderate (0.5<OR<2.0 for alleles with AF>1%, with the exception of DRB1*15:02 (OR=2.87; 95% CI: 2.46-3.36 in the Japanese population) and previously described DRB1*01:03.

Analysis of beta estimates also showed that Japanese and Korean effects estimates were most similar, while Iranian and Indian effects estimates correlated better with those of the Caucasian population (**Supplementary Figure 5**). The deviation from non-additivity of effects at the HLA locus observed in Goyette *et al.*[4] could not be replicated in this study (data not shown).

To reduce the complexity of the HLA signal further, we aimed to identify common peptide binding patterns of peptides that the herein detected significant HLA alleles could potentially bind **(Figure 4)** as well as to identify shared physico-chemical properties within UC risk and protective HLA proteins **(Figure 5)**. For this analysis, we only selected proteins for which the corresponding alleles had a significant P-value in the meta-analysis (RE2Cp*<0.05) and focused on the results of the DRB1 proteins **(DQ shown in Supplementary Figure 6).** First, we predicted the binding affinities for 5 sets of 200,000 random unique peptides sampled from the human proteome to the DRB1 proteins using NetMHCIIpan-3.2[27]. Next, we performed clustering analysis on the alleles using the top 2% ranked preferentially binding peptides for each allele based on pairwise observed complete observations. In general, we found DRB1-clustering **(Figure 4)** to be more informative regarding separation of protective and risk alleles than DQ-clustering. Additionally, DRB1-clustering was more stable across the sets of random peptides **(Supplementary Figure 6).** Larger "risk clusters" were identified for DRB1 including DRB1*15:01 and the newly identified DRB1*15:03. We defined 2 risk clusters including DRB1*11:01/04 and DRB1*13:01 (RISK 1) DRB1*12:01, DRB1*14:04 and DRB1*15:01/03 (RISK 2), and 2 protective clusters including DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and DRB1*10:01 (PROT 1) and DRB1*04:03/06 (PROT 2). Within each cluster, we calculated a combined peptide binding motif by combining the top 2% of binders for each allele in the groups **(Figure 4).** The peptide binding motifs of the two risk groups were enriched for basic amino acids (K and R) and depleted for acidic amino acids, while the peptide binding motifs of the protective group were enriched for hydrophobic and polar amino acids. Interestingly, DRB1*01:03 clustered with protective alleles DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and DRB1*10:01, however, a more detailed analysis of its physico-chemical properties resulted in a predominant clustering with DRB1*15 **(Figure 5)**. Equally, DRB1*15:02 clustered with DRB1*13:02, while physico-chemical properties resulted in a predominant clustering with the DRB1*15 group. In **Supplementary Figure 7**

we show that this may be an artefact of NetMHCIIpan-3.2 caused by extrapolation of the DRB1*15:02 signal for unknown peptides from DRB1*13:02.

DISCUSSION

Several conclusions can be drawn from this trans-ethnic HLA fine mapping study in UC:

(I) HLA allele associations are mostly consistent across different populations regarding effects direction, we herein highlight the DRB1*15 allele group, which is represented by three alleles DRB1*15:01 (most frequent in the Caucasian population), DRB1*15:02 (most frequent in the Korean and Japanese population) and DRB1*15:03 (most frequent in the African American population), as well as the DQA1*01:02/03- DQB1*06:01/02 located on the same haplotype as the DRB1*15 alleles. Incidentally, DRB1*15:01/03 have previously also been described in other immune related diseases. Amongst others in Multiple Sclerosis[34] and leprosy, indicating a shared disease aetiology of these diseases. Other DRB1*15, for instance DRB1*15:06, which did not show association with UC, were likely too infrequent in the analysed populations. Interestingly, however DRB1*15:06 has the same amino acid sequence as DRB1*15:01 and may therefore biologically indeed play a role in IBD. This is also true for other alleles listed in **Supplementary Table 7**. Within this fine mapping study, we identify several population specific signals **(Figure 3)**, amongst others DRB1*14:04 and surprisingly also DRB1*01:03, which is highly associated in Caucasian IBD. Notably, DRB1*01:03 was not present in the Asian populations and was only observed with a frequency of <0.1% in the African American and Puerto Rican populations. Detailed analysis of the geographic distribution of the DRB1*01:03 allele showed, that I it seemingly occurs in Western Europe (Great Britain, Ireland, France, Spain) and former Western colonies with AF >1%, while it seems to be infrequent in the Eastern parts of Europe. We therefore hypothesize that this allele is linked to the history Western European countries. **(Figure 6)**. Within this study, the frequency of DRB1*01:03 in the Caucasian population is likely underestimated and therefore not the top associated signal in the Caucasian analysis (i.e. DRB1*01:03 was imputed as DRB1*01:01 or DRB1*01:02 due to similarities in SNP haplotype between these alleles) due to applying a reference panel, containing

20

mostly non-Caucasian individuals and European individuals from Germany only. Indeed, using the European HLARES imputation panel, which contains a more diverse Caucasian population, we re-established the signal. The frequency of the remaining alleles imputed with our transethnic reference dataset highly correlated with our original study in the Caucasian population **(Supplementary Figure 8)**.

(II) Allele associations across the different genes are highly correlated between *HLA-DR* and *HLA-DQ* and can be explained by population haplotype structures. Here we phased whole HLA haplotypes and could therefore directly infer correlations between the two loci. Overall, unexpectedly, trans-ethnic analysis does not aid the assignment of causal alleles within the HLA. Thus, the HLA-*DQ* locus, and HLA-DRB3/4/5 cannot be fully ruled out as being significant in UC disease aetiology. However, overall, the analysis of physico-chemical properties and preferential peptide binding was more informative for alleles of the HLA-*DR* locus. For DRB1*13:01 and DRB1*13:02, which interestingly share the same amino acid properties across binding pockets P1-P9 of the HLA, the associated DQA1-DQB1 haplotypes DQA1*01:03-DQB1*06:03 and DQA1*01:02-DQB1*06:04 do not. This may point to *HLA-DQ* for these alleles to be causative, since DRB1*13:01 and DRB1*13:02 have opposite effect sizes in the association analysis.

(III) Analysis of peptide binding motifs showed that protective and risk alleles cluster stably and that risk and protective groups have distinct peptide binding motifs, which is supported by their showing similar physico-chemical properties. In the future, this should be followed up by suitable peptidomics experiments, which will enlarge datasets available for peptide binding predictions and increase confidence in the risk and protective motifs that may be indicative of culprit antigens in UC having distinct features. Larger, per-population patient collections will be needed in future studies to confirm our results and to obtain even more precise effect estimates of associated HLA alleles. In addition, we hope that IBD patient panels from other ethnicities will become available for genetic fine mapping studies. With typing of HLA alleles now being possible using next-generation sequencing methods, real typing rather than imputation analyzes should become standard, thereby avoiding possible imputation artefacts. The construction of haplotype maps will then likely be even more accurate.

21

FIGURES

Figure 1 – HLA regional association plots. Association analysis results for imputed and genotyped single nucleotide variants (grey) and 4-digit HLA alleles (yellow) are shown for **(a)** 373 African American cases and 590 controls (AA), **(b)** 13,927 Caucasian cases and 26,764 controls (EUR), and **(c)** 709 Japanese cases 3,169 and controls (JPN) as well as **(d)** the meta-analysis (META) results from the analysis with RE2C (Lee *et al.*[26]) at variants with a MAF > 1% in the respective cohorts (including 17,276 cases and 32,975 controls from 9 different cohorts). The association plots for the remaining populations are provided in **Supplementary Figure 2**. SNP. The curves in (a)-(c) show the P-value of the meta-analysis (REC2p*). In (d) the overlying curve shows the I2 as a measure of heterogeneity in the meta-analysis indicating the heterogeneity of effects and allele frequencies in that region. Dashed lines indicate the thresholds of genome-wide ($P=5x10^{-8}$) and nominal significance ($P=10^{-5}$) The association analyses indicate HLA class II as the most associated susceptibility region across the different populations. In the Korean and the Japanese populations, a strong association signal is also seen for B*52:01 and C*12:02, both alleles being in strong linkage disequilibrium with the HLA class II loci DRB1*15:02, DQA1*01:02 and DQB1*06:01, i.e. another population-specific haplotype association in these ethnicities exists.

Figure 2 – HLA single allele association analysis results at 2- and 4-digit resolution for MHC class II loci *-DRB3/4/5, -DRB1, -DQA1-DQB1.* (AF; common defined as AF>1%), odds ratio (OR), P-value (P) and whether an allele had a P-value<0.05 (circle symbol) is shown for the respective population (e.g. circles with black boundary and red color represent an allele that is common and associated with risk). We depict association results of the analysis of the African American (AA), Puerto Rican (PRI), Caucasian (EUR), Maltese (MLT), Iranian (IRN), North Indian (IND), Chinese (CHN), Korean (KOR) and Japanese (JPN) cohorts and the meta-analysis (META) with RE2C's I2 as an indicator of allelic heterogeneity and the P-value of association (RE2Cp*, combined here with single study P-values P).

Only HLA alleles which are significant in the meta-analysis, that have an AF>1% in at least one population and that have a marginal post imputation probability >0.6 are shown. The strongest association signals in the meta-analysis are for risk alleles of the DRB1*15 group, i.e. DRB1*15:01, DRB*15:02 and DRB1*15:03 and the alleles located on the same respective haplotype **(Figure 3).** Alleles with OR>5.0 or OR<0.2 (rare and non-significant alleles may have larger/smaller OR) values were "ceiled" at 5.0 and 0.2 respectively. The "consistent alleles" that are highlighted in Figure 3 are highlighted in bold type on the left side.

Figure 3 – Haplotypes for associated HLA alleles.

For a selection of associated HLA alleles, we show the most frequently observed risk **(a)** and protective **(b)** haplotypes in the respective populations. (African American (AA), Puerto Rican (PRI), Caucasian (EUR), Maltese (MLT), Iranian (IRN), North Indian (IND), Chinese (CHN), Korean (KOR) and Japanese (JPN)). Here we show only DRB1-DQA1-DQB1 haplotypes with a frequency >1% in the case individuals in each respective population. The most frequently observed C-B alleles in each population were then added if the C-B-DRB1-DQA1-DQB1 haplotype occurred in more than or equal to 5 individuals. HLA-*DRB3/4/5* alleles were taken from Degenhardt *et al.*[14] and calculated based on individuals hemizygous for HLA- *DRB3/4/5* (i.e. carrying only one HLA-*DRB1* observed with either HLA-*DRB3*, -*DRB4* or -*DRB5* and one DRB1*01, DRB1*08 or DRB1*10 which are not observed with any of the HLA- *DRB3/4/5*.)

Figure 4 – Clustering of DRB1 proteins according to preferential peptide binding and combined peptide binding motifs. (MIDDLE CLUSTER): For 5 sets of 200,000 unique random human peptides the percentile rank scores of preferential peptide binding were calculated using NetMHCIIpan-3.2.[27] for all DRB1 proteins that were significant in the meta-analysis of genetic analysis of the HLA with and AF > 1% in at least one cohort. We additionally included DRB1*01:03. Within each set, the top 2% binders (according to NetMHCIIpan-3.2 threshold) were used to perform a clustering on the pairwise correlations between two alleles using complete observations only. We show clustering results for peptide set 2. Labels were colored according to risk (red) or protective (blue). (BINDING MOTIFS): Top 2% binders

were combined for proteins (RISK 1) DRB1*11:01/04 and DRB1*13:01 DRB1*12:01, DRB1*14:04 and DRB1*15:01/03 (RISK 2), DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and DRB1*10:01 (PROT 1) and DRB1*04:03/04/06 (PROT 2). For this analysis shared peptides (10% top binders) between at least two of the groups where deleted from the set. Here we depict the results for human peptide set 2. Peptide motifs were plotted using Seq2Logo.[28]. The color scheme shows the chemistry of the amino acids. *Red*: positively charged amino acids, *blue*: negatively charged amino acids, *green*: polar amino acid, *purple*: neutral amino acid and *black*: hydrophobic amino acid.

Figure 5 – Cluster according to chosen physico-chemical properties of amino acids within the peptide binding pockets.

We only show sites with variable information in pockets (P) 1, 4, 6, 7 and 9 and only proteins for which the genetic analysis was significant (meta-analysis RE2Cp* <0.05) and for which at least 1 cohort had AF >1%. We additionally show DRB1*01:03. Clustering was performed using the hclust function of the R package stats. The box below the cluster plot shows positions of P1, 4, 6, 7 and 9 of the beta (B) chain of the molecules (as defined in **Supplementary Table 4**). Here we show combined scores F1 **(a)** and F3 **(b)** derived from a factor analysis of 54 unique amino acid properties (Atchley *et al*.[29]). F1 captures polarity and hydrophobicity of the amino acid, while factor F3 captures amino acid size and bulkiness. For F1, high values indicate larger hydrophobicity, polarity and hydrogen donor abilities while low values indicate non-polar amino acids. For F3, high values indicate larger and bulkier amino acids while low values indicate smaller, more flexible amino acids. We additionally show the residue-volume **(c)** as a measure of pocket size and defined a score "hydrogen acceptor" (HB-acceptor) **(d)**, which defines the ability of amino acids to participate in hydrogen bonds and corresponds to the number of atoms within amino acid sidechains accept a hydrogen. Additional information for the "charge" parameter and the analysis for DQA1-DQB1 can be found in **Supplementary Figures 9,10**.

Figure 6 – Frequency of DRB1*01:03 across populations available in the allele frequency net database. (a) "Worldmap", **(b)** zoom into European continent. Frequencies are shown within different ranges noted by AF. Allele frequencies of DRB1*01:03 are lower across central Europe than in the UK, Spain, India, South Africa, United States, and coastal regions of South America. Frequencies were binned according to allele frequency. The figures were created using the R-package rworldmap. Frequencies were extracted from the allele frequency network database[33] for populations larger than 100 individuals. To plot the geographic locations, we converted assigned degree and minutes to decimal numbers. We deleted all non-Caucasian populations with USA coordinates prior to plotting.

25

REFERENCES

1. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
2. 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).
3. 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).
4. 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).
5. Stokkers, P. C., Reitsma, P. H., Tytgat, G. N. & van Deventer, S. J. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* **45**, 395–401 (1999).
6. Lappalainen, M. *et al.* Association of IL23R, TNFRSF1A, and HLA-DRB1*0103 allele variants with inflammatory bowel disease phenotypes in the Finnish population. *Inflamm Bowel Dis* **14**, 1118–1124 (2008).
7. Lu, M. & Xia, B. Polymorphism of HLA-DRB1 gene shows no strong association with ulcerative colitis in Chinese patients. *Int J Immunogenet* **33**, 37–40 (2006).
8. Okada, Y. *et al.* HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* **141**, 864–865 (2011).
9. Myung, S. J. *et al.* HLA-DRB1*1502 confers susceptibility to ulcerative colitis, but is negatively associated with its intractability: a Korean study. *Int J Color. Dis* **17**, 233–237 (2002).
10. Mohammadi, M. *et al.* Association of HLA-DRB1 Alleles with Ulcerative Colitis in the City of Kerman, South Eastern Iran. *Iran J Allergy Asthma Immunol* **14**, 306–312 (2015).
11. Gao, F. *et al.* Association of HLA-DRB1 alleles and anti-neutrophil cytoplasmic antibodies in Han and Uyghur patients with ulcerative colitis in China. *J Dig Dis* **15**, 299–305 (2014).
12. Uyar, F. A. *et al.* The distribution of HLA-DRB alleles in ulcerative colitis patients in Turkey. *Eur*

J Immunogenet **25**, 293–296 (1998).

13. Han, B. *et al.* Amino acid position 37 of HLA-DRβ1 affects susceptibility to Crohn's disease in Asians. *Hum. Mol. Genet.* (2018). doi:10.1093/hmg/ddy285
14. 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.* (2019). doi:10.1093/hmg/ddy443
15. Huang, C. *et al.* Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans. *Gastroenterology* **149**, 1575–1586 (2015).
16. Ye, B. D. *et al*. Identification of Ten Additional Susceptibility Loci for Ulcerative Colitis Through Immunochip Analysis in Koreans. *Inflamm. Bowel Dis.* (2016). doi:10.1097/MIB.0000000000000584
17. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
18. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**, 116–126 (2016).
19. Zheng, X. *et al*. HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
20. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2011).
21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
22. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457–470 (2011).
23. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
24. Gourraud, P. A. *et al*. HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282 (2014).
25. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol*

35, 809–822 (2011).

26. Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379–i388 (2017).

27. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* (2018). doi:10.1111/imm.12889

28. 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.* (2012). doi:10.1093/nar/gks469

29. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0408677102

30. Goldsack, D. E. & Chalifoux, R. C. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *J. Theor. Biol.* (1973). doi:10.1016/0022-5193(73)90075-1

31. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28**, 1598–1603 (2012).

32. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

33. Gonzalez-Galarza, F. F. *et al*. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res* **43**, D784-8 (2015).

34. Hollenbach, J. A. & Oksenberg, J. R. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun* **64**, 13–25 (2015).

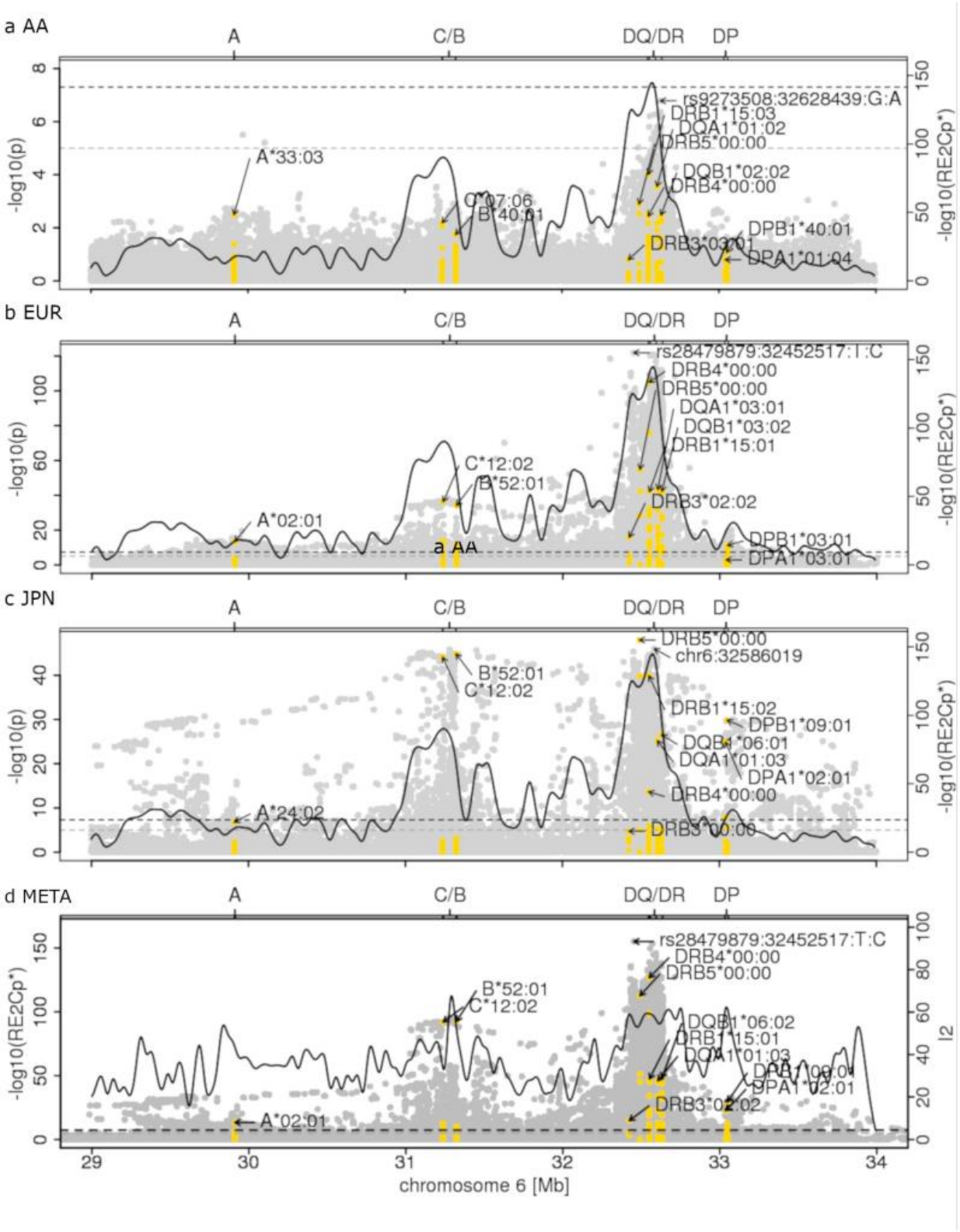

a AA
A
C/B
DQ/DR
DP
-log10(p)
-log10(RE2Cp*)
rs9273508:32628439:G:A
DRB1*15:03
DQA1*01:02
DRB5*00:00
DQB1*02:02
DRB4*00:00
A*33:03
C*07:06
B*40:01
DPB1*40:01
DRB3*03:01
DPA1*01:04
b EUR
rs28479879:32452517:T:C
DRB4*00:00
DRB5*00:00
DQA1*03:01
DQB1*03:02
DRB1*15:01
C*12:02
B*52:01
DRB3*02:02
A*02:01
DPB1*03:01
DPA1*03:01
a AA
c JPN
DRB5*00:00
chr6:32586019
B*52:01
C*12:02
DRB1*15:02
DPB1*09:01
DQB1*06:01
DQA1*01:03
DPA1*02:01
DRB4*00:00
A*24:02
DRB3*00:00
d META
I2
rs28479879:32452517:T:C
DRB4*00:00
DRB5*00:00
B*52:01
C*12:02
DQB1*06:02
DRB1*15:01
DQA1*01:03
DPB1*09:01
DPA1*02:01
DRB3*02:02
A*02:01
29
30
31
32
33
34
chromosome 6 [Mb]

Figure 1

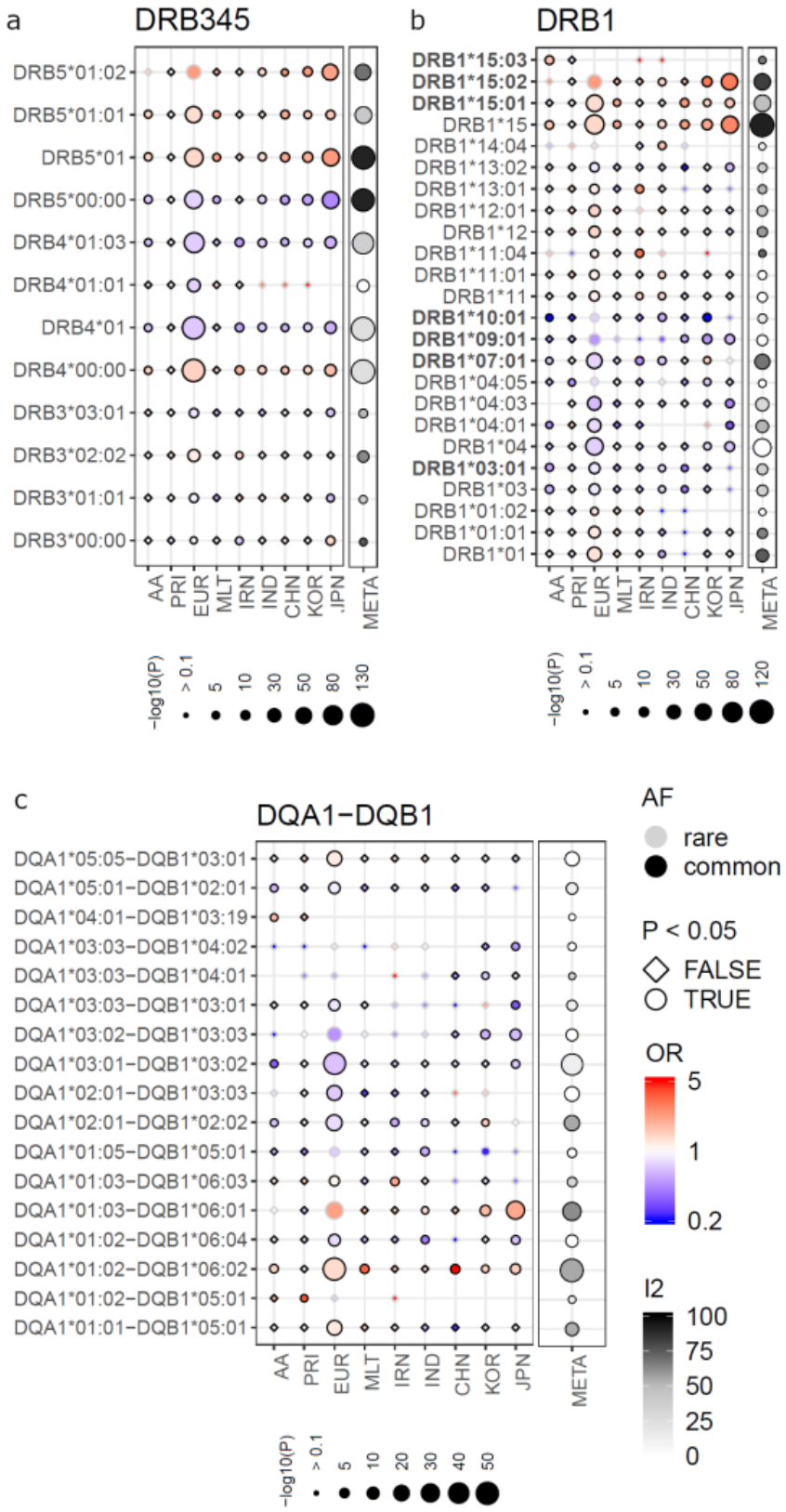

a
DRB345
DRB5*01:02
DRB5*01:01
DRB5*01
DRB5*00:00
DRB4*01:03
DRB4*01:01
DRB4*01
DRB4*00:00
DRB3*03:01
DRB3*02:02
DRB3*01:01
DRB3*00:00
AA
PRI
EUR
MLT
IRN
IND
CHN
KOR
JPN
META
−log10(P)
> 0.1
5
10
30
50
80
130
b
DRB1
DRB1*15:03
DRB1*15:02
DRB1*15:01
DRB1*15
DRB1*14:04
DRB1*13:02
DRB1*13:01
DRB1*12:01
DRB1*12
DRB1*11:04
DRB1*11:01
DRB1*11
DRB1*10:01
DRB1*09:01
DRB1*07:01
DRB1*04:05
DRB1*04:03
DRB1*04:01
DRB1*04
DRB1*03:01
DRB1*03
DRB1*01:02
DRB1*01:01
DRB1*01
AA
PRI
EUR
MLT
IRN
IND
CHN
KOR
JPN
META
−log10(P)
> 0.1
5
10
30
50
80
120
c
DQA1−DQB1
DQA1*05:05−DQB1*03:01
DQA1*05:01−DQB1*02:01
DQA1*04:01−DQB1*03:19
DQA1*03:03−DQB1*04:02
DQA1*03:03−DQB1*04:01
DQA1*03:03−DQB1*03:01
DQA1*03:02−DQB1*03:03
DQA1*03:01−DQB1*03:02
DQA1*02:01−DQB1*03:03
DQA1*02:01−DQB1*02:02
DQA1*01:05−DQB1*05:01
DQA1*01:03−DQB1*06:03
DQA1*01:03−DQB1*06:01
DQA1*01:02−DQB1*06:04
DQA1*01:02−DQB1*06:02
DQA1*01:02−DQB1*05:01
DQA1*01:01−DQB1*05:01
AA
PRI
EUR
MLT
IRN
IND
CHN
KOR
JPN
META
−log10(P)
> 0.1
5
10
20
30
40
50
AF
rare
common
P < 0.05
FALSE
TRUE
OR
5
1
0.2
I2
100
75
50
25
0

Figure 2

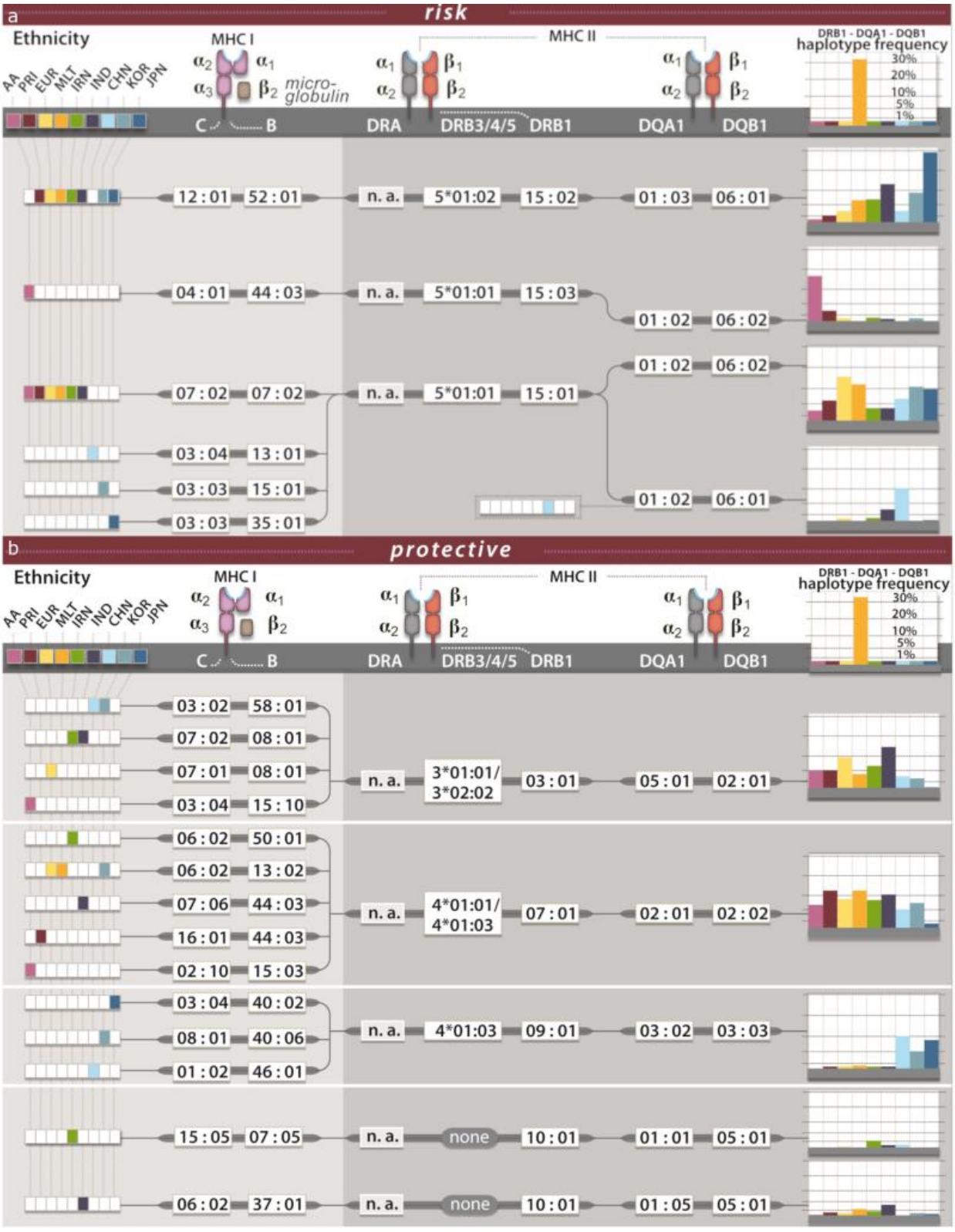

a
risk
Ethnicity
AA PRI EUR MLT IRN IND CHN KOR JPN
MHC I
MHC II
micro-globulin
C B
DRA
DRB3/4/5
DRB1
DQA1
DQB1
DRB1 - DQA1 - DQB1 haplotype frequency
30%
20%
10%
5%
1%
12 : 01 52 : 01 n. a. 5*01:02 15 : 02 01 : 03 06 : 01
04 : 01 44 : 03 n. a. 5*01:01 15 : 03 01 : 02 06 : 02
07 : 02 07 : 02 n. a. 5*01:01 15 : 01 01 : 02 06 : 02
03 : 04 13 : 01
03 : 03 15 : 01
03 : 03 35 : 01
01 : 02 06 : 01
b
protective
Ethnicity
AA PRI EUR MLT IRN IND CHN KOR JPN
MHC I
MHC II
C B
DRA
DRB3/4/5
DRB1
DQA1
DQB1
DRB1 - DQA1 - DQB1 haplotype frequency
30%
20%
10%
5%
1%
03 : 02 58 : 01
07 : 02 08 : 01
07 : 01 08 : 01
03 : 04 15 : 10
n. a. 3*01:01/ 3*02:02 03 : 01 05 : 01 02 : 01
06 : 02 50 : 01
06 : 02 13 : 02
07 : 06 44 : 03
16 : 01 44 : 03
02 : 10 15 : 03
n. a. 4*01:01/ 4*01:03 07 : 01 02 : 01 02 : 02
03 : 04 40 : 02
08 : 01 40 : 06
01 : 02 46 : 01
n. a. 4*01:03 09 : 01 03 : 02 03 : 03
15 : 05 07 : 05 n. a. none 10 : 01 01 : 01 05 : 01
06 : 02 37 : 01 n. a. none 10 : 01 01 : 05 05 : 01

