Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals shared but also ethnicity-specific disease associations

3

Frauke Degenhardt^{1,#}, Gabriele Mayr¹, Mareike Wendorff¹, Gabrielle Boucher², Eva Ellinghaus³, David 4 Ellinghaus^{1,4}, Hesham ElAbd¹, Elisa Rosati¹, Matthias Hübenthal^{1,5}, Simonas Juzenas¹, Shifteh 5 Abedian^{6,7}, Homayon Vahedi⁷, Thelma BK⁸, Suk-Kyun Yang⁹, Byong Duk Ye⁹, Jae Hee Cheon¹⁰, Lisa 6 Wu Datta¹¹, Naser Ebrahim Daryani¹², Pierre Ellul¹³, Motohiro Esaki¹⁴, Yuta Fuyuno^{14,15}, Dermot PB 7 8 McGovern¹⁶, Talin Haritunians¹⁶, Myhunghee Hong¹⁷, Garima Juyal¹⁸, Eun Suk Jung^{1,10}, Michiaki Kubo¹⁹, Subra Kugathasan^{20,21}, Tobias L. Lenz²², Stephen Leslie²³, Reza Malekzadeh⁷, Vandana 9 Midha²⁴, Allan Motyer²³, Siew C Ng²⁵, David T Okou²⁶, Soumya Raychaudhuri^{27,28,29,30,31}, John 10 11 Schembri¹³, Stefan Schreiber^{1,32}, Kyuyoung Song¹⁷, Ajit Sood²⁴, Atsushi Takahashi³³, Esther A Torres³⁴, Junji Umeno¹⁴, Behrooz Z. Alizadeh⁶, Rinse K Weersma³⁵, Sunny H Wong²⁵, Keiko 12 Yamazaki¹⁵, Tom H Karlsen^{4,36}* John D Rioux^{2,*}, Steven R Brant^{11,37,*} for the MAAIS Recruitment 13 Center, Andre Franke^{1,*} for the International IBD Genetics Consortium 14

- 15 *joint senior authors
- 16

17 1 Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany.

18 2 Université de Montréal and the Montréal Heart Institute, Research Center, Montréal Heart
19 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.

24 5 Department of Dermatology, Venerology and Allergy, University Hospital Schleswig-Holstein,

25 Campus Kiel, Kiel, Germany.

- 26 6 Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands.
- 27 7 Digestive Disease Research Center, Digestive Disease Research Institute, Tehran University of
- 28 Medical Sciences, Tehran, Iran.
- 29 8 Department of Genetics, University of Delhi South Campus, New Delhi, India.
- 30 9 Department of Gastroenterology, Asan Medical Center, University of Ulsan College of Medicine,
- 31 Seoul, Korea.
- 32 10 Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of
 33 Medicine, Seoul, Korea.
- 11 Harvey M. and Lyn P. Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine,
- 35 John Hopkins University School of Medicine, Baltimore, USA.
- 36 12 Department of Gastroenterology, Emam Hospital, Tehran University of Medical Sciences, Tehran,
 37 Iran.
- 38 13 Department of Gastroenterology, Mater Dei Hospital, Msida, Malta.
- 14 Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu
 40 University, Fukuoka, Japan.
- 41 15 Laboratory for Genotyping Development, Center for Integrative Medical Sciences, Riken,
 42 Yokohama, Japan.
- 43 16 F.Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai
 44 Medical Center, Los Angeles, California, USA.
- 45 17 Department of Biochemistry and Molecular Biology, University of Ulsan College of Medicine,
 46 Seoul, Korea.
- 47 18 School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.
- 48 19 RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.
- 49 20 Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA.
- 50 21 Pediatric Institute, Children's Healthcare of Atlanta, Atlanta, USA.

- 51 22 Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology,
- 52 Plön, Germany.
- 53 23 Schools of Mathematics and Statistics and BioSciences and Melbourne Integrative Genomics,
- 54 University of Melbourne, Australia.
- 55 24 Dayanand Medical College and Hospital, Ludhiana, India.
- 56 25 Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of
- 57 Hong Kong, Hong Kong.
- 58 26 Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Emory 59 University School of Medicine, Atlanta, USA.
- 60 27 Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA,

61 USA.

- 62 28 Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA,63 USA.
- 64 29 Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
- 65 30 Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.
- 66 31 Centre for Genetics and Genomics Versus Arthritis, Division of Muscolosceletal and
- 67 Dermatological Sciences, School of Biological Sciences, University of Manchester, Manchester, UK.
- 68 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.
- 73 35 Department of Gastroenterology and Hepatology, University of Groningen and University Medical
 74 Center Groningen, Groningen, The Netherlands.
- 75 36 Research Institute for Internal Medicine, Division of Surgery, Inflammatory Diseases and

- 76 Transplantation, Oslo University Hospital Rikshospitalet and University of Oslo, Oslo, Norway.
- 77 37 Department of Medicine, Rutgers Robert Wood Johnson School of Medicine and Department of
- 78 Genetics, Rutgers University Brunswick and Piscataway, New Jersey, USA.

79

81 **ABBREVIATIONS**

- 82
- 83 (A) Alpha chain of an HLA protein
- 84 (B) Beta chain of an HLA protein
- 85 AA African American population of this study

86 AFR African American population of the 1000 Genomes/HapMap population see also 87 https://www.internationalgenome.org/category/population/)

88 AMR Admixed American population of the 1000 Genomes/HapMap population (see also

- 89 https://www.internationalgenome.org/category/population/)
- 90 AF Allele Frequency

91 CEU Utah Residents (CEPH) with Northern and Western European Ancestry of the 1000

- 92 Genomes/HapMap population (see also https://www.internationalgenome.org/category/population/)
- 93 CI Confidence Interval

94 EAS East Asian population of the 1000 Genomes/HapMap population (see also 95 https://www.internationalgenome.org/category/population/)

96 EUR Caucasian population of this population or (mentioned within the context of the 97 1000Genomes/HapMap population European data of the latter; see also 98 https://www.internationalgenome.org/category/population/)

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

- 100 HLA Human Leukocyte Antigen
- 101 HLA-A Human Leukocyte Antigen gene locus A
- 102 HLA-B Human Leukocyte Antigen gene locus B
- 103 HLA-C Human Leukocyte Antigen gene locus C
- 104 HLA-DRA Human Leukocyte Antigen gene locus DRA
- 105 HLA-DRB1 Human Leukocyte Antigen gene locus DRB1

