nature genetics

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

Philippe Goyette[1,30], Gabrielle Boucher[1,30], Dermot Mallon[2,3], Eva Ellinghaus[4], Luke Jostins[5,6], Hailiang Huang[7,8], Stephan Ripke[7,8], Elena S Gusareva[9,10], Vito Annese[11,12], Stephen L Hauser[13], Jorge R Oksenberg[13], Ingo Thomsen[4], Stephen Leslie[14,15], International Inflammatory Bowel Disease Genetics Consortium[16], Mark J Daly[7,8], Kristel Van Steen[9,10], Richard H Duerr[17,18], Jeffrey C Barrett[19], Dermot P B McGovern[20], L Philip Schumm[21], James A Traherne[22,23], Mary N Carrington[24,25], Vasilis Kosmoliaptsis[2,3], Tom H Karlsen[26–28,31], Andre Franke[4,31]& John D Rioux[1,29,31]

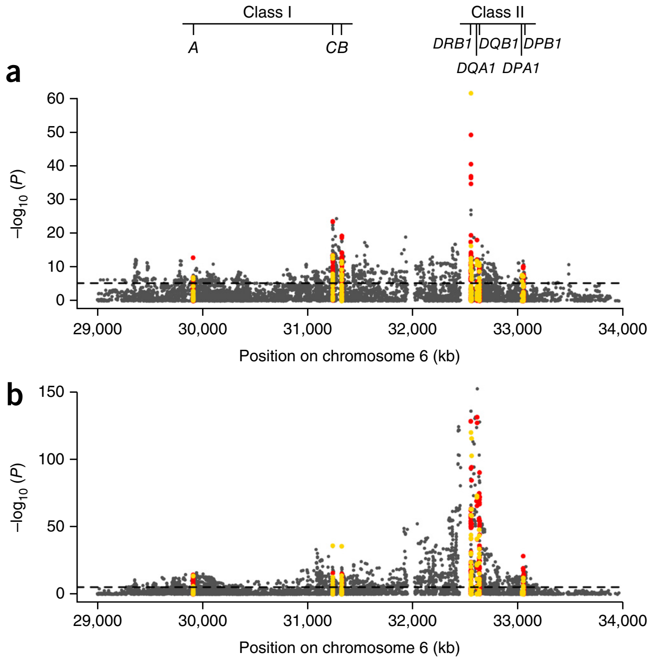
Genome-wide association studies of the related chronic inflammatory bowel diseases (IBD) known as Crohn's disease and ulcerative colitis have shown strong evidence of association to the major histocompatibility complex (MHC). This region encodes a large number of immunological candidates, including the antigen-presenting classical human leukocyte antigen (HLA) molecules[1]. Studies in IBD have indicated that multiple independent associations exist at HLA and non-HLA genes, but they have lacked the statistical power to define the architecture of association and causal alleles[2,3]. To address this, we performed high-density SNP typing of the MHC in >32,000 individuals with IBD, implicating multiple HLA alleles, with a primary role for HLA-DRB1*01:03 in both Crohn's disease and ulcerative colitis. Noteworthy differences were observed between these diseases, including a predominant role for class II HLA variants and heterozygous advantage observed in ulcerative colitis, suggesting an important role of the adaptive immune response in the colonic environment in the pathogenesis of IBD.

Meta-analyses of genome-wide association studies (GWAS) have recently shown that Crohn's disease (MIM 266600) and ulcerative colitis (MIM 191390) share the majority of the 163 known genetic risk factors for IBD, with the MHC region being one of the notable exceptions[4]. Data from these GWAS, however, have had insufficient variant density to define the association signals within the MHC. Targeted studies of IBD with higher variant density within the MHC region but with modest sample sizes have indicated that multiple independent associations are likely to exist in HLA and non-HLA genes, with the most consistent associations being with HLA class II loci, mainly *HLA-DRB1* and *HLA-DQB1*, and with some reports of association at the *HLA-C* class I locus and potentially also at non-HLA genes[2,3,5–8]. In the current study, we generated high-quality genotypes for 7,406 SNPs within the MHC region for a total of 18,405 individuals with Crohn's disease, 14,308 individuals with ulcerative colitis and 34,241 control subjects. Using these SNP data, we imputed and benchmarked the genetic variation within the class I (*HLA-B*, *HLA-C* and *HLA-A*) and class II (*HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*,