Figure 3

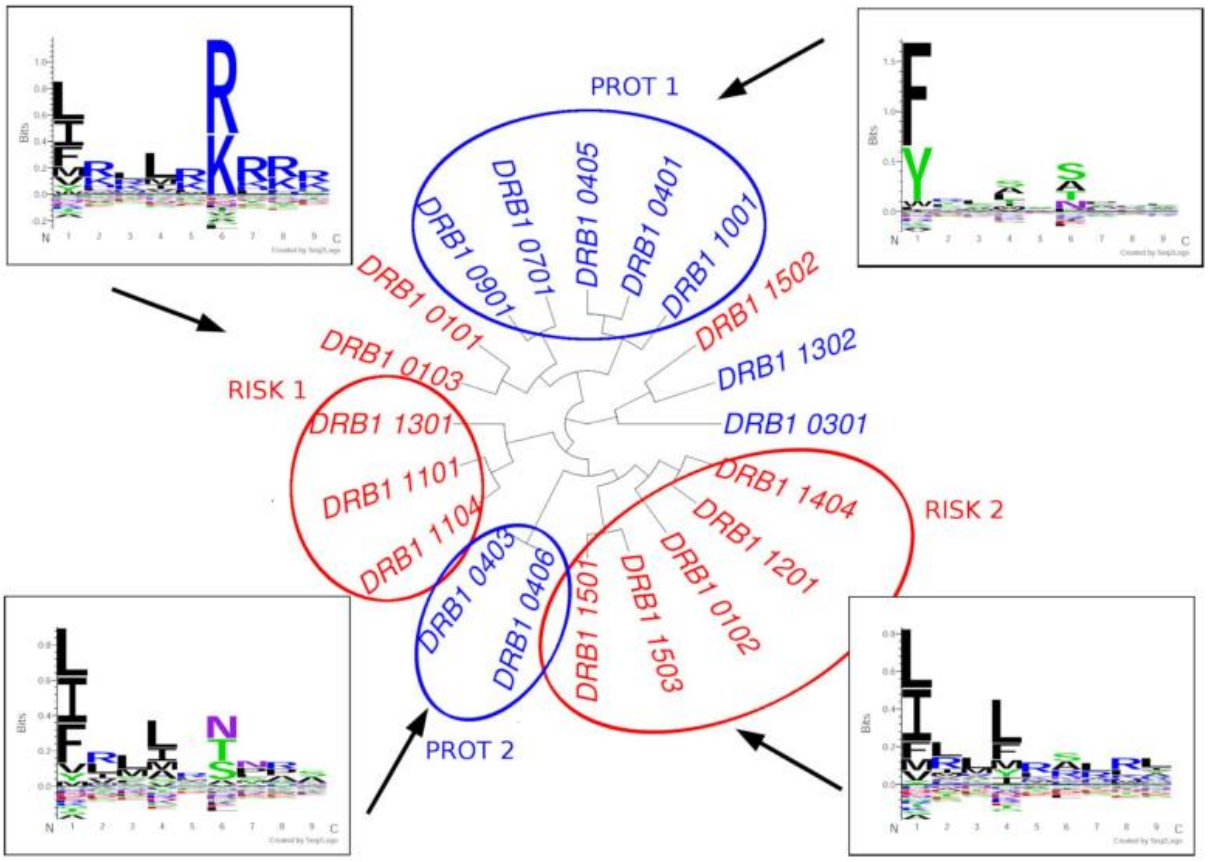

PROT 1
DRB1 0901
DRB1 0701
DRB1 0405
DRB1 0401
DRB1 1001
DRB1 1502
DRB1 1302
DRB1 0301
DRB1 0101
DRB1 0103
RISK 1
DRB1 1301
DRB1 1101
DRB1 1104
DRB1 1404
RISK 2
DRB1 1201
DRB1 0102
DRB1 1503
DRB1 1501
DRB1 0403
DRB1 0406
PROT 2

Figure 4

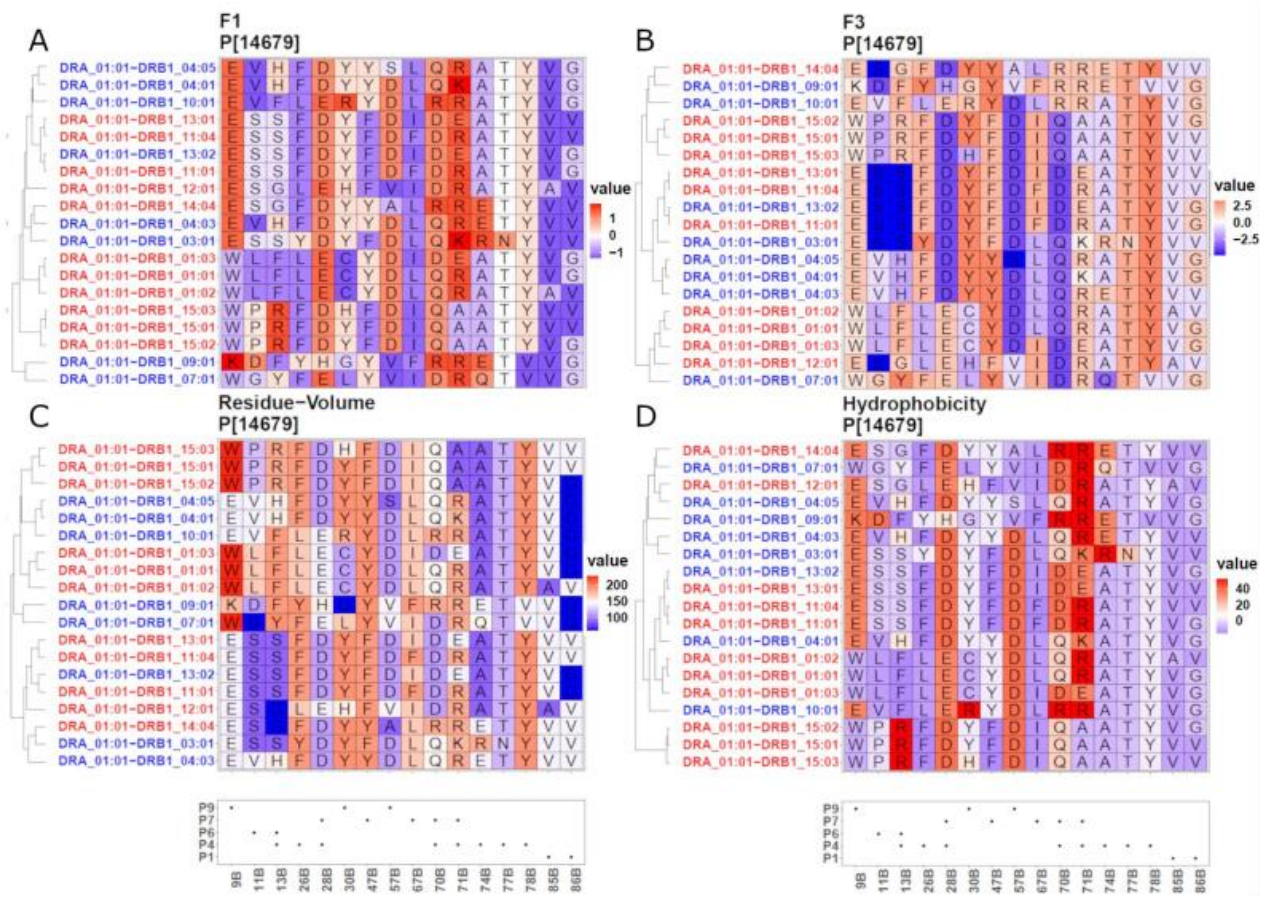

A
F1
P[14679]
B
F3
P[14679]
C
Residue−Volume
P[14679]
D
Hydrophobicity
P[14679]
value

Figure 5

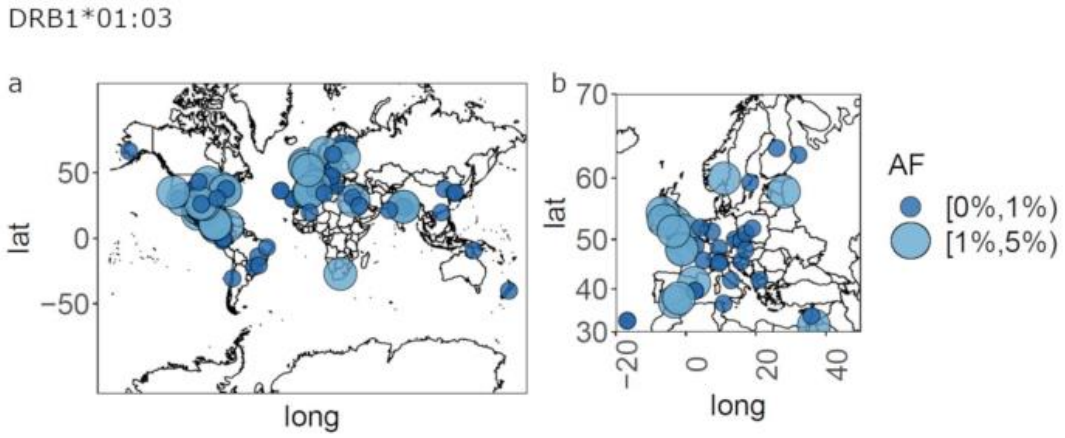


Figure 6

5.3. Additional results

5.3.1. HLA haplotype maps

I developed a graphical illustration of HLA haplotypes (HLA haplotype maps) in the context of **Paper C** but did not include these maps due to the limitations of the journal. Correlation between multiallelic variants cannot be adequately captured by LD. This is shown also in Table 6. In addition to using classical LD measures for the HLA alleles, the relative frequency h of co-occurrence of two HLA alleles was calculated as stated in Equation 5.

Equation 5

$$\mathrm{h}(\mathrm{A},\mathrm{B}) = \frac{\mathrm{p_{AB}}}{\mathrm{p_B}} \qquad \text{and} \qquad \mathrm{h}(\mathrm{B},\mathrm{A}) = \frac{\mathrm{p_{AB}}}{\mathrm{p_A}}$$

Here, the relative frequency of the co-occurrence of an allele at locus A with an allele at locus B is calculated conditionally on the allele at locus B and vice versa. $\mathrm{p_{AB}}$ denotes the true haplotype frequency of the haplotype formed by the allele at locus A and the allele at locus B. $\mathrm{p_A}$ and $\mathrm{p_B}$ denote the frequencies of each allele.

Table 6 – Explanation of Equation 5 with an example.
As the allele of reference DPA1*01:03-DPB1*01:01 was chosen. The LD (r.sq) and the 95% confidence interval, given as the 2.50% and 97.50% percentiles, are shown. The frequency of the HLA-*DRB1* allele is given as “a”, and the frequency of DPA1*01:03-DPB1*01:01 is given as “b”. “ab” denotes the haplotype frequency of both alleles. Data was calculated for the Caucasian (EUR) population including both UC cases and healthy controls and a total of 31,112 samples with a phasing probability >0.8 across all classical HLA loci (excluding HLA-*DRB3/4/5*). ab/a and ab/b are the conditional frequencies.

DRB1	r.sq	2.50%	97.50%	a	b	ab	ab/a	ab/b
01:01	3.02E-03	3.31E-03	2.71E-03	1.13E-01	2.70E-02	2.41E-04	2.13E-03	8.92E-03
01:02	3.55E-04	3.84E-04	2.69E-04	1.38E-02	2.70E-02	1.61E-05	1.17E-03	5.95E-04
03:01	**1.93E-01**	2.03E-01	1.83E-01	1.05E-01	2.70E-02	2.47E-02	2.35E-01	**9.14E-01**
04:01	1.91E-03	2.25E-03	1.58E-03	9.65E-02	2.70E-02	5.14E-04	5.33E-03	1.90E-02
04:02	1.38E-04	2.06E-04	5.27E-05	8.12E-03	2.70E-02	4.82E-05	5.94E-03	1.78E-03
04:03	6.36E-04	7.89E-04	4.71E-04	3.00E-02	2.70E-02	1.13E-04	3.76E-03	4.16E-03
04:06	1.01E-07	1.09E-04	2.79E-09	6.40E-04	2.70E-02	1.61E-05	2.50E-02	5.95E-04
04:07	8.97E-05	1.03E-04	2.78E-05	4.32E-03	2.70E-02	1.61E-05	3.72E-03	5.95E-04
07:01	2.80E-03	3.15E-03	2.42E-03	1.20E-01	2.70E-02	4.50E-04	3.76E-03	1.66E-02
08:01	4.96E-04	6.53E-04	3.37E-04	2.60E-02	2.70E-02	1.29E-04	4.94E-03	4.76E-03
09:01	1.37E-04	2.22E-04	4.37E-05	9.02E-03	2.70E-02	6.43E-05	7.13E-03	2.38E-03
10:01	1.46E-04	1.90E-04	6.44E-05	7.39E-03	2.70E-02	3.21E-05	4.35E-03	1.19E-03
11:01	2.49E-03	2.75E-03	2.23E-03	9.38E-02	2.70E-02	1.77E-04	1.88E-03	6.54E-03
11:04	3.43E-04	3.72E-04	2.60E-04	1.34E-02	2.70E-02	1.61E-05	1.20E-03	5.95E-04
12:01	5.32E-04	6.21E-04	4.17E-04	2.21E-02	2.70E-02	4.82E-05	2.18E-03	1.78E-03
13:01	1.53E-03	1.77E-03	1.29E-03	6.78E-02	2.70E-02	2.41E-04	3.55E-03	8.92E-03
13:02	1.03E-03	1.18E-03	8.72E-04	4.14E-02	2.70E-02	8.04E-05	1.94E-03	2.97E-03
14:54	6.60E-04	7.31E-04	5.45E-04	2.55E-02	2.70E-02	3.21E-05	1.26E-03	1.19E-03
15:01	4.34E-03	4.59E-03	4.07E-03	1.39E-01	2.70E-02	6.43E-05	4.62E-04	2.38E-03
16:01	2.08E-04	2.55E-04	1.25E-04	9.64E-03	2.70E-02	3.21E-05	3.33E-03	1.19E-03

To compare the relative conditional frequencies of the co-occurrence of two HLA alleles to the standard LD measure, the LD was calculated as described in Equation 1. The 95% confidence interval of the LD measure was calculated on 1,000 random subsamples of alleles at locus A and alleles at locus B. Unlike the LD, the percentage calculated with Equation 5 is not symmetrical and depends on the allele used as reference. Comparing the values computed for the LD of DRB1*03:01 and DPA1*01:03-DPB1*01:01 to their conditional frequency in Table 6 shows that DRB1*03:01 is located on the same haplotype as DPA1*01:03-DPB1*01:01 in 91.4% of all haplotypes the DRB1*03:01 is observed in. The LD of 0.193 [0.183, 0.203] is much lower (0.193 compared to 0.914), because DPA1*01:03-DPB1*01:01 is also observed with other HLA-*DRB1* alleles.

Figures 25 to 27 show the graphical interpretation of results as displayed in Table 6 for the allele group DRB1*15, which was observed to be associated with UC across ethnicities in **Paper C**. Here, only individuals for which all loci (HLA-*A*, -*B*, -*C*, HLA loci-*DPA1*, -DPB1, -*DQA1*, -*DQB1* as well as -*DRB1*) had a phasing certainty of >0.8 were considered. The analysis was carried out irrespective of case-control status. In total 641 African American (AA), 198 Chinese (CHN), 31,112 Caucasian (EUR), 1,203 Indian (IND), 484 Iranian (IRN), 2,862 Japanese (JPN), 1,021 Korean (KOR), 236 Maltese (MLT) and 472 Puerto Rican (PRI) individuals had a phasing probability >0.8 across all analysed loci.

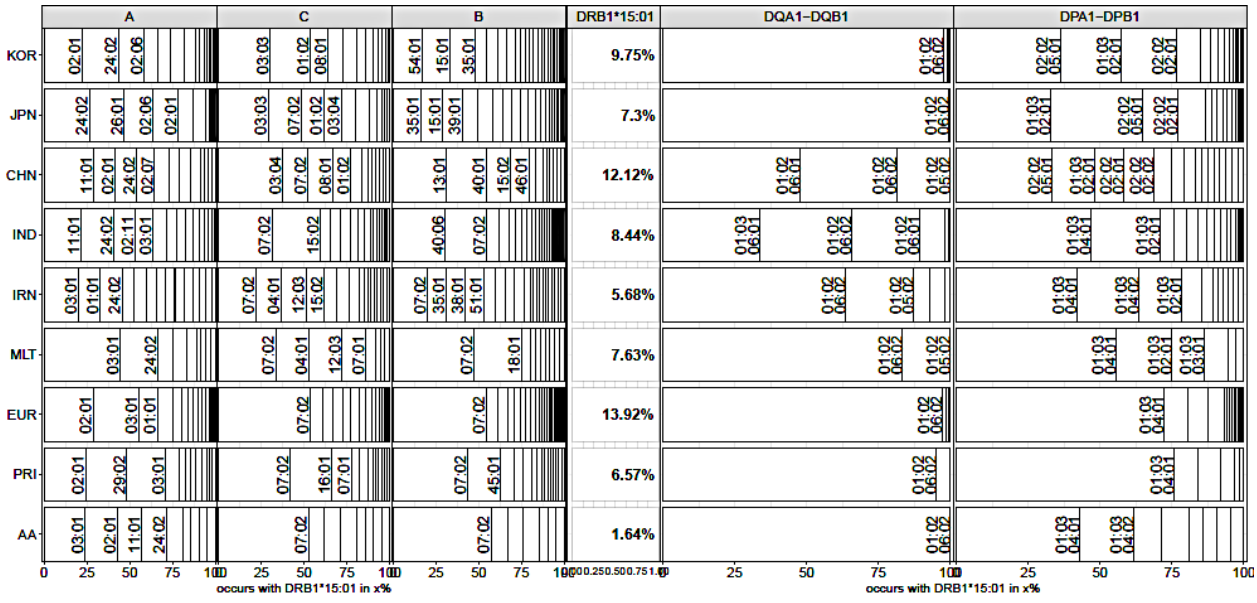


Figure 25 – Example haplotype plot for DRB1*15:01.
The given DRB1*15:01 frequencies were calculated across all individuals of the population irrespective of case-control status. Loci are sorted by genomic location. (AA: African American, PRI: Puerto Rican, EUR: Caucasian, MLT: Maltese, IND: Indian, IRN: Iranian, CHN: Chinese, JPN: Japanese, KOR: Korean).

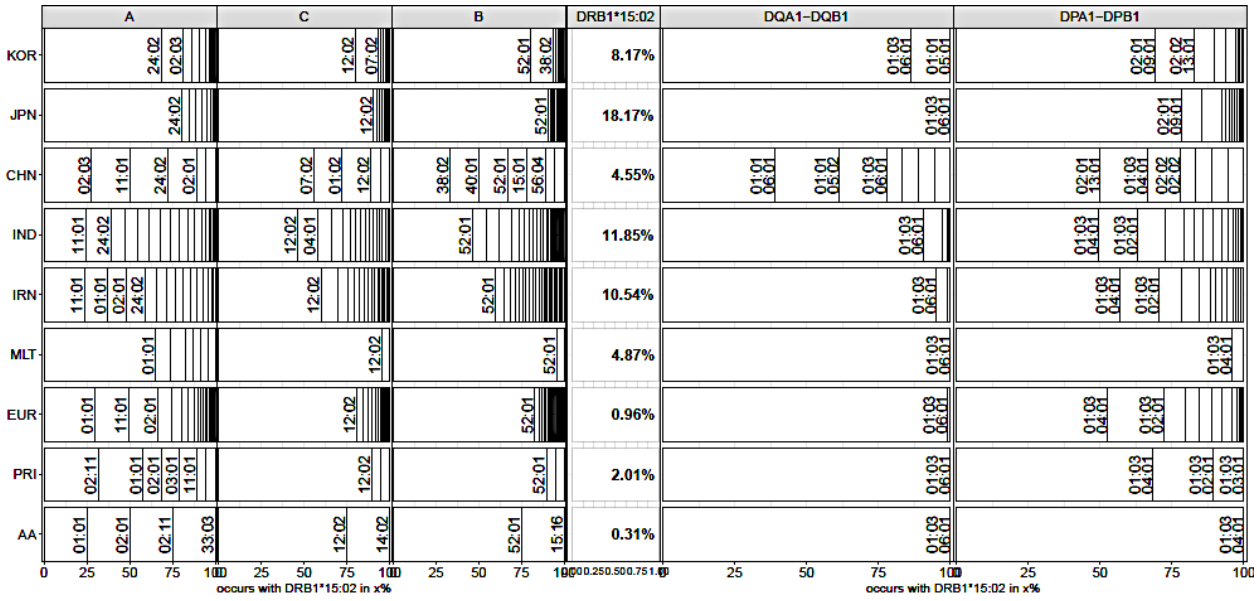


Figure 26 – Example haplotype plot for DRB1*15:02.
The given DRB1*15:02 frequencies were calculated across all individuals of the population irrespective of case-control status. Loci are sorted by genomic location. (AA: African American, PRI: Puerto Rican, EUR: Caucasian, MLT: Maltese, IND: Indian, IRN: Iranian, CHN: Chinese, JPN: Japanese, KOR: Korean).

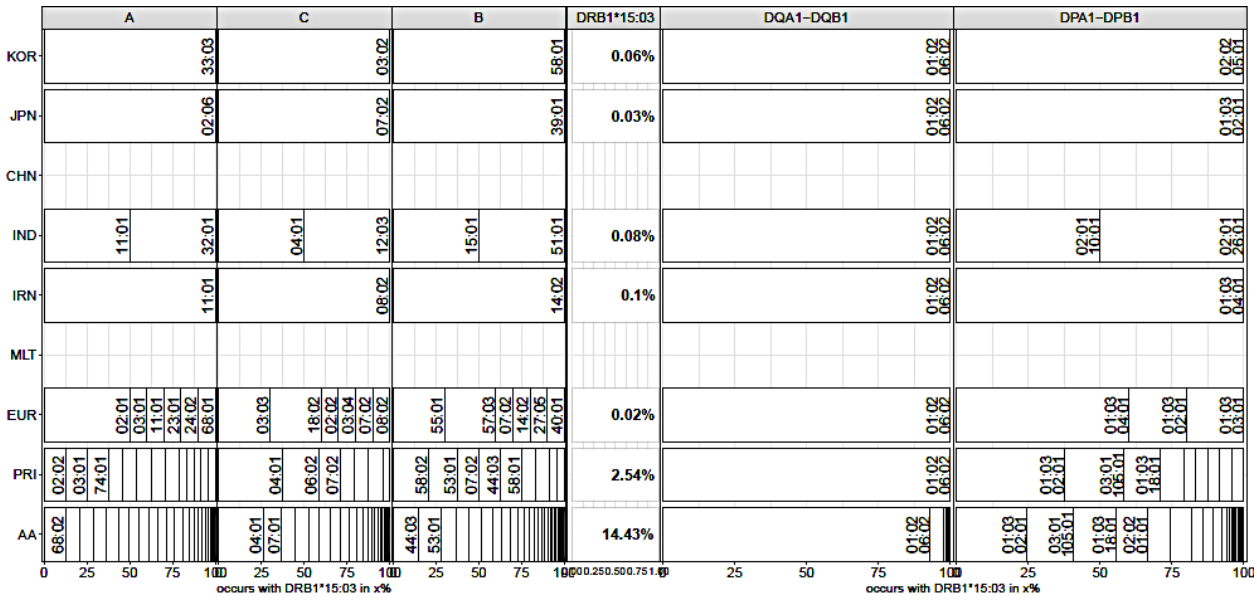


Figure 27 – Example haplotype plot for DRB1*15:03.
The given DRB1*15:03 frequencies were calculated across all individuals of the population irrespective of case-control status. Loci are sorted by genomic location. (AA: African American, PRI: Puerto Rican, EUR: Caucasian, MLT: Maltese, IND: Indian, IRN: Iranian, CHN: Chinese, JPN: Japanese, KOR: Korean).

In these plots, the haplotype co-occurrence of alleles at non-reference loci with the reference allele, is indicated by the size of the boxes along the x-axes. The allele of reference is shown in the middle of

the plot, together with its frequency across all analysed individuals, respective to the individual cohort. Note that these frequencies may be different from the frequencies calculated in the single association statistics of the reference allele, as only individuals with complete haplotype information across all loci are considered. These frequencies, in addition to the marginal probabilities calculated for individual alleles (i.e. a measure how accurately each allele could be imputed) in **Paper C**, can be used to aid interpretation of the maps (i.e. the rarer an allele in a population, the more likely imputation artefacts will impact the results shown in the map). In addition, HLA-*DRB3/4/5* alleles located on the same haplotype as specific HLA-*DRB1* alleles are shown in Supplementary Table 10 of **Paper B**. These alleles cannot be accurately phased, because they can either be present or absent in an individual, which is something phasing cannot accommodate for.

5.3.2. HLA imputation pipeline

Scripts written for the HLA imputation and phasing in **Paper C** were incorporated into a single HLA imputation pipeline. In collaboration with Mareike Wendorff and Marc Höppner, this pipeline was written to be used with *Nextflow*, a bioinformatic workflow manager that is publicly available as a github repository. A flowchart detailing the steps of the pipeline is shown in Figure 28.

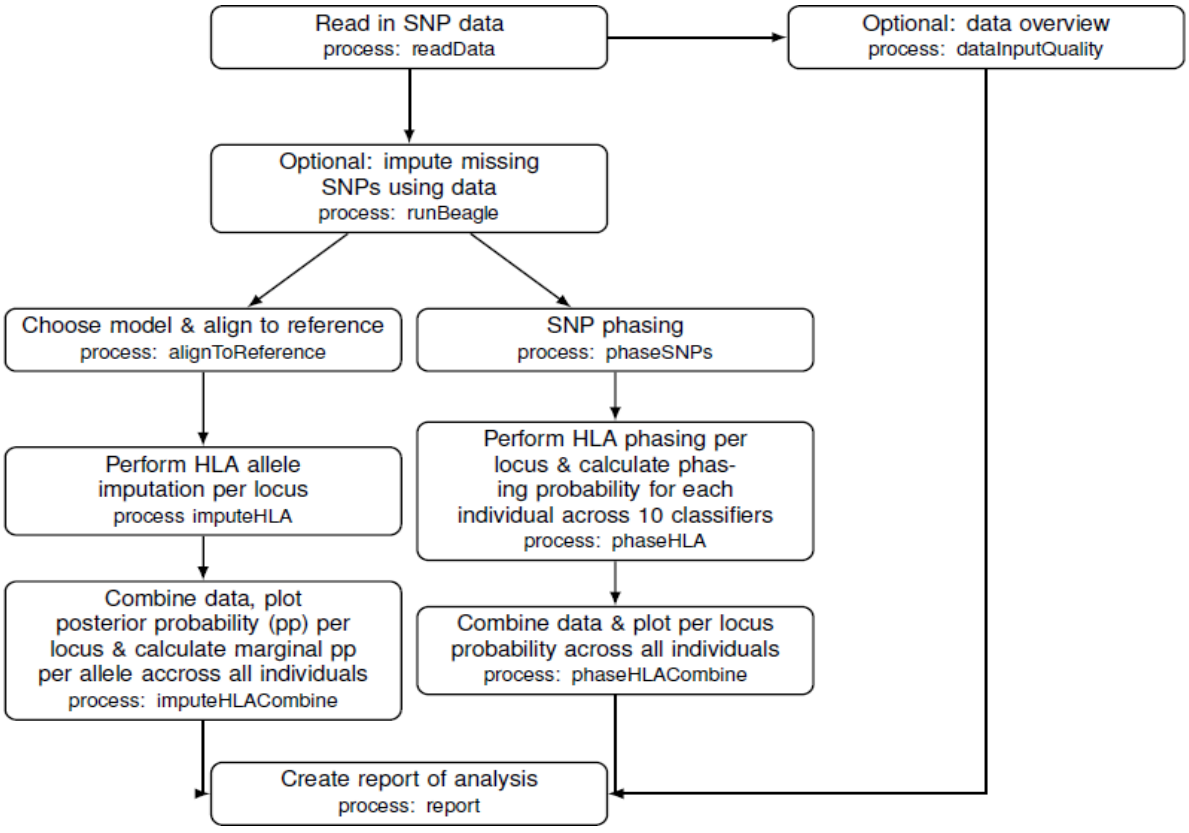


Figure 28 – HLAPipe workflow.

Detailed information on methods used in this pipeline are described in **Paper C**.
The main scripts of this pipeline were written in R and modified from scripts used in **Paper C**. This pipeline also incorporates some of the R scripts written for other projects during the time of the production of this thesis, that were written to complement a general GWAS quality control pipeline. The *Nextflow* workflow will be available on github as ikmb/HLAPipe.

Across the workflow, plus-strand annotation and genome build hg19 are assumed. In the workflow, the input of a dataset (readData) is followed by the optional evaluation of the quality of the submitted dataset (dataInputQuality). This includes a check for genotyping call rate (per variant statistics that gives information on the success of genotyping of this variant across all analysed individuals), sample call rate (per individual statistic calculated across all variants) and concordance of observed and genotyped gender (if chromosomes X and Y are present in the study dataset). dataInputQuality is only performed if at least information for chromosomes 1 to 22 is present in the input data. A second optional step includes the imputation of missing SNPs into the study dataset using the publicly available software Beagle[106] (runBeagle). Here SNPs are imputed for individuals that were not assigned a genotype at a respective SNP from information that is available for other individuals in the study dataset. Next, genotype data are aligned to a chosen HLA reference (alignToReference). Variants that do not match between reference and input dataset are excluded. ATCG variants (i.e. A/T and C/G variations) are excluded if the alleles do not match to the reference (on assumed +-strand notation) and the frequency of the ATCG variant is >0.4. ATCG variants are also excluded if their frequencies between reference and study differ by more than a value of 0.2. HLA alleles are inferred using HIBAG[92] and a chosen reference model that can be used with HIBAG. Only alleles at loci that are present in the reference can be imputed (imputeHLA). The user can also select specific loci present in the reference. Imputation is performed for each locus in parallel. Data are combined and a plot detailing the posterior probability of imputation, as shown in Figure 3 of **Paper B**, is generated. Additionally, marginal imputation probabilities per allele are calculated as described in the Methods section of **Paper C** (impute HLACombine). SNPs across the HLA are phased (phaseSNPs) and phasing certainty is calculated based on 100 random samples of the haplotype graph (*hgraph) as produced by SHAPEIT2[107]. Variants with a median certainty of 0.8 across the 100 subsamples are excluded from the data. HLA allele haplotypes are phased based on SNP haplotypes generated with HIBAG and SHAPEIT2 as described in the Methods section of **Paper C**. Phasing probabilities are calculated across 10 classifiers of the HIBAG model and subsequently plotted (phaseHLA on gene level, phaseHLACombine combines these data). Plots and detailed information on the analysis and tool version are than submitted to a report HTML file.

6. Discussion

The aim of this thesis was the characterisation of the genetic association of UC with the highly complex HLA region in a trans-ancestry context. By leveraging the differences in LD, one goal was also to assign causal variants to the association signal. To achieve this aim, an HLA imputation approach was chosen. In **Paper B**, we compiled a trans-ancestry HLA imputation reference to optimally cover the allelic diversity of the cohorts we analysed in the fine mapping of the HLA in **Paper C**. In **Paper C**, we observed that whole HLA-*DR-DQ* haplotypes were associated with UC even across different ancestries and were therefore unable to pinpoint causal variation to either locus. However, other interesting and valuable conclusions could be drawn from this analysis as discussed in detail the respective publications (Chapter 5.1 and Chapter 5.2) and summarized in the overview table in Chapter 3. As one of the main results of **Paper C**, alleles of the DRB1*15 group were associated with UC across populations of different ancestries. The previously described DRB1*01:03 was identified to be population-specific and not present in non-Caucasian populations. We also showed that proteins translated from risk alleles are predicted to bind peptides that contain the amino acids arginine and lysine. Here, I want to briefly address some additional points regarding **Paper B** and **Paper C**.

In **Paper C**, we analysed individuals of African American, Caucasian, Chinese, Indian, Iranian, Japanese, Korean and Puerto Rican descent. The Caucasian individuals were recruited in Germany and Malta. We used a subgroup of samples from the same study populations to construct our imputation reference panel. For the Puerto Rican population, no DNA samples were available for NGS typing in **Paper B**. Therefore, the imputation reference was constructed for 8 of the 9 analysed populations. In **Paper B**, we showed that our imputation reference could be used for highly accurate HLA imputation within these cohorts and at good quality also for the Puerto Rican population of the 1000 Genomes population (Table 2 of **Paper B**).

In **Paper C**, we used the constructed reference, to impute HLA alleles into a large case-control dataset for fine mapping of the HLA in UC. One of the main findings of **Paper C** is, that alleles of the DRB1*15 group are associated with UC across different ethnicities. For South Western Asian (Iranian, Indian) individuals other HLA associations were observed to be more dominant (i.e. HLA-DRB1*11 and HLA-DRB1*14). Additionally, we identified HLA-DRB1*01:03, previously reported as the most associated risk allele, as specific to subpopulations of the Caucasian population. This indicates that there is genetic variation of the HLA even among the Caucasian population. DRB1*01:03 was not present in non-Caucasian populations. This warrants the need for either stratification by subpopulations or the consideration of HLA allele frequencies, as reported for instance in the allele frequency net database. While HLA fine mapping studies have concentrated on the analysis of single amino acid changes in

the past, we analysed the amino acid properties of the whole peptide binding groove in **Paper C**. This approach better captures the actual biology of HLA-peptide binding and was also applied by Goyette *et al.*[1], who conducted an HLA fine mapping in the Caucasian population also included in **Paper C**. In their analysis, alleles associated with both CD and UC as well as common alleles not associated with the disease were analysed in respect to the electrostatic properties within the binding groove of the resulting proteins. In **Paper C**, we performed a similar analysis on alleles associated with UC only in a trans-ancestry setting, using a more diverse set of electro-chemical properties of amino acids across 5 pockets of the binding groove. The considered pockets 1, 4, 6, 7 and 9 come into close contact with the peptide. We largely reproduced results by Goyette *et al.*[1] and likewise found that risk and protective proteins cluster regarding their amino acid properties. To take this analysis a step further, we additionally analysed the properties of the peptides that are preferentially bound by risk proteins and showed for the first time that these peptides are enriched for the negatively charged amino acids lysine and arginine. This finding could aid the identification of culprit antigens in the future.

Below, the potential roles of arginine and lysine in UC are discussed (Chapter 6.1). What is currently known on the actual expression of HLA alleles in the gut is described in Chapter 6.2. Alleles of the DRB1*15 group were associated with UC across different ancestries. Though little is known on functional implications on the genetic level, alleles of these groups have also been associated with other diseases, which is described in Chapter 6.3. DRB1*15 has not been associated with CD, the other subform of IBD. Potential reasons for this are discussed below (Chapter 6.3). This Chapter ends with the description of some limitations of the methods used in **Paper B** and **Paper C** (Chapter 6.4).

6.1. Potential roles of arginine and lysine in UC

In **Paper C**, we showed that peptides bound preferentially by HLA proteins derived from HLA risk alleles to be rich in arginine (R) and lysine (K) after excluding peptides that were HLA proteins derived from both protective and risk allele groups. Incidentally, peptides rich in these amino acids have been shown to be good T cell epitopes, i.e. can elicit an immune response[108,109]. Dhanda *et al.*[106] analysed 1,032 known epitopes from 14 different sources (including *Mycobacterium tuberculosis*, Dengue Fever Virus, Zika Fever Virus, house mite and other allergens) and known non-epitopes from the same dataset. They compared amino acid distributions between these epitopes and non-epitopes of 15 amino acids length. The results of this analysis, taken from the publication, are shown in Figure 29. The displayed motif is remarkably similar to the combined peptide binding motifs of DRB1*11:01/04 and DRB1*13:01, specifically (Figure 4 in **Paper C**).