- 106 HLA-DRB3 Human Leukocyte Antigen gene locus DRB3
- 107 HLA-DRB4 Human Leukocyte Antigen gene locus DRB4
- 108 HLA-DRB5 Human Leukocyte Antigen gene locus DRB5
- 109 HLA-DQA1 Human Leukocyte Antigen gene locus DQA1
- 110 HLA-DQB1 Human Leukocyte Antigen gene locus DQB1
- 111 HLA-DPA1 Human Leukocyte Antigen gene locus DPA1
- 112 HLA-DPB1 Human Leukozyten Antigen gene locus DPB1
- 113 IND Indian population
- 114 IRN Iranian population
- 115 JPN Japanese population
- 116 KOR Korean population
- 117 MAF Minor Allele Frequency
- 118 MLE Maximum Likelihood Estimator
- 119 MLT Maltese population
- 120 PRI Puerto Rican population
- 121 P1-P9 Pockets 1 to 9 of the HLA protein within the HLA peptide binding site
- 122 QC Quality Control
- 123 SAS South Asian population of the 1000 Genomes/HapMap population (see also
- 124 https://www.internationalgenome.org/category/population/)
- 125 SNP Single Nucleotide Polymorphism (MAF >= 1%)
- 126 SNV Single Nucleotide Variation (MAF < 1%)
- 127 xHLA extended HLA region
- 128 YRI Yoruba in Ibadan, Nigeria population of the 1000 Genomes/HapMap population (see also
- 129 https://www.internationalgenome.org/category/population/)

130 CORRESPONDENCE

- 131 [#]Corresponding Author:
- 132 Frauke Degenhardt
- 133 Institute of Clinical Molecular Biology
- 134 Christian-Albrechts-University of Kiel
- 135 Rosalind-Franklin-Street 12
- 136 D-24105 Kiel
- 137 Germany
- 138 Tel.: +49 431 500 15147
- 139 E-mail: f.degenhardt@ikmb.uni-kiel.de

141 ABSTRACT

142 Inflammatory bowel disease (IBD) is a chronic inflammatory disease of the gut. Genetic association 143 studies have identified the highly variable human leukocyte antigen (HLA) region as the strongest 144 susceptibility locus for IBD, and specifically DRB1*01:03 as a determining factor for ulcerative colitis 145 (UC). 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 146 147 characterize the HLA signal using a trans-ethnic approach. We performed a comprehensive fine 148 mapping of single HLA alleles in UC in a cohort of 9,272 individuals with African American, East Asian, 149 Puerto Rican, Indian and Iranian descent and 40,691 previously analyzed Caucasians, additionally 150 analyzing whole HLA haplotypes. We computationally characterized the binding of associated HLA 151 alleles to human self-peptides and analysed the physico-chemical properties of the HLA proteins and 152 predicted self-peptidomes. Highlighting alleles of the HLA-DRB1*15 group and their correlated HLA-153 DQ-DR haplotypes, we identified consistent associations across different ethnicities but also identified 154 population-specific signals. We observed that DRB1*01:03 is mostly present in individuals of Western 155 European descent and hardly present in non-Caucasian individuals. We found peptides predicted to 156 bind to risk HLA alleles to be rich in positively charged amino acids such. We conclude that the HLA 157 plays an important role for UC susceptibility across different ethnicities. This research further 158 implicates specific features of peptides that are predicted to bind risk and protective HLA proteins.

159

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

163 **INTRODUCTION**

164

165 Ulcerative colitis (UC) is a chronic inflammatory disease of the gut. Like Crohn's disease (CD), the 166 other main subphenotype of inflammatory bowel disease (IBD), it is most likely caused by an abnormal 167 reaction of the immune system to microbial stimuli with environmental factors also playing a role. 168 Currently more than 240 genetic susceptibility loci have been associated with IBD in Caucasians the 169 majority of which are shared between UC and CD (Jostins et al., Liu et al., Ellinghaus et al., de Lange 170 et al.¹⁻⁴). Strong genetic association signals with both diseases have been identified in the human 171 leukocyte antigen (HLA) region. The HLA is mapped to the long arm of chromosome 6 between 29 172 and 34 Mb and moderates complex functions within the immune system. One of the major tasks of the 173 HLA is the presentation of antigens to the host immune system. While HLA class I proteins usually 174 present peptides derived the cytosol (i.e. peptides derived from intracellularly replicating viruses), HLA 175 class II proteins present peptides from extracellular pathogens that have entered the cell e.g. by 176 phagocytosis. In Caucasian IBD patients a large percentage of the phenotypic variation is explained 177 by variants within the HLA class II locus, with DRB1*01:03 being the most significant risk allele for UC (P [P-value]=2.68×10⁻¹¹⁹, OR [odds ratio]=3.59; 95% CI [confidence interval]=3.22-4.00)⁵, specifically 178 179 by alleles of the HLA-DR and -DQ loci, though tight structures of linkage disequilibrium (LD) have 180 hindered the assignment of the causal variants. Additionally, a systematic comparison across 181 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 sets⁵⁻¹⁴. 182 183 Recently, we created such a trans-ethnic HLA imputation reference panel including dense single 184 nucleotide polymorphism (SNP) fine mapping data typed on Illumina's ImmunoChip, covering a large proportion of the HLA, within 8 populations of different ethnicities¹⁵. Here we report the first trans-185 186 ethnic fine mapping study of the HLA in UC and some biological implications of the results.

187

188 METHODS

189

190 **Cohort description**

191 A detailed description of the cohorts and recruitment sites can be found in the **Supplementary** 192 Methods and Supplementary Table 1. In brief, a total of 52,550 individuals (including 18,142 UC 193 patients and 34,408 controls) were used in this study, of which 10,063 (3,517 UC cases and 6,546 194 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. 195 196 while individuals of African American (Huang et al.¹⁶), Korean (Ye et al.¹⁷), Maltese, and Puerto Rican 197 descent were added. The recruitment of study subjects was approved by the ethics committees or 198 institutional review boards of all individual participating centers or countries. Written informed consent 199 was obtained from all study participants.

200

201 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**.

205

206 **Phasing of single nucleotide variants**

Using SHAPEIT2¹⁸ 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 III¹⁹ (October 2014) alleles (published with the SNP imputation tool IMPUTE2^{20,21}) 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.

218

219 Imputation of single nucleotide variants

220 To increase the density of single nucleotide variants (SNVs, including variants with minor allele 221 frequency (MAF)<1%) within the HLA region, we used publicly available nucleotide sequences of HLA 222 alleles and further imputed SNVs based on the HLA alleles imputed for each individual using 223 IMPUTE2²² with the 1000 Genomes Phase III¹⁹ individuals as reference (October 2014) using 224 parameters: -Ne 20,000, -buffer 250, -burnin 10, -k 80, -iter 30, -k_hap 500, -outdp 3, -pgs_miss, -os 0 225 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 226 227 performed post-imputation excluding variants with an IMPUTE2 info score <0.8. For the Caucasian 228 data set, we excluded variants with a median IMPUTE2 info score <0.8 and a minimum IMPUTE2 info 229 score <0.3. Additionally, we imputed SNPs into the data set using imputed HLA allele information (i.e. 230 translated imputed HLA information into real nucleotide information at each position of the allele) 231 (Supplementary Methods).

232

233 HLA Imputation

QC-ed genotype data for each cohort were imputed using Beagle version 4.1^{22,23} based on the corresponding genotype variants 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 4digit level using the IKMB reference published in Degenhardt *et al.*¹⁵ and the imputation tool HIBAG²⁴. 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.*¹⁵) into consideration during interpretation.