[1] Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [2] Department of Surgery, University of Cambridge, Cambridge, UK. [3] National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, Cambridge, UK. [4] Institute of Clinical Molecular Biology, Christian Albrechts University, Kiel, Germany. [5] Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK. [6] Christ Church, University of Oxford, St Aldates, UK. [7] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [8] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [9] Systems and Modeling Unit, Montefiore Institute, University of Liege, Liege, Belgium. [10] Bioinformatics and Modeling, GIGA-R (Groupe Interdisciplinaire de Génoprotéomique Appliquée) Research Center, University of Liege, Liege, Belgium. [11] Unit of Gastroenterology, IRCCS-CSS (Istituto di Ricovero e Cura a Carattere Scientifico–Casa Sollievo della Sofferenza) Hospital, San Giovanni Rotondo, Italy. [12] Unit of Gastroenterology SOD2 (Strutture Organizzative Dipartimentali), Azienda Ospedaliero Universitaria (AOU) Careggi, Florence, Italy. [13] Department of Neurology, University of California, San Francisco, San Francisco, California, USA. [14] Murdoch Children's Research Institute, Parkville, Victoria, Australia. [15] Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [16] A full list of members and affiliations appears at the end of the paper. [17] Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [18] Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA. [19] Wellcome Trust Sanger Institute, Hinxton, UK. [20] F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [21] Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA. [22] Cambridge Institute for Medical Research, Cambridge, UK. [23] Department of Pathology, University of Cambridge, Cambridge, UK. [24] Cancer and Inflammation Program, Laboratory of Experimental Immunology, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA. [25] Ragon Institute of Massachusetts General Hospital, MIT and Harvard, Cambridge, Massachusetts, USA. [26] Research Institute of Internal Medicine, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [27] Norwegian Primary Sclerosing Cholangitis Research Center, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [28] K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [29] Faculté de Médecine, Université de Montréal, Montreal, Quebec, Canada. [30] These authors contributed equally to this work. [31] These authors jointly supervised this work. Correspondence should be addressed to J.D.R. (john.david.rioux@umontreal.ca).

Received 8 July 2014; accepted 4 December 2014; published online 5 January 2015; doi:10.1038/ng.3176

Figure 1 Primary univariate association analyses of Crohn's disease and ulcerative colitis. Univariate association analysis results for 8,939 SNPs (dark gray) (**Supplementary Table 12**) and 90 2-digit and 138 4-digit HLA alleles (yellow) (**Supplementary Table 13**), as well as 741 single–amino acid variants (red) (**Supplementary Table 4**) in the MHC region are shown for 18,405 Crohn's disease cases and 14,308 ulcerative colitis cases (with 34,241 shared control subjects). Given that previous genetic analyses have identified distinct effects in the MHC region for Crohn's disease and ulcerative colitis, with different non-correlated alleles identified in each disease, we opted to perform the fine-mapping analyses separately for each disease. (**a**) The primary univariate association analysis in Crohn's disease identified over 1,789 markers showing study-wide significant association ($P < 5 \times 10^{-6}$, dashed line) across the MHC region, including 32 4-digit classical HLA alleles (**Fig. 3** and **Supplementary Table 2**). The single most significant variant for Crohn's disease is HLA-DRB1*01:03 ($P = 3 \times 10^{-62}$, OR = 2.51). (**b**) The primary univariate association analysis in ulcerative colitis identified over 2,762 markers showing study-wide significant association across the MHC region, including 50 4-digit classical HLA alleles (**Fig. 3** and **Supplementary Table 3**). The single most significant variant for ulcerative colitis is rs6927022 ($P = 8 \times 10^{-154}$, OR = 1.49), whereas the best HLA allele is HLA-DRB1*01:03 ($P = 3 \times 10^{-119}$, OR = 3.59), each acting independently. Twenty-nine SNPs and nine amino acid variants surpassed HLA-DRB1*01:03 as the next most significant variants in the primary analysis; however, all of these are correlated with rs6927022, and their significance is dramatically reduced by conditional logistic regression.