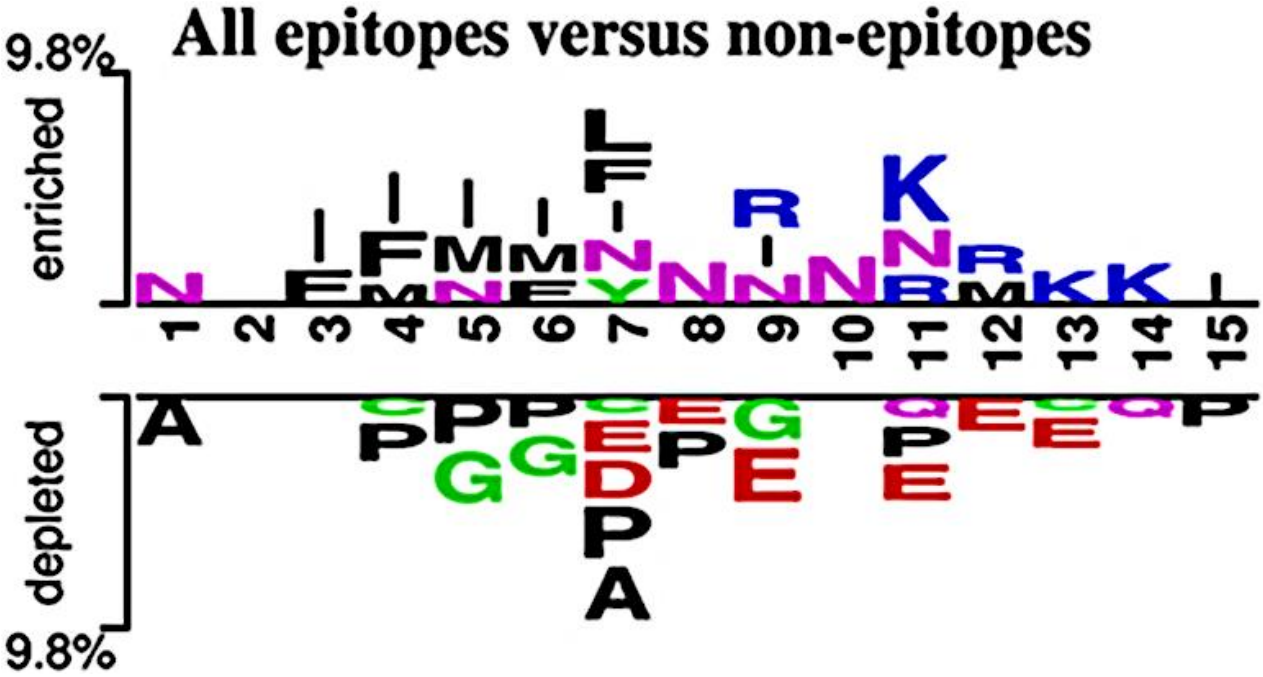


Figure 29 – Epitope vs. non epitope properties.
Figure taken from Dhanda *et al.*[108]. Original description: Figure 4. Two-sample logo created using epitopes and non-epitopes in all the data (p-value <0.01). The immunogenicity motifs for epitopes and non-epitopes were derived from the combination of all the datasets.

What occurs at T cell level when T cells bind to peptides rich in arginine and lysine (i.e. which types of responses are produced in the host) as opposed to peptides depleted in these amino acids, has to my knowledge not been studied so far. Arginine has been reported to be an important agent in the immune system and has the capacity to dampen inflammation[110]. Though this effect is most likely independent of possible arginine-rich peptides driving a pro-inflammatory signal in IBD, a controversial hypothesis may be the formation of short RRR or KKK peptides during inflammation. Huan *et al.* showed, that shorter fragments of peptides could combine in the HLA peptide binding groove to form a full length peptide[111]. Research on the length requirements of peptides to elicit T cell response and survival, has demonstrated that peptides as small as 3 amino acids can elicit a response. Though I also carefully point out, that this finding has not been reproduced.

In the literature more information can be found on arginine than lysine in the context of IBD. Arginine is an essential amino acid that can be delivered in the diet but can also be synthesized in the body. It is a substrate for different enzymes, including nitric oxide (NO) synthetases (NOS1 – also termed neuronal NOS (nNOS), NOS2 – also termed inducible (iNOS), and NOS3 – also termed endothelial NOS (eNOS)). NO synthetases catalyse the generation of NO and citrulline from arginine, oxygen and NADPH. Arginine is also a substrate of arginine synthases ARG1 and ARG2 which catalyse the hydrolysis of arginine to ornithine and urea[110]. NO is, amongst others, a major agent in the defence against pathogens in the immune response and is toxic to bacteria. It is produced, amongst others, by macrophages. In animal models, arginine supplementation has been shown to attenuate tissue

damage and improve clinical parameters of dextran sulfate sodium (DSS)-induced colitis as well as enhance cell migration, which has been implicated in improved wound repair[110,112]. In the same study, arginine supplementation in DSS-induced colitis reduced blood levels of inflammatory cytokines and showed a positive effect (i.e. enhanced microbial diversity) on the gut microbiota in mice[110]. Arginine has been shown to be elevated in serum levels and diminished in colonic tissues of patients with active UC[113,114]. Expression analysis of genes related to arginine synthesis and uptake have shown a decreased expression of arginine transporter Solute Carrier Family Member 2 (*SLC7A2*) and an increased expression of *NOS2* and *ARG2* in humans[114]. Overall, the diminished arginine concentration in the tissue was attributed to a "decreased cellular uptake and increased consumption by NOS2"[114]. In addition to arginine, the same publication also showed decreased serum levels of lysine. *Candida albicans*, which is a fungus that produces arginine as a response to host-derived reactive oxygen species (ROS), has been shown to be present in IBD patients at a higher level[40,115]. Arginine is also found at an increased level in antimicrobial peptides[116–118]. Antimicrobial peptides are made of cationic residues and are part of the innate immunity. They target the cell wall of bacteria or structures in the cytosol of bacteria[119]. A hypothesis therefore could also be that an auto-reactive process in affected individuals leads to a T cell response after production of antimicrobial peptides upon breach of the mucosal barrier.

6.2. HLA-*DRB1* expression

While research, including this thesis, has shown that variants of the HLA-*DRB1* gene are associated with UC, the (differential) expression of these variants at an allelic level has not been investigated in the context of IBD. This is likely due the challenges in the primer design for quantitative real time PCR measurements of expression at these highly polymorphic and complex genes. While method establishment for HLA genotyping has been driven to a large part by the need to match patients to donors in the context of tissue transplantation, measuring the HLA expression has not been an important step in this process. Appropriate tools for the analysis of HLA expression from available RNA-sequencing (RNA-Seq) data at a higher resolution have been emerging in the recent years. Still, the problems of correctly matching the RNA-Seq reads to HLA reference genomes sequences remain[120,121]. The sequence-redundancy across genes like the HLA-*A* and HLA-*H* genes and the HLA-*DRB1/3/4/5* genes, which have arisen by gene duplication, further complicate this endeavour. However, some research on HLA expression at the gene level has been previously published, dating back as far as 30 years ago. These studies are based on antibody staining using fluorescently labelled anti-sera and visualization using phase-microscopy[44,122–124] or antibody staining using gold-conjugated anti-sera and electron microscopy[125]. They specifically evaluated expression on epithelial cells. A study of HLA expression in tissue across the whole human body was conducted by Boegel *et al.*[122] in publicly available datasets obtained mostly by RNA bulk sequencing, such that the origin of the cells that

expressed HLA class I and class II molecules in the intestine was not addressed. Bulk RNA sequencing, as opposed to single cell sequencing in which transcriptomes are evaluated at the single cell level, measures expression across various cell types and tissues. Contrarily to the common opinion on HLA class II genes only being expressed on APCs, this early research has pointed towards HLA class II genes also being expressed by epithelial cells of the gut. However, here reports are contradictory. While some (earlier) research reported that DRB1 proteins are only presented on epithelial cells of the colon during active inflammation and absent in non-inflamed healthy tissue, more recent analysis of HLA expression along the gut points towards HLA class II proteins also being expressed in the healthy intestine. This finding is also supported by NGS-based analysis of RNA levels in the tissue[44,122–124]. Selby *et al.*[44] showed diminished expression of the HLA-*DP* gene on intestinal epithelial cells, while no expression of HLA-*DQ* was observed (though transcripts for these proteins were present in the cytosol, no HLA-*DQ* proteins were observed on the cell surface). This observation was not replicated in other studies. Boegel *et al.* showed expression, albeit low, of the HLA-*DR*, -*DP* and -*DQ* genes in healthy individuals (Supplementary Table 2 of Boegel *et al.*[122]). Theleman *et al.* showed that MHC class II expression can be induced by interferon (INF) gamma[126]. Koyama *et al.* showed that gnotobiotic mice lack MHC class II expression on intestinal cells and that exposition to microbiota in the gut is required for expression, hence indicating a cross-talk between the microbiota and the mucosa of the gut also in the healthy individuals[127], which is supported by data from Häsler *et al.*[128]. The latter study investigated correlations between expression patterns and abundance of microbiota in the human gut and showed that the cross-talk between the microbiota in the gut is reduced in individuals affected by IBD. They also reported HLA expression to be constitutively present in the thinner epithelial linings of the small intestine, while constitutive expression in the colon was not observed.

Since HLA-*DRB1* variants are located on the same haplotype as specific HLA-*DQ* and HLA-*DRB3/4/5* alleles (Figure 3 of **Paper C**), and their correlation cannot be dissolved using genetic association studies, neither of these genes can be specifically ruled out as disease driving. It may well be that HLA proteins expressed from all three loci are involved. In the context of Multiple Sclerosis, which is also associated with DRB1*15:01, peptide repertoires from both DRB1*15:01 and the correlated DRB5*01:01 were analysed by Scholz *et al.*[129]. The authors showed that “quantitatively, both molecules contributed to the peptide repertoire presented by cells expressing the HLA-DR15 haplotype”, but that the peptide binding repertoire of both proteins were distinct. This is also further supported by DRB1*15:01 and DRB5*01:01 peptide binding motifs generated by NetMHCIIpan-3.2 – which are differing especially in pockets 6 through 9, shown in Figure 30.

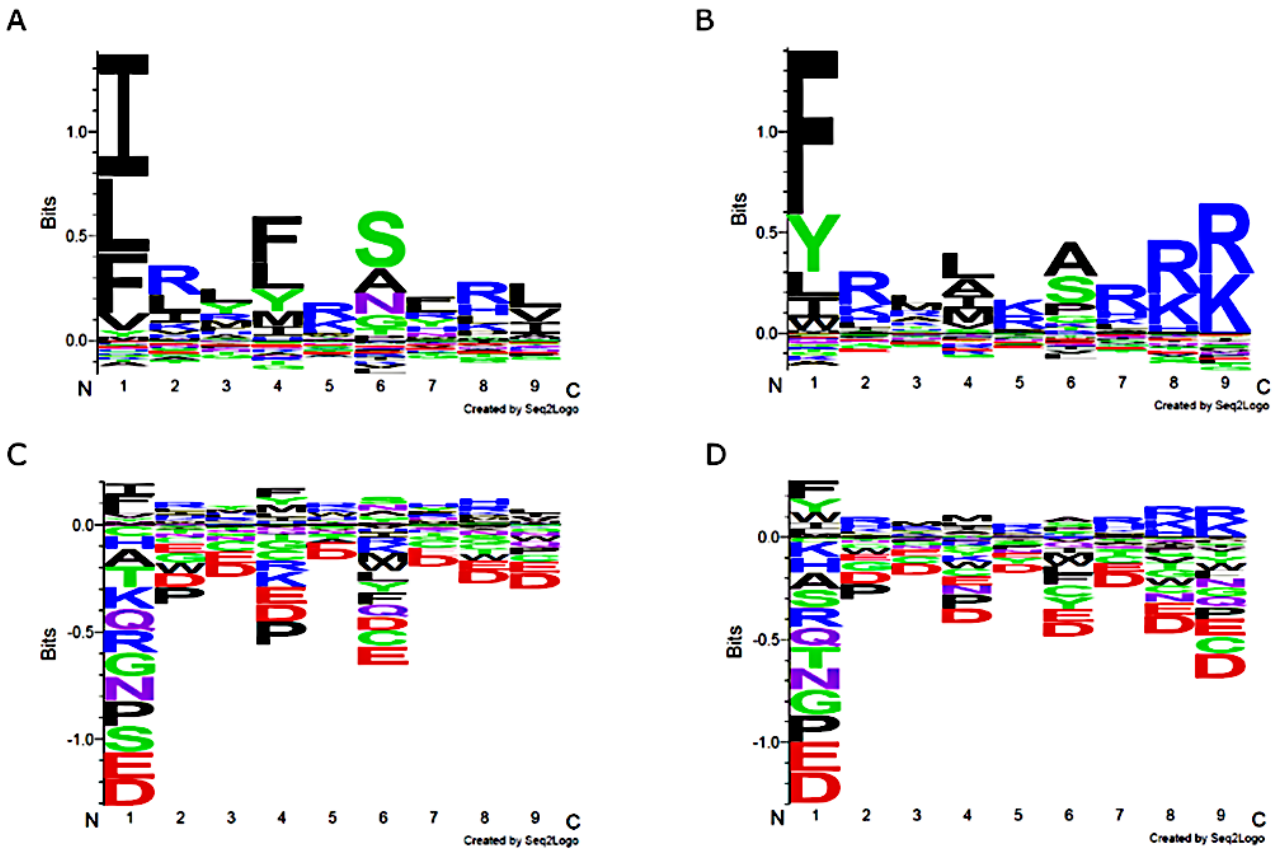


Figure 30 – Weighted Kullback-Leibler motifs of DRB1*15:01 and DRB5*01:01.
For 200,000 unique random human peptides the percentile rank scores of preferential peptide binding were calculated using NetMHCIIpan-3.2[101] (Set 2 of **Paper C**). Peptide motifs were plotted using Seq2Logo[130] on the top 2% binders using the NetMHCII3.2-pan rank score. Here amino acids enriched (A, B) and depleted (C, D) in the binding motifs are shown. The colour scheme shows the chemistry of the amino acids. *Red:* positively charged amino acids, *blue*: negatively charged amino acids, *green*: polar amino acids, *purple*: neutral amino acids and *black*: hydrophobic amino acids.

Finally, it is possible that the extent of gene expression is not related to disease development. It may well be that increased expression levels of HLA variants are part of a healthy immune response and thus independent of disease. Furthermore, one could also imagine that an individual susceptible to UC possesses lowly expressed HLA variants (i.e. DRB1*15:01 or DRB1*01:03 (in the British-descended population)) that can still bind detrimental peptides, which then leads to disease.

6.3. DRB1*15 in Crohn's disease and other immune diseases

Though as we show in **Paper C**, there is clear association of DRB1*15 alleles with UC, such a clear signal is not seen in CD. It was not identified in the largest fine mapping study by Goyette *et al*. of Caucasian CD as associated at a genome-wide significance level (DRB1*15: P = 0.00173, OR = 0.94, DRB1*15:01 P=2.69x10^{-5}, OR=0.92) and was also not included in the genetic map published with the article[1]. Stokkers *at al.* reported an association of the DR2 specificity (alleles of the DRB1*15 and

DRB1*16 groups belong to this specificity) and observed that DR2 was slightly protective in CD (OR=0.83, 95%CI = 0.70 - 0.98. no P-value given). Neither studies included in Stokkers *et al.* nor the Goyette *et al*. study stratified by disease location (i.e. ileal vs. colonic CD). Genetically, ileal CD has been associated with *NOD2* while colonic CD has been associated with the HLA. It is therefore possible that DRB1*15 is not associated in the more prevalent ileal CD but increases susceptibility to colonic CD. This would need further investigation, as described in the outlook (Chapter 6.5).

Alleles of the DRB1*15 group also play a role as risk factors in other immune-related diseases including Multiple Sclerosis[131–134] (a chronic inflammatory neurological disorder), Systemic Lupus Erythematosus[135] and Dupuytrien's disease[97,136] (both are disorders of the connective tissue). It has also been reported to be associated with adult onset Still's disease[137] (a systemic inflammatory disease), Graves disease (an autoimmune disease that affects the thyroid), pulmonary tuberculosis and Lepra[138,139] (a disease caused by *Mycobacterium Leprae* that affects the skin). For Multiple Sclerosis, DRB1*15:03, like in our study, was observed to be specific for African American populations[131]. DRB1*15 alleles have been reported to be strongly protective in type 1 (T1) diabetes (in which the autoimmune system attacks insulin producing beta cells of the pancreas), and pemphigus vulgaris (a skin blistering disease). However, the functional consequences of HLA-DRB1*15 association with these diseases have not been addressed and for most of them the potential disease driving antigens are not known. Exceptions are lepra, in which *Mycobacterium Leprae* causes the disease and pemphigus vulgaris, in which the skin protein desmoglein is targeted. Krause-Kyora *et al.* found that DRB1*15:01, among 18 contemporary DRB1 proteins, was predicted to "bind the second smallest number of potential *Mycobacterium Leprae* antigens" and further hypothesized that limited presentation of the *Mycobacterium Leprae*, may impair the immune response against this pathogen[138].

6.4. Limitations

6.4.1. Limitations of statistical power to detect specific HLA alleles

Statistical power describes the probability of a statistical test to detect a true effect, i.e. to reject the null hypothesis when a specific alternative hypothesis is true. In the context of HLA fine mapping studies (and GWAS) specifically, the power to detect an effect is dependent on the number of samples investigated and the frequency of a genetic variant in the population. Other factors like the significance level, case-control ratio and LD also play a role but are not discussed here. The lower the frequency and the closer the OR to 1, the more samples are needed to detect a statistical association. This is one important limitation of **Paper C** that is discussed further below. The power to detect an effect in **Paper C** with respect to the OR and population size is shown in Figure 31.

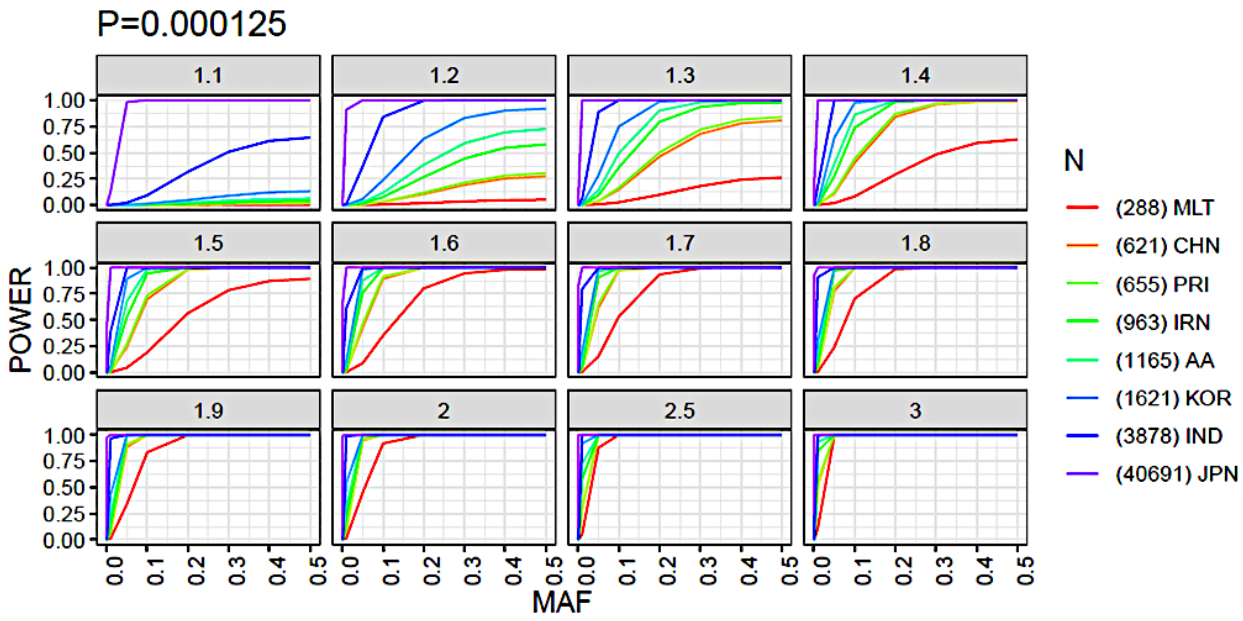


Figure 31 – Power to detect a true effect in our study.
The power to detect a statistical association was calculated for the sample sizes N of our study cohort, OR (denoted by number in the grey box), allele frequencies (MAF) and for a significance level of 0.000125. To generate these data the R script https://github.com/kaustubhad/gwas-power, which implements a formula for the calculation of power by Visscher *et al.*[140], was used. AA: African American, CHN: Chinese, EUR: European ancestry, IND: Indian, IRN: Iranian, JPN: Japanese, KOR: Korean, MLT: Maltese, PRI: Puerto Rican.

I chose a significance level of 0.000125 to indicate a P-value threshold corrected for multiplicity using the Bonferroni method (0.05 divided by the maximal observed allele number per population of ~ 400 alleles). Note that the power to detect an effect is identical for the OR (here risk) and the reciprocal OR (1/OR, here: protective). These plots show that the power to detect a variant with a "small" OR (<1.5) is decreased when compared to a variant with a "large" OR (e.g. >3), dependent on the size of the studied population and the minor allele frequency of the variant. To take this into account during interpretation of the data, we performed a meta-analysis across association results obtained from single statistical tests for each of the 9 analysed populations. We also visualized the effects directions and magnitudes across them using a plot specially designed for purposes of interpretation in **Paper C** (Figure 3, **Paper C**).

In **Paper C**, we described that DRB1*15:01 and DRB1*15:06 share the same amino acid properties within the HLA peptide binding groove. With low overall global frequency of the DRB1*15:06 allele, it was not statistically associated with UC. It was most frequent in the Indian population (AF= 2.1%, OR= 1.27, 95%CI [0.77, 2.11]). In Figure 32, I show that the OR and frequency of the allele may have been too low to detect an effect in the Indian population at the given sample size (N=1,621). The theoretical power to detect an effect at the given sample size of 1,621, OR of 1.27 and AF of 2.1% at a significance level of 0.05 is 0.50 and even less at the Bonferroni corrected significance level of 0.000125.

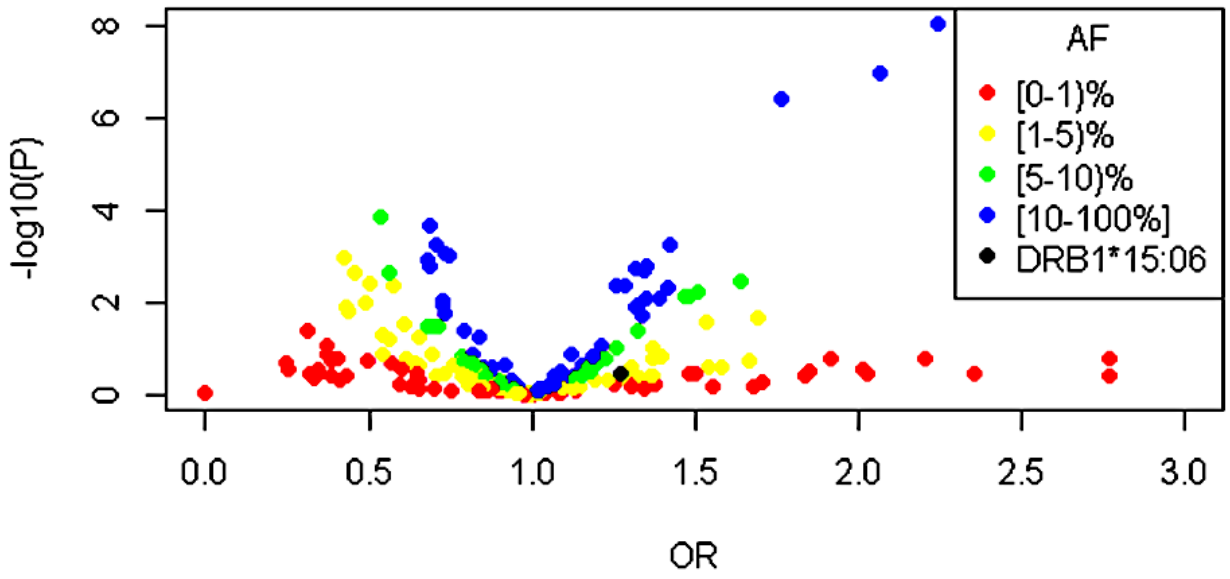


Figure 32 – P-values vs. OR in the Indian population for different allele frequencies.
OR and P-values were taken from the association analysis of HLA alleles within the Indian population in **Paper C**. The allele frequency of the HLA alleles is shown for different ranges. The OR and P-value of DRB1*15:06 is marked as a black dot. This plot shows the relationship between the OR, frequency and P-value of an allele in this study and indicates that at the estimated OR and frequency of DRB1*15:06, an association was likely not observed due to reasons of statistical power.

A list of other HLA alleles with sequence identity of amino acids in P1-P9 is shown in Supplementary Table 7 of **Paper C**. This table can be used to identify possible low-frequency HLA alleles associated with UC. Note however, that though candidates can be selected for further investigation, sequence identity of two alleles does not necessarily have to mean that both have an effect. By comparing sequence identity only, other modifications – like glycosylation – of the HLA alleles binding pockets and T cell interaction are not considered.

6.4.2. Limitations of HLA imputation

Limitations of HLA imputation are also discussed in **Paper B**. The quality of HLA imputation is dependent on the diversity and size of the HLA imputation reference as well as the frequency of individual HLA alleles. HLA imputation is most accurate for highly frequent HLA alleles and less accurate for infrequent alleles. This is simply because, adding more available data points, machine learning algorithms can better differentiate between similar alleles. When in doubt, the more frequent allele is assigned to an individual. In **Paper B**, we created a highly accurate HLA reference panel that can be used for the imputation of HLA alleles in populations of diverse ancestries. Here we focused on a 2-field resolution. Even better resolution would require a larger reference dataset to increase the number of each 3- or 4-field HLA allele. Since we were most interested in the HLA binding site considering 2-field resolution was however sufficient in our case. As also described in **Papers B** and

C, some HLA alleles are difficult to impute using genotypes measured on genotyping arrays. With gaps in the sequences of the HLA alleles, their underlying SNP haplotypes are similar. This is especially true for HLA-*DRB1* alleles. The extent of this is of course dependent on the genotyping or sequencing strategy (i.e. different HLA alleles can be differentiated better if more genomic variants are covered), the diversity of the HLA imputation panel and the number of samples included. This directly also impacts the quality of the HLA imputation map. This is one of the reasons, we opted not to include HLA-DRB1*04 and DRB1*11 in Figure 3 of **Paper C**. With similarities between the HLA-DRB1*04:01/03/04 alleles, we were not sufficiently able to differentiate between these alleles in our benchmark. This is also true for HLA*DRB11:01/04. Apart from this, the imputation quality may also have an impact on the phasing results – for which again very similar alleles are difficult to distinguish, if observed in the same individual. The phasing results in turn will only be as correct as the imputation results, i.e. although the phasing may be correct, the alleles themselves may only be correct on the group level. As shown in **Paper C**, our phasing approach was highly accurate in a test panel of the HapMap Phase3 CEU trio individuals.

6.4.3. Limitations of peptide prediction

In **Paper C**, we used the machine learning tool NetMHCIIpan-3.2 to predict HLA peptide binding between UC risk alleles and random sets of human peptides. This is still one of the most comprehensive methods for the prediction of HLA-peptide binding. Experimental data is still lacking for a lot of (infrequent, non-Caucasian) alleles. This is also due to the fact, that data on HLA-peptide binding has been obtained by manually and laboriously measuring the interaction of certain HLA alleles with single peptides (affinity measurement) or lower-throughput methods based on Mass Spectrometry. One advantage of NetMHCIIpan-3.2 in this setting is that it can extrapolate information on unknown HLA-peptide binding from known HLA-peptide binding data across sequentially similar alleles (i.e. similar in their amino acid sequences). In general, the quality of HLA-peptide binding prediction using machine learning algorithms is highly dependent on the quality of the data that the machine learning algorithm is trained on. Again, the sample size matters, i.e. more diverse datasets giving more information to the machine, will result in better HLA-peptide predictions. The machine learning tool NetMHCIIpan-3.2 extrapolates information that is known for a specific HLA allele to other less represented HLA alleles by generating and comparing pseudo-amino acid sequences of the HLA alleles. Based on this, HLA-peptide interactions are predicted. Though, as also stated above, this is a great advantage of this tool, in some instances this may introduce a bias. In **Paper C**, we hypothesized that the motif of peptides predicted to bind to DRB1*15:02, for which only little data was available to NetMHCIIpan-3.2 during training, was influenced by data available for DRB1*13:02, specifically by falsely introducing an asparagine (N) at position 4 of the allele. Indeed, this hypothesis has since been supported by experimental data produced by Wendorff *et al*.[141]. Data in Wendorff *et al*. were produced

using a peptide array. This array measured HLA-peptide binding of fixed peptides to experimentally introduced HLA molecules. It contained about 70,000 peptides. Amongst others, Wendorff *et al.* produced large-scale experimental data on DRB1*15:02, which they used to train machine learning algorithms, NNAlign and PIA (Figure 33).

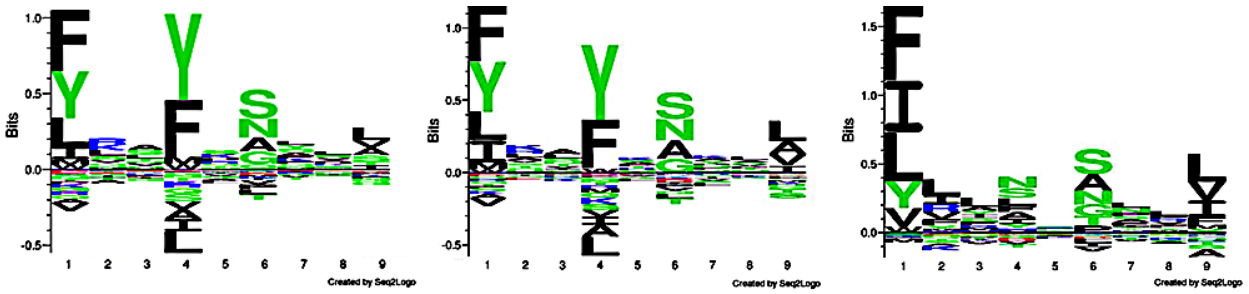


Figure 33 – Motif plots of DRB1 *15:02 from experimental data.
Figure taken and modified from Wendorff *et al.*[141]. HLA-peptide binding was predicted using different machine learning algorithms and a selection of 100,000 random peptides; (A) deep learning model (PIA) first described in Wendorff *et al.*, (B) NNAlign[142] and (C) NetMHCIIpan-3.2[101]. Prior to prediction, PIA and NNAlign had been trained on peptides obtained from experimental analysis of HLA-peptide binding using a high-throughput microarray technology. NetMHCIIpan-3.2 was trained by the authors of the NetMHCIIpan-3.2. tool on experimental data obtained from HLA-peptide affinity measurements. Peptide motifs were plotted using Seq2Logo[130] on the top 1% binders. The colour scheme shows the chemistry of the amino acids. *Red*: positively charged amino acids, *blue*: negatively charged amino acids, *green*: polar amino acids, *purple*: neutral amino acids and *black*: hydrophobic amino acids.

They compared their results to predictions produced by NetMHCIIpan-3.2 which had been trained by the authors of the NetMHCIIpan-3.2 paper on another publicly available dataset[101]. Asparagine (N) at position 4 is not present in the experimental data for DRB1*15:02 as shown in Wendorff *et al.*.

6.5. Concluding remarks & Outlook

Within the work process of this thesis, I generated a workflow for the analysis of the HLA in a trans-ancestry cohort that is applicable to the analysis of the HLA in the context of other immune-related diseases. It can also be used in studies in which not only HLA alleles but also their phase is of interest (i.e. knowledge on which HLA alleles are located on the same haplotype can be incorporated). Little is currently known on the association of UC subphenotypes with the HLA across different ancestries. HLA associations have also not been analysed in trans-ancestry CD cohorts. Hence, the analysis of subphenotype associations with the HLA for both UC and CD in general would be a next logical step. A first unstratified (disease location was not considered) analysis of trans-ancestry CD data, that I performed using the workflow in **Paper C**, indicated that the signals seen across different ancestries for CD are less consistent. However, especially in CD, disease location may play a vital role in the analysis of HLA association. Cleynen *et al.*[19] observed a more significant association with variants in HLA in colonic CD, while variants in the *NOD2* gene were more associated with disease in CD patients

with small-bowel involvement. This study was conducted in the Caucasian population. Overall small-bowel involvement, and not colonic CD, is more prevalent among CD patients. Stratification may strengthen the HLA signal for colonic CD, reducing signals that are derived from possibly non-associated ileal CD. However, the weaker HLA association of CD with the HLA as well as the diminished sample numbers resulting from splitting the data, remain major bottlenecks of this type of analysis

Current HLA fine mapping studies focus on the classical HLA class I and class II alleles, while non-classical HLA alleles (i.e. HLA-*E*, -*F* and -*G*) are mostly overlooked. Research into the non-classical HLA-*G* has suggested differential expression of this gene in IBD cases versus healthy controls. Whether this is linked to specific HLA variation in this gene, i.e. related to specific HLA-*G* alleles, has not been studied. 6 to 19 different proteins are recorded for the non-classical HLA class I genes HLA-*E*, -*F* and -*G* in the IMGT/HLA database version 3.39.0. This is about a factor of 150-800 less than known for the more commonly studied classical HLA class I alleles HLA-*A*, -*B* and -*C.* Though I do not expect these numbers to increase drastically in the future, more HLA-*E*, -*F* and -*G* alleles likely await discovery especially in the non-Caucasian population – i.e. the reference databases for these alleles are likely incomplete. This directly influences the quality of NGS typing for these alleles. A reference panel used specifically for the HLA imputation of these genes should currently therefore best only be constructed in the Caucasian population.

In this thesis I have made a first step towards identifying how culprit antigens may look like. To continue along these lines, limitations of current HLA-peptide predictions using machine learning algorithms need to be overcome. The major limitation in this effort is the number of available data points to train the algorithms. This limitation can be addressed by performing HLA-peptide binding analyses in a high-throughput manner. This will increase the number of known HLA-peptide binding and thereby create larger databases that can be used for the training of suitable machine learning algorithms and prediction of more refined peptide motifs for HLA proteins associated with IBD. Current higher throughput experiments include peptide arrays as proposed by Wendorff e*t al.*[141] and Buus *et al.*[143], as well as more computationally expensive peptide elution experiments, in which peptides that bind to predefined HLA proteins are eluted from these proteins and then identified using Mass Spectrometry. These technologies are more feasible for this application than the manual measurement of HLA-peptide affinities of single predefined peptides against known HLA proteins. Still, one of the main challenges in the analysis of HLA-peptide-binding, even when done *in silico,* is the vast number of bacterial, fungal – down to the strain levels – and food-borne proteomes. Some hints towards which bacterial species may be implicated in IBD can be deduced from microbiome studies, though results are in parts still inconsistent and strain differences are usually not considered. The human proteome, which we subsampled in **Paper C**, can be screened comparably fast. However, challenges in

considering processes that play a role *in vivo,* like peptide processing or differential HLA expression, remain.