242

244

243 Generation of HLA haplotypes

HLA haplotypes were generated by comparing SNP haplotypes generated by SHAPEIT2¹⁸ for each 245 246 individual and SNP haplotypes stored for the alleles within the classifiers of the HLA reference model¹⁵ 247 for the alleles that were imputed for each individual at a given locus. For 10 random classifiers, we 248 calculated the minimal distance between the SNP haplotypes stored for the allele of interest in the 249 HLA reference model and the SNP haplotypes generated by SHAPEIT2. We assigned alleles to a 250 parental haplotype based on how often this allele had minimal difference to the haplotype. Phasing 251 certainty was calculated as the percentage of times an allele was correctly assigned to the chosen 252 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 253 254 certainty was set to 1.

255

256 **HLA haplotype benchmark**

257 We tested the generation of HLA haplotypes with the above method, using genotype information of trio 258 samples (Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) and Yoruba 259 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 panel²⁵ using the most common 260 261 allele for ambiguous calls. In total, 27 CEU samples and 24 YRI samples and their parents were 262 analyzed. Genotype data were downloaded from the HapMap Phase 3 Server (version 2015-05) and 263 positions present on the Illumina ImmunoChip were extracted. We applied the procedure described 264 above for phasing of HLA alleles. The results are shown in Supplementary Table 2.

265

266 Calculation of marginal probabilities for each allele

Since HIBAG stores a matrix of all posterior probability values of each (biallelic) 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.

271

272 Association analysis

Subsequently, we performed a standard logistic regression association analysis on single alleles and SNVs. 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. SNPs imputed with IMPUTE2 were included as dosages. SNVs inferred from HLA alleles were coded as 0,1,2 on the minor allele. Additive association analyses for each marker were performed using

278

279
$$\log(\text{odds}_i) = \beta_0 + \beta_1 x_i + \beta_2 U_{1i} + \beta_3 U_{2i} + \beta_4 U_{3i} + \beta_5 U_{4i} + \beta_6 U_{5i} + (\beta_7 b_i)$$

280

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).

284

285 Meta-analysis

We performed a meta-analysis of association statistics from the 9 analysed cohorts using the tool RE2C²⁶. 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.*²⁷) 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. Here we either report the REC2p* p-value or the REC2 p-value if exceeding
heterogeneity was observed for an association. For the original definition see Lee *et al.*²⁶.

295

296 Clustering according to preferential peptide binders

297 Using NetMHCIIpan-3.2²⁸, we predicted binding affinities for five sets of the 200,000 unique random 298 15mer peptides (Supplementary Methods) for all alleles that were significant in the meta-analysis 299 and had a frequency of >1% in at least one of the 9 populations. We give the amino acid distribution of 300 these sets in Supplementary Table 3. We selected the top 2% (strong binders (SB)) preferential 301 peptide binders as given by the NetMHCIIpan-3.2 software for each allele and calculated the pairwise 302 Pearson correlation between alleles based on complete observations for the respective allele combinations²⁸ using R (version 3.3.1), creating a matrix of correlations. Clustering was performed on 303 304 this matrix using hclust of the R package stats. Correlation between the clusters was calculated using 305 corrplot (version 0.84) and dendextend (version 1.12). Here, the correlation between cluster 306 dendrograms (i.e. the concordance of the tree-structure) is calculated with a value of 0 signifying 307 dissimilar tree-structures and 1 signifying highly similar tree-structures. Dendrograms were plotted 308 using the ape (version 5.3) package, for DQ and DRB1.

309

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.²⁹ for each of the groups and also plotted the (Position-specific

317 scoring matrix) PSSM scores for chosen amino acids within a group.

318 Clustering according to physico-chemical properties

319 Clustering of HLA proteins was performed using 5 different numerical scores: the Atchley scores F1 and F3³⁰, residue-volume³¹ and self-defined parameters charge and hydrogen-acceptor capability 320 321 (Supplementary Table 4). The amino acid sequence of each respective allele was extracted at 322 positions noted in Supplementary Table 5 for Pockets 1, 4, 6, 7 and 9 from HLA allele protein 323 sequences that were retrieved from the IMGT/HLA database (version3.37.0)³² and aligned using 324 MUSCLE³³. The alpha chain, of the HLA-DR locus is invariable and was not considered in the analysis 325 of this locus. For the respective analysis, each amino acid was assigned its numerical score. 326 Clustering was then performed on the scores using the hclust function of the R (version 3.3.1) 327 package stats and Euclidian distances.

328

330 **RESULTS**

331 Here we imputed HLA alleles for a total of 9 cohorts (Supplementary Figure 1, Supplementary 332 Table 1) within 3 HLA class I (HLA-A, -C and -B) and 8 class II loci (HLA- DRB3, -DRB5, -DRB4, -333 DRB1, -DQA1, -DQB1, -DPA1 and -DPB1) at full context 4-digit level utilizing a median of 8,555 SNP 334 genotypes (located within extended HLA between 25 and 34 Mb on chromosome 6p21) from 335 Illumina's ImmunoChip. After QC, a total of 17,276 UC cases and 32,975 controls remained. 13,927 cases and 26,764 controls were previously reported Caucasians⁵ and 3,251 cases and 6,021 controls 336 337 were non-Caucasian individuals (Liu et al.²). After SNP imputation and respective quality control a 338 median of 88,087 SNVs with INFO score > 0.8 were additionally analysed.

339 In line with our previous study in Caucasians⁵, we observed strong, consistent association signals for SNPs and HLA alleles within the HLA class II region, featuring HLA-DRB1, HLA-DQA1, and HLA-340 341 DQB1, for all UC case-control panels except the small-sized Puerto Rican and Maltese cohorts 342 (Figure 1 and Supplementary Figure 2). The strongest association signal was seen for SNP rs28479879 (RE2Cp=5.41×10⁻¹⁵⁷, RE2Cp*=8.87×10⁻¹⁵⁶, I²=79), located in the HLA-DR locus, including 343 344 HLA-DRB1 and HLA-DRB3/4/5. In the Japanese and Korean panels, we further observed a "roof-top"-345 like association signal spanning the HLA class I and II loci (Figure 1) that, as we subsequently 346 demonstrated, was caused by strong LD between the most disease-associated class II alleles 347 DRB1*15:02, DQA1*01:03, and DQB1*06:01, and the class I alleles B*52:01 and C*12:02. The "roof-348 top"-like signal disappeared when conditioning on class I and class II alleles separately 349 (Supplementary Figure 3). Likely due to lack of statistical power, e.g. for the Maltese data set, and/or 350 diversity of the population, e.g. for the Puerto Ricans, association P-values for these populations did not achieve the genome-wide significance threshold ($P<5\times10^{-8}$). 351

The most strongly and consistently associated class II risk alleles within the meta-analysis were alleles of the DRB1*15 group (RE2Cp*=1.87×10⁻¹¹⁶, RE2Cp=1.31×10⁻¹¹⁷, I²=92) (Figure 2, Supplementary Table 6), observed to be located on the same haplotype as DQA1*01:02/03 and DQB1*06:01/02 (Figure 3, Supplementary Table 7). DRB1*15:02 was most frequent in the Asian populations