HLA-DRB5, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) HLA genes at the level of classical HLA alleles and amino acid positions (Online Methods).

As a first step to defining the nature of the association with Crohn's disease and ulcerative colitis within the MHC, we performed univariate analyses of the SNPs, classical HLA alleles and HLA amino acids. These analyses identified a very large number of variants across the MHC region with significant association ($P < 5 \times 10^{-6}$) with these disease phenotypes (**Fig. 1**), with major peaks of association centered in and around the classical HLA genes, suggesting a role for classical HLA alleles in risk of Crohn's disease and ulcerative colitis. This observation is consistent with gene-based analyses, which show strong association in the HLA genes for both ulcerative colitis and Crohn's disease (for example, $P < 1 \times 10^{-300}$ for *HLA-DRB1* in ulcerative colitis) (**Supplementary Table 1**). In particular, these analyses demonstrated a role for *HLA-DRB1* that cannot be attributed to other HLA genes, with evidence of residual association in class I and class II regions (**Supplementary Table 1**). To be more quantitative, we calculated the variance explained by the class I and class II alleles. Whereas the contributions of the class I and class II alleles were relatively equivalent in Crohn's disease, not only was the overall impact of the HLA region on disease risk greater in ulcerative colitis, but the alleles in the class II region had nearly threefold greater impact than the class I alleles (**Fig. 2**). Moreover, these analyses showed that classical HLA alleles explained three- to tenfold more of the disease variance than that explained by the index SNPs that were previously identified (~3% versus ~0.3% in Crohn's disease; ~6% versus ~2% in ulcerative colitis) (**Fig. 2**).

Specifically, in our univariate analyses, the most significant association in Crohn's disease was with HLA-DRB1*01:03 ($P < 4 \times 10^{-62}$, odds ratio (OR) = 2.53), with a P value over ten orders of magnitude more significant than the next best associated variants in the region. Notably, HLA-DRB1*01:03 had an effect in Crohn's disease that was statistically independent from the effects of the other most strongly associated variants in the MHC region, as shown by reciprocal conditional logistic regression (**Supplementary Fig. 1**). In the ulcerative colitis univariate analysis, the single most significant variant was a noncoding SNP (rs6927022: $P < 5 \times 10^{-153}$, OR = 1.49) near *HLA-DQA1*, identified in the recent meta-analysis of GWAS[4]; although multiple additional variants showed highly significant association, most were correlated with this top signal (**Fig. 1** and **Supplementary Fig. 2**). Strikingly, the next strongest independent association was with HLA-DRB1*01:03, having a much larger OR ($P < 1 \times 10^{-120}$, OR = 3.63; conditional P value (P_{cond}) $< 2 \times 10^{-89}$,