Another general goal in deciphering IBD is also to understand what triggers the deterioration of the mucosal barrier in the first place – which then leads to the overstimulation of the immune system by exposition of the tissue to the gut microbes. Genetical association and sequencing studies have implicated several genes important for maintaining the intestinal epithelial barrier homeostasis. Expression studies of these genes have shown differences in the expression between individuals affected by IBD and unaffected individuals. The genes associated with differential expression were especially enriched in genes whose products are involved in the formation of the mucus layer (mucin genes (*MUC*), including *MUC1* and *MUC4*) and included genes involved in the formation of tight junctions and the extracellular matrix[144]. Van der Post *et al.* analysed the proteome of the IBD and healthy gut and implicated amongst others mucin proteins, like the MUC2 to be differentially present in the gut proteome, while Visschedijk et *al.* showed that rare variants in the MUC2 are associated with UC[145,146]. Loss of cell polarity and cell dissociation has been linked to highly glycosylated MUC proteins in cancer[147]. Functional follow up and further characterisation of genes involved in the formation of the epithelial barrier in the gut may give deeper insight into the processes leading to the mucosal barrier deterioration. With respect to the results of this thesis, HLA-peptide binding of the identified risk alleles to peptides derived from proteins that are involved in the epithelial barrier homeostasis, like the *MUC* genes, would be of interest. Especially the wild-type proteins derived from these genes could be compared to proteins derived from known (coding) IBD variations, to investigate whether self-tolerance is disrupted in IBD patients.

7. References

1. 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).
2. *Molecular Genetics of Inflammatory Bowel Disease*. (Springer, Boston, M, 2019).
3. Crohn's & Colitis UK. CROHN'S DISEASE - YOUR GUIDE. Available at: www.crohnsandcolitis.org.uk.
4. Crohn's & Colitis UK. ULCERATIVE COLITIS - YOUR GUIDE. Available at: www.crohnsandcolitis.org.uk.
5. Shi, H. Y. *et al.* Ethnicity Influences Phenotype and Outcomes in Inflammatory Bowel Disease: A Systematic Review and Meta-analysis of Population-based Studies. *Clinical Gastroenterology and Hepatology* (2017). doi:10.1016/j.cgh.2017.05.047
6. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
7. Lee, J. M. & Lee, K. M. Endoscopic diagnosis and differentiation of inflammatory bowel disease. *Clinical Endoscopy* (2016). doi:10.5946/ce.2016.090
8. Vavricka, S. R. *et al.* Extraintestinal manifestations of inflammatory bowel disease. *Inflammatory Bowel Diseases* (2015). doi:10.1097/MIB.0000000000000392
9. Ferreira, S. da C., de Oliveira, B. B. M., Morsoletto, A. M., Martinelli, L. C. & Troncon, L. E. de A. Extraintestinal manifestations of inflammatory bowel disease: clinical aspectis and pathogenesis. *J Gastroenterol Dig Dis* (2018).
10. Hsu, Y. C., Wu, T. C., Lo, Y. C. & Wang, L. S. Gastrointestinal complications and extraintestinal manifestations of inflammatory bowel disease in Taiwan: A population-based study. *J. Chinese Med. Assoc.* (2017). doi:10.1016/j.jcma.2016.08.009
11. Mahid, S. S., Mulhall, A. M., Gholson, R. D., Eichenberger, M. R. & Galandiuk, S. Inflammatory bowel disease and African Americans: A systematic review. *Inflammatory Bowel Diseases* (2008). doi:10.1002/ibd.20389
12. Duricova, D., Burisch, J., Jess, T., Gower-Rousseau, C. & Lakatos, P. L. Age-related differences in presentation and course of inflammatory bowel disease: An update on the population-based literature. *Journal of Crohn's and Colitis* (2014). doi:10.1016/j.crohns.2014.05.006
13. Kedia, S. & Ahuja, V. Epidemiology of Inflammatory Bowel Disease in India: The Great Shift East. *Inflamm. Intest. Dis.* (2017). doi:10.5771/0175-274X-2017-2-53
14. Lahad, A. Current therapy of pediatric Crohn's disease. *World J. Gastrointest. Pathophysiol.* (2015). doi:10.4291/wjgp.v6.i2.33
15. Degenhardt, F. *et al.* Serologic Anti-GP2 Antibodies Are Associated with Genetic

Polymorphisms, Fibrostenosis, and Need for Surgical Resection in Crohn's Disease. *Inflamm Bowel Dis* **22**, 2648–2657 (2016).

16. Kuna, A. T. Serological markers of inflammatory bowel disease. *Biochemia Medica* (2013). doi:10.11613/BM.2013.006
17. Satsangi, J., Silverberg, M. S., Vermeire, S. & Colombel, J. F. The Montreal classification of inflammatory bowel disease: Controversies, consensus, and implications. *Gut* (2006). doi:10.1136/gut.2005.082909
18. Mohammadi, M. *et al.* Association of HLA-DRB1 Alleles with Ulcerative Colitis in the City of Kerman, South Eastern Iran. *Iran J Allergy Asthma Immunol* **14**, 306–312 (2015).
19. Cleynen, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**, 156–167 (2016).
20. Best, W. R., Becktel, J. M., Singleton, J. W. & Kern, F. Development of a Crohn's Disease Activity Index: National Cooperative Crohn's Disease Study. *Gastroenterology* (1976). doi:10.1016/S0016-5085(76)80163-1
21. 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* **390**, 2769–2778 (2018).
22. Schroeder, K. W., Tremaine, W. J. & Ilstrup, D. M. Coated Oral 5-Aminosalicylic Acid Therapy for Mildly to Moderately Active Ulcerative Colitis. *N. Engl. J. Med.* (1987). doi:10.1056/NEJM198712243172603
23. Tontini, G. E., Vecchi, M., Pastorelli, L., Neurath, M. F. & Neumann, H. Differential diagnosis in inflammatory bowel disease colitis: State of the art and future perspectives. *World J. Gastroenterol.* (2015). doi:10.3748/wjg.v21.i1.21
24. Everhov, Å. H. *et al.* Changes in inflammatory bowel disease subtype during follow-up and over time in 44,302 patients. *Scand. J. Gastroenterol.* (2019). doi:10.1080/00365521.2018.1564361
25. Lee, H. S. *et al.* Change in the diagnosis of inflammatory bowel disease: A hospital-based cohort study from Korea. *Intest. Res.* (2016). doi:10.5217/ir.2016.14.3.258
26. De Vries, L. C. S., Wildenberg, M. E., De Jonge, W. J. & D'Haens, G. R. The Future of Janus Kinase Inhibitors in Inflammatory Bowel Disease. *J Crohns Colitis* **11**, 885–893 (2017).
27. Alatab, S. *et al.* The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2011. Alatab, S. et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a s. *Lancet Gastroenterol. Hepatol.* (2020). doi:10.1016/S2468-1253(19)30333-4
28. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* (2017). doi:10.1038/nature22969
29. 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).
30. Brant, S. R. *et al.* Genome-Wide Association Study Identifies African-Specific Susceptibility Loci

in African Americans With Inflammatory Bowel Disease. *Gastroenterology* **152**, 206-217 e2 (2017).

31. Zumelzu, C. *et al.* Black patients of african descent and HLA-DRB115:03 frequency overrepresented in epidermolysis bullosa acquisita. *J. Invest. Dermatol.* (2011). doi:10.1038/jid.2011.231
32. Fuyuno, Y. *et al.* Genetic characteristics of inflammatory bowel disease in a Japanese population. *J. Gastroenterol.* (2016). doi:10.1007/s00535-015-1135-3
33. Yang, S. K. *et al.* Identification of Loci at 1q21 and 16q23 That Affect Susceptibility to Inflammatory Bowel Disease in Koreans. *Gastroenterology* (2016). doi:10.1053/j.gastro.2016.08.025
34. Yang, S. K. *et al.* Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut* (2014). doi:10.1136/gutjnl-2013-305193
35. Yang, S. K. *et al.* Genome-wide association study of ulcerative colitis in koreans suggests extensive overlapping of genetic susceptibility with caucasians. *Inflamm. Bowel Dis.* (2013). doi:10.1097/MIB.0b013e3182802ab6
36. 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).
37. Petersen, B. S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D. & Franke, A. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet* **18**, 14 (2017).
38. Zeissig, Y. *et al.* XIAP variants in male Crohn's disease. *Gut* (2015). doi:10.1136/gutjnl-2013-306520
39. Zeissig, S. *et al.* Early-onset Crohn's disease and autoimmunity associated with a variant in CTLA-4. *Gut* (2015). doi:10.1136/gutjnl-2014-308541
40. Sokol, H. *et al.* Fungal microbiota dysbiosis in IBD. *Gut* (2017). doi:10.1136/gutjnl-2015-310746
41. Stange, E. F. & Schroeder, B. O. Microbiota and mucosal defense in IBD: an update. *Expert Review of Gastroenterology and Hepatology* (2019). doi:10.1080/17474124.2019.1671822
42. Alam, M. T. *et al.* Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathog.* (2020). doi:10.1186/s13099-019-0341-6
43. Sommer, F. *et al.* Microbiomarkers in inflammatory bowel diseases: Caveats come with caviar. *Gut* (2017). doi:10.1136/gutjnl-2016-313678
44. Selby, W. S., Janossy, G., Mason, D. Y. & Jewell, D. P. Expression of HLA-DR antigens by colonic epithelium in inflammatory bowel disease. *Clin Exp Immunol* (1983).
45. Wosen, J. E., Mukhopadhyay, D., MacAubas, C. & Mellins, E. D. Epithelial MHC class II expression and its role in antigen presentation in the gastrointestinal and respiratory tracts. *Frontiers in Immunology* (2018). doi:10.3389/fimmu.2018.02144
46. Rosati, E. *et al.* Identification of Disease-associated Traits and Clonotypes in the T Cell Receptor

Repertoire of Monozygotic Twins Affected by Inflammatory Bowel Diseases. *J. Crohn's Colitis* (2020). doi:10.1093/ecco-jcc/jjz210

47. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* (2017). doi:10.1186/s12896-017-0379-9
48. Murphy, K. *Janeway's immunobiology*. (Garland Science/Taylor & Francis Group, LLC, 2012).
49. Perera, L. *et al.* Expression of nonclassical class I molecules by intestinal epithelial cells. *Inflamm. Bowel Dis.* (2007). doi:10.1002/ibd.20026
50. Murphy K, W. C. Janeway's Immunobiology. **9**, (2016).
51. Torres, M. I. *et al.* Expression of HLA-G in inflammatory bowel disease provides a potential way to distinguish between ulcerative colitis and Crohn's disease. *Int. Immunol.* (2004). doi:10.1093/intimm/dxh061
52. Gomes, R. G. *et al.* HLA-G is expressed in intestinal samples of ulcerative colitis and Crohn's disease patients and HLA-G5 expression is differentially correlated with TNF and IL-10 cytokine expression. *Hum. Immunol.* (2018). doi:10.1016/j.humimm.2018.03.006
53. Zhang, Q. *et al. Practical Atals of Transplan Pathology - Histocompatibility and Immunogenetics for Solid Organ Transplantation*. (Springer, Cham, 2016).
54. Jurewicz, M. M. & Stern, L. J. Class II MHC antigen processing in immune tolerance and inflammation. *Immunogenetics* (2018). doi:10.1007/s00251-018-1095-x
55. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423-31 (2015).
56. Nomenclature for Factors of the HLA System, 1975. *Bull World Heal. Organ.* **52**, 261–265 (1975).
57. Albert, E. *et al.* Nomenclature for Factors of the HLA System-1977. *Tissue Antigens* (1978). doi:10.1111/j.1399-0039.1978.tb01231.x
58. Bodmer, W. F. *et al.* Nomenclature for factors of the HLA system 1984. *Hum. Immunol.* (1984). doi:10.1016/0198-8859(84)90068-5
59. Dupont, B. Nomenclature for factors of the HLA system, 1987. Decisions of the Nomenclature Committee on Leukocyte Antigens, which met in New York on November 21-23, 1987. in *Human immunology* (1989). doi:10.1016/0198-8859(89)90027-X
60. Bodmer, J. G. *et al.* Nomenclature for factors of the HLA system, 1989. *Hum. Immunol.* (1990). doi:10.1016/0198-8859(90)90060-3
61. Hampe, J. *et al.* Linkage of Inflammatory Bowel Disease to Human Chromosome 6p. *Am. J. Hum. Genet.* (1999). doi:10.1086/302677
62. Stokkers, P. C., Reitsma, P. H., Tytgat, G. N. & van Deventer, S. J. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* **45**, 395–401 (1999).
63. Ahmad, T., Marshall, S. E. & Jewell, D. Genetics of inflammatory bowel disease: The role of the HLA complex. *World Journal of Gastroenterology* (2006). doi:10.3748/wjg.v12.i23.3628

64. Gao, F. *et al.* Association of HLA-DRB1 alleles and anti-neutrophil cytoplasmic antibodies in Han and Uyghur patients with ulcerative colitis in China. *J Dig Dis* **15**, 299–305 (2014).
65. Han, B. *et al.* Amino acid position 37 of HLA-DRβ1 affects susceptibility to Crohn's disease in Asians. *Hum. Mol. Genet.* (2018). doi:10.1093/hmg/ddy285
66. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
67. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput. Biol.* **9**, e1002877 (2013).
68. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* (2012). doi:10.1371/journal.pcbi.1002822
69. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* (2002). doi:10.1038/nrg777
70. Chen, J. M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics* (2007). doi:10.1038/nrg2193
71. Gaunt, T. R., Rodríguez, S. & Day, I. N. M. Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-428
72. Single, R. M. *et al.* Asymmetric linkage disequilibrium: Tools for assessing multiallelic LD. *Hum. Immunol.* (2016). doi:10.1016/j.humimm.2015.09.001
73. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28**, 1598–1603 (2012).
74. Jostins, L. Using next-generation genomic datasets in disease association. (2012).
75. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Research and Therapy* (2011). doi:10.1186/ar3204
76. Clarke, G. M. *et al.* Basic statistical analysis in genetic case-control studies. *Nat. Protoc.* (2011). doi:10.1038/nprot.2010.182
77. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* (2011). doi:10.1038/ng.764
78. Sachs, L. & Hedderich, J. *Angewandte Statistik: Methodensammlung mit R*. (Springer Spektrum, 2018).
79. Chen, H., Cohen, P. & Chen, S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Commun. Stat. Simul. Comput.* (2010). doi:10.1080/03610911003650383
80. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* (2016). doi:10.1038/ejhg.2015.269

81. Saffari, A. *et al.* Estimation of a significance threshold for epigenome-wide association studies. *Genet. Epidemiol.* (2018). doi:10.1002/gepi.22086
82. Panagiotou, O. A. *et al.* What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* (2012). doi:10.1093/ije/dyr178
83. Michael, B., Hedges, L. V., Higgins, J. & Rothstein, H. R. *Introduction to Meta-Analysis*. (Wiley, 2009).
84. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809–822 (2011).
85. Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379–i388 (2017).
86. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* (2011). doi:10.1016/j.ajhg.2011.04.014
87. Faculté de médecine de Genève. HLA typing. (2015). Available at: medweb4.unige.ch/immunologie/home/HSC/donor/HLA_typing/%0D.
88. Wittig, M. *et al.* Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.* **43**, e70 (2015).
89. *HLA typing - Methods and Protocols*. (Springer, Boston, M, 2018).
90. Leslie, S., Donnelly, P. & McVean, G. A Statistical Method for Predicting Classical HLA Alleles from SNP Data. *Am. J. Hum. Genet.* (2008). doi:10.1016/j.ajhg.2007.09.001
91. Dilthey, A. T., Moutsianas, L., Leslie, S. & McVean, G. HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics* **27**, 968–972 (2011).
92. Zheng, X. *et al.* HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
93. Levin, A. M. *et al.* Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. *BMC Genet.* (2014). doi:10.1186/1471-2156-15-72
94. Karnes, J. H. *et al.* Comparison of HLA allelic imputation programs. *PLoS One* (2017). doi:10.1371/journal.pone.0172444
95. Pappas, D. J. *et al.* Significant variation between SNP-based HLA imputations in diverse populations: The last mile is the hardest. *Pharmacogenomics J.* (2018). doi:10.1038/tpj.2017.7
96. Venkateswaran, S. *et al.* Enhanced Contribution of HLA in Pediatric Onset Ulcerative Colitis. *Inflamm Bowel Dis* **24**, 829–838 (2018).
97. Replication, D. Ia. G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234–244 (2014).
98. Yeturu, K., Utriainen, T., Kemp, G. J. L. & Chandra, N. An automated framework for understanding structural variations in the binding grooves of MHC class II molecules. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-S1-S55

99. Chen, B. *et al.* Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0280-2
100. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* (2017). doi:10.1016/j.immuni.2017.02.007
101. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* (2018). doi:10.1111/imm.12889
102. Gonzalez-Galarza, F. F. *et al.* Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res* **43**, D784-8 (2015).
103. Gourraud, P. A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282 (2014).
104. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
105. International HapMap, C. The International HapMap Project. *Nature* **426**, 789–796 (2003).
106. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
107. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2011).
108. Dhanda, S. K. *et al.* Predicting HLA CD4 immunogenicity in human populations. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.01369
109. Paul, S. *et al.* Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. *J. Immunol. Methods* (2015). doi:10.1016/j.jim.2015.03.022
110. Singh, K. *et al.* Dietary arginine regulates severity of experimental colitis and affects the colonic microbiome. *Front. Cell. Infect. Microbiol.* (2019). doi:10.3389/fcimb.2019.00066
111. Huan, X., Zhuo, Z., Xiao, Z. & Ren, E. C. Crystal structure of suboptimal viral fragments of Epstein Barr Virus Rta peptide-HLA complex that stimulate CD8 T cell response. *Sci. Rep.* (2019). doi:10.1038/s41598-019-53201-6
112. Coburn, L. A. *et al.* L-arginine supplementation improves responses to injury and inflammation in dextran sulfate sodium colitis. *PLoS One* (2012). doi:10.1371/journal.pone.0033546
113. Hong, S. K. S. *et al.* Increased serum levels of L-arginine in ulcerative colitis and correlation with disease severity. *Inflamm. Bowel Dis.* (2010). doi:10.1002/ibd.21035
114. Coburn, L. A. *et al.* L -Arginine Availability and Metabolism Is Altered in Ulcerative Colitis. *Inflamm. Bowel Dis.* (2016). doi:10.1097/MIB.0000000000000790
115. Jiménez-López, C. *et al.* Candida albicans induces arginine biosynthetic genes in response to host-derived reactive oxygen species. *Eukaryot. Cell* (2013). doi:10.1128/EC.00290-12
116. Nguyen, L. T. *et al.* Serum stabilities of short tryptophan- and arginine-rich antimicrobial peptide analogs. *PLoS One* (2010). doi:10.1371/journal.pone.0012684

117. Chan, D. I., Prenner, E. J. & Vogel, H. J. Tryptophan- and arginine-rich antimicrobial peptides: Structures and mechanisms of action. *Biochimica et Biophysica Acta - Biomembranes* (2006). doi:10.1016/j.bbamem.2006.04.006

118. Cutrona, K. J., Kaufman, B. A., Figueroa, D. M. & Elmore, D. E. Role of arginine and lysine in the antimicrobial mechanism of histone-derived antimicrobial peptides. *FEBS Lett.* (2015). doi:10.1016/j.febslet.2015.11.002

119. Brogden, K. A. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nature Reviews Microbiology* (2005). doi:10.1038/nrmicro1098

120. Boegel, S. *et al.* HLA typing from RNA-Seq sequence reads. *Genome Med.* (2012). doi:10.1186/gm403

121. Orenbuch, R. *et al.* arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btz474

122. Boegel, S. *et al.* HLA and proteasome expression body map. *BMC Med. Genomics* (2018). doi:10.1186/s12920-018-0354-x

123. Horie, Y., Chiba, M., Iizuka, M. & Masamune, O. Class II (HLA-DR, -DP, and -DQ) antigens on intestinal epithelia in ulcerative colitis, Crohn's disease, colorectal cancer and normal small intestine. *Gastroenterol. Jpn.* (1990). doi:10.1007/BF02779357

124. Mayer, L. *et al. Expression of Class II Molecules on Intestinal Epithelial Cells in Humans Differences Between Normal and Inflammatory Bowel Disease*. *GASTROENTEROLOGY* (1991). doi:10.5555/URI:PII:0016508591905756

125. Bär, F. *et al.* Inflammatory bowel diseases influence major histocompatibility complex class I (MHC I) and II compartments in intestinal epithelial cells. *Clin. Exp. Immunol.* (2013). doi:10.1111/cei.12047

126. Thelemann, C. *et al.* Interferon-γ induces expression of MHC class II on intestinal epithelial cells and protects mice from colitis. *PLoS One* (2014). doi:10.1371/journal.pone.0086844

127. Koyama, M. *et al.* MHC Class II Antigen Presentation by the Intestinal Epitahelium Initiates Graft-versus-Host Disease and Is Influenced by the Microbiota. *Immunity* (2019). doi:10.1016/j.immuni.2019.08.011

128. Häsler, R. *et al.* Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut* (2017). doi:10.1136/gutjnl-2016-311651

129. Scholz, E. M. *et al.* Human leukocyte antigen (hla)-DrB1*15:01 and hla-DrB5*01:01 present complementary peptide repertoires. *Front. Immunol.* (2017). doi:10.3389/fimmu.2017.00984

130. 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.* (2012). doi:10.1093/nar/gks469

131. Hollenbach, J. A. & Oksenberg, J. R. The immunogenetics of multiple sclerosis: A

comprehensive review. *J Autoimmun* **64**, 13–25 (2015).

132. Alcina, A. *et al.* Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. *PLoS One* **7**, e29819 (2012).
133. Prat, E. *et al.* HLA-DRB5*0101 and -DRB1*1501 expression in the multiple sclerosis-associated HLA-DR15 haplotype. *J. Neuroimmunol.* (2005). doi:10.1016/j.jneuroim.2005.04.027
134. Patsopoulos, N. A. *et al.* Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS Genet.* **9**, e1003926 (2013).
135. Hanscombe, K. B. *et al.* Genetic fine mapping of systemic lupus erythematosus MHC associations in Europeans and African Americans. *Hum. Mol. Genet.* (2018). doi:10.1093/hmg/ddy280
136. Brown, J. J., Ollier, W., Thomson, W. & Bayat, A. Positive association of HLA-DRB1*15 with Dupuytren's disease in Caucasians. *Tissue Antigens* (2008). doi:10.1111/j.1399-0039.2008.01082.x
137. Yap, L. M., Ahmad, T. & Jewell, D. P. The contribution of HLA genes to IBD susceptibility and phenotype. *Best Practice and Research: Clinical Gastroenterology* (2004). doi:10.1016/j.bpg.2004.01.003
138. Krause-Kyora, B. *et al.* Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat. Commun.* (2018). doi:10.1038/s41467-018-03857-x
139. Zhang, F. *et al.* Evidence for an association of HLA-DRB115 and DRB109 with leprosy and the impact of DRB109 on disease onset in a Chinese Han population. *BMC Med. Genet.* (2009). doi:10.1186/1471-2350-10-133
140. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* (2017). doi:10.1016/j.ajhg.2017.06.005
141. Wendorff, M. *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. *Front. Immunol.*
142. Nielsen, M. & Andreatta, M. NNAlign: A platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx276
143. Buus, S. *et al.* High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol. Cell. Proteomics* (2012). doi:10.1074/mcp.M112.020800
144. Vancamelbeke, M. *et al.* Genetic and Transcriptomic Bases of Intestinal Epithelial Barrier Dysfunction in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* (2017). doi:10.1097/MIB.0000000000001246
145. Van Der Post, S. *et al.* Structural weakening of the colonic mucus barrier is an early event in ulcerative colitis pathogenesis. *Gut* (2019). doi:10.1136/gutjnl-2018-317571
146. Visschedijk, M. C. *et al.* Pooled resequencing of 122 ulcerative colitis genes in a large Dutch

cohort suggests population-Specific associations of rare variants in MUC2. *PLoS One* (2016). doi:10.1371/journal.pone.0159609

147. Pelaseyed, T. & Hansson, G. C. Membrane mucins of the intestine at a glance. *J. Cell Sci.* **133**, (2020).

8. Appendix

8.1. Appendix 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.

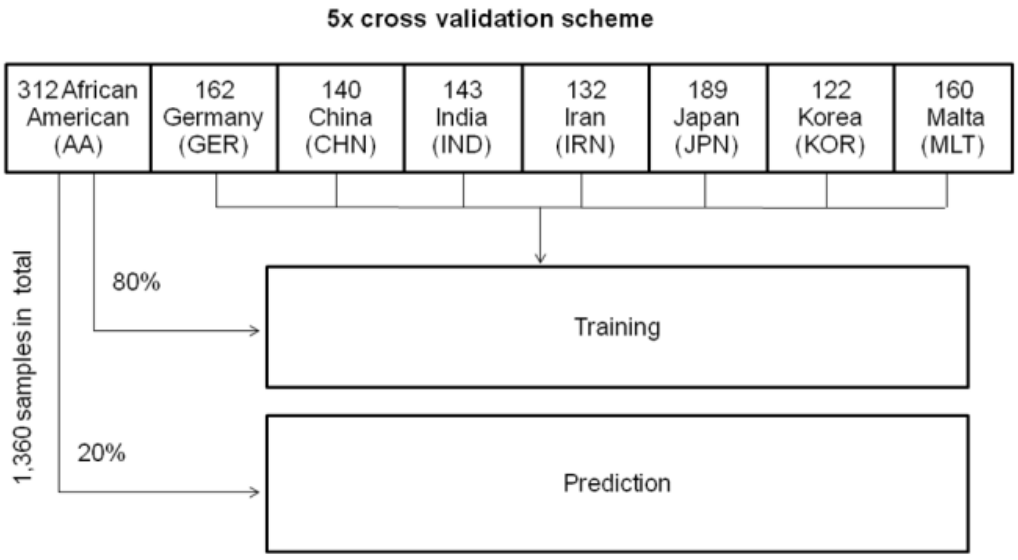
SUPPLEMENTARY METHODS

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.

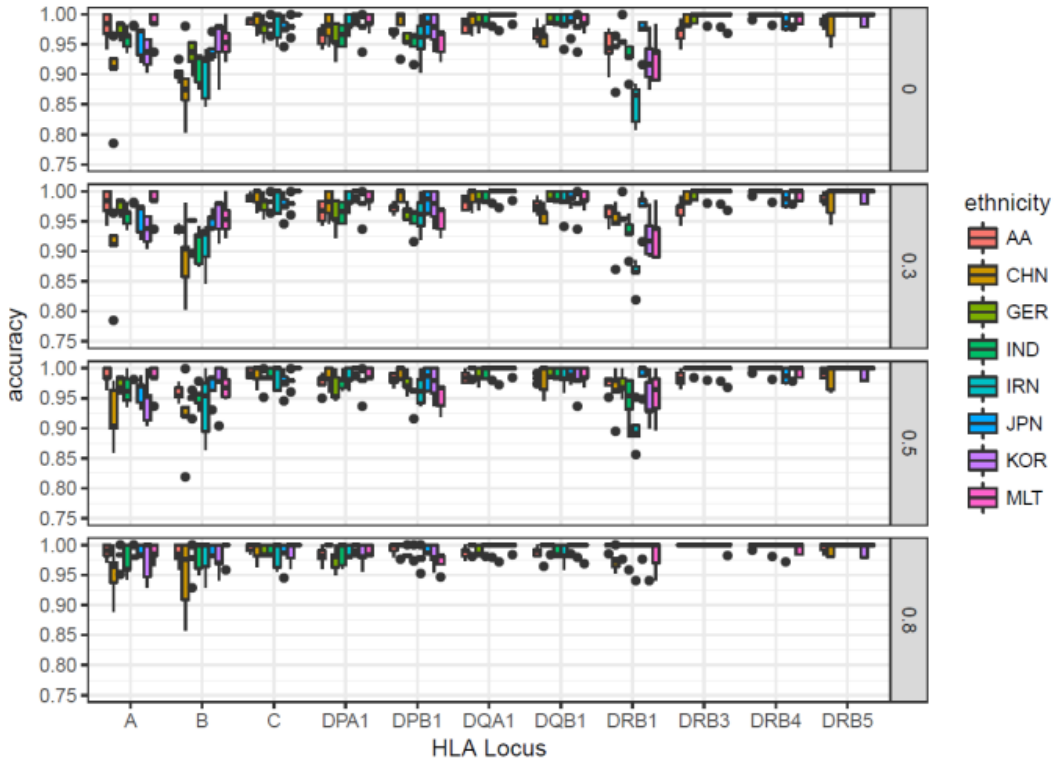
SUPPLEMENTARY FIGURES

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.

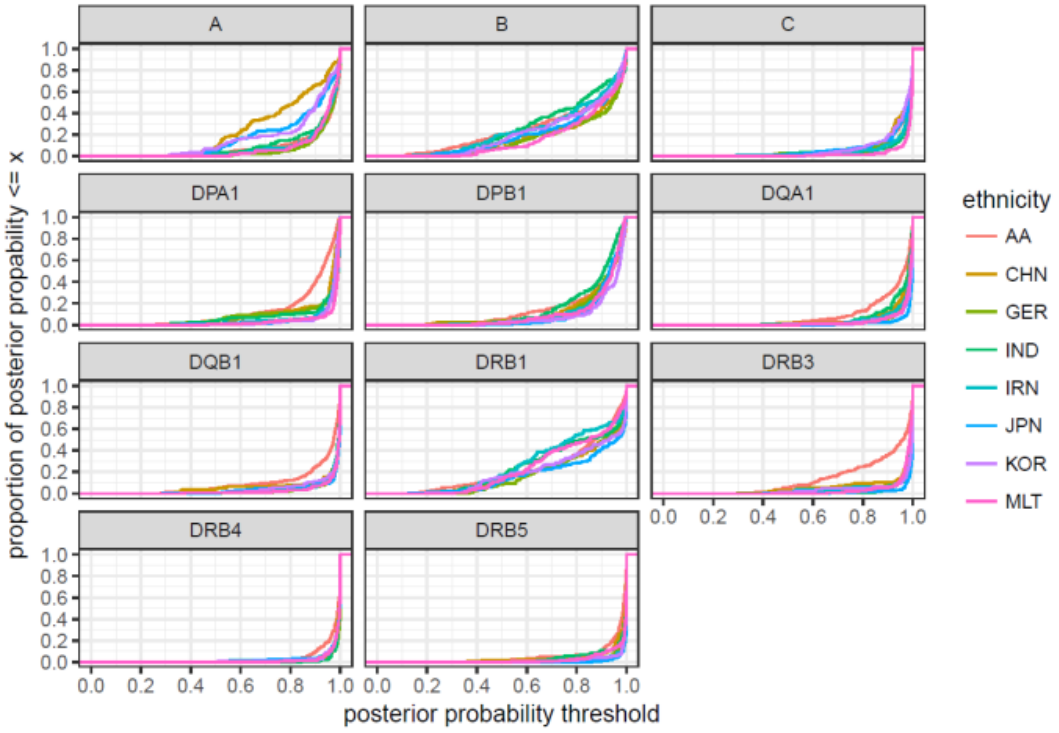


Supplementary Figure 2 – Imputation accuracies employing the multi-ethnic reference combined with the 1000 Genomes dataset: Accuracies and post-imputation probabilities of HLA imputation with HIBAG (1) using a 5x cross validation scheme and the trans-ethnic and 1000 Genomes dataset (4) with 4-digit G 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)

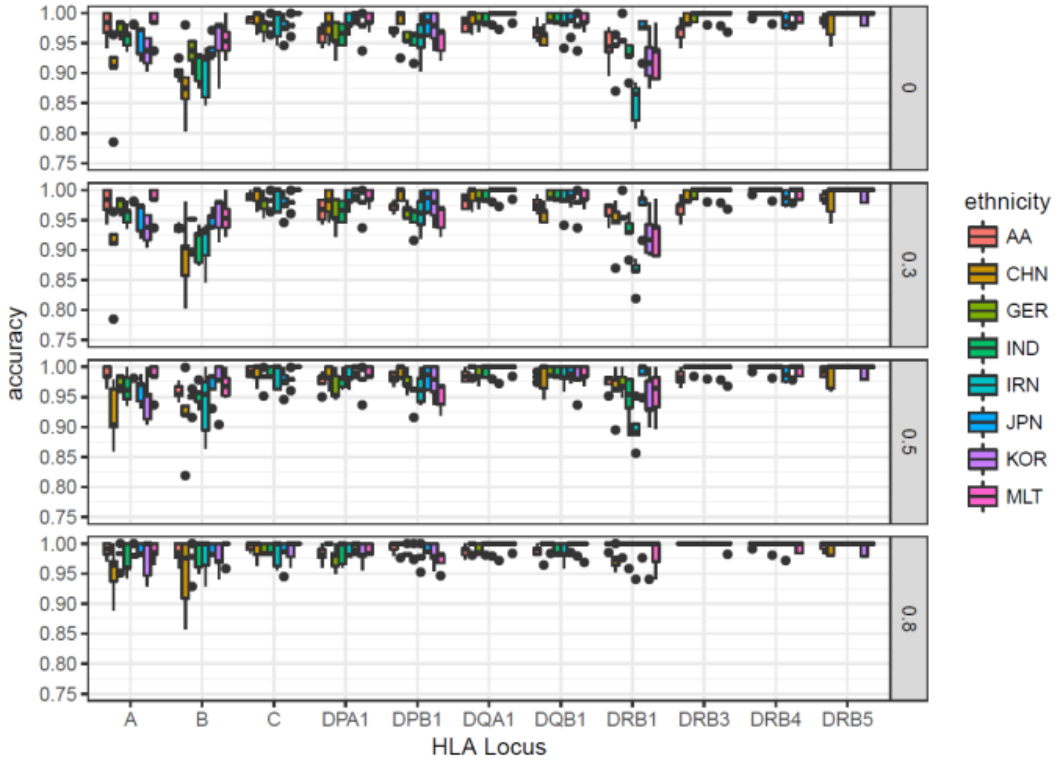


(b)