356 (Japanese, Korean), while DRB1*15:03 was specific to the African American population and 357 DRB1*15:01 had the stronger association and higher allele frequency in the Chinese and Caucasian 358 population (Figure 2, Supplementary Table 6), which is consistent to data published in the HLA allele frequency database³⁴. Since effect sizes were heterogeneous across populations, we did not compute 359 360 a combined score, but rather show the OR in Supplementary Figure 4. Other associated class II alleles included DQA1*03 alleles (RE2Cp*=5.83×10⁻⁸¹) that were observed to be located on a 361 362 DRB1*04 $(RE2Cp^*=2.36\times10^{-55}),$ DRB1*07:01 $(REC2p^*=5.99\times10^{-35})$ haplotype with RE2Cp=6.485.99×10⁻³⁶, I²=68) or DRB1*09:01 (RE2Cp*=2.73×10⁻¹²). DRB1*04/07/09 alleles are all 363 located on the same haplotype as HLA-DRB4 alleles¹⁵, therefore absence of HLA-DRB4, hereafter 364 named DRB4*00:00, was significantly associated with high risk (RE2Cp*=2.35×10⁻¹²⁷). Along the 365 366 same line HLA-DRB5 is located on the same haplotype as DRB1*15. Its absence was therefore 367 observed to be protective. We identified DRB1*10:01 as a novel association signal (RE2Cp*=1.03×10⁻ ⁶). It was observed to be most frequent in the Iranian (3.2% controls and 1.6% cases) and Indian 368 369 (6.7% controls and 3.3% cases) populations and rare in other populations (Supplementary Table 6), 370 which is most likely why it has not been described before. Among population-specific signals, we also 371 observed significant association of UC with DRB1*14:04 (P=0.004, OR=1.64 95%CI: 1.18-2.29) in the 372 Indian population (Figure 2). Overall, alleles of 11 of the 13 known HLA-DRB1 2-digit groups and all 5 373 known -DQB1 groups were associated with UC across the different cohorts (Figure 2, 374 Supplementary Table 6, Supplementary Figure 5), with more HLA-DRB1 alleles conferring 375 protection than risk. Effect sizes in the larger Caucasian and Japanese populations were observed to 376 be moderate (0.5<OR<2.0 for alleles with AF>1%, with the exception of DRB1*15:02 (OR=2.87; 95%) 377 CI: 2.46-3.36 in the Japanese population). The comparison of beta estimates also showed that Japanese and Korean effects estimates were most similar (weighted correlation of 0.84, P=1.3×10⁻³⁰), 378 379 while Iranian and Indian effects estimates correlated better with those of the Caucasian population (weighted correlation of 0.65, $P=1.0\times10^{-16}$ and 0.69, $P=2.0\times10^{-17}$, Supplementary Figure 6, 380 381 Supplementary Methods). Notably, we identified DRB1*01:03, which was identified as the strongest 382 association signal for IBD in our previous fine mapping analysis⁵ as being population specific. It was 17

383 not present in the Asian populations and was only observed with a frequency of <0.1% in the African 384 American and Puerto Rican populations. Detailed analysis of the geographic distribution of the 385 DRB1*01:03 allele showed, that it seemingly occurs in Western Europe (Great Britain, Ireland, France, 386 Spain) and former Western colonies with AF >1%, while it seems to be infrequent in the Eastern parts 387 of Europe. We therefore hypothesize that this allele is linked to the history of Western European 388 countries. (Figure 6). Within this study, the frequency of DRB1*01:03 in the Caucasian population is 389 likely underestimated and therefore not the top associated signal in the Caucasian analysis (i.e. 390 DRB1*01:03 was imputed as DRB1*01:01 or DRB1*01:02 due to similarities in SNP haplotype 391 between these alleles) due to applying a reference panel containing mostly non-Caucasian individuals 392 and European individuals from Germany only. Indeed, using the European HLARES imputation panel, 393 which contains a more diverse Caucasian population, we re-established the signal. The frequency of 394 the remaining alleles imputed with our transethnic reference dataset highly correlated with our original 395 study in the Caucasian population (Supplementary Figure 7). Other DRB1*15, for instance 396 DRB1*15:06 did not show association with UC. Interestingly, however DRB1*15:06 has the same amino acid sequence as DRB1*15:01 in the peptide binding groove and may therefore biologically 397 398 indeed play a role in IBD. With low overall global frequency of the DRB1*15:06 allele, it was not 399 statistically associated with UC. It was most frequent in the Indian population (AF= 2.1%, OR= 1.27, 400 95%CI [0.77, 2.11]. The theoretical power to detect an effect at the given sample size 1.621, with OR 401 1.27 and AF 2.1% is estimated to be 0.50 for a significance level of 0.05. This is also true for other 402 alleles listed in **Supplementary Table 8**. The deviation from non-additivity of effects at the HLA locus 403 observed in Goyette et al.⁵ could not be replicated in this study (data not shown).

To reduce the complexity of the HLA signal further and to identify the properties of potential culprit antigens leading to disease, we analysed peptides preferentially bound by proteins, attributed risk and protection on the genetic level **(Figure 4)**. Additionally, we tried to identify shared physico-chemical properties of these proteins. 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

409 DRB1 proteins (DQ shown in Supplementary Figure 8). First, we predicted the binding affinities for 410 5 sets of 200,000 random unique peptides sampled from the human proteome to the DRB1 proteins 411 using NetMHCIIpan-3.2²⁸ (Supplementary Table 3, amino acid distribution). Next, we performed 412 clustering analysis on the alleles using the top 2% ranked preferentially binding peptides for each 413 allele based on pairwise observed complete observations. In general, we found DRB1-clustering 414 (Figure 4) to be more informative regarding separation of protective and risk alleles than DQ-415 clustering. Additionally, DRB1-clustering was more stable across the sets of random peptides 416 (Supplementary Figure 8). Larger "risk clusters" were identified for DRB1 including DRB1*15:01 and 417 the newly identified DRB1*15:03. We defined 2 risk clusters including DRB1*11:01/04 and 418 DRB1*13:01 (RISK 1) DRB1*12:01, DRB1*14:04 and DRB1*15:01/03 (RISK 2), and 2 protective 419 clusters including DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and DRB1*10:01 (PROT 1) and 420 DRB1*04:03/06 (PROT 2). Within each cluster, we calculated a unique peptide binding motif by 421 combining the top 2% of binders for each allele in the groups (Figure 4). The peptide binding motifs of 422 the two risk groups were enriched for basic amino acids (K and R) and depleted for acidic amino 423 acids, while the peptide binding motifs of the protective group were enriched for hydrophobic and polar 424 amino acids. Interestingly, DRB1*01:03 clustered with protective alleles DRB1*04:01/05, DRB1*07:01, 425 DRB1*09:01 and DRB1*10:01, however, a more detailed analysis of its physico-chemical properties 426 resulted in a predominant clustering with DRB1*15 (Figure 5). Equally, DRB1*15:02 clustered with 427 DRB1*13:02, while physico-chemical properties resulted in a predominant clustering with the 428 DRB1*15 group. In Supplementary Figures 9-12 we show that this may be an artefact of 429 NetMHCIIpan-3.2 caused by extrapolation of the DRB1*15:02 signal for unknown peptides from 430 DRB1*13:02.