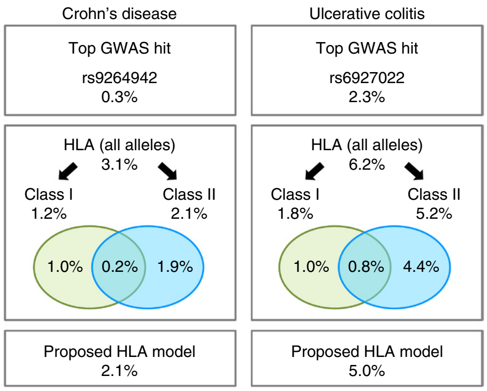


Figure 2 Variance explained by four-digit HLA alleles in Crohn's disease and ulcerative colitis. Proportion of variance explained on a logit scale (McKelvey and Zavoina's pseudo R^2; Online Methods) for different models in Crohn's disease (left) and ulcerative colitis (right). The top boxes show the variance explained by previously identified GWAS index SNPs within the MHC region[4]. The middle boxes illustrate the variance explained by HLA models including all 4-digit alleles of frequency > 0.5% (126 alleles in Crohn's disease and ulcerative colitis) and models restricted to 4-digit alleles within either the class I (63 alleles) or class II (63 alleles) region, respectively. The Venn diagram illustrates the proportion of variance explained that is unique to class I or class II alleles or is shared. The bottom boxes indicate the variance explained by the proposed HLA models (15 and 16 alleles in Crohn's disease and ulcerative colitis, respectively). Note that these estimations of variance explained were performed on the logit scale for practical reasons and should not be directly compared to heritability estimates computed on the (Gaussian) liability scale.