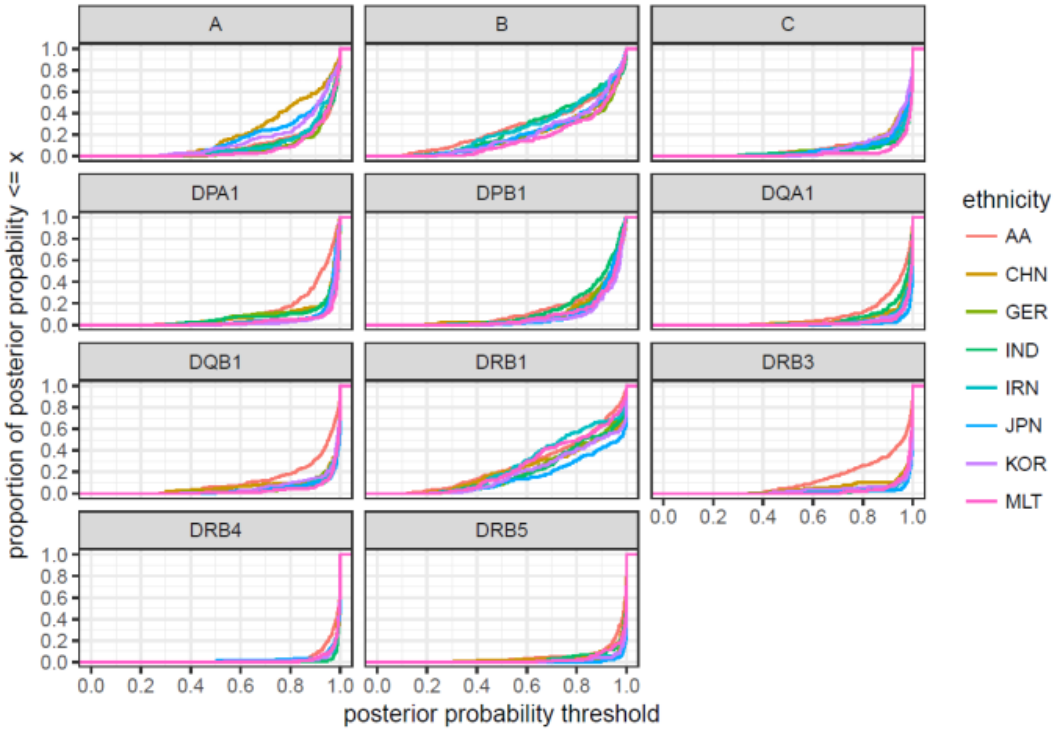

A
B
C
DPA1
DPB1
DQA1
DQB1
DRB1
DRB3
DRB4
DRB5
proportion of posterior propability <= x
posterior probability threshold
0.0
0.2
0.4
0.6
0.8
1.0
ethnicity
AA
CHN
GER
IND
IRN
JPN
KOR
MLT

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)

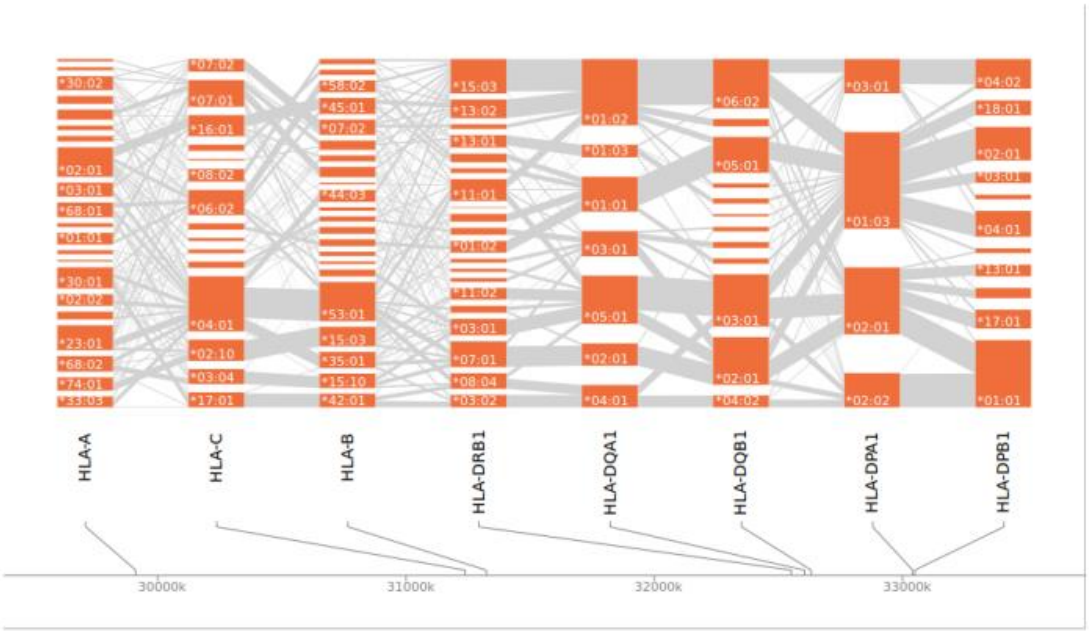


(b)

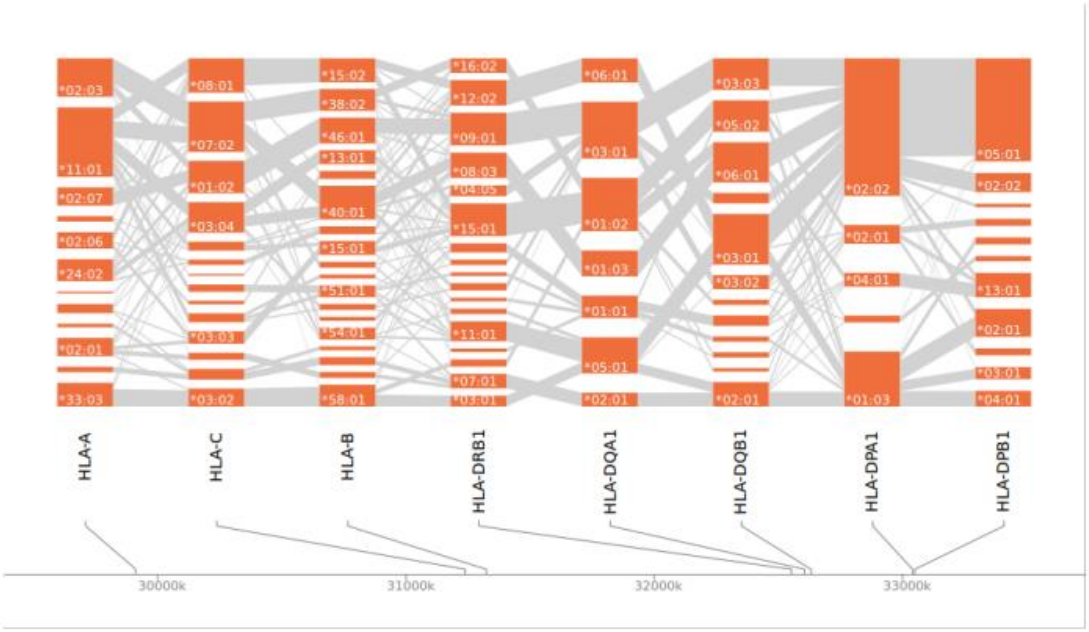

A
B
C
DPA1
DPB1
DQA1
DQB1
DRB1
DRB3
DRB4
DRB5
proportion of posterior propability <= x
posterior probability threshold
0.0
0.2
0.4
0.6
0.8
1.0
ethnicity
AA
CHN
GER
IND
IRN
JPN
KOR
MLT

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.

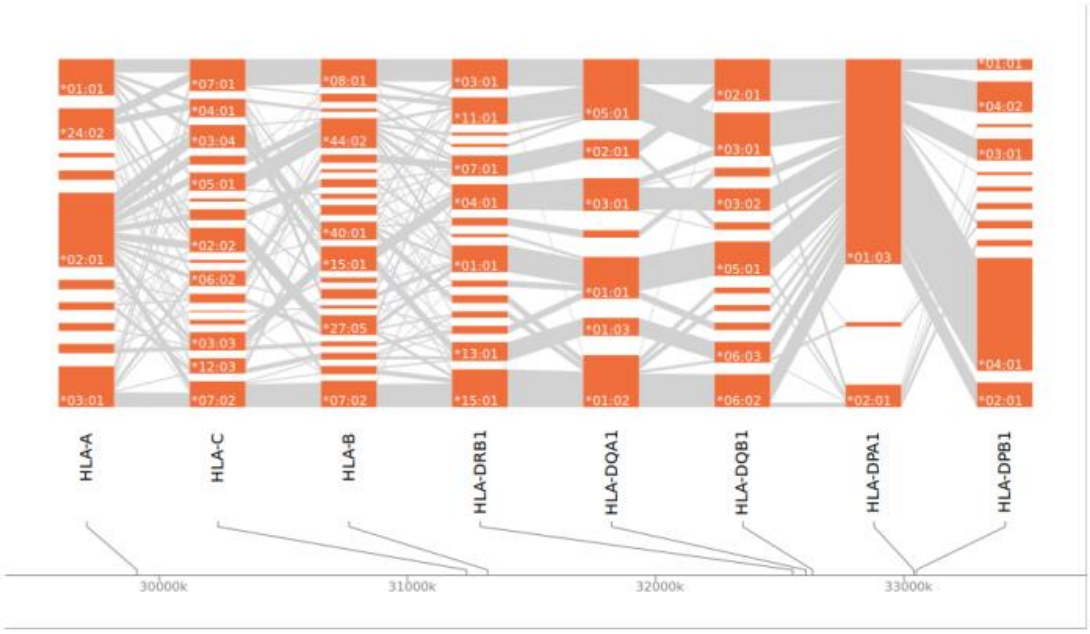
(a) AA



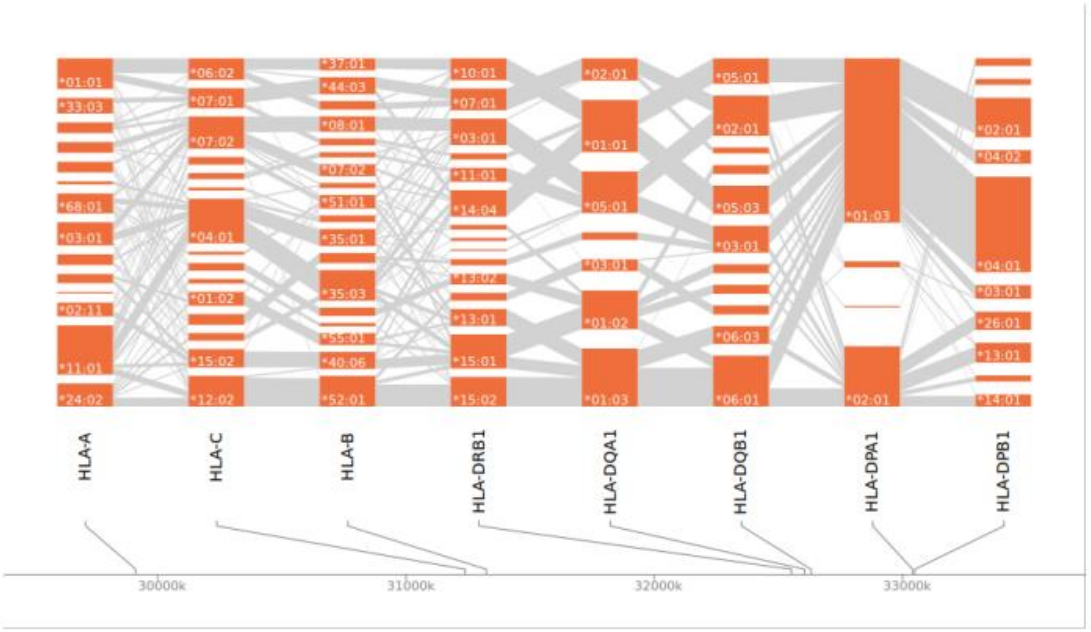
(b) CHN


*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
*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

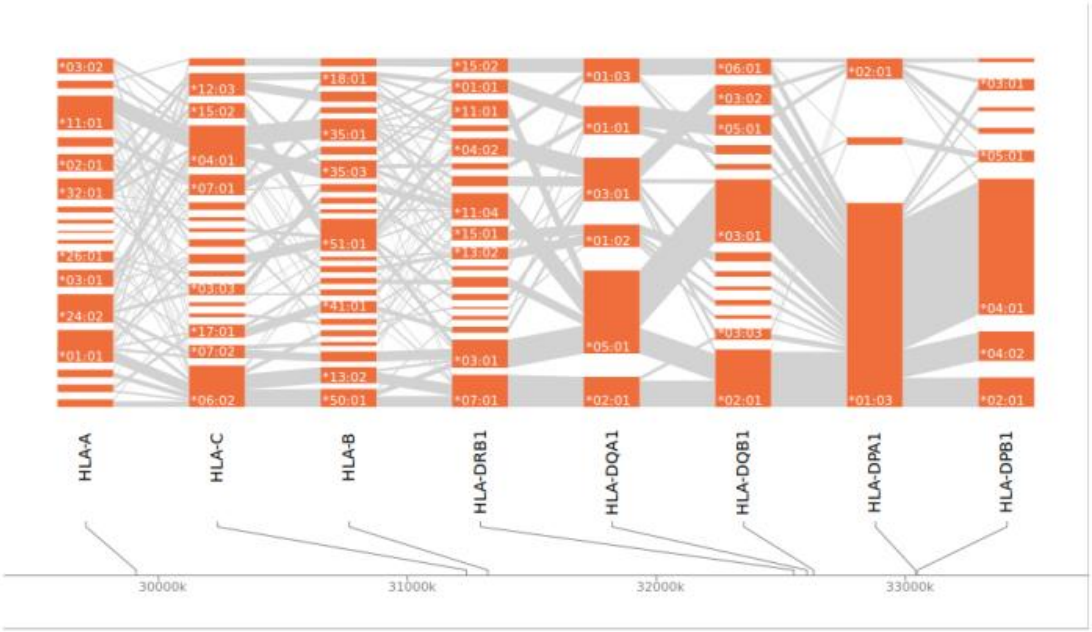
(c) GER


*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

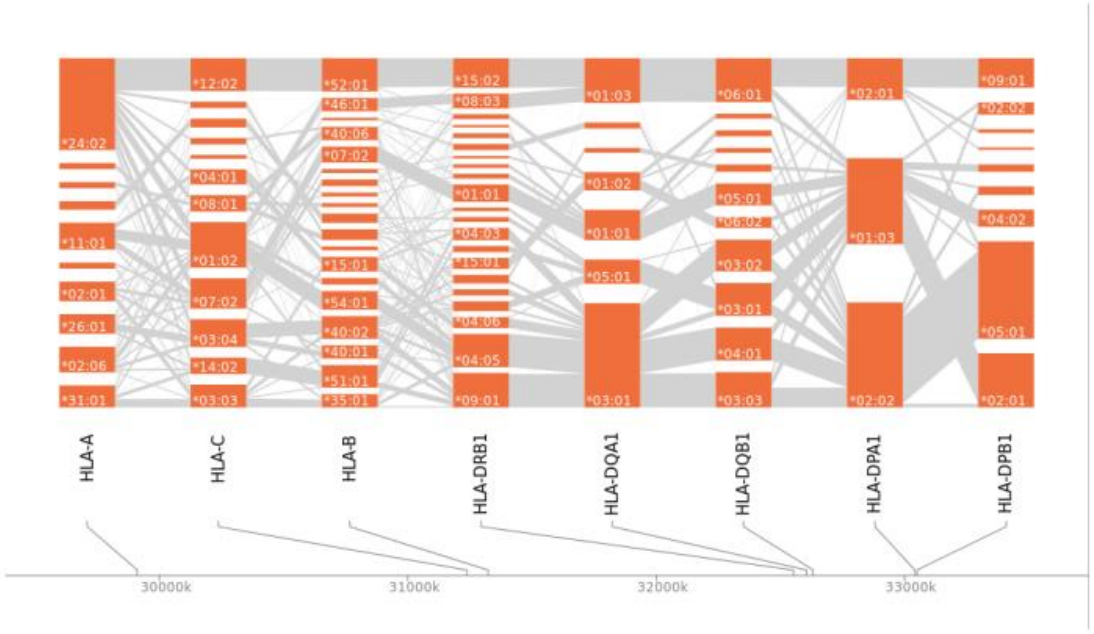
(d) IND


HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

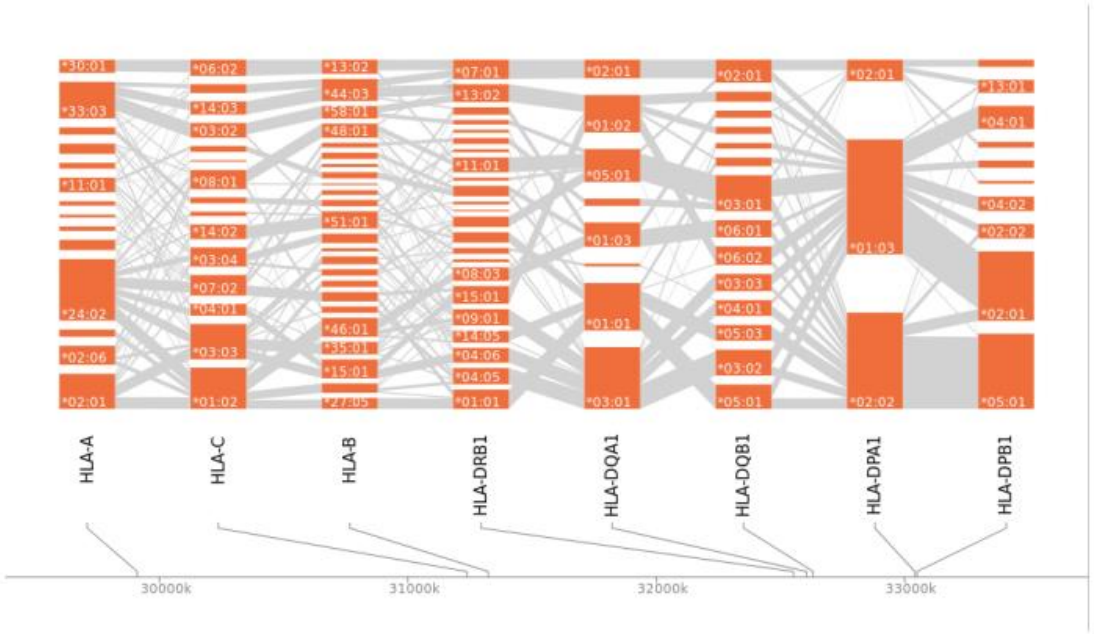
(e) IRN


HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

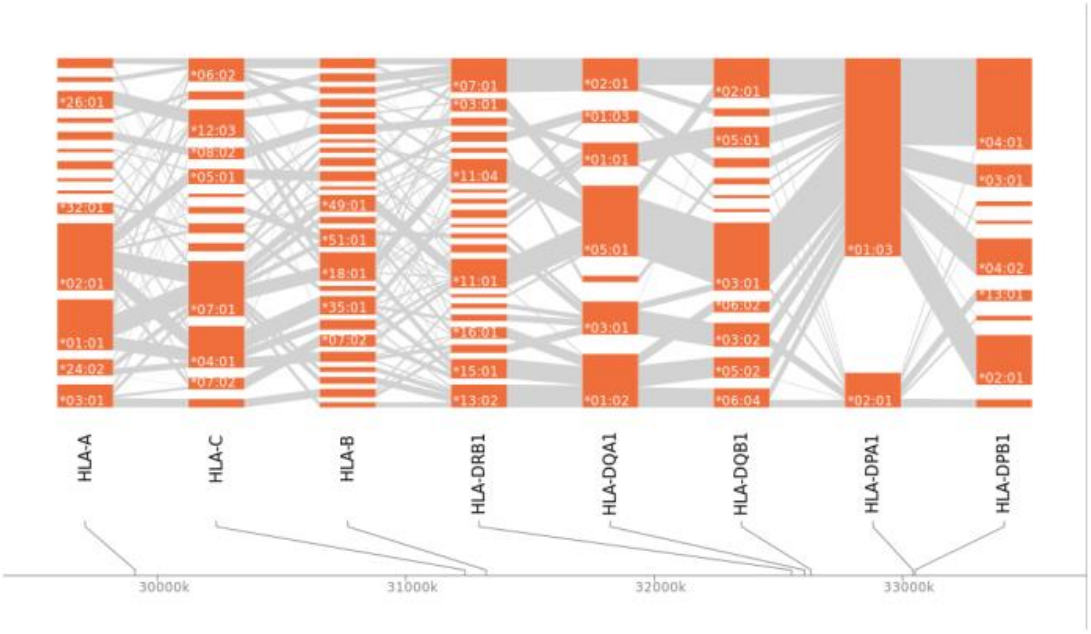
(f) JPN


HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

(g) KOR


HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1
30000k
31000k
32000k
33000k

(h) MLT


*26:01
*24:01
*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
*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.

8.2. Appendix B

SUPPLEMENT

Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals common disease signatures

CONTENTS

SUPPLEMENTARY METHODS 1
SUPPLEMENTARY TABLES 4
SUPPLEMENTARY FIGURES 6
COMPARISON TO PREVIOUS STUDY (GOYETTE *ET AL.*[5]) 27
LIMITATIONS OF HLA IMPUTATION 27
ANALYSIS OF GENDER EFFECTS 27
SUPPLEMENTARY METHODS 28
REFERENCES 37

SUPPLEMENTARY METHODS

Supplementary cohort description

Individuals contained in this study are part of the data freeze published in Liu *et al.*[1], in Ye *et al.*[2] and Huang *et al.*[3].

Asian samples: Chinese samples were collected in and around Hong Kong (Chinese University of Hong Kong), Korean samples in Seoul, Korea (Asan Medical Centre, Ye *et al.*[2]), Japanese samples in Tokyo (Institute of Medical Science, University of Tokyo, RIKEN Yokohama Institute and Japan Biobank),
Iranian samples: Iranian samples were collected in Tehran (Tehran University of Medical Science).
Indian samples: Indian samples were collected in North India (Dayanand Medical College and Hospital, Ludhiana), all self-reported North Indian which was consistent with their genetically determined background.
Caucasian samples: The European samples were collected from different study centers in Europe and North America and are described in Liu *et al.*[1] Maltese samples were collected in Malta (Department of Gastroenterology, Mater Dei Hospital, Msida, Malta).
African American samples: The African American samples, all self-described as African American and determined genetically as consistent with African American ethnicity by their each having an admixture of West African and European ancestry, were subjects collected 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, the Emory University GENESIS study and at Cedars Sinai Medical Center (Huang *et al.*[3], **Supplementary Note**).
Puerto Rican samples: The Puerto Rico samples were collected from third generation Puerto Ricans recruited for the study Mapping genes for IBD in Puerto Ricans for the NIDDK IBD Genetics Consortium, in collaboration with Cedars Sinai Medical Center. Puerto Ricans are a genetically admixed population having European, African and Native American ancestry, Both parents and all four grandparents of each subject were required to be Puerto Ricans[4]. [1].

Sample numbers are shown in **Supplementary Table 1**. The herein used Caucasian data freeze did not include the Multiple Sclerosis controls published with Goyette *et al.*[5]. These missing control data could not be accessed due to a change in the use and access policy. The East Asian data set was split into Chinese and Japanese populations, for the Indian data set we used only samples with reported Indian ancestry. Additional data sets include data for African American, Korean, Maltese and Puerto Rican ulcerative colitis patients and controls.

Supplementary information on genotyping & genotype calling

All individuals were typed on the Illumina HumanImmuno v.1.0 (ImmunoChip) or the Illumina Infimum ImmunoArray 24 v2.0 (Malta only). Both chips interrogate a wide proportion of the extended HLA with a total of 196,524 and 253,702 variants respectively. Illumina IDAT files were provided for the African American, Korean, Maltese and Puerto Rican data sets. Genotypes for these data sets were called using Opticall version 0.7.0[6] with default parameter settings (-hwep 10-15, -minp 0.7). For the remaining cohorts pre-called data were used from Liu *et al.*[1].

All data were exported on the on the + strand and SNPs matching with missing strand or chromosome annotation, InDels coded as I and D and duplicated positions with duplicated allelic information were removed.

1

Quality control of study genotypes

The quality control was performed separately in each population using PLINK[7,8]. Individuals with missing gender information, genotype missingness > 10%, outlying heterozygosity (not within interval [mean + 3*sd, mean -3*sd]) (sd: standard deviation), with a kinship coefficient (R-package: SNPRelate version 1.4.2; gdsfmt version 1.6.2) > 0.185 (IBD > 0.37) and those failing gender check were removed from the data set. SNPs with missingness of > 5%, differential missingness between cases and controls ($P < 10^{-5}$) and a deviation from Hardy-Weinberg equilibrium (HWE) $P < 10^{-5}$ in controls were excluded. We additionally excluded SNPs that failed one of the following criteria in one or more batches: per batch missingness > 0.1, HWE in controls ($P < 10^{-5}$) or differential missingness between cases and controls ($P < 10^{-5}$). For the analysis of population stratification by PCA, we merged each data set with genotype data of the 1000 Genomes Phase III population and calculated PCs as defined below. 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 as outliers. Next, we performed PCA analysis on each cohort separately (without merging it to the 1000 Genomes Phase III population) to identify batch outliers, that we defined as above.

For the Indian and the Puerto Rican data set we identified 3 and 2 distinct subpopulations during QC. For the Indian population this has been observed before (Negi *et al.*[10]) and represents population substructure within the Indian population. The Puerto Rican data were split into two distinct genotyping batches and had been typed at different time points. We kept these data and adjusted for batch in the later association analyzes. For the Liu *et al.* data set, the samples we received had previously undergone QC as described in the original publication. Post-QC sample and SNP statistics are listed in **Supplementary Table 1**.

Calculation of Principal Components (PCs)

Principal components were calculated SNPs pruned for linkage disequilibrium (LD) data with MAF > 5% for each population separately. SNPs were pruned using PLINK's -indep-pairwise 50 5 0.2. Additionally, several regions of high LD were excluded (chr5:44Mb-51.5Mb, chr6:25Mb-33.5Mb, chr8:8Mb-12Mb, chr11:45Mb-57Mb). For calculation of PCs, FlashPCA2[9] was used. Genomic inflation factors were calculated on 3,120 SNPs "null SNPs" not associated with immune disease (Liu *et al.*[1]) as λ_{GC}=0.98, 0.98, 1.33, 0.98, 0.97, 1.01, 0.95, 0.95, 1.01 and 1.05 for the analysis in the African American, Chinese, Caucasian, Iranian, Indian, Japanese, Korean, Maltese and Puerto Rican populations, respectively, using the first five PCs.

Calculation of kinship

First, possible related samples were identified using PLINK[7,8] and its method of moments (MOM; --genome) estimator. Related samples were then further analyzed with the R-package SNPRelate version 1.4.20 based on the more computationally intensive calculation of a maximum likelihood estimator (MLE) and related samples were then finally identified using this MLE estimation.

Gender check

The gender check was conducted by counting heterozygous calls on X and Y separately for each individual and plotting the sum for each individual. A cluster analysis (kmeans) was performed on the counts using the kmeans-function (2 centers, 10 iterations, nstart 1) of R (version 3.3.1) and incorrectly assigned samples (gender and cluster assignment not identical) were removed.

2

Supplementary information on SNP imputation: Imputation from HLA alleles

For this, we generated a library using FASTA files of HLA alleles published with the IMGT/HLA database (version 3.30.0)[11]. The human reference genome (hg19, GRCh37 Genome Reference Consortium Human Reference 37) contains contigs of 7 different cell lines for the HLA region of chromosome 6. The main contig is derived from the PGF cell line which carries A*03:01:01:01, B*07:02:01, C*07:02:01:03, DPA1*01:03:01:02, DPB1*04:01:01:01, DQA1 01:02:01:05, DQB1 06:02:01:01 and DRB1 15:01:01:01[12]. For these alleles, we generated genomic position information for each single variant position by aligning the FASTA sequences against the hg19.2bit (09112017) published with the UCSC hg19 assembly using Blat version 35 (Kent *et al.*[13]) with default parameters. We then proceeded to aligning all alleles present in our multi-ethnic reference (Degenhardt *et al.*[14]) per locus to the respective PGF allele using MAFFT (Katoh *et al.*[15]) on a 4-digit basis, using full length sequences where possible. For allele sequences located on the minus strand of the reference, we generated reverse complement sequences for assignation to the genomic position and concatenated insertions into single string information. For each position we then checked for allelic consistency with alleles names in the NCBI's dbSNP Human Build 150 (i.e. to determine if alleles generated for specific positions in our library were consistent with known alleles).

HLA pocket definitions

Our HLA-II pocket definitions are based on direct sidechain-sidechain interactions between the HLA protein and the nine-peptide residue that contacts the HLA binding groove. First, we defined all HLA residues that locate within 7Å of the peptide. Second, members of the same pocket P1 to P9 were only included if the interaction was based on sidechain-sidechain contacts to the same peptide sidechain. By this strict definition we aimed for the highest specificity regarding HLA-peptide interaction, because only a few HLA residues contact more than one peptide sidechain directly. The protein structure of HLA-DRA*01:01-DRB1*01:01 (PBD ID 1dlh) in complex with an influenza virus peptide was used as reference[16].

Generation of human peptides

The human reference proteome, available at https://www.uniprot.org/proteomes/UP000005640, was downloaded together with its additional isoforms. We fragmented this dataset using a sliding window with size 15 and a step size of 1 into 15-mer peptides using a custom C++ code and Python version 3.7.4 and the Pandas library version 0.24.1. The fragmentation process generated 36,964,665 15-mers. We deleted peptide sequences containing unconventional or unknown amino acids and sampled 200,000 unique random peptide sequences.

Clustering of HLA proteins

Based on the results shown in **Figure 4** of the Main Manuscript, we clustered both risk and protective alleles into two groups each. Risk group 1 (RISK 1) included DRB1*11:01/04 and DRB1*13:01, risk group 2 (RISK 2), DRB1*12:01, DRB1*14:04 DRB1*15:01/03. We also separated the protective alleles DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and DRB1*10:01 (PROT 1) and alleles DRB1*04:03/06 (PROT 2) into two protective clusters. With DRB1*03:01 forming its own cluster (prominent D at position 4 of the binding motif), we did not analyze this protein further. DRB1*15:02 and DRB1*13:02 were also not considered with their clustering probably being an artefact of NetMHCIIpan-3.2 **(Supplementary Figure 7).**

3

SUPPLEMENTARY TABLES

Supplementary Table 3 - Applied numerical scores for amino acid properties.

We show the Atchley scores F1 and F3, charge, residue-volume and hydrogen-acceptor capability. Atchley *et al.*[17]. generated multidimensional values from 54 different amino acid properties as factors F1 to F5. F1 is also named polarity index and reflects intramolecular characteristics of amino acids, such as polarity, hydrophobicity and hydrogen donor capacity. F3 combines values of amino acid size and bulkiness. Additionally, we defined two scores "hydrogen acceptor", for the ability of amino acids to participate in hydrogen bonds, and charge. For the former, the value applied simply corresponds to the number of atoms within amino acid sidechains that can accept a hydrogen. The AAI score "charge" used here, describes the presence of positive and negative charge of an amino acid sidechain by integers (positive charge: +1, negative charge: -1 and absence of charge: 0). The parameter "residue volume" (GOLD730102; based on Goldsack *et al.*[18]) was selected from the AAIndex database (https://www.genome.ad.jp/aaindex/) (Kawashima *et al.*[19]) as an indirect measure of pocket volume.

Amino acids	Charge	HB-Acceptor	Residue-Volume	F1	F3
Ala (A)	0	0	88.3	-0.591	-0.733
Cys (C)	0	0	112.4	-1.343	-0.862
Asp (D)	-1	4	110.8	1.05	-3.656
Glu (E)	-1	4	140.5	1.357	1.477
Phe (F)	0	0	189	-1.006	1.891
Gly (G)	0	0	60	-0.384	1.33
His (H)	0	2	152.6	0.336	-1.673
Ile (I)	0	0	168.5	-1.239	2.131
Lys (K)	1	0	175.6	1.831	0.533
Leu (L)	0	0	168.5	-1.019	-1.505
Met (M)	0	0	162.2	-0.663	2.219
Asn (N)	0	2	125.1	0.945	1.299
Pro (P)	0	0	122.2	0.189	-1.628
Gln (Q)	0	2	148.7	0.931	-3.005
Arg (R)	1	0	181.2	1.538	1.502
Ser (S)	0	2	88.7	-0.228	-4.76
Thr (T)	0	2	118.2	-0.032	2.213
Val (V)	0	0	141.4	-1.337	-0.544
Trp (W)	0	0	227	-0.595	0.672
Tyr (Y)	0	1	193	0.26	3.097

4

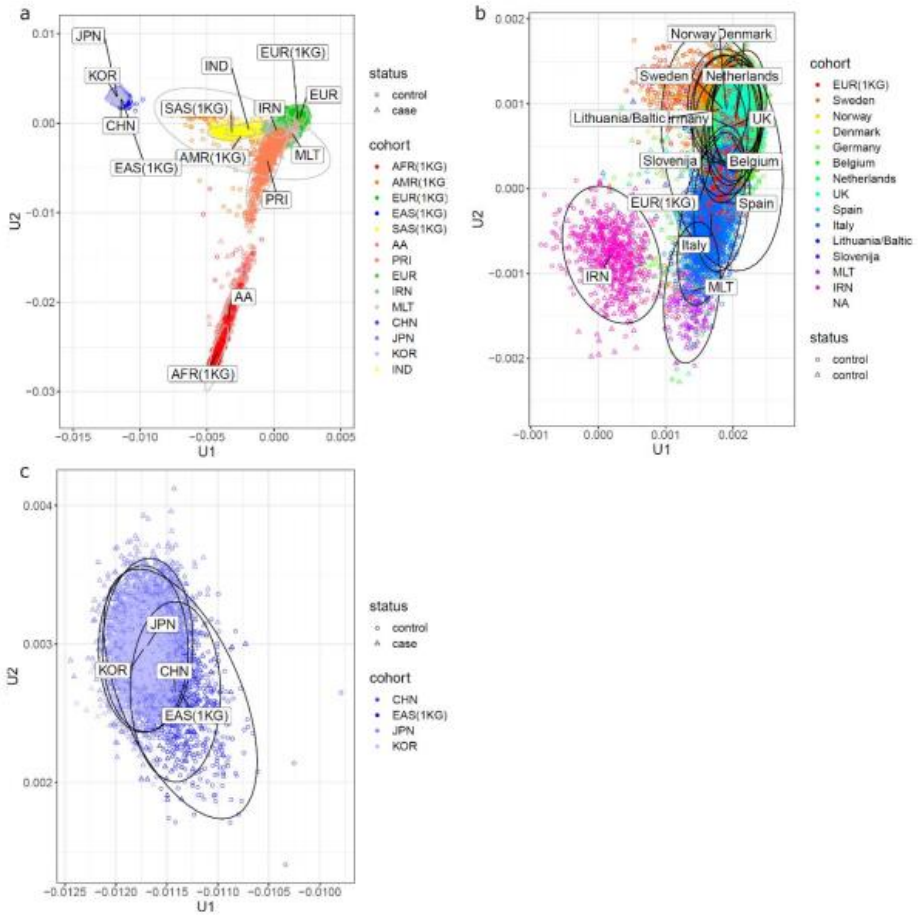
Supplementary Table 4 - HLA pocket definitions for the alpha (DQA1) and beta chains (DQB1, DRB1). Our HLA-II pocket definitions are based on direct sidechain-sidechain interactions between the HLA protein and the nine-peptide residue that contact the HLA binding groove. First, we defined all HLA residues that locate within 7Å of the peptide. Second, members of the same pocket P1 to P9 were only included if the interaction was based on sidechain-sidechain contacts to the same peptide sidechain. By this strict definition we aimed for the highest specificity regarding HLA-peptide interaction, because only a few HLA residues contact more than one peptide sidechain directly. The protein structure of HLA-DRA*01:01-DRB1*01:01 (PBD ID 1dlh) in complex with an influenza virus peptide was used as reference[16].