431

432 **DISCUSSION**

433 Several conclusions can be drawn from this trans-ethnic HLA fine mapping study in UC: HLA allele 434 associations and their effect directions are broadly consistent across the different populations

435 analysed in this study and signals previously observed in a Caucasian-only approach can be replicated in this context⁵. While not in every case the same HLA allele is implicated across the 436 437 different populations, alleles of the same HLA allele group are associated with UC, as is the case for 438 the HLA allele group DRB1*15, of which DRB1*15:01, DRB1*15:02 and DRB1*15:03 are all 439 associated with the disease dependent on the HLA allele frequencies in each respective population. 440 The frequencies for these alleles computed in this study were consistent to the frequencies stored in 441 the allele frequency net database (AFND) ³⁴. Heterogeneity of effect sizes was observed, however the 442 accuracy of estimation of the effect sizes would increase with larger per-population sample sizes. As 443 observed also in the Caucasian-only approach, HLA associations are correlated across different HLA 444 genomic loci, especially for HLA-DRB1, -DRB3/4/5 and -DQ alleles, such that neither locus can be 445 ruled out as disease-relevant. Overall a high conservation of HLA-DQ-DR haplotypes was observed 446 across different ethnicities. In the Japanese and Korean population and entire haplotype spanning 447 class I and class II was observed for C*12:01-B*52:02-DRB5*01:02-DRB1*15:02-DQA1*01:03-448 DQB1*06:01. For South Western Asian (Iranian, Indian) individuals, other HLA associations were 449 observed to be more dominant (i.e. HLA-DRB1*11 and HLA-DRB1*14). Overall, for HLA-association 450 was dependent on the frequency of the allele and the size of the study cohort (i.e. alleles with a 451 sufficiently high frequency at the DRB1 locus were usually also associated with the disease, except for 452 alleles of the HLA-DRB1*08 and HLA-DRB1*16 groups which had frequency of 2.9% and 1.7% 453 respectively in the Caucasian population). Associations at DPA1-DPB1 can most likely be ruled out 454 and associations may merely result from correlation with HLA-DQ-DRB1. In the analysis of peptide-455 binding preferences for HLA-DRB1 alleles, we observed clustering according to the effect's direction in 456 the genetic analysis, i.e. protective or risk, which may point more to HLA-DRB1 playing a role. 457 However, an important limitation for the analogous DQ analysis is the limited availability of data 458 present in models for HLA-peptide binding prediction. The highly similar binding pockets of HLA-459 DRB1*13:01 and HLA-DRB1*13:02 suggest HLA-DQ alleles to mediate disease risk. DRB1*13:01, 460 which was estimated to confer risk is correlated with DQA1*01:03-DQB1*06:03 while DRB1*13:02, 461 which was estimated to be protective, is correlated with DQA1*01:02-DQB1*06:04.

462 Alleles of the DRB1*15 group also play a role as risk factors in other immune-related diseases including Multiple Sclerosis³⁵⁻³⁸ (a chronic inflammatory neurological disorder), Systemic Lupus 463 Erythematosus³⁹ and Dupuytrien's disease^{40,41} (both are disorders of the connective tissue). They 464 have also been reported to be associated with adult onset Still's disease⁴² (a systemic inflammatory 465 466 disease), Graves disease (an autoimmune disease that affects the thyroid), pulmonary tuberculosis and leprosy^{43,44} (a disease caused by *Mycobacterium leprae* that affects the skin). For Multiple 467 468 Sclerosis, DRB1*15:03, like in our study, was observed to be specific for African American 469 populations³⁵. DRB1*15 alleles have been reported to be strongly protective in type 1 diabetes (T1D) 470 (in which the autoimmune system attacks insulin producing beta cells of the pancreas), and 471 pemphigus vulgaris (a skin blistering disease). However, the functional consequences of HLA-472 DRB1*15 association with these diseases have not been addressed and for most of them the potential 473 disease driving antigens are not known. Exceptions are leprosy, in which Mycobacterium leprae 474 causes the disease and pemphigus vulgaris, in which the skin protein desmoglein is targeted. Krause-475 Kyora et al. found that DRB1*15:01, among 18 contemporary DRB1 proteins, was predicted to "bind 476 the second smallest number of potential *Mycobacterium leprae* antigens" and further hypothesized 477 that limited presentation of the Mycobacterium leprae antigens, may impair the immune response against this pathogen⁴³. Here an important note should be, that DRB1*15:01 is on average also the 478 479 most frequent HLA-DRB1 allele in the most analysed British/Central American European populations 480 as such has a higher statistical power to be detected in an association analysis.

Analysis of peptide binding motifs showed that protective and risk alleles cluster stably and that risk and protective groups have peptide binding motifs which are distinguishable by their physico-chemical properties. The arginine (R) and lysine (K) content was observed to be increased in peptides bound by HLA-proteins that were assigned to confer risk on a genetic level. This was more prominent for risk cluster 1 than risk cluster 2. Interestingly, Dhanda *et al.*, who compared 1,032 known T-cell epitopes from 14 different sources (including Mycobacterium tuberculosis, Dengue Fever, Virus, Zika Virus, house mite and other allergens) and known non-epitopes from the same data set, showed that T-cell

epitope amino acid motifs also are enriched in lysine and arginine content. The established motif is especially similar to the binding motif of risk cluster 1. Arginine is also found at an increased level in antimicrobial peptides^{45–47}. 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 bacteria⁴⁸. If and how this plays a role in the etiology of UC is however only to be speculated about.

493 One important limitation of the analysis of preferential HLA-peptide binding is the amount of data that 494 is used to train machine learning algorithms, which was especially limited for the HLA-DQ proteins. In 495 the future, larger datasets from peptidomics experiment will likely increase the accuracy of these 496 predictions and increase confidence in the risk and protective motifs that may be indicative of culprit 497 antigens in UC due to distinct features. Larger per-population patient collections will be needed in 498 future studies to confirm our results and to obtain even more precise effect estimates of associated 499 HLA alleles. In addition, we hope that IBD patient panels from other ethnicities will become available 500 for genetic fine mapping studies. With typing of HLA alleles now being possible using next-generation 501 sequencing methods, real typing rather than imputation analyzes should become standard, thereby 502 avoiding possible imputation artefacts. The construction of haplotype maps will then likely be even 503 more accurate.

505 **Description of Supplemental Data**

506

507 Supplementary Data contain **9 Tables** and **12 Figures**.

508

- 509 **Declaration of interests**
- 510 The authors declare no competing interests.