Pocket	Positions alpha chain	Positions beta chain
1	7,24,31,32,43,52,53,54	82,85,86,89
2	55	77,78,81,82
3	22,54,58,61,62	
4	9,62	13,26,28,70,71,74,77,78
5	61,62	70,71
6	11,62,65,66,69	11,13
7		28,47,61,64,67,70,71
8	65,68,69,72	60
9	69,72,73,76	9,30,57,61

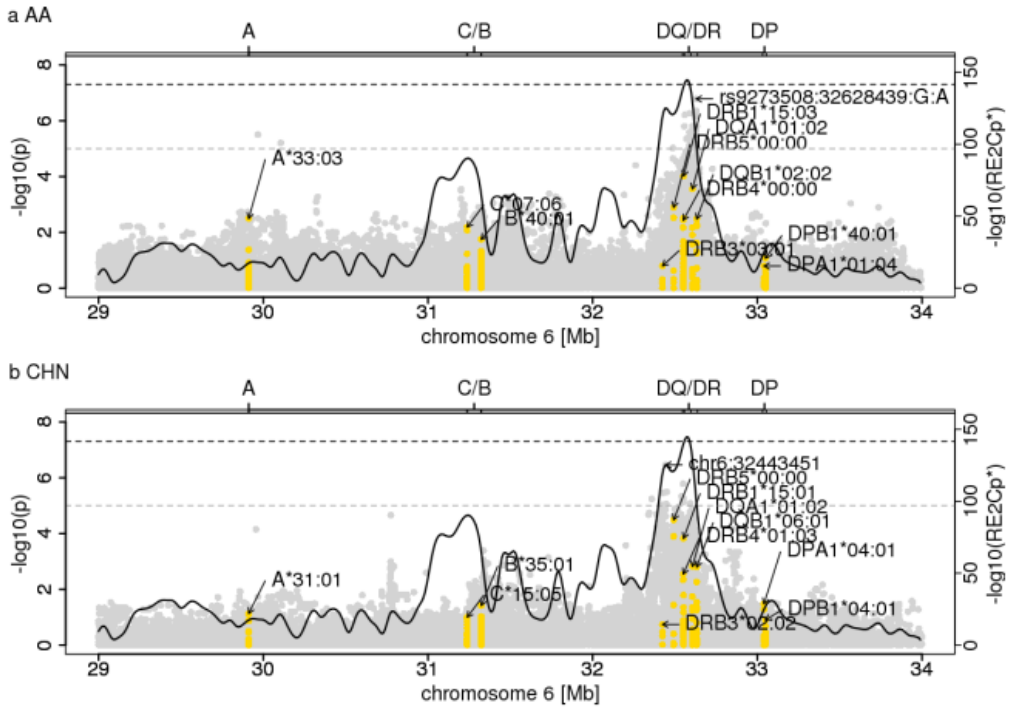
5

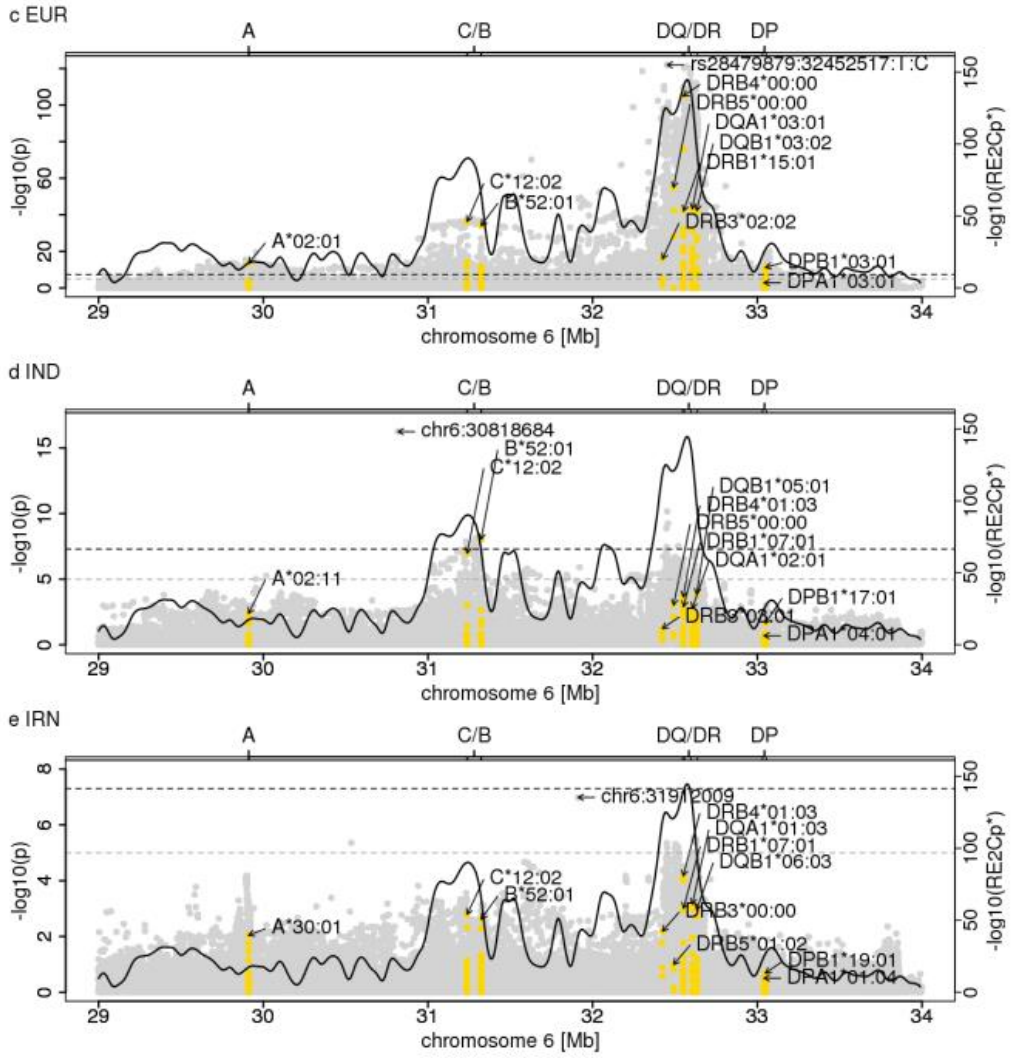
SUPPLEMENTARY FIGURES

Supplementary Figure 1 – Cluster analysis. Cluster analysis of quality-controlled SNP genotypes (a)-(c). **(a)** Analysis of the whole study cohort with African American (AA), Puerto Rican (PRI), European (EUR), Maltese (MLT), Iranian (IRN), North Indian (IND), Chinese (CHN), Korean (KOR) and Japanese (JPN) individuals merged together with the 1000 Genomes Phase III dataset of Admixed American (AMR), African (AFR), Caucasian (EUR), East Asian (EAS) and South Asian (SAS) individuals. **(b)** Zoom into the datasets that match to the European region with detailed patient origins. **(c)** Zoom into the datasets that match to the East Asian region. The Chinese, Korean and Japanese datasets cluster well with the East Asian 1000 Genomes samples, the Indian dataset clusters best with the South Asian 1000 Genomes samples, with the Iranian dataset clustering closer to the European samples. The African American samples cluster closest to the African 1000 Genomes samples and towards the 1000 Genomes European samples, while the Puerto Rican dataset clusters closest to the 1000 Genomes European samples and towards the 1000 Genomes African samples. Both datasets contain admixed populations. The Maltese samples cluster closest to the Italian population within our European data.

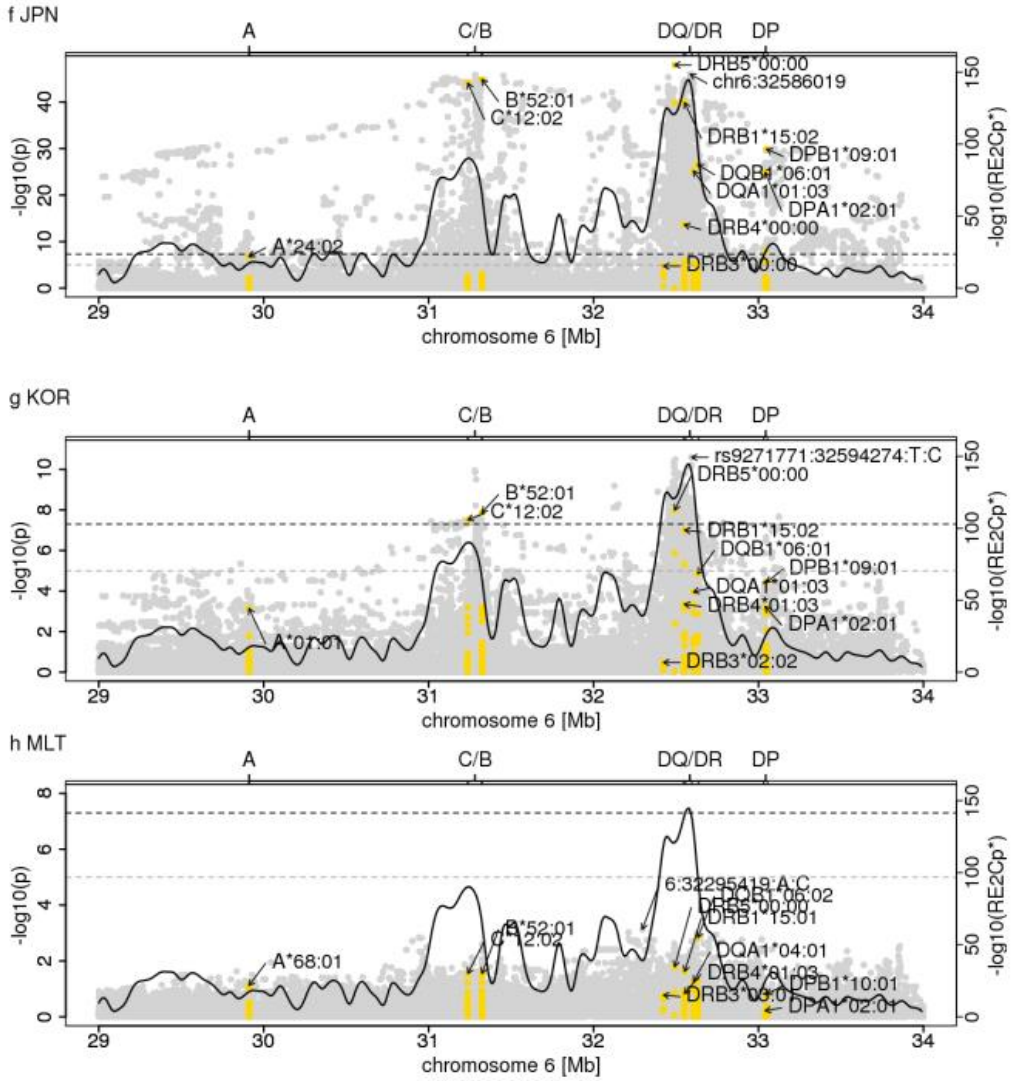


Supplementary Figure 2 – Association analysis on single HLA alleles Univariate association analysis. **(a-f)** Association analysis results for imputed and genotyped SNVs (grey) and 4-digit HLA alleles **(Supplementary Table 5)** (yellow) are shown for **(a)** 373 African American cases and 590 controls (AA), **(b)** 141 Chinese cases and 228 controls (CHN), **(c)**13,927 Caucasian cases and 26,764 controls (EUR) **(d)** 779 Indian cases and 842 controls, (IND), **(e)** 343 Iranian cases and 312 controls (IRN), and **(f)** 709 Japanese cases 69 and controls (JPN) as well as **(g)** 704 Korean cases and 461 controls (KOR), **(h)** 98 Maltese cases and 190 controls (MLT), **(i)** 202 Puerto Rican cases and 419 controls (PRI) at variants with MAF >1% and **(i)** the meta-analysis (META) results from the analysis with RE2C (Lee *et al.*[20]) at variants with a MAF > 1% in the respective cohorts (including 17,276 cases and 32,975 controls from 9 different cohorts). Arrows indicate the top associated alleles for each locus and the top associated SNP. The overlying curves show the P-value of the meta-analysis (REC2p*), including 17,276 cases and 32,975 controls from 9 different cohorts. Dashed lines indicate the thresholds of genome-wide ($P=5x10^{-8}$) and nominal significance ($P=10^{-5}$).

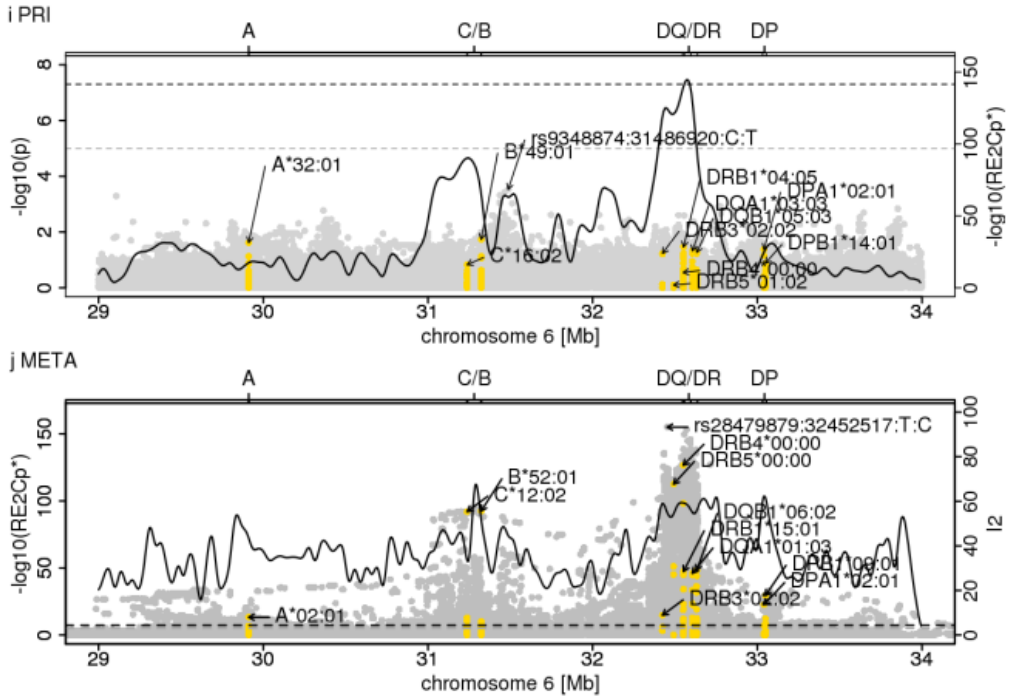



c EUR
A
C/B
DQ/DR
DP
rs28479879:32452517:T:C
DRB4*00:00
DRB5*00:00
DQA1*03:01
DQB1*03:02
DRB1*15:01
C*12:02
B*52:01
DRB3*02:02
A*02:01
DPB1*03:01
DPA1*03:01
-log10(p)
-log10(RE2Cp*)
chromosome 6 [Mb]
d IND
chr6:30818684
B*52:01
C*12:02
DQB1*05:01
DRB4*01:03
DRB5*00:00
DRB1*07:01
DQA1*02:01
A*02:11
DPB1*17:01
DPA1*04:01
e IRN
chr6:31972009
DRB4*01:03
DQA1*01:03
DRB1*07:01
DQB1*06:03
C*12:02
B*52:01
DRB3*00:00
A*30:01
DRB5*01:02
DPB1*19:01
DPA1*01:04

8

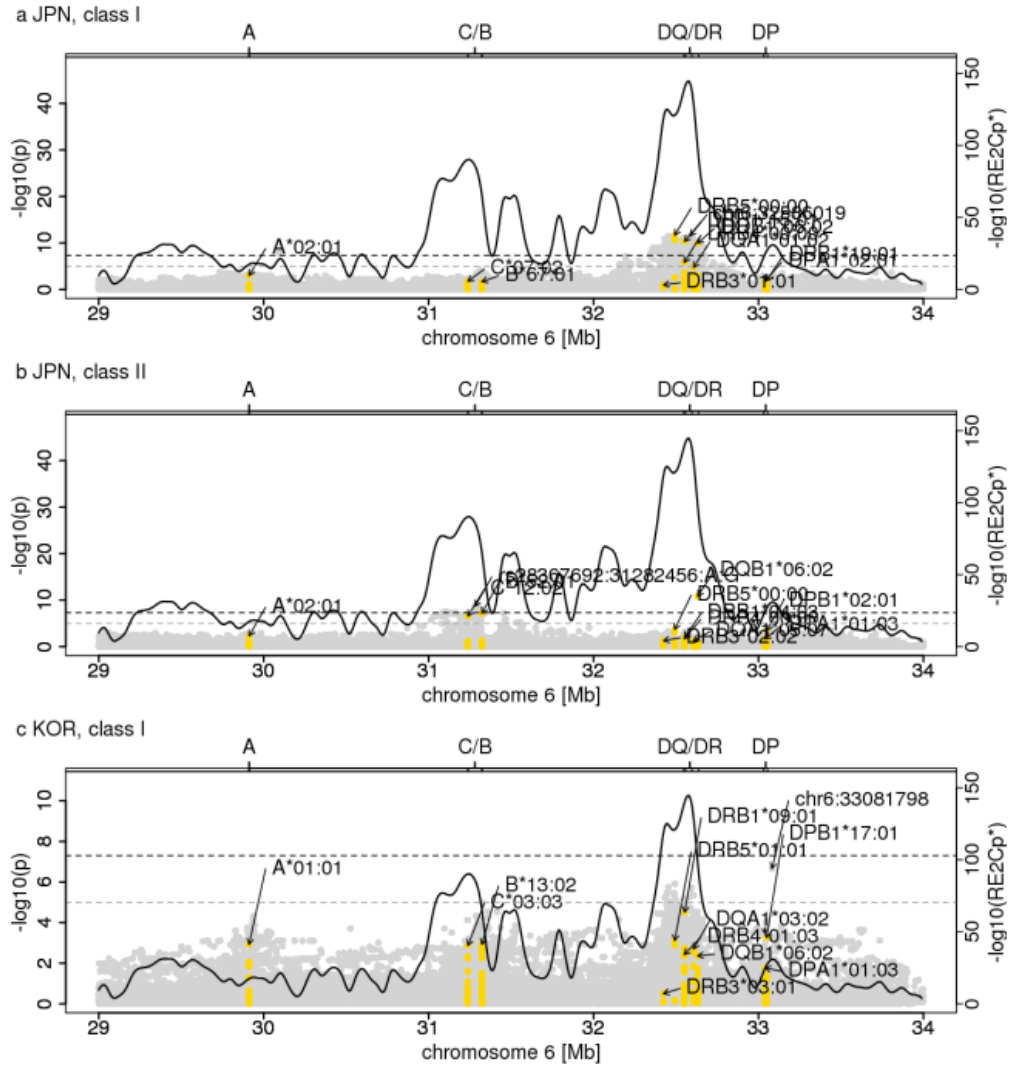

f JPN
A
C/B
DQ/DR
DP
DRB5*00:00
chr6:32586019
B*52:01
C*12:02
DRB1*15:02
DPB1*09:01
DQB1*06:01
DQA1*01:03
DPA1*02:01
DRB4*00:00
A*24:02
DRB3*00:06
-log10(p)
-log10(RE2Cp*)
chromosome 6 [Mb]
g KOR
rs9271771:32594274:T:C
DRB5*00:00
B*52:01
C*12:02
DRB1*15:02
DQB1*06:01
DPB1*09:01
DQA1*01:03
DRB4*01:03
DPA1*02:01
A*01:01
DRB3*02:02
h MLT
6:32295419:A:C
DQB1*06:02
DRB5*00:00
DRB1*15:01
B*52:01
C*12:02
DQA1*04:01
A*68:01
DRB4*01:03
DPB1*10:01
DRB3*00:01
DPA1*02:01

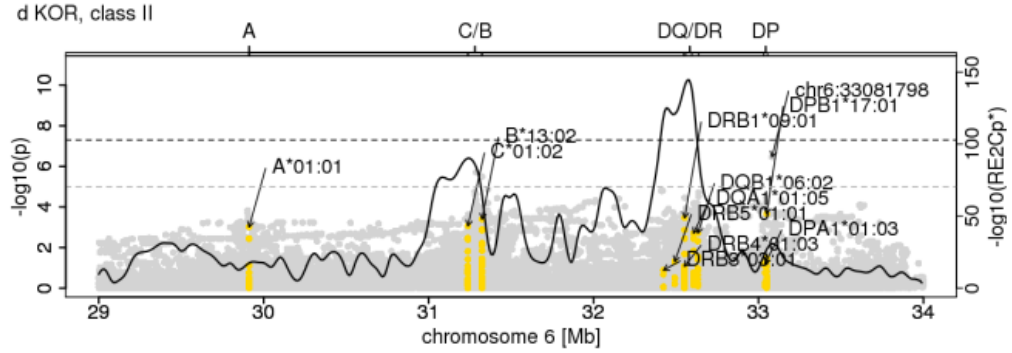
9


i PRI
A
C/B
DQ/DR
DP
-log10(p)
-log10(RE2Cp*)
rs9348874:31486920:C:T
A*32:01
B*49:01
C*16:02
DRB1*04:05
DPA1*02:01
DQA1*03:03
DQB1*05:03
DRB3*02:02
DPB1*14:01
DRB4*00:00
DRB5*01:02
chromosome 6 [Mb]
j META
rs28479879:32452517:T:C
DRB4*00:00
DRB5*00:00
B*52:01
C*12:02
DQB1*06:02
DRB1*15:01
DQA1*01:03
DPA1*02:01
DRB3*02:02
A*02:01
I2

10

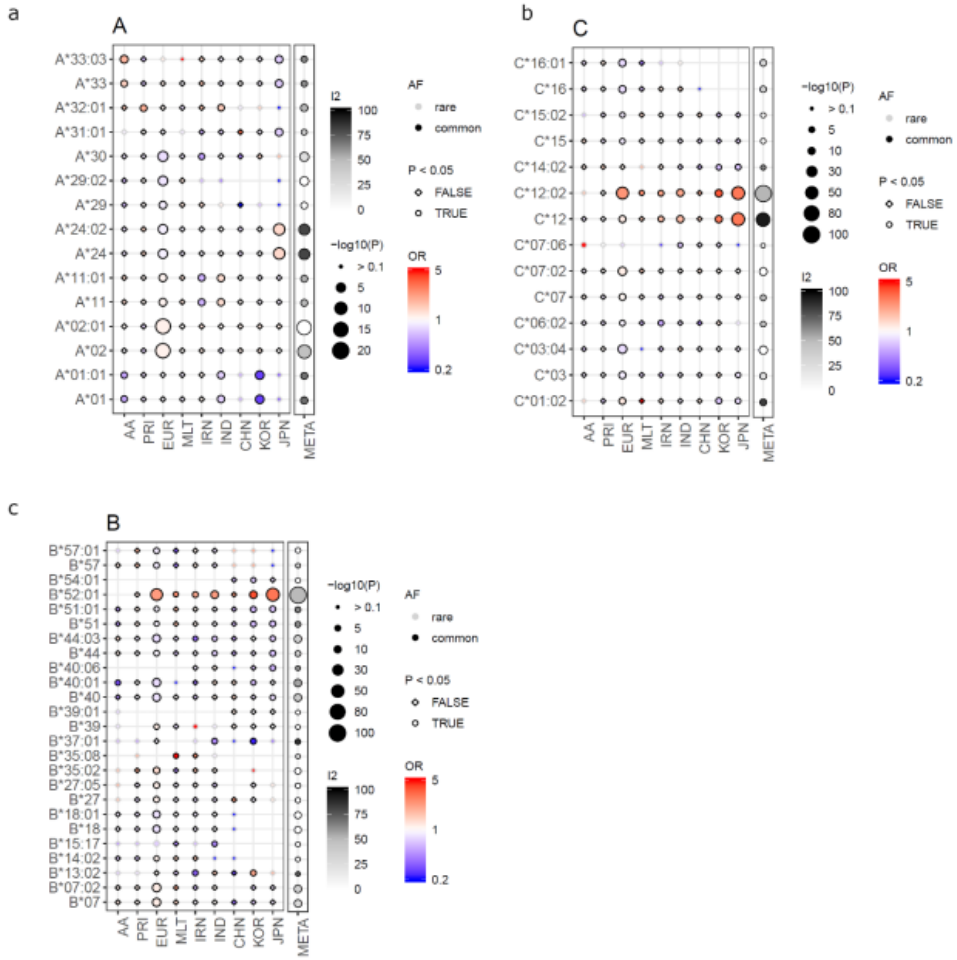
Supplementary Figure 3 - Conditional analysis for Korean and Japanese cohorts.
(**a-d**) Association analysis of UC conditioned on B*52:01, C*12:02, DQA1*01:03, DQB1*06:01 and DRB1*15:02 alleles. Association analysis results for imputed and genotyped SNVs (grey) and 4-digit HLA alleles in the extended HLA region are shown for **(a,b)** 709 Japanese cases and 69 controls (JPN), **(c,d)** 704 Korean cases and 461 controls (KOR) at variants with MAF >1%. Arrows indicate the remaining top associated alleles for each locus and the top associated SNPs. The overlying curve in shows the P-value of the meta-analysis (REC2p*) including 17,276 cases and 32,975 controls from 9 different cohorts. **a** and **c** show the results of conditioning on HLA class I B*52:01 and C*12:02 while **b** and **d** show the results of conditioning on HLA class II DRB1*15:02 and DQA1*01:03, DQB1*06:01.

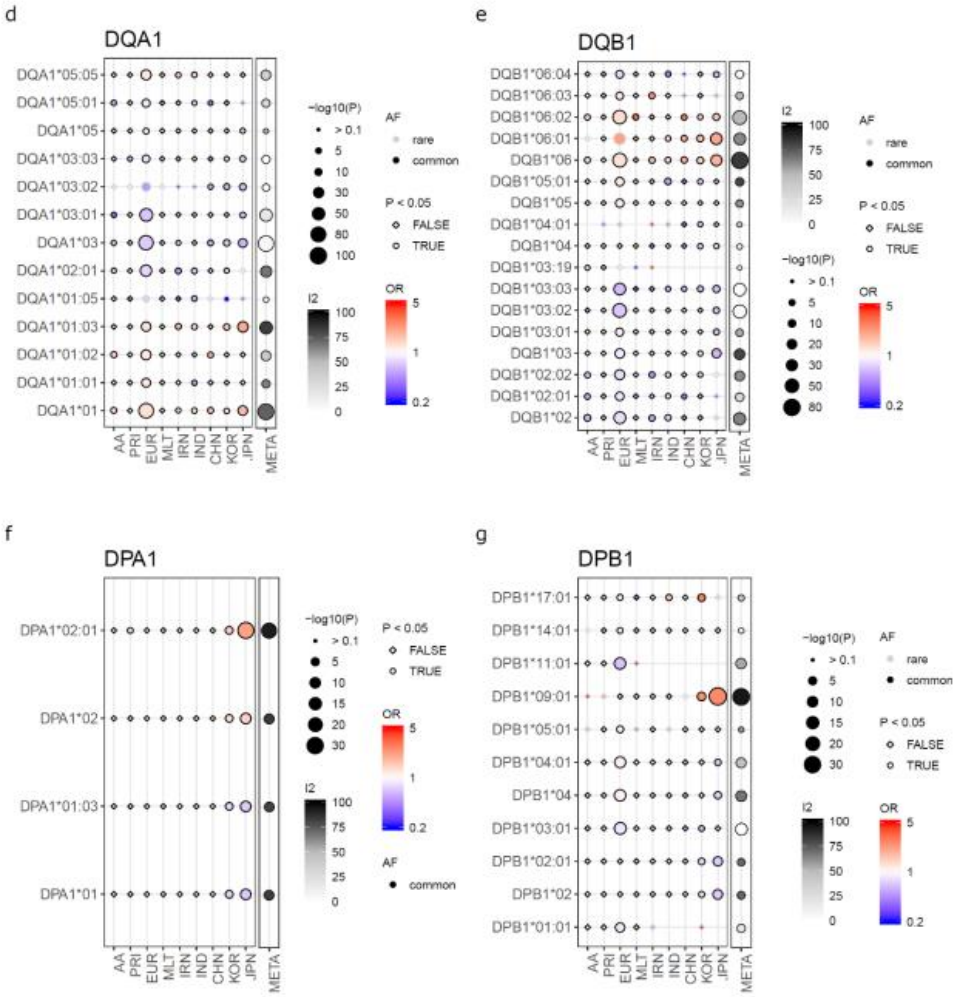



d KOR, class II
A
C/B
DQ/DR
DP
-log10(p)
-log10(RE2Cp*)
chr6:33081798
DPB1*17:01
DRB1*09:01
B*13:02
C*01:02
A*01:01
DQB1*06:02
DQA1*01:05
DRB5*01:01
DPA1*01:03
DRB4*01:03
DRB3*02:01
chromosome 6 [Mb]

12

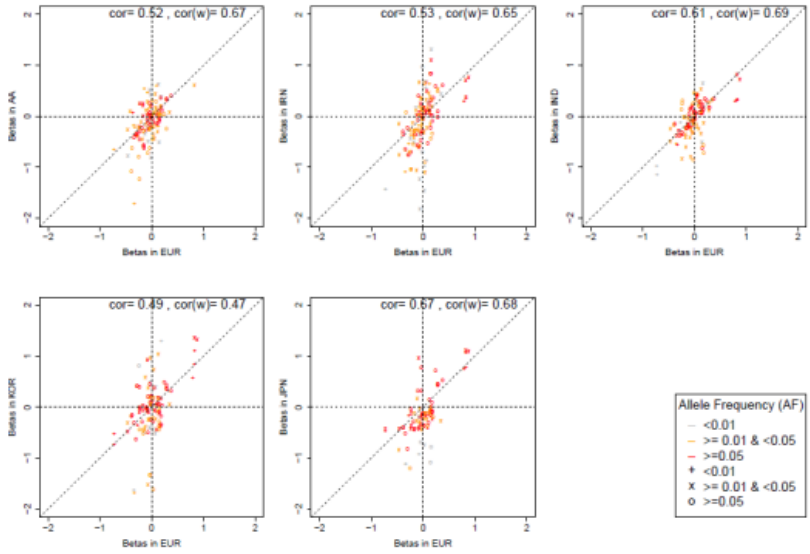
Supplementary Figure 4 – Association analysis for single HLA alleles. Results at 2- and 4-digit resolution for HLA class I loci HLA-*A, -C, -B* **(a-c)** and class II loci HLA-*DQA1, -DQB1, -DPA1, -DPB1* **(d-g).** (AF; common defined as AF>1%), odds ratio (OR), P-value (P) and whether an allele had a P-value<0.05 (circle symbol) is shown for the respective population (e.g. circles with black boundary and *red* color represent an allele that is common and associated with risk). We depict association results of the analysis of the African American (AA), Puerto Rican (PRI), Caucasian (EUR), Maltese (MLT), Iranian (IRN), Indian (IND), Chinese (CHN), Korean (KOR) and Japanese (JPN) cohorts and the meta-analysis (META) with RE2C's I2 as an indicator of allelic heterogeneity and the P-value of association (RE2Cp*, combined here with single study P-values P). Only HLA alleles which are significant in the meta-analysis, that have an AF>1% in at least one population and that have a marginal post imputation probability >0.6 are shown. Alleles with OR>5 or OR<0.2 (rare and non-significant alleles may have larger/smaller OR) values were "ceiled" at 5.0 and 0.2 respectively. The alleles of the HLA-*A*, *-DPA1*, *-DPB1* genes do not show consistently strong association signals across the different ethnicities. C*12:02 and B*52:01 are located on a haplotype with DRB1*15:02.



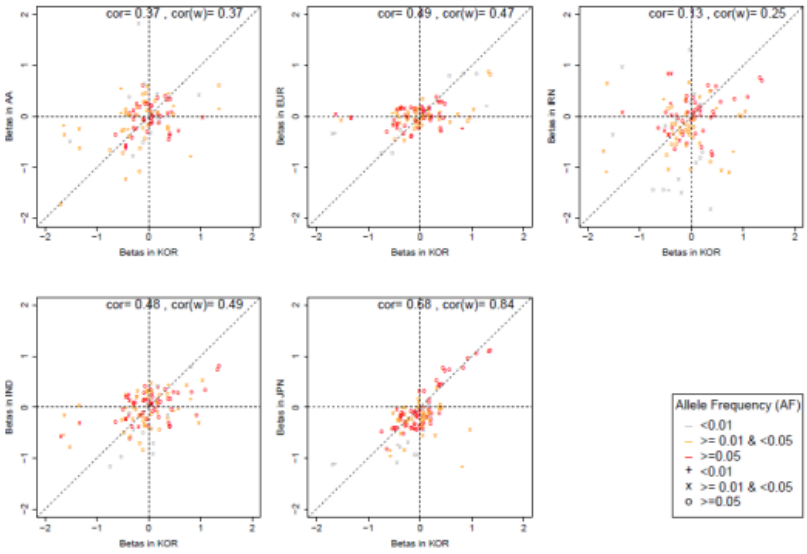

d
DQA1
DQA1*05:05
DQA1*05:01
DQA1*05
DQA1*03:03
DQA1*03:02
DQA1*03:01
DQA1*03
DQA1*02:01
DQA1*01:05
DQA1*01:03
DQA1*01:02
DQA1*01:01
DQA1*01
AA PRI EUR MLT IRN IND CHN KOR JPN META
−log10(P)
> 0.1
5
10
30
50
80
100
I2
100
75
50
25
0
AF
rare
common
P < 0.05
FALSE
TRUE
OR
5
1
0.2
e
DQB1
DQB1*06:04
DQB1*06:03
DQB1*06:02
DQB1*06:01
DQB1*06
DQB1*05:01
DQB1*05
DQB1*04:01
DQB1*04
DQB1*03:19
DQB1*03:03
DQB1*03:02
DQB1*03:01
DQB1*03
DQB1*02:02
DQB1*02:01
DQB1*02
AA PRI EUR MLT IRN IND CHN KOR JPN META
I2
100
75
50
25
0
AF
rare
common
P < 0.05
FALSE
TRUE
−log10(P)
> 0.1
5
10
20
30
50
80
OR
5
1
0.2
f
DPA1
DPA1*02:01
DPA1*02
DPA1*01:03
DPA1*01
AA PRI EUR MLT IRN IND CHN KOR JPN META
−log10(P)
> 0.1
5
10
15
20
30
P < 0.05
FALSE
TRUE
OR
5
1
0.2
I2
100
75
50
25
0
AF
common
g
DPB1
DPB1*17:01
DPB1*14:01
DPB1*11:01
DPB1*09:01
DPB1*05:01
DPB1*04:01
DPB1*04
DPB1*03:01
DPB1*02:01
DPB1*02
DPB1*01:01
AA PRI EUR MLT IRN IND CHN KOR JPN META
−log10(P)
> 0.1
5
10
15
20
30
I2
100
75
50
25
0
AF
rare
common
P < 0.05
FALSE
TRUE
OR
5
1
0.2

Supplementary Figure 5 – Correlation between effects in different populations. We show the pairwise correlation between the beta estimates of the analysis in two different populations. Betas in analysis of African American (AA), Caucasian (EUR), Iranian (IRN), North Indian (IND), Korean (KOR) and Japanese (JPN) are shown. Correlation of effects were calculated on alleles that had a frequency > 0.5% in each population. Both Pearson correlation (cor) and weighted correlation (cor(w)) were calculated. The weighted correlation estimates were weighted by the inverse of the larger variance for an allele in the populations that were analyzed (which might originate from either low AF or low sample size). Colors represent AF in the population on the x-axis, symbol represent AF in the population on the y-axis. Alleles with frequencies of <1% in both cohorts were omitted.