511 Acknowledgements & Grant support

512 This project received infrastructure support from the DFG Excellence Cluster No. 306 "Inflammation at 513 Interfaces". M.W. and H.E. are supported by the German Research Foundation (DFG) through the 514 Research Training Group 1743, "Genes, Environment and Inflammation". E.E. received funding from 515 the European Union Seventh Framework Program (FP7-PEOPLE-2013-COFUND; grant agreement 516 No. 609020 (Scientia Fellows)). S.A. is supported by joint funding from the University Medical Center 517 Groningen, Groningen, The Netherlands, and Institute for Digestive System Disease, Tehran 518 University of Medical Sciences, Tehran, Iran. Funding for the Multicenter African American IBD Study 519 (MAAIS) samples, for the GENESIS samples, and for the African Americans recruited by Cedars Sinai 520 was provided by the U.S.A. National Institutes of Health (NIH) grants DK062431 (S.R.B.), DK 087694 521 (S.K.), and DK062413 (D.P.B.M), respectively. This work was supported by a grant from the BioBank 522 Japan Project and, in part, by a Grant-in-Aid for Scientific Research (B) (26293180) funded by the 523 Ministry of Education, Culture, Sports, Science, and Technology, Japan. This research was supported 524 by a Mid-career Researcher Program grant through the National Research Foundation of Korea to 525 K.S. (2017R1A2A1A05001119), funded by the Ministry of Science, Information & Communication 526 Technology and Future Planning, and a grant of the Korea Health Technology R&D Project through 527 the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare 528 (grant number: HI18C0094), Republic of Korea. Funding for the Indian samples was provided by the 529 Centre Excellence Genome Predictive Medicine (Grant of in Sciences and #

530	BT/01/COE/07/UDSC/2008) from the Department of Biotechnology, Government of India). The
531	funders had no role in study design, data collection and analysis, decision to publish, or preparation of
532	the manuscript.
533	
534	
535	Author information
536	Steve R Brant, Tom H Karlsen, John D Rioux and Andre Franke: These authors jointly supervised this
537	work.
538	
539	International IBD Genetics Consortium
540	A full list of members and affiliations appears in the Supplementary Note.
541	
542	MAAIS Recruitment center
543	A full list of members and affiliations appears in the Supplementary Note.
544	
545	Authors contributions
546	F.D. performed statistical and computational analysis, G.B. contributed to statistical analysis. M.W.
547	and H.E. performed computational analysis with contributions from D.E., M.Hü, S.L., A.M., T.L. and
548	S.R G.M. performed protein structure analysis and analysis of physico-chemical properties with
549	contributions from F.D. F.D. and M.W. set up the HLA imputation pipeline. S.J. performed HLA typing
550	in contribution to the HLA reference panel. F.D., G.M., E.E., E.R. wrote or revised this manuscript.
551	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.,
552	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.,

553 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.

555 All authors reviewed, edited and approved the final manuscript.

556

557 Data availability

558 The ImmunoChip data used in this study are proprietary to the IIBGDC genetics consortium and may 559 be requested from the consortium. Any data produced within this study, may be requested from the 560 corresponding authors upon reasonable request including association statistics of imputed and 561 genotyped SNVs.

562 **Code availability**

563 Code used for analysis of data within this study is available from <u>f.degenhardt@ikmb.uni-kiel.de</u> upon 564 reasonable request. Part of the code has been incorporated into a github project and is available at 565 ikmb/HLApipe. This pipeline was developed by Frauke Degenhardt and Mareike Wendorff. It is based 566 on the HLA imputation tool HIBAG published by Zheng *et al.*²⁴.

567

568

569

570

572 **FIGURES**

573

574 Figure 1 – HLA regional association plots. Association analysis results for imputed and genotyped 575 single nucleotide variants (grev) and 4-digit HLA alleles (vellow) are shown for (a) 373 African 576 American cases and 590 controls (AA), (b) 13,927 Caucasian cases and 26,764 controls (EUR), and 577 (c) 709 Japanese cases 3,169 and controls (JPN) as well as (d) the meta-analysis (META) results 578 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 579 580 remaining populations are provided in **Supplementary Figure 2**. SNP. The curves in (a)-(c) show the 581 P-value of the meta-analysis (REC2p* or REC2p). In (d) the overlying curve shows the l² as a 582 measure of heterogeneity in the meta-analysis indicating the heterogeneity of effects and allele 583 frequencies in that region. Dashed lines indicate the thresholds of genome-wide (P=5x10⁻⁸) and 584 nominal significance (P=10⁻⁵) The association analyses indicate HLA class II as the most associated 585 susceptibility region across the different populations. In the Korean and the Japanese populations, a 586 strong association signal is also seen for B*52:01 and C*12:02, both alleles being in strong linkage 587 disequilibrium with the HLA class II loci DRB1*15:02, DQA1*01:02 and DQB1*06:01, i.e. another 588 population-specific haplotype association in these ethnicities exists.

589 Figure 2 – HLA single allele association analysis results at 2- and 4-digit resolution for MHC 590 class II loci -DRB3/4/5, -DRB1, -DQA1-DQB1. (AF; common defined as AF>1%), odds ratio (OR), P-591 value (P) and whether an allele had a P-value<0.05 (circle symbol) is shown for the respective 592 population (e.g. circles with black boundary and red color represent an allele that is common and 593 associated with risk). We depict association results of the analysis of the African American (AA), 594 Puerto Rican (PRI), Caucasian (EUR), Maltese (MLT), Iranian (IRN), North Indian (IND), Chinese 595 (CHN), Korean (KOR) and Japanese (JPN) cohorts and the meta-analysis (META) with RE2C's I2 as 596 an indicator of allelic heterogeneity and the P-value of association (RE2Cp or RE2Cp*, combined here 597 with single study P-values P). Only HLA alleles which are significant in the meta-analysis, that have an 598 AF>1% in at least one population and that have a marginal post imputation probability >0.6 are shown. 599 The strongest association signals in the meta-analysis are for risk alleles of the DRB1*15 group, i.e. 600 DRB1*15:01, DRB*15:02 and DRB1*15:03 and the alleles located on the same respective haplotype 601 (Figure 3). Alleles with OR>5.0 or OR<0.2 (rare and non-significant alleles may have larger/smaller 602 OR) values were "ceiled" at 5.0 and 0.2 respectively. The "consistent alleles" that are highlighted in 603 Figure 3 are highlighted in bold type on the left side.

604

605 Figure 3 – Haplotypes for associated HLA alleles.

606 For a selection of associated HLA alleles, we show the most frequently observed risk (a) and 607 protective (b) haplotypes in the respective populations. (African American (AA), Puerto Rican (PRI), 608 Caucasian (EUR), Maltese (MLT), Iranian (IRN), North Indian (IND), Chinese (CHN), Korean (KOR) 609 and Japanese (JPN)). Here we show only DRB1-DQA1-DQB1 haplotypes with a frequency >1% in the 610 case individuals in each respective population. The most frequently observed C-B alleles in each 611 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.¹⁵ and calculated based on 612 613 individuals hemizygous for HLA- DRB3/4/5 (i.e. carrying only one HLA-DRB1 observed with either 614 HLA-DRB3, -DRB4 or -DRB5 and one DRB1*01, DRB1*08 or DRB1*10 which are not observed with

615 any of the HLA- DRB3/4/5.)

616

617 Figure 4 – Clustering of DRB1 proteins according to preferential peptide binding and combined 618 peptide binding motifs. (MIDDLE CLUSTER): For 5 sets of 200,000 unique random human peptides 619 the percentile rank scores of preferential peptide binding were calculated using NetMHCIIpan-3.2.²⁸ for 620 all DRB1 proteins that were significant in the meta-analysis of genetic analysis of the HLA with and AF 621 > 1% in at least one cohort. We additionally included DRB1*01:03. Within each set, the top 2% 622 binders (according to NetMHCIIpan-3.2 threshold) were used to perform a clustering on the pairwise 623 correlations between two alleles using complete observations only. We show clustering results for 624 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, 625 626 DRB1*14:04 and DRB1*15:01/03 (RISK 2), DRB1*04:01/05, DRB1*07:01, DRB1*09:01 and 627 DRB1*10:01 (PROT 1) and DRB1*04:03/04/06 (PROT 2). For this analysis shared peptides (10% top 628 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 Seg2Logo.²⁹. The color scheme shows the 629 630 chemistry of the amino acids. Red: positively charged amino acids, blue: negatively charged amino 631 acids, green: polar amino acid, purple: neutral amino acid and black: hydrophobic amino acid.

633 Figure 5 – Cluster according to chosen physico-chemical properties of amino acids within the

634 peptide binding pockets.

We only show sites with variable information in pockets (P) 1, 4, 6, 7 and 9 and only proteins for which 635 636 the genetic analysis was significant (meta-analysis RE2Cp/RE2Cp* <0.05) and for which at least 1 637 cohort had AF >1%. We additionally show DRB1*01:03. Clustering was performed using the hclust 638 function of the R package stats. The box below the cluster plot shows positions of P1, 4, 6, 7 and 9 of 639 the beta (B) chain of the molecules (as defined in Supplementary Table 5). Here we show combined 640 scores F1 (a) and F3 (b) derived from a factor analysis of 54 unique amino acid properties (Atchley et al.³⁰). F1 captures polarity and hydrophobicity of the amino acid, while factor F3 captures amino acid 641 642 size and bulkiness. For F1, high values indicate larger hydrophobicity, polarity and hydrogen donor 643 abilities while low values indicate non-polar amino acids. For F3, high values indicate larger and 644 bulkier amino acids while low values indicate smaller, more flexible amino acids. We additionally show 645 the residue-volume (c) as a measure of pocket size and defined a score "hydrogen acceptor" (HB-646 acceptor) (d), which defines the ability of an amino acid to participate in hydrogen bonds and 647 corresponds to the number of atoms within the sidechain that can accept a hydrogen. Additional 648 information for the "charge" parameter and the analysis for DQA1-DQB1 can be found in 649 Supplementary Figures 9,10.

650

651 Figure 6 – Frequency of DRB1*01:03 across populations available in the allele frequency net 652 database. (a) "Worldmap", (b) zoom into European continent. Frequencies are shown within different 653 ranges noted by AF. Allele frequencies of DRB1*01:03 are lower across central Europe than in the 654 UK, Spain, India, South Africa, United States, and coastal regions of South America. Frequencies 655 were binned according to allele frequency. The figures were created using the R-package rworldmap. Frequencies were extracted from the allele frequency network database³⁴ for populations larger than 656 657 100 individuals. To plot the geographic locations, we converted assigned degree and minutes to 658 decimal numbers. We deleted all non-Caucasian populations with USA coordinates prior to plotting.

659 **REFERENCES**

- 1. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm,
- L.P., Sharma, Y., Anderson, C.A., et al. (2012). Host-microbe interactions have shaped the genetic
 architecture of inflammatory bowel disease. Nature *491*, 119–124.
- 2. 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.
- 3. Ellinghaus, D., Jostins, L., Spain, S.L., Cortes, A., Bethune, J., Han, B., Park, Y.R., Raychaudhuri,
- 667 S., Pouget, J.G., Hubenthal, M., et al. (2016). Analysis of five chronic inflammatory diseases identifies
- 668 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 48, 510–518.
- 4. de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice,
- D.L., Gutierrez-Achury, J., Ji, S.G., et al. (2017). Genome-wide association study implicates immune
- activation of multiple integrin genes in inflammatory bowel disease. Nat Genet *49*, 256–261.
- 5. 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
- 674 for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis.
- 675 Nat. Genet. 47, 172–179.
- 676 6. Stokkers, P.C., Reitsma, P.H., Tytgat, G.N., and van Deventer, S.J. (1999). HLA-DR and -DQ 677 phenotypes in inflammatory bowel disease: a meta-analysis. Gut *45*, 395–401.
- 7. Lappalainen, M., Halme, L., Turunen, U., Saavalainen, P., Einarsdottir, E., Farkkila, M., Kontula, K.,
 and Paavola-Sakki, P. (2008). 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.
- 682 8. Lu, M., and Xia, B. (2006). Polymorphism of HLA-DRB1 gene shows no strong association with
- 683 ulcerative colitis in Chinese patients. Int J Immunogenet 33, 37–40.

- 9. 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
- 686 for ulcerative colitis but reduces risk for Crohn's disease. Gastroenterology 141, 864–865.
- 10. Myung, S.J., Yang, S.K., Jung, H.Y., Chang, H.S., Park, B., Hong, W.S., Kim, J.H., and Min, I.
- 688 (2002). HLA-DRB1*1502 confers susceptibility to ulcerative colitis, but is negatively associated with its
- 689 intractability: a Korean study. Int J Color. Dis *17*, 233–237.
- 11. Mohammadi, M., Rastin, M., Rafatpanah, H., Abdoli Sereshki, H., Zahedi, M.J., Nikpoor, A.R.,
- 691 Baneshi, M.R., and Hayatbakhsh, M.M. (2015). Association of HLA-DRB1 Alleles with Ulcerative
- 692 Colitis in the City of Kerman, South Eastern Iran. Iran J Allergy Asthma Immunol 14, 306–312.
- 12. Gao, F., Aheman, A., Lu, J.J., Abuduhadeer, M., Li, Y.X., and Kuerbanjiang, A. (2014). Association
- of HLA-DRB1 alleles and anti-neutrophil cytoplasmic antibodies in Han and Uyghur patients with
- 695 ulcerative colitis in China. J Dig Dis *15*, 299–305.
- 13. Uyar, F.A., Imeryuz, N., Saruhan-Direskeneli, G., Ceken, H., Ozdogan, O., Sahin, S., and Tozun,
- N. (1998). The distribution of HLA-DRB alleles in ulcerative colitis patients in Turkey. Eur J
 Immunogenet 25, 293–296.
- 699 14. Han, B., Akiyama, M., Kim, K.-K., Oh, H., Choi, H., Lee, C.H., Jung, S., Lee, H.-S., Kim, E.E.,
- 700 Cook, S., et al. (2018). Amino acid position 37 of HLA-DRβ1 affects susceptibility to Crohn's disease
- in Asians. Hum. Mol. Genet.
- 15. Degenhardt, F., Wendorff, M., Wittig, M., Ellinghaus, E., Datta, L.W., Schembri, J., Ng, S.C.,
- Rosati, E., Hübenthal, M., Ellinghaus, D., et al. (2019). Construction and benchmarking of a multi-
- ethnic reference panel for the imputation of HLA class I and II alleles. Hum. Mol. Genet.
- 16. Huang, C., Haritunians, T., Okou, D.T., Cutler, D.J., Zwick, M.E., Taylor, K.D., Datta, L.W.,
- 706 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.
- 708 17. Ye, B.D., Choi, H., Hong, M., Yun, W.J., Low, H.Q., Haritunians, T., Kim, K.J., Park, S.H., Lee, I.,
- 709 Bang, S.Y., et al. (2016). Identification of Ten Additional Susceptibility Loci for Ulcerative Colitis