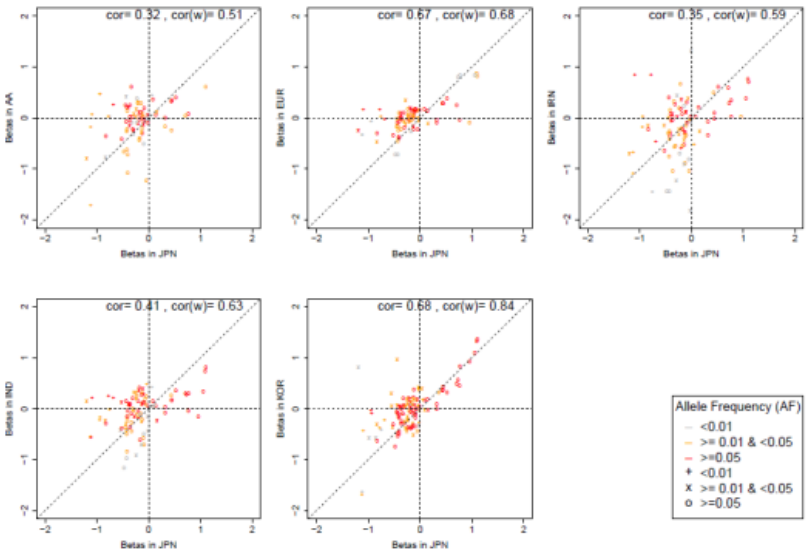
(a, EUR)



(b, KOR)


cor= 0.37 , cor(w)= 0.37
cor= 0.49 , cor(w)= 0.47
cor= 0.13 , cor(w)= 0.25
cor= 0.48 , cor(w)= 0.49
cor= 0.68 , cor(w)= 0.84
Betas in AA
Betas in EUR
Betas in IRN
Betas in IND
Betas in JPN
Betas in KOR
Allele Frequency (AF)
<0.01
>= 0.01 & <0.05
>=0.05
<0.01
>= 0.01 & <0.05
>=0.05

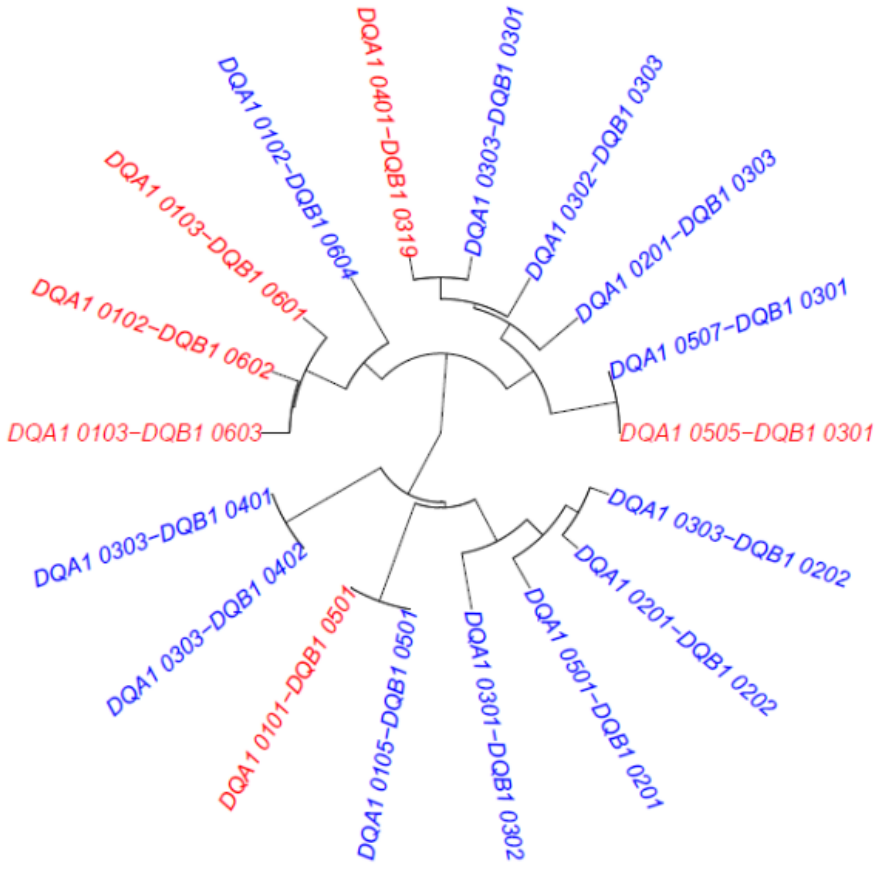
(c, JPN)


cor= 0.32 , cor(w)= 0.51
cor= 0.67 , cor(w)= 0.68
cor= 0.35 , cor(w)= 0.59
cor= 0.41 , cor(w)= 0.63
cor= 0.68 , cor(w)= 0.84
Betas in AA
Betas in EUR
Betas in IRN
Betas in IND
Betas in KOR
Betas in JPN
Allele Frequency (AF)
<0.01
>= 0.01 & <0.05
>=0.05
<0.01
>= 0.01 & <0.05
>=0.05

16

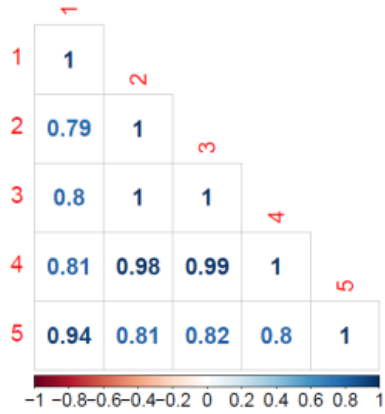
Supplementary Figure 6 – Clustering of DRB1 and DQA1-DQB1 proteins according to preferential peptide binding.

For 5 replicate sets of 200.000 unique random human peptides, the percentile rank scores of preferential peptide binding were calculated using NetMHCIIpan-3.2 (Jensen *et al.*[21]). Within each set, the top 2% binders (following default NetMHCIIpan threshold) were used to perform a clustering on the pairwise correlations between two proteins using complete observations (i.e. data points present in the top 2% of both respective proteins) only. **(a)** Exemplary clustering for peptide set 2 and DQBA1-DBQ1. Labels were colored according to risk (*red*) or protective (*blue*). We show all DQA1-DQB1 that were significant in the meta-analysis of genetic association of the HLA. **(b)** Cluster stability between the 5 sets of 200,000 random peptides. Cluster stability was evaluated using the corrplot (version 0.84) and dendextend (version 1.12) packages in R for DRB1 (1) and DQA1-DQB1 (2). Here, the correlation between cluster dendrograms (i.e. the concordance of the tree-structure) is calculated with a value of 0 signifying dissimilar tree-structures and 1 signifying highly similar tree-structures.

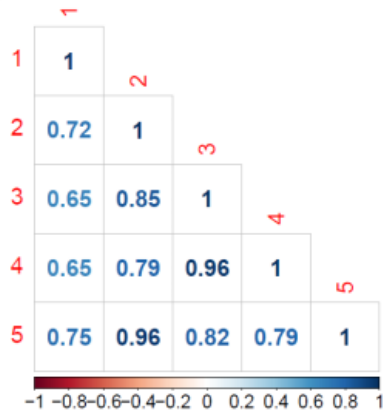


17

(b1, DRB1)



(b2, DQ)

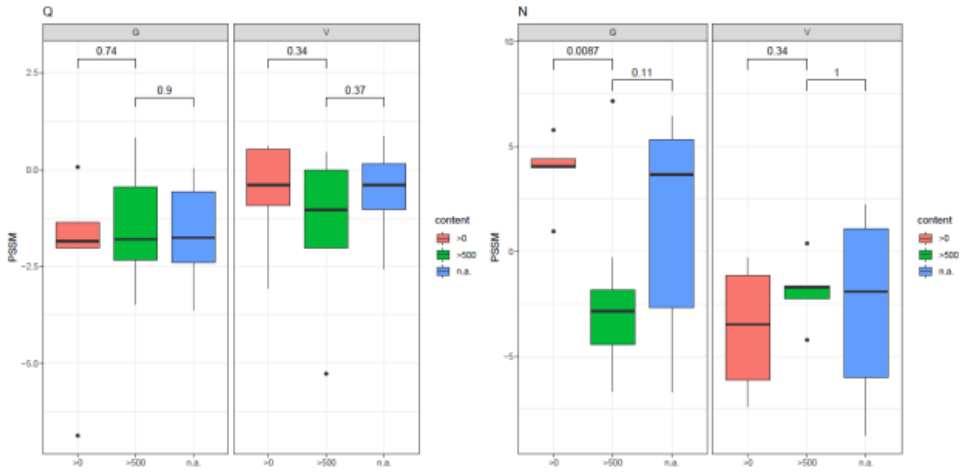


18

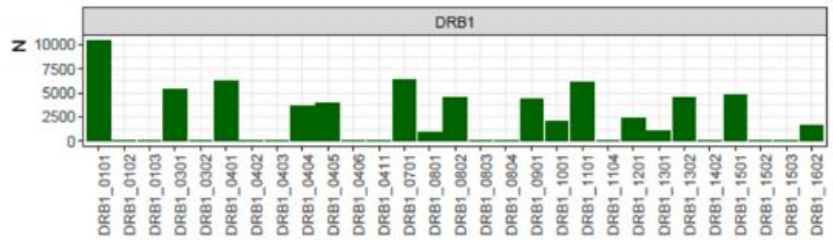
Supplementary Figure 7 – Why DRB1*15:02 and DRB1*13:02 may cluster together

Within the peptide binding pockets P1-P9 (**Methods**), DRB1*15:01/03 and DRB1*15:02 differ in their amino acid composition only in position 86 (86) within P4. Here, DRB1*15:02 displays a glycine (G) instead of valine (V), which it shares, amongst others, with DRB1*13:02. Peptides predicted to bind to DRB1*13:02 are frequently showing asparagine (N) at position 4 (position binding to P4 of the HLA protein). We observed that N is also present in peptides predicted to bind to DRB1*15:02 and other alleles that have a G at position 86B of the HLA protein. Screening through the peptide binding motifs of the DRB1 alleles with a G at position 86B, we observed that, among these alleles, only the alleles which are underrepresented or not present in the dataset have a higher chance of an N occurring at position 4 of their peptide binding motifs. Though the amount of data present for DRB1*13:02 in the training dataset is comparably lower or the same as other alleles in this group, DRB1*13:02 may have a greater impact on the extrapolation of peptide binding HLA proteins with few or no training data, because among its bound peptides, N is very prominent at position 4.

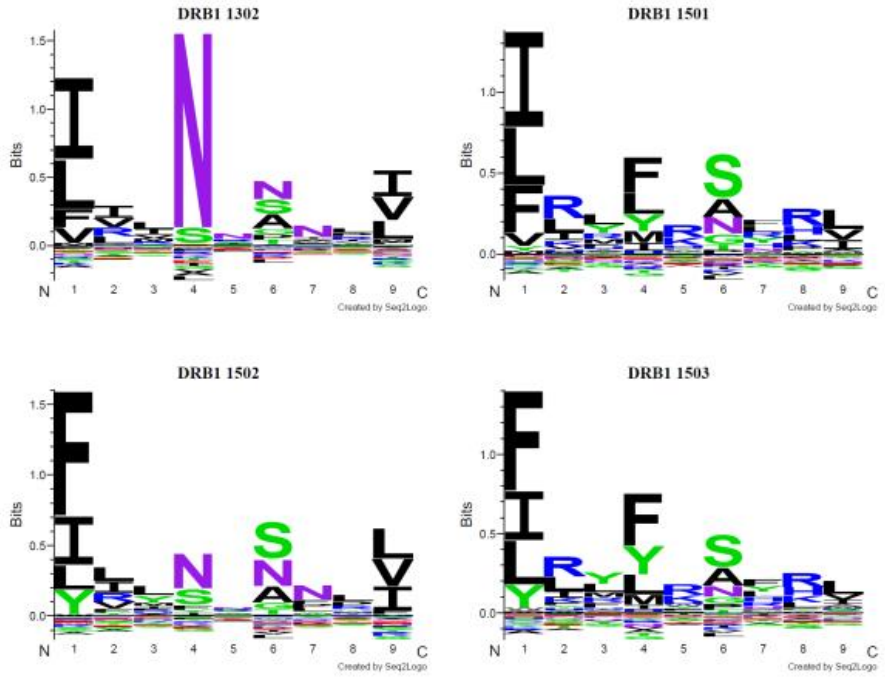
(a) Exemplary boxplots showing positions-specific scores within the position-specific scoring matrix (PSSM) of glutamine (Q) and asparagine (N) at position 4 of the 9 amino acid peptide. PSSM scores were calculated with Seq2Logo[22] based on preferential binders (top 2% as given by NetMHCIIpan-3.2[21]) of set 2 of the 200,000 unique random human peptides. PSSM values are given in halfbits and indicate which amino acids are typically over – (positive) or under- (negative) represented at a specific position. In total 8/5/17 proteins with a valine (V) at position 86B (beta-chain at position 86) in pocket 4 had <=500/>500/no (n.a.) recorded peptide-protein affinities, respectively. In total 5/11/11 proteins with a glycine (G) at position 85B had <=500/>500/no recorded peptide-protein affinities



(b) Number of peptide-protein affinity measurements used for the training in NetMHCIIpan-3.2 (based on *train1* and *test1* datasets downloaded from the NetMHCIIpan-3.2 server).



(c) Peptide motifs of DRB1*13:02 and DRB1*15 alleles.

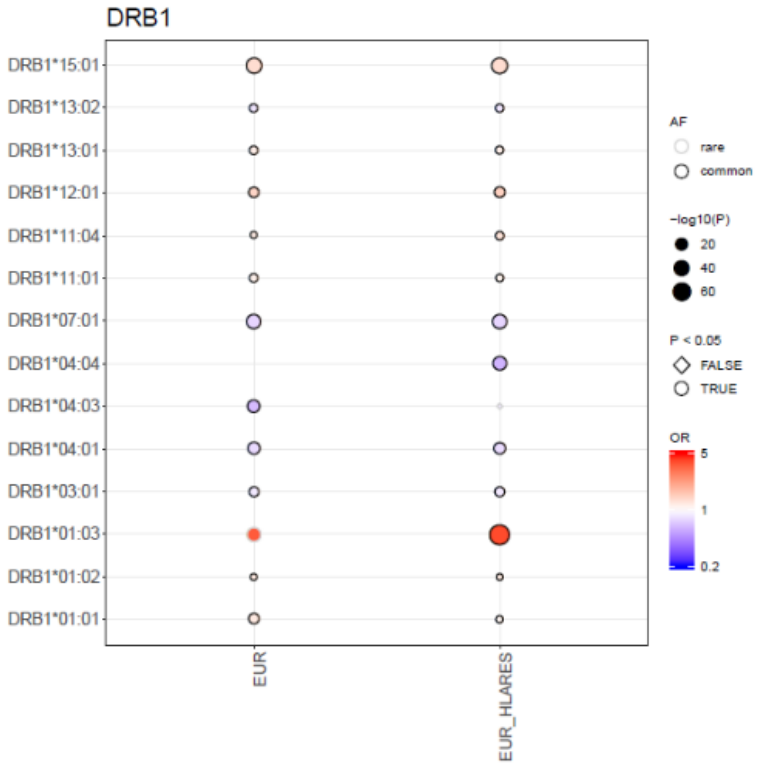


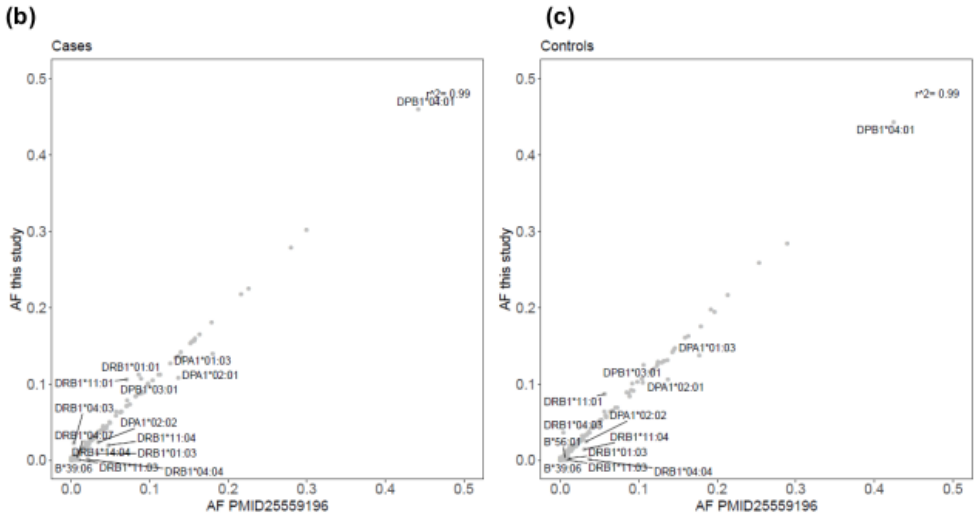
20

Supplementary Figure 8 – Analysis of the previously published Caucasian population.

(a) Reestablishing the DRB1*01:03 signal using the HLARES panel We show allele frequency (AF), odds ratio (OR), P-value (P) and whether an allele had a P-value of < 0.05 for the HLA imputation with our own reference panel and the HLARES panel published with HIBAG[23] (ImmunoChip-EuropeanHLARES-HLA4-hg19.RData). Only HLA alleles which are significant at a P-value threshold of 0.05 and have an AF >1% in at least one cohort are shown. The HLARES imputation panel was obtained from publicly available pre-trained model of Zheng *et al.*[23].**(b-c)** Comparison to previous study. Allele frequencies are compared between controls and cases separately. Pearson correlation coefficient denoted as r^2. Labels are shown for alleles for which the deviation between studies is within the upper 95% percentile of the positive absolute or relative deviation of alleles with an AF > 0.5% in our previous study (Goyette *et al.*[5], PMID25559196).

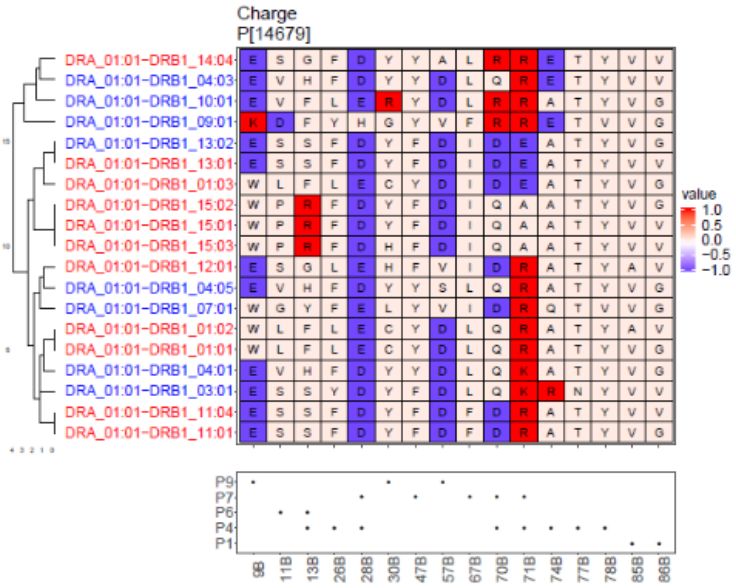
(a)




(b)
Cases
AF this study
AF PMID25559196
r^2= 0.99
DPB1*04:01
DRB1*01:01
DPA1*01:03
DPA1*02:01
DRB1*11:01
DPB1*03:01
DRB1*04:03
DPA1*02:02
DRB1*04:07
DRB1*11:04
DRB1*14:04
DRB1*01:03
B*39:06
DRB1*11:03
DRB1*04:04
(c)
Controls
AF this study
AF PMID25559196
r^2= 0.99
DPB1*04:01
DPA1*01:03
DPB1*03:01
DPA1*02:01
DRB1*11:01
DPA1*02:02
DRB1*04:03
B*56:01
DRB1*11:04
DRB1*01:03
B*39:06
DRB1*11:03
DRB1*04:04

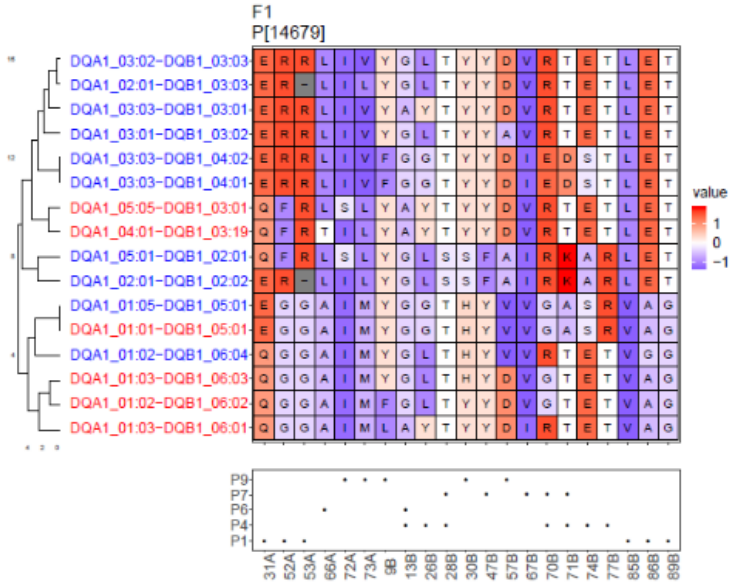
22

Supplementary Figure 9 - Results of physico-chemical analysis of DRB1 - Charge. We only show sites with variable information in pockets (P) 1, 4, 6, 7 and 9 and only proteins for which the genetic analysis was significant (meta-analysis RE2Cp* <0.05) and for which at least 1 cohort had AF >1% are depicted. (**Methods**). Clustering was performed using the hclust function of the R package stats. Box below cluster plot shows positions of P1, 4, 6, 7, 9 at the alpha (A) and beta (B) chains of the molecules. The score "charge" used here, describes the presence of positive and negative charge of an amino acid sidechain by integers (positive charge: +1, negative charge: -1 and absence of charge:0).

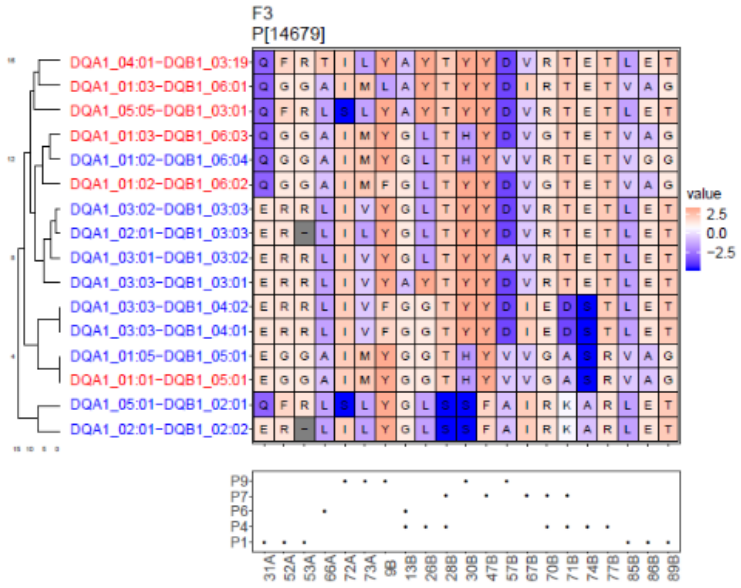


Supplementary Figure 10 - Results of physico-chemical analysis of DQA1-DQB1 properties. We show the analysis on (a) F1, (b) F2, (c) HB-Acceptor, (d) residue volume and (e) charge. We only show sites with variable information in pockets (P)1, 4 ,6, 7 and 9 (**Methods**) and only proteins for which the genetic analysis was significant (meta-analysis RE2Cp* <0.05) and for which at least 1 cohort had AF >1% are depicted. Clustering was performed using the hclust function of the R package stats. Box below cluster plot shows positions of P1, 4, 6, 7 and 9 of the alpha (A) and beta (B) chains of the molecules. Here we show parameters F1, F3 which were taken from Atchley et al.[17] and were calculated in a factor analysis from 54 unique amino acid properties. F1 captures polarity and hydrophobicity of the amino acid, while factor F3 captures amino acid size and bulkiness. For F1, high values indicate larger hydrophobicity, polarity and hydrogen donor abilities while low values indicate non-polar amino acids. For F3, high values indicate larger and bulkier amino acids while low values indicate smaller, more flexible amino acids. Additionally, we defined two scores and "hydrogen acceptor" (HB-acceptor) and "charge". The score "charge" used here, describes the presence of positive and negative charge of an amino acid sidechain by integers (positive charge: +1, negative charge: -1 and absence of charge:0). The value applied for HB-acceptor defines the ability of amino acids to participate in hydrogen bonds and corresponds to the number of atoms within amino acid sidechains accept a hydrogen. We additionally show the residue-volume as a measure of pocket size.

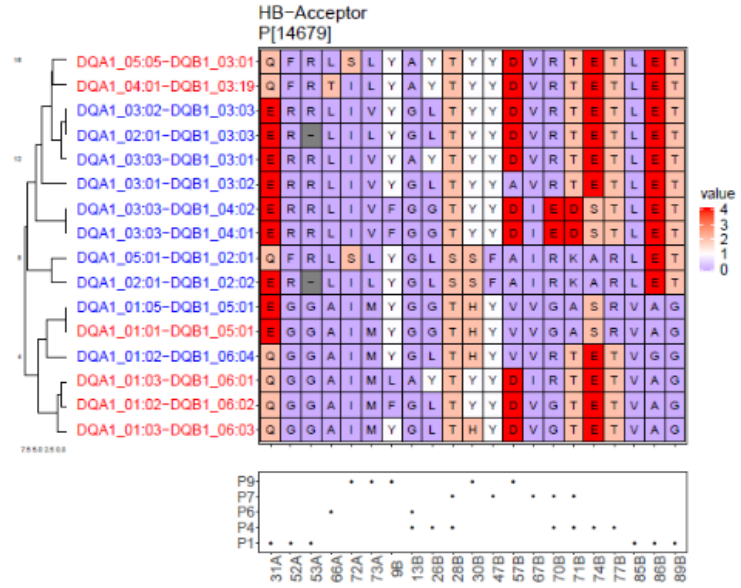
(a)



(b)

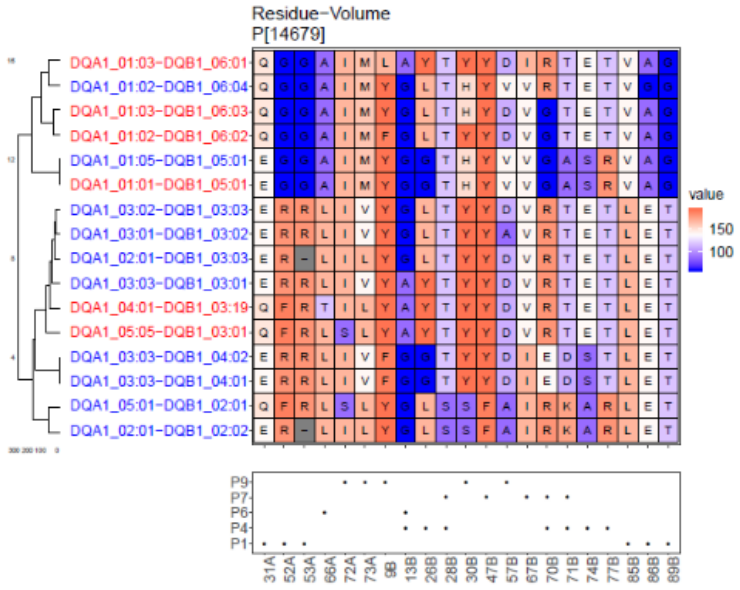

F3
P[14679]
DQA1_04:01-DQB1_03:19
DQA1_01:03-DQB1_06:01
DQA1_05:05-DQB1_03:01
DQA1_01:03-DQB1_06:03
DQA1_01:02-DQB1_06:04
DQA1_01:02-DQB1_06:02
DQA1_03:02-DQB1_03:03
DQA1_02:01-DQB1_03:03
DQA1_03:01-DQB1_03:02
DQA1_03:03-DQB1_03:01
DQA1_03:03-DQB1_04:02
DQA1_03:03-DQB1_04:01
DQA1_01:05-DQB1_05:01
DQA1_01:01-DQB1_05:01
DQA1_05:01-DQB1_02:01
DQA1_02:01-DQB1_02:02
value
2.5
0.0
−2.5
P9
P7
P6
P4
P1
31A 52A 53A 66A 72A 73A 9B 13B 26B 28B 30B 47B 57B 67B 70B 71B 74B 77B 85B 86B 89B

(c)

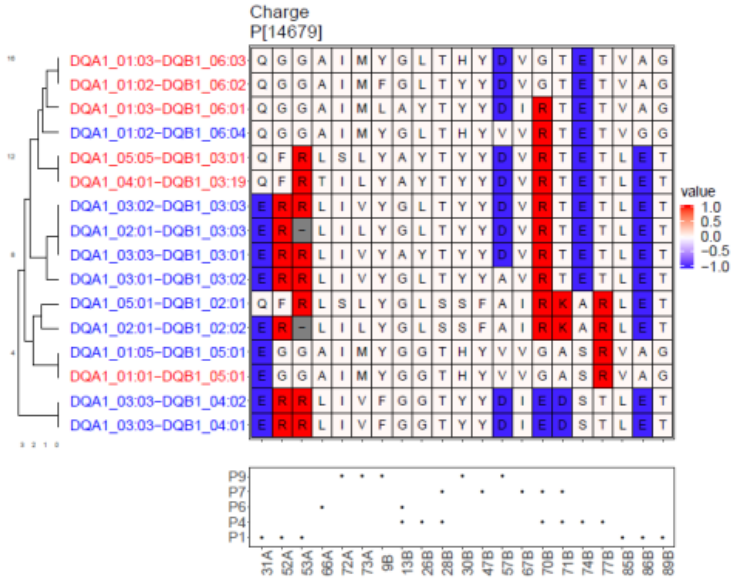

HB−Acceptor
P[14679]
DQA1_05:05-DQB1_03:01
DQA1_04:01-DQB1_03:19
DQA1_03:02-DQB1_03:03
DQA1_02:01-DQB1_03:03
DQA1_03:03-DQB1_03:01
DQA1_03:01-DQB1_03:02
DQA1_03:03-DQB1_04:02
DQA1_03:03-DQB1_04:01
DQA1_05:01-DQB1_02:01
DQA1_02:01-DQB1_02:02
DQA1_01:05-DQB1_05:01
DQA1_01:01-DQB1_05:01
DQA1_01:02-DQB1_06:04
DQA1_01:03-DQB1_06:01
DQA1_01:02-DQB1_06:02
DQA1_01:03-DQB1_06:03
value
4
3
2
1
0
P9
P7
P6
P4
P1
31A 52A 53A 66A 72A 73A 9B 13B 26B 28B 30B 47B 57B 67B 70B 71B 74B 77B 85B 86B 89B

25

(d)


Residue−Volume
P[14679]
DQA1_01:03−DQB1_06:01
DQA1_01:02−DQB1_06:04
DQA1_01:03−DQB1_06:03
DQA1_01:02−DQB1_06:02
DQA1_01:05−DQB1_05:01
DQA1_01:01−DQB1_05:01
DQA1_03:02−DQB1_03:03
DQA1_03:01−DQB1_03:02
DQA1_02:01−DQB1_03:03
DQA1_03:03−DQB1_03:01
DQA1_04:01−DQB1_03:19
DQA1_05:05−DQB1_03:01
DQA1_03:03−DQB1_04:02
DQA1_03:03−DQB1_04:01
DQA1_05:01−DQB1_02:01
DQA1_02:01−DQB1_02:02
value
150
100
P9
P7
P6
P4
P1
31A 52A 53A 66A 72A 73A 9B 13B 26B 28B 30B 47B 57B 67B 70B 71B 74B 77B 85B 86B 89B

(e)


Charge
P[14679]
DQA1_01:03−DQB1_06:03
DQA1_01:02−DQB1_06:02
DQA1_01:03−DQB1_06:01
DQA1_01:02−DQB1_06:04
DQA1_05:05−DQB1_03:01
DQA1_04:01−DQB1_03:19
DQA1_03:02−DQB1_03:03
DQA1_02:01−DQB1_03:03
DQA1_03:03−DQB1_03:01
DQA1_03:01−DQB1_03:02
DQA1_05:01−DQB1_02:01
DQA1_02:01−DQB1_02:02
DQA1_01:05−DQB1_05:01
DQA1_01:01−DQB1_05:01
DQA1_03:03−DQB1_04:02
DQA1_03:03−DQB1_04:01
value
1.0
0.5
0.0
−0.5
−1.0
P9
P7
P6
P4
P1
31A 52A 53A 66A 72A 73A 9B 13B 26B 28B 30B 47B 57B 67B 70B 71B 74B 77B 85B 86B 89B

26

COMPARISON TO PREVIOUS STUDY (GOYETTE *ET AL.*[5])

Previously, we imputed HLA alleles into the Caucasian population using the T1DGC reference (Mychaleckyj *et al.*[24]) panel and SNP2HLA (Jia *et al.*[25]) as well as HLA*IMP (Dilthey *et al.*[26,27]). Imputation results achieved with the T1DGC reference panel are based on custom G groups (for class I alleles with identical sequences at exons 2 and 3 are grouped together; for class II alleles with identical sequences at exon 2 are grouped together). Additionally, there was a major update of allele nomenclature in 2010. This combined mainly has an impact on the alleles of the *HLA-DQ* and *HLA-DP* genes, i.e. frequencies of alleles in the same group are combined. In this study we imputed our cohort using our full context (no G grouping) multi-ethnic reference panel with the most recent nomenclature. To account for this in comparing the results of this study to the results of Goyette *et al.*[5], we estimated summed allele frequencies of the current study and the previous study according to G-group annotation (IPD-IMGT/HLA 3.29.0). Though different imputation panels were used, the association results correlate well **(Supplementary Figure 8).**

LIMITATIONS OF HLA IMPUTATION

We would like to point out a few drawbacks of HLA imputation, since this is also important to understand some of the results. Especially rare and low-frequency (AF <1%) alleles, as well as alleles that are very similar based on the variants that were used to construct the HLA reference are important to note here. Previously, we constructed a reference panel for the imputation of HLA alleles in a multi ethnic context (Degenhardt *et al.*[14]), including sensitivity and specificity values for each single allele in each population. We saw that apart from rare alleles being challenging to impute, there are also some allele groups, comprising more than one allele, for which alleles are challenging to impute from ImmunoChip data to different extents across divergent populations. This is true for alleles of the DRB1*04 group, where DRB1*04:04 is misclassified as DRB1*04:03, for DRB1*11, where DRB1*11:04 is often misclassified as DRB1*11:01 and DQA1*01 where DQA1*01:01 and DQA1*01:05 are similar. These challenges are consistent with those of previous imputation panels and may be overcome by using panels resulting from sequencing approaches.