- 710 Through Immunochip Analysis in Koreans. Inflamm. Bowel Dis.
- 18. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for
- thousands of genomes. Nat Methods 9, 179–181.
- 19. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L.,
- 714 McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic
- 715 variation. Nature *526*, 68–74.
- 20. Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of
- 717 genomes. G3 1, 457–470.
- 21. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation
- 719 method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529.
- 22. Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference
- 721 Samples. Am J Hum Genet *98*, 116–126.
- 23. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-
- data inference for whole-genome association studies by use of localized haplotype clustering. Am. J.
- 724 Hum. Genet. *81*, 1084–1097.
- 725 24. 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.
- 25. 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,
- 729 e97282.
- 730 26. Lee, C.H., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide
- association studies to detect heterogeneous effects. Bioinformatics 33, i379–i388.
- 732 27. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet
- 733 Epidemiol 35, 809–822.
- 28. Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters,

- B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II
- 736 molecules. Immunology.
- 29. Thomsen, M.C.F., and Nielsen, M. (2012). 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.
- 30. Atchley, W.R., Zhao, J., Fernandes, A.D., and Drüke, T. (2005). Solving the protein sequence
- 741 metric problem. Proc. Natl. Acad. Sci. U. S. A.
- 742 31. Goldsack, D.E., and Chalifoux, R.C. (1973). Contribution of the free energy of mixing of
- hydrophobic side chains to the stability of the tertiary structure of proteins. J. Theor. Biol.
- 32. Shah, T.S., Liu, J.Z., Floyd, J.A., Morris, J.A., Wirth, N., Barrett, J.C., and Anderson, C.A. (2012).
- optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants.
- 746 Bioinformatics 28, 1598–1603.
- 33. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
- throughput. Nucleic Acids Res *32*, 1792–1797.
- 749 34. Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., da Silva, A.L.,
- 750 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, D784-8.
- 753 35. Hollenbach, J.A., and Oksenberg, J.R. (2015). The immunogenetics of multiple sclerosis: A
- comprehensive review. J Autoimmun *64*, 13–25.
- 36. Alcina, A., Abad-Grau Mdel, M., Fedetz, M., Izquierdo, G., Lucas, M., Fernandez, O., Ndagire, D.,
- 756 Catala-Rabasa, A., Ruiz, A., Gayan, J., et al. (2012). Multiple sclerosis risk variant HLA-DRB1*1501
- associates with high expression of DRB1 gene in different human populations. PLoS One 7, e29819.
- 758 37. Prat, E., Tomaru, U., Sabater, L., Park, D.M., Granger, R., Kruse, N., Ohayon, J.M., Bettinotti,
- 759 M.P., and Martin, R. (2005). HLA-DRB5*0101 and -DRB1*1501 expression in the multiple sclerosis-

- associated HLA-DR15 haplotype. J. Neuroimmunol.
- 38. 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.
- 39. Hanscombe, K.B., Morris, D.L., Noble, J.A., Dilthey, A.T., Tombleson, P., Kaufman, K.M.,
- Comeau, M., Langefeld, C.D., Alarcon-Riquelme, M.E., Gaffney, P.M., et al. (2018). Genetic fine
- 766 mapping of systemic lupus erythematosus MHC associations in Europeans and African Americans.
- 767 Hum. Mol. Genet.
- 40. Brown, J.J., Ollier, W., Thomson, W., and Bayat, A. (2008). Positive association of HLA-DRB1*15
- 769 with Dupuytren's disease in Caucasians. Tissue Antigens.
- 41. Replication, D.Ia.G., Meta-analysis, C., Asian Genetic Epidemiology Network Type 2 Diabetes, C.,
- South Asian Type 2 Diabetes, C., Mexican American Type 2 Diabetes, C., Type 2 Diabetes Genetic
- Exploration by Nex-generation sequencing in muylti-Ethnic Samples, C., Mahajan, A., Go, M.J.,
- Zhang, W., Below, J.E., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into
- the genetic architecture of type 2 diabetes susceptibility. Nat Genet *46*, 234–244.
- 42. Yap, L.M., Ahmad, T., and Jewell, D.P. (2004). The contribution of HLA genes to IBD susceptibility
 and phenotype. Best Pract. Res. Clin. Gastroenterol.
- 43. Krause-Kyora, B., Nutsua, M., Boehme, L., Pierini, F., Pedersen, D.D., Kornell, S.C., Drichel, D.,
- Bonazzi, M., Möbus, L., Tarp, P., et al. (2018). Ancient DNA study reveals HLA susceptibility locus for
- 779 leprosy in medieval Europeans. Nat. Commun.
- 44. Zhang, F., Liu, H., Chen, S., Wang, C., Zhu, C., Zhang, L., Chu, T., Liu, D., Yan, X., and Liu, J.
- 781 (2009). Evidence for an association of HLA-DRB115 and DRB109 with leprosy and the impact of
- 782 DRB109 on disease onset in a Chinese Han population. BMC Med. Genet.
- 45. Nguyen, L.T., Chau, J.K., Perry, N.A., de Boer, L., Zaat, S.A.J., and Vogel, H.J. (2010). Serum
- stabilities of short tryptophan- and arginine-rich antimicrobial peptide analogs. PLoS One.

- 46. Chan, D.I., Prenner, E.J., and Vogel, H.J. (2006). Tryptophan- and arginine-rich antimicrobial
- 786 peptides: Structures and mechanisms of action. Biochim. Biophys. Acta Biomembr.
- 47. Cutrona, K.J., Kaufman, B.A., Figueroa, D.M., and Elmore, D.E. (2015). Role of arginine and
- 788 Iysine in the antimicrobial mechanism of histone-derived antimicrobial peptides. FEBS Lett.
- 48. Brogden, K.A. (2005). Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? Nat.
- 790 Rev. Microbiol.
- 791



Figure 1



Figure 2



Figure 3



Figure 4



Figure 5



Figure 6