ANALYSIS OF GENDER EFFECTS

Because former linkage studies for IBD hinted towards sex-specific signals in the HLA region(Fisher et al., Venkateswaran et al.[28,29]), we performed an HLA-wide association analysis within the Caucasian, Japanese, Indian and Korean populations for case only and control only as well as a meta-analysis using RE2C (Lee at al.[30]). We additionally calculated stratified odds ratios (OR) for male and female UC vs. male and female controls. With a larger sample size, the control data (except for the Korean dataset) were predicted to have a greater power of discovering an effect. Overall the association with gender was observed to be moderate, none of the non-Caucasian associations remained significant after P-value correction with Bonferroni-Holm (**Supplementary Table 8**). We could not reproduce the previously observed[29] stronger association with DRB1*01:03 with female gender in the Caucasian population when imputing the HLARES imputation panel published with HIBAG (Zheng et al [23]) (male vs. female UC: OR=1.12 p=0.17). The most associated (though still moderately associated) HLA-*DRB1* alleles in the meta-analysis were DRB1*07:01 (RE2Cp* = 7.01×10^{-5}) and DRB1*04 (RE2Cp*=9.38×10^{-4}) as well as alleles located on the same haplotypes (**Main Manuscript**). For these alleles effect estimates were observed to be slightly smaller in male than in female UC patients As already discussed in the main paper, DRB1*07 and DRB1*04 are both located on the same haplotype as HLA-*DRB4* alleles, which increases power for these alleles show association. Hence, DRB4*00:00 (not having DRB4) and DRB4*01:03 are among the strongest associated alleles in the meta-analysis

(RE2Cp*=6.45x10^{-8} and RE2Cp*=4.69x10^{-6}). The opposite is true for alleles located on the same haplotype as HLA-*DRB3* (DRB3*00:00; RE2Cp* =1.89x10^{-6}), namely alleles located on a haplotype with DRB1*11:01 and DRB1*03:01, for which effect estimates of the alleles were observed to be slightly larger in males than in females.

SUPPLEMENTARY METHODS

In the analysis of gender effects, we performed association analysis in a case only and control only populations with gender as the dependent variable (male was coded as 0 and female as 1) according to

$$\log(\text{odds}_i) = \beta_0 + \beta_1 x_i + \beta_2 U_{1i} + \beta_3 U_{2i} + \beta_4 U_{3i} + \beta_5 U_{4i} + \beta_6 U_{5i} + (\beta_7 b_i)$$

with g as gender and b as batch. Gender specific odds ratios were calculated in the female only case-control data and in the male only case-control data.

Members of the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)

Shifteh Abedian[6,7], Clara Abraham[38], Jean-Paul Achkar[39,40], Tariq Ahmad[41], Rudi Alberts[42], Behrooz Alizadeh[6], Leila Amininejad[43,44], Ashwin N Ananthakrishnan[45,46], Vibeke Andersen[47,48], Carl A Anderson[49], Jane M Andrews[50], Vito Annese[51,52], Guy Aumais[53,54], Leonard Baidoo[55], Robert N Baldassano[56], Peter A Bampton[57], Murray Barclay[58], Jeffrey C Barrett[49], Johannes Bethge[59], Claire Bewshea[41], Joshua C Bis[60], Alain Bitton[61], Thelma BK[8], Gabrielle Boucher[62], Oliver Brain[63], Stephan Brand[64], Steven R Brant[11,37], Jae Hee Cheon[10], Angela Chew[65,66], Judy H Cho[67], Isabelle Cleynen[68], Ariella Cohain[69], Rachel Cooney[70], Anthony Croft[71], Mark J Daly[72,73], Mauro D'Amato[74,75], Silvio Danese[76], Naser Ebrahim Daryani[12], Lisa Wu Datta[11], Frauke Degenhardt[1], Goda Denapiene[77], Lee A Denson[78], Kathy L Devaney[45], Olivier Dewit[79], Renata D'Inca[80], Hazel E Drummond[81], Marla Dubinsky[82], Richard H Duerr[55,83], Cathryn Edwards[84], David Ellinghaus[1], Pierre Ellul[13], Motohiro Esaki[14], Jonah Essers[85,86], Lynnette R Ferguson[87], Eleonora A Festen[42], Philip Fleshner[16], Tim Florin[88], Denis Franchimont[43,44], Andre Franke[1], Yuta Fuyuno[14,15], Richard Gearry[58,89], Michel Georges[90,91], Christian Gieger[92], Jürgen Glas[61], Philippe Goyette[93], Todd Green[73,85], Anne M Griffiths[94], Stephen L Guthery[95], Hakon Hakonarson[96,97], Jonas Halfvarson[98], Katherine Hanigan[71], Talin Haritunians[16], Ailsa Hart[99], Chris Hawkey[100], Nicholas K Hayward[101], Matija Hedl[38], Paul Henderson[102], Georgina L Hold[103], Myhunghee Hong[17], Xinli Hu[104], Hailiang Huang[105,72], Jean-Pierre Hugot[106], Ken Y Hui[96], Marcin Imielinski[16], Omid Jazayeri[107], Laimas Jonaitis[108], Luke Jostins[109], Garima Juyal[18], Ramesh Chandra Juyal[110], Rahul Kalla[81], Tom H Karlsen[36,4], Nicholas A Kennedy[111], Mohammed Azam Khan[112], Won Ho Kim[113], Takanari Kitazono[14], Gediminas Kiudelis[108], Michiaki Kubo[19], Subra Kugathasan[20], Limas Kupcinskas[114], Christopher A Lamb[115], Katrina M de Lange[49], Anna Latiano[51], Debby Laukens[116], Ian C Lawrance[66], James C Lee[117], Charlie W Lees[81], Marcis Leja[118], Nina Lewis[100], Johan Van Limbergen[94], Paolo Lionetti[119], Jimmy Z Liu[49], Edouard Louis[120], Yang Luo[49], Gillian Mahy[121], Masoud Mohammad Malekzadeh[122,123], Reza Malekzadeh[7], John Mansfield[124], Suzie Marriott[41], Dunecan Massey[125], Christopher G Mathew[126], Toshiyuki Matsui[127], Dermot PB McGovern[16], Andrea van der Meulen[128], Vandana Midha[24], Raquel Milgrom[129], Samaneh Mirzaei[122,123], Mitja Mitrovic[107,130], Grant W Montgomery[101], Craig Mowat[131], Christoph Müller[132], William G Newman[112], Aylwin Ng[133,45], Siew C Ng[25], Sok Meng Evelyn Ng[38], Susanna Nikolaus[59], Kaida Ning[38], Markus Nöthen[134], Ioannis Oikonomou[38], David Okou[20], Timothy R Orchard[135], Orazio Palmieri[51], Miles Parkes[125], Anne Phillips[131], Cyriel Y Ponsioen[136], Urõs Potocnik[137,138], Hossein Poustchi[122,123], Natalie J Prescott[126], Deborah D Proctor[38], Graham Radford-Smith[139,71], Jean-Francois Rahier[140], Miguel Regueiro[55], Walter Reinisch[141], Florian Rieder[39], John D Rioux[62,93], Rebecca Roberts[58], Gerhard Rogler[142], Richard K Russell[143], Jeremy D Sanderson[144], Miquel Sans[145], Jack Satsangi[81], Eric E Schadt[69], Michael Scharl[142], John Schembri[13], Stefan Schreiber[1,59], L Philip Schumm[146], Regan Scott[55], Mark Seielstad[147,148], Tejas Shah[49], Yashoda Sharma[38], Mark S Silverberg[129], Alison Simmons[63], Lisa A Simms[71], Abhey Singh[41], Jurgita Skieceviciene[108], Suzanne van Sommeren[42], Kyuyoung Song[17], Ajit Sood[24], Sarah L Spain[126], A. Hillary Steinhart[129], Joanne M Stempak[129], Laura Stronati[149], Joseph JY Sung[25], Stephan R Targan[16], Kirstin M Taylor[144], Emilie Theatre[90,91], Leif Torkvist[150], Esther A Torres[34], Mark Tremelling[151], Holm H Uhlig[152], Junji Umeno[14], Homayon Vahedi[153], Eric Vasiliauskas[16], Anje ter Velde[136], Nicholas T Ventham[81], Severine Vermeire[154,155], Hein W Verspaget[128], Martine De Vos[116], Thomas Walters[156,94], Kai Wang[96], Ming-Hsi Wang[157,39], Rinse K Weersma[42], Zhi Wei[158], David Whiteman[101], Cisca Wijmenga[107], David C Wilson[102,143], Juliane Winkelmann[159,160], Sunny H Wong[25], Ramnik J Xavier[45,73], Keiko Yamazaki[15,161], Suk-Kyun Yang[9], Byong Duk Ye[9], Sebastian Zeissig[162], Bin Zhang[69], Clarence K Zhang[163], Hu Zhang[164,165], Wei Zhang[38], Hongyu Zhao[163], Zhen Z Zhao[101], Australia and New Zealand IBDGC, Belgium IBD Genetics Consortium, Italian Group for IBD Genetic Consortium, NIDDK Inflammatory Bowel Disease Genetics Consortium,

Quebec IBD Genetics Consortium, United Kingdom IBDGC, Wellcome Trust Case Control Consortium

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany.
[4]Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway.
[6]Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands.
[7]Digestive Disease Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran.
[8]Department of Genetics, University of Delhi South Campus, New Delhi, India.
[9]Department of Gastroenterology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea.
[10]Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of Medicine, Seoul, Korea.
[11]Harvey M. and Lyn P. Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, John Hopkins University School of Medicine, Baltimore, USA.
[12]Department of Gastroenterology, Tehran University of Medical Sciences Emam Hospital, Tehran, Iran.
[13]Department of Gastroenterology, Mater Dei Hospital, Msida, Malta.
[14]Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan.
[15]Laboratory for Genotyping Development, Center for Integrative Medical Sciences, Riken, Yokohama, Japan.
[16]F.Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.
[17]Department of Biochemistry and Molecular Biology, University of Ulsan College of Medicine, Seoul, Korea.
[18]School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.
[19]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.
[20]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA.
[24]Dayanand Medical College and Hospital, Ludhiana, India.
[25]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong.
[34]Department of Medicine, University of Puerto Rico Center for IBD, University of Puerto Rico School of Medicine, Rio Piedras, Puerto Rico.
[36]Research Institute for Internal Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet and University of Oslo, Oslo, Norway.
[37]Department of Medicine, Rutgers Robert Wood Johnson School of Medicine and Department of Genetics, Rutgers University Brunswick and Piscataway, New Jersey, USA.
[38]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, NewHaven, Connecticut, USA.
[39]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA.
[40]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA.
[41]Peninsula College of Medicine and Dentistry, Exeter, UK.
[42]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, The Netherlands.
[43]Department of Gastroenterology, Erasmus Hospital, Brussels, Belgium.
[44]Department of Gastroenterology, Free University of Brussels, Brussels, Belgium.

[45]Gastroenterology Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[46]Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA.
[47]Focused Research Unit for Molecular Diagnostic and Clinical Research, IRS-Center Soenderjylland, University Hospital of Southern Denmark, Denmark.
[48]Institute of Molecular Medicine University of Southern Denmark, Denmark.
[49]Wellcome Trust Sanger Institute, Hinxton, UK.
[50]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, Australia.
[51]Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCSCSS) Hospital, San Giovanni Rotondo, Italy.
[52]Strutture Organizzative Dipartimentali (SOD) Gastroenterologia 2, Azienda Ospedaliero Universitaria (AOU) Careggi, Florence, Italy.
[53]Department of Gastroenterology, Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada.
[54]Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada.
[55]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.
[56]Center for Applied Genomics, The Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
[57]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, Australia.
[58]Department of Medicine, University of Otago, Christchurch, New Zealand.
[59]Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Germany.
[60]Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA.
[61]Division of Gastroenterology, Royal Victoria Hospital, Montréal, Québec, Canada.
[62]Université de Montréal and the Montréal Heart Institute, Research Center, Montreal Heart Institute, Montréal, Québec, Canada.
[63]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK.
[64]Department of Medicine II, Ludwig-Maximilians-University Hospital Munich- Grosshadern, Munich, Germany.
[65]IBD Unit, Fremantle Hospital, Fremantle, Australia.
[66]School of Medicine and Pharmacology, University of Western Australia, Fremantle, Australia.
[67]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA.
[68]Department of Human Genetics, KU Leuven, Leuven, Belgium.
[69]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA.
[70]Department of Gastroenterology, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK.
[71]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[72]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[73]Broad Institute of MIT and Harvard, Cambridge, 24 Massachusetts, USA.
[74]School of Biological Sciences, Monash University, Clayton VIC, Australia.
[75]Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden.
[76]IBD Center, Department of Gastroenterology, Istituto Clinico Humanitas, Milan, Italy.
[77]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.
[78]Pediatric Gastroenterology, Cincinnati Childrens Hospital Medical Center, Cincinnati, Ohio, USA.

[79]Department of Gastroenterology, Université Catholique de Louvain (UCL) Cliniques Universitaires Saint-Luc, Brussels, Belgium.
[80]Division of Gastroenterology, University Hospital Padua, Padua, Italy.
[81]Gastrointestinal Unit, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK.
[82]Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California, USA.
[83]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA.
[84]Department of Gastroenterology, Torbay Hospital, Torbay, Devon, UK.
[85]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[86]Pediatrics, Harvard Medical School, Boston, Massachusetts, USA.
[87]Faculty of Medical & Health Sciences, School of Medical Sciences, The University of Auckland, Auckland, New Zealand.
[88]Department of Gastroenterology, Mater Health Services, Brisbane, Australia.
[89]Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand.
[90]Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) Research Center, University of Liege, Liege, Belgium.
[91]Faculty of Veterinary Medicine, University of Liege, Liege, Belgium.
[92]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.
[93]Research Center, Montreal Heart Institute, Montréal, Québec, Canada.
[94]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada.
[95]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah, USA.
[96]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
[97]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104.
[98]Department of Gastroenterology, Faculty of Medicine and Health, Örebro University, SE 702 81, Örebro, Sweden.
[99]IBD Unit, St Marks Hospital, Harrow, Middlesex, UK.
[100]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK.
[101]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[102]Child Life and Health, University of Edinburgh, Edinburgh, Scotland, UK.
[103]Gastrointestinal Research Group, Division of Applied Medicine, University of Aberdeen, Aberdeen, UK.
[104]Division of Rheumatology Immunology and Allergy, Brigham and Womens Hospital, Boston, Massachusetts, USA.
[105]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[106]Centre de recherche sur l'inflammation, UMR1149 INSERM et Université de Paris, Paris, France.
[107]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands.
[108]Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania.
[109]Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK.
[110]National Institute of Immunology, Aruna Asaf Ali Road, New Delhi, India.
[111]IBD Pharmacogenetics Group, University of Exeter, Exeter, UK.
[112]Manchester Centre for Genomic Medicine, University of Manchester and Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK.
[113]Yonsei University College of Medicine, Seoul, Korea.

[114]Department of Gastroenterology, Kaunas University of Medicine, Kaunas, Lithuania.
[115]Institute of Cellular Medicine, Newcastle University, Newcastle-upon-Tyne, UK.
[116]Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium.
[117]Department of Medicine, University of Cambridge, UK.
[118]Faculty of Medicine, University of Latvia, Riga, Latvia.
[119]Dipartimento di Neuroscienze, Psicologia, Area del Farmaco e Salute del Bambino (NEUROFARBA), Università di Firenze Strutture Organizzative Dipartimentali (SOD) Gastroenterologia e Nutrizione Ospedale pediatrico Meyer, Firenze, Italy.
[120]Division of Gastroenterology, Centre Hospitalier Universitaire (CHU) de Liège, Liege, Belgium.
[121]Department of Gastroenterology, The Townsville Hospital, Townsville, Australia.
[122]Digestive Disease Research Institute, Shariati Hospital, Tehran, Iran.
[123]Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran.
[124]Department of Gastroenterology & Hepatology, Royal Victoria Infirmary, Newcastle-upon-Tyne, UK.
[125]Inflammatory Bowel Disease Research Group, Addenbrookes Hospital, Cambridge, UK.
[126]Department of Medical and Molecular Genetics, Kings College London School of Medicine, Guys Hospital, London, UK.
[127]Department of Gastroenterology, Fukuoka University Chikushi Hospital, Fukuoka, Japan.
[128]Department of Gastroenterology, Leiden University Medical Center, Leiden, The Netherlands.
[129]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada.
[130]Center for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia.
[131]Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK.
[132]Institut für Pathologie, Universität Bern.
[133]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[134]Department of Genomics Life & Brain Center, University Hospital Bonn, Bonn, Germany.
[135]St Marys Hospital, London, UK.
[136]Department of Gastroenterology, Academic Medical Center, Amsterdam, The Netherlands.
[137]Icahn School of Medicine, Mount Sinai New York, New York, USA.
[138]Faculty for Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia.
[139]Department of Gastroenterology, Royal Brisbane and Womens Hospital, Brisbane, Australia.
[140]Department of Gastroenterology, Université Catholique de Louvain (UCL) Centre Hospitalier Universitaire (CHU) Mont- Godinne, Mont-Godinne, Belgium.
[141]Department Internal Medicine III, Division Gastroenterology & Hepatology, Medical University Vienna, Vienna, Austria.
[142]UniversitätsSpital Zürich, Klinik für Gastroenterologie und Hepatologie, Zürich.
[143]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK.
[144]Department of Gastroenterology, St Thomas Hospital, London, UK.
[145]Department of Digestive Diseases, Hospital Quiron Teknon, Barcelona, Spain.
[146]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA.
[147]Human Genetics, Genome Institute of Singapore, Singapore.
[148]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA.
[149]Department of Biology of Radiations and Human Health, Agenzia nazionale per le nuove tecnologie lenergia e lo sviluppo economico sostenibile (ENEA), Rome, Italy.
[150]Department of Clinical Science Intervention and Technology, Karolinska Institutet, Stockholm, Sweden.
[151]Gastroenterology & General Medicine, Norfolk and Norwich University Hospital, Norwich, UK.
[152]Translational Gastroenterology Unit, Nuffield.
[153]Digestive Disease Research Center, Tehran University of Medical Sciences, Tehran, Iran.

[154]Department of Chronic Diseases, Metabolism & ageing, Translational Research in GastroIntestinal Disorders (TARGID), Katholieke Universiteit (KU) Leuven, Leuven, Belgium.
[155]Division of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium.
[156]Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.
[157]Meyerhoff Inflammatory Bowel Disease Center, Department of medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
[158]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA.
[159]Institute of Human Genetics, Technische Universität München, Munich, Germany.
[160]Department of Neurology, Technische Universität München, Munich, Germany.
[161]Department of Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan.
[162]Department of Medicine I, University Medical Center Dresden, Technische Universität (TU) Dresden, Dresden, Germany.
[163]Department of Biostatistics, School of Public Health, Yale University, NewHaven, Connecticut, USA.
[164]Department of Gastroenterology, West China Hospital, Chengdu, Sichuan, China.
[165]State Key Laboratory of Biotherapy, Sichuan University West China University of Medical Sciences (WCUMS), Chengdu, Sichuan, China.

Members of the Multicenter African American IBD Study (MAAIS) Recruitment Center,[A] NIDDK IBD Genetics Consortium (IBDGC),[B] Emory University GENESIS study[C], and Cedars Sinai Medical Center[D] responsible for African American case and control ascertainment and SNP genotyping

(A) Lisa W. Datta[1], Themistocles Dassopoulos[2], Jason K. Hou[3], Chengrui Huang[1], Kim L. Isaacs[4], Howard Kader[5], John F. Kuemmerle[6], John H. Kwon[7], Mark Lazarev[1], Peter Mannon[8], Ellen J. Scherl[9], Ann Silverman[10], Claire Simpson[11,12], John Valentine[13,14], Ming-Hsi Wang[1,15] and Steven R. Brant[1,16].

(B) Steven R. Brant[1,16], Judy H. Cho[17], Richard Duerr[18], Dermot P.B. McGovern[19], John D. Rioux[20], Mark S. Silverberg[21]

(C) Jonathan S. Alexander[22], Robert N. Baldassano[23], Pankaj Chopra[24], Raymond K. Cross[25], David J. Cutler[24], Tanvi A. Dhere[26], John S. Hanson[27], Sunny Z. Hussain[28], Kelly E. Kachelries[23], Michael D. Kappelman[29], Jeffrey Katz[30], Richard Kellermayer[31], Barbara S. Kirschner[32], Dedrick E. Moulton[33], David T. Okou[34], Bankole O. Osuntokun[35], Ashish S. Patel[36], Shehzad Saeed[37], Michael E. Zwick[24] and Subra Kugathasan[34]

(D) Talin Haritunians[19], Jerome I. Rotter[38], Kent D. Taylor[38], Stephan R. Targan[19] and Dermot P.B. McGovern[19]

[1]Department of Medicine, Meyerhoff Inflammatory Bowel Disease Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
[2]Department of Medicine, Washington University School of Medicine, St Louis, Missouri, USA.
[3]Department of Medicine, Baylor College of Medicine; Veterans Affairs Health Services Research and Development Service, Center for Innovations in Quality Effectiveness and Safety; Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA.
[4]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.
[5]Department of Pediatrics, University of Maryland School of Medicine, Baltimore, Maryland, USA.

[6]Medicine and Physiology and Biophysics, Medical College of Virginia Campus of Virginia Commonwealth University, Richmond, Virginia, USA.
[7]Section of Gastroenterology, Department of Medicine, University of Chicago, Chicago, Illinois USA.
[8]Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA.
[9]Department of Medicine, Weill Cornell Medical College, New York, New York, USA.
[10]Department of Internal Medicine, Henry Ford Health System, Detroit, Michigan, USA.
[11]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, USA.
[12]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA.
[13]Departments and Physiology and Biophysics of Division of Gastroenterology, Hepatology & Nutrition, University of Florida, Gainesville, Florida, USA.
[14]University of Utah, Health Sciences, Salt Lake City, Utah, USA.
[15]Department of Medicine, Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, USA.
[16]Crohn's Colitis Center of New Jersey, Division of Gastroenterology and Hepatology, Department of Medicine, Rutgers Robert Wood Johnson Medical School and Department of Genetics, The Human Genetics Institute of New Jersey, Rutgers University, New Brunswick and Piscataway, New Jersey, USA.
[17]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
[18]Department of Medicine and Clinical and Translational Science Institute, School of Medicine and Department of Human Genetics, Graduate School of Public Health; University of Pittsburgh, Pittsburgh, Pennsylvania, USA.
[19]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.
[20]Department of Medicine, Université de Montréal and the Montreal Heart Institute Research Center, Montreal, Quebec, Canada.
[21]Department of Medicine, Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, University of Toronto, Toronto, Toronto, Ontario, Canada.
[22]Department of Molecular and Cellular Physiology, Louisiana State University Health Sciences Center, Shreveport, Louisiana, USA.
[23]Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
[24]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA.
[25]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA.
[26]Department of Medicine, Emory University School of Medicine, Atlanta, Georgia, USA.
[27]Charlotte Gastroenterology and Hepatology, Charlotte, North Carolina, USA.
[28]Department of Pediatrics, Willis-Knighton Physician Network, Shreveport, Louisiana, USA.
[29]Department of Pediatrics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.
[30]Division of Gastroenterology, Case Western Reserve University, Cleveland, Ohio, USA.
[31]Section of Pediatric Gastroenterology, Baylor College of Medicine, Texas Children's Hospital, Houston, Texas, USA.
[32]Department of Pediatrics, University of Chicago Comer Children's Hospital, Chicago, Illinois, USA.
[33]Division of Gastroenterology, Vanderbilt Children's Hospital, Nashville, Tennessee, USA.
[34]Division of Pediatric Gastroenterology, Department of Pediatrics, Emory University School of Medicine & Children's Healthcare of Atlanta, Atlanta, Georgia, USA.
[35]Department of Pediatrics, Cook Children's Medical Center, Fort Worth, Texas, USA.
[36]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas,

USA.
[37]Dayton Children's Hospital, Dayton, Ohio, USA.
[38]Institute for Translational Genomics and Population Sciences and Division of Genomic Outcomes, Departments of Pediatrics and Medicine, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California, USA.

REFERENCES

1. 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).
2. Ye, B. D. *et al*. Identification of Ten Additional Susceptibility Loci for Ulcerative Colitis Through Immunochip Analysis in Koreans. *Inflamm. Bowel Dis.* (2016). doi:10.1097/MIB.0000000000000584
3. Huang, C. *et al*. Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans. *Gastroenterology* **149**, 1575–1586 (2015).
4. Ballester, V. *et al.* Association of NOD2 and IL23R with inflammatory bowel disease in Puerto Rico. *PLoS One* (2014). doi:10.1371/journal.pone.0108204
5. 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).
6. Replication, D. Ia. G. *et al*. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234–244 (2014).
7. Purcell, S. PLINK 1.9.
8. 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).
9. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
10. Negi, S. *et al.* A genome-wide association study reveals ARL15, a novel non-HLA susceptibility gene for rheumatoid arthritis in North Indians. *Arthritis Rheum.* (2013). doi:10.1002/art.38110
11. Robinson, J. *et al*. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423-31 (2015).
12. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
13. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656–664 (2002).
14. 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.* (2019). doi:10.1093/hmg/ddy443
15. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
16. Stern, L. J. *et al*. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* (1994). doi:10.1038/368215a0
17. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0408677102
18. Goldsack, D. E. & Chalifoux, R. C. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *J. Theor. Biol.* (1973). doi:10.1016/0022-5193(73)90075-1
19. Kawashima, S. *et al*. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* (2008). doi:10.1093/nar/gkm998
20. Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379–i388 (2017).
21. Jensen, K. K. *et al*. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* (2018). doi:10.1111/imm.12889
22. 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.* (2012). doi:10.1093/nar/gks469

23. Zheng, X. *et al.* HIBAG-HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
24. Mychaleckyj, J. C. *et al.* HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin. Trials* **7**, 75–87 (2010).
25. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
26. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput. Biol.* **9**, e1002877 (2013).
27. Dilthey, A. T., Moutsianas, L., Leslie, S. & McVean, G. HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics* **27**, 968–972 (2011).
28. Fisher, S. A. *et al.* Sex stratification of an inflammatory bowel disease genome search shows male-specific linkage to the HLA region of chromosome 6. *Eur J Hum Genet* **10**, 259–265 (2002).
29. Venkateswaran, S. *et al.* Enhanced Contribution of HLA in Pediatric Onset Ulcerative Colitis. *Inflamm Bowel Dis* **24**, 829–838 (2018).
30. Brant, S. R. *et al*. Genome-Wide Association Study Identifies African-Specific Susceptibility Loci in African Americans With Inflammatory Bowel Disease. *Gastroenterology* **152**, 206-217 e2 (2017).

9. Declaration

Declaration

I herewith confirm that this thesis is completely the result of my own work. Apart from the advice of my supervisors, all sources are listed in the references. The results of this thesis (**Paper A** and **B**) have already been published in the journals *Gastroenterologe* and *Human Molecular Genetics.* **Paper C** is currently under review. This thesis was prepared in accordance with the rules of the DFG and has not been submitted to any other board for another qualification. I further confirm that I have not been stripped of any academic title.

Erklärung

Hiermit versichere ich, dass diese Dissertation nach Inhalt und Form das Resultat meiner eigenen Arbeit ist. Abgesehen vom Rat meiner akademischen Lehrer sind sämtliche Quellen in den Referenzen aufgeführt. Die Ergebnisse dieser Dissertation (**Publikation A** und **B)** wurden bereits in dem Zeitschriften *Gastroenterologe* und *Human Molecular Genetics* veröffentlicht. **Publikation C** ist im Moment zur Begutachtung. Diese Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden und hat weder im Ganzen noch zum Teil an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen. Ich versichere weiterhin, dass mir kein akademischer Grad entzogen wurde.

Kiel, den ....

(Frauke Degenhardt)

10. Acknowledgements

This study would not have been possible without the contribution of my research colleagues, family, and friends. I would therefore like to thank

Prof. Andre Franke for giving me this project and for his support and helpful discussions during the thesis, his understanding and great kindness especially but not only in more challenging times. His continued efforts in extending international research collaborations, maintaining a good research infrastructure, and supporting me in extending my knowledge by visiting conferences and the Harvard Medical University in Boston. These have given me the opportunity to meet many fellow researches and enabled fruitful inter-group collaborations.

Mareike Wendorff and Elisa Rosati for their support in many research related questions and their continuous friendship. I would especially like to thank Mareike for her support during the writing of our HLA imputation reference paper, for computing our model as many times as it needed to be computed and her open mindedness and general happy attitude. I would like to thank Elisa as well as Eva Ellinghaus for critically reading and editing the trans-ethnic imputation reference manuscript.

Prof. David Ellinghaus for his support and helpful discussion on genotyping and matters related.

To Prof. Peta Bacher and our immunogenetics group for helpful discussions towards the end of my thesis.

Michael Wittig for spending some months on manually calling a large part of the HLA alleles which were included in the HLA imputation reference panel, for answering my questions related to the HLA and introducing me to the concepts of HLA typing by sequencing with clear and helpful communication.

Our collaborators from the International IBD Genetics Consortium, especially Steven R. Brant, Tom H. Karlsen, John D. Rioux, Gabrielle Boucher, their teams and anyone involved in the international efforts involving the acquisition and maintenance of the cross-ethnic dataset of the IIBDGC. The large number of people involved can be deduced from the Supplementary Notes of Paper C, showing the intense and laborious task of gathering these large datasets, enabling high quality bioinformatic research.

Dr. Silke Szymzcak for introducing me to the finer nuances of statistics and for her supervision in the first months of my time at the IKMB. For her continued help in answering questions, helpful discussions in matters of statistics and for introducing me to experts of the field. Prof. Michael Krawczak of the IMIS for allowing me to be part of their weekly group meetings.

Prof. Soumya Raychaudhuri for giving me the opportunity of an extended visit with his lab at the Harvard Medical University in Boston, USA and interesting discussions. I also thank his team for the kind reception and helpful discourse.

I thank our technical assistants from the IKMB DNA, genotyping and NGS labs. In the context of this thesis, especially Tanja Wesse and Sanaz Sabet for their help in many little and big challenges related to genotyping, and all members of the DNA lab, for carrying out the wetlab part of the genotype analysis in this and other projects. Special thanks to Tanja for her competent and fast help in locating samples and manifest files of genotyping chips many times over. Thanks to Sanaz for introducing me to manual genotype calling and helping me in the interpretation of the genotype calls.

Jan Kässens, Iacopo Torre, Georg Hemmrich-Stanisak, Marc Höppner and Michael Kisiela as well as Karsten Balzer from the HPC Cluster for their continuous support in all questions related to using the HPC and my desktop computer. For solving one or the other problem regarding installations and broken computers or disks. Special thanks to Jan and Iacopo for their patience and friendly attitudes in solving mysteries surrounding software installation and deciphering the sometimes cryptic problems of my computer. Special thanks also to Marc Höppner for helping with the scripting of the Nextflow pipeline.

Prof. Philip Rosenstiel for giving me a space in the office of the cell biology lab during the first 4 years of my work at the IKMB. My co-workers from the cell biology lab including doctoral students and lab technicians alike, especially Anne Luzius, Marlene Jentzsch, Helene Geese, Stefanie Stengel, Anna Groth and Berith Messner for welcoming me in their group/office and for many light and fun moments spent together. Thanks for including me in your group specially during the first years of my time at the IKMB and for many helpful discussions about work and life. Special thanks to Berith for keeping me company in writing the thesis and her helpful comments on my thesis.

Matthias Barann and Anupam Sinha for keeping me company as my direct – seated next to me – coworkers in the time spent in the office of the cell biology lab. Matthias for teaching me a lot of things on bioinformatics especially during my first years at the institute.

My co-workers from the 1st floor bioinformatics office: Elisa Rosati, Mareike Wendorff, Matthias Hübenthal, Sören Mucha, Louise Tingholm, Malte Rühlemann, Michael Forster, Simonas Juzenas,

Lars Wienbrandt, Jan Kässens and Hesham ElAbd for interesting and helpful discussions in and out of the office and many delicious cakes.

I would like to thank Natalie Tepling, Eike Zell, Christiane Wolf-Schwerin and Tosca Heinrich for their competent and helpful support in matters of organization in and around the IKMB as well as all of my co-workers at the IKMB in and outside the Genetics & Bioinformatics group and technicians from all the labs, who have enabled many interesting research projects across the years. With the deepest apologies if I have forgotten anyone here.

I would like to thank everyone who read the drafts of my thesis and gave me valuable input.

Lastly, I would like to thank my parents, sisters and friends for their continuous support, love and warmth and for being part of my life for the largest part of my life. For your support especially in times that were difficult. I hope you know how much I appreciate you and how much you all enrich my life.