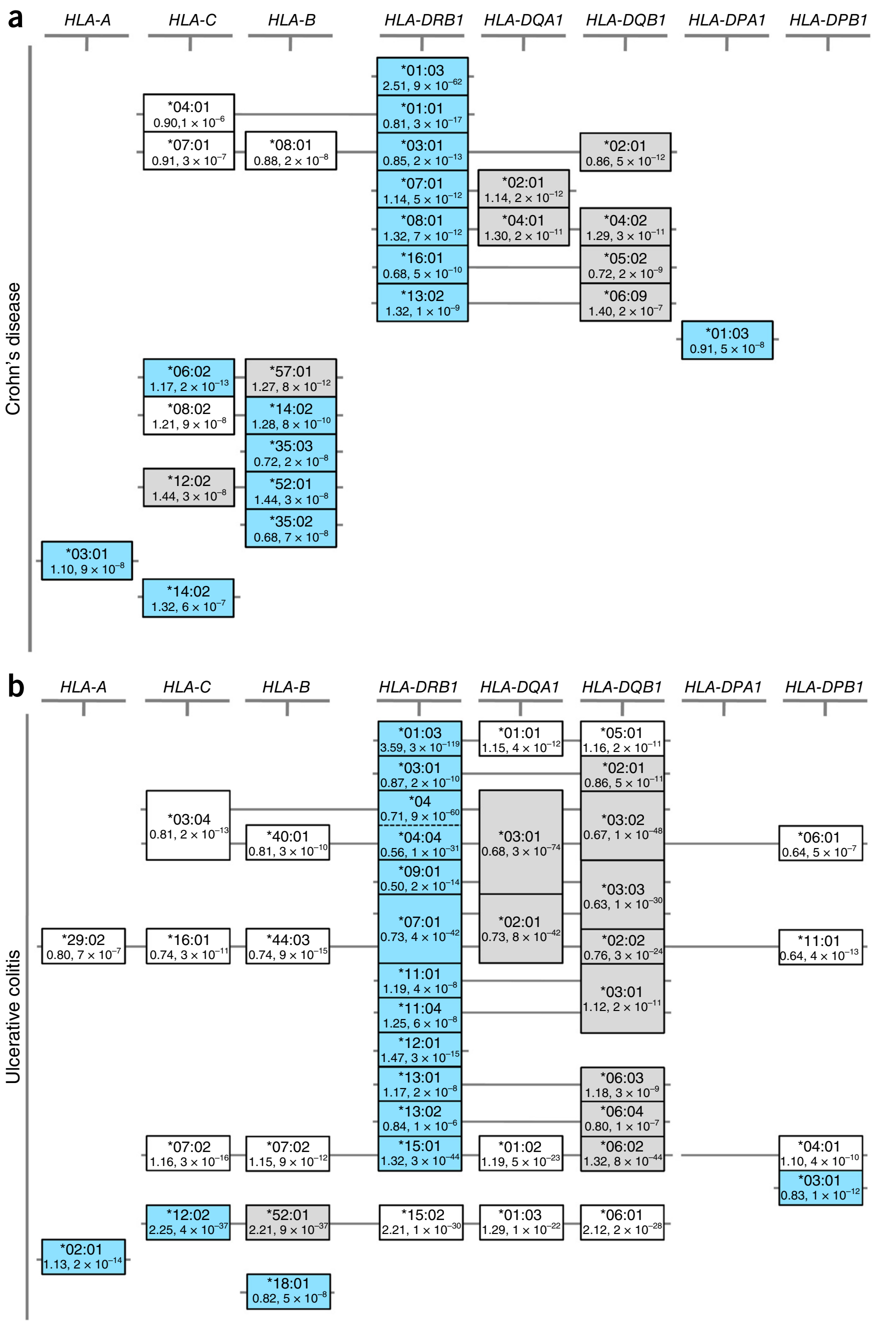


Figure 3 Correlated association signals at HLA alleles support potential alternate association models for both Crohn's disease and ulcerative colitis. (**a**,**b**) Equivalence of effect for the different study-wide significant associated four-digit HLA alleles is shown for Crohn's disease (**a**) and ulcerative colitis (**b**). 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 (**Supplementary Tables 2** and **3**), although such correlations are clearly dependent on the underlying haplotypic structure of the region. 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. Alleles at the *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* genes were omitted to simplify the display; many of the alleles for these genes show high frequency and, as such, are correlated to many different alleles (both risk and protective) at the other class II genes. Of note, the HLA-DRB4*null allele is the second strongest associated allele in ulcerative colitis (**Supplementary Table 3**).

OR = 3.06) (**Supplementary Fig. 2**). Reciprocal conditioning on HLA-DRB1*01:03 did not abolish the effect seen at rs6927022 ($P < 9 \times 10^{-123}$, OR = 1.43), indicating that these have mostly statistically independent effects in ulcerative colitis. Taken together, our analyses point to HLA-DRB1*01:03 as likely being causal in both diseases, with additional causal alleles in the class II and class I regions. Given this observation, it is probable that additional alleles within *HLA-DRB1* contribute to IBD risk.

We thus examined an *HLA-DRB1*–centric model and identified seven *HLA-DRB1* alleles with independent effects on risk of Crohn's disease (study-wide significance threshold of 5×10^{-6}) (**Supplementary Table 2**). Moreover, when controlling for these seven *HLA-DRB1* alleles, we identified only a single additional class II allele (HLA-DPA1*01:03) independently associated with Crohn's disease. Using the same conditional logistic regression framework for analysis of the class I locus, we identified seven class I HLA alleles that were significantly associated with Crohn's disease after conditioning on the eight class II alleles (**Fig. 3** and **Supplementary Table 2**). This *HLA-DRB1*–centric model explained about 2% of disease variance (**Fig. 2**). In ulcerative colitis, we identified a total of 12 *HLA-DRB1* alleles, 1 *HLA-DPB1* allele and 3 class I alleles (**Supplementary Table 3**) that could explain the association with the MHC region and that accounted for about 5% of disease variance (**Fig. 2**).

For many of the alleles identified in the *HLA-DRB1*–centric model, a few other candidate alleles in class I or class II genes could be considered (**Fig. 3**). In particular, multiple *HLA-DRB1* alleles had associations equivalent to those at *HLA-DQA1* and *HLA-DQB1* (for example, HLA-DQA1*03:01 was equivalent to the HLA-DRB1*04 and HLA-DRB1*09 alleles in ulcerative colitis), equally supporting a role for genetic variation within *HLA-DQA1* and *HLA-DQB1* in disease susceptibility, particularly for ulcerative colitis (**Fig. 3**). However, several of the alleles in these models, including HLA-DRB1*01:03, did not have any such proxies and thus are strong candidates for being causal (**Fig. 3**). Further dissection of these class II correlated signals to identify potential causal alleles may only be feasible in admixed populations or ones of diverse ancestry[9]. Further refinement might also be possible by examining the impact of clinical subphenotype and associated autoimmune comorbidities on observed associations, although functional studies will be needed to infer causality. For the present analysis, we were able to assess the impact of colonic versus non-colonic inflammation and found that HLA-DRB1*01:03 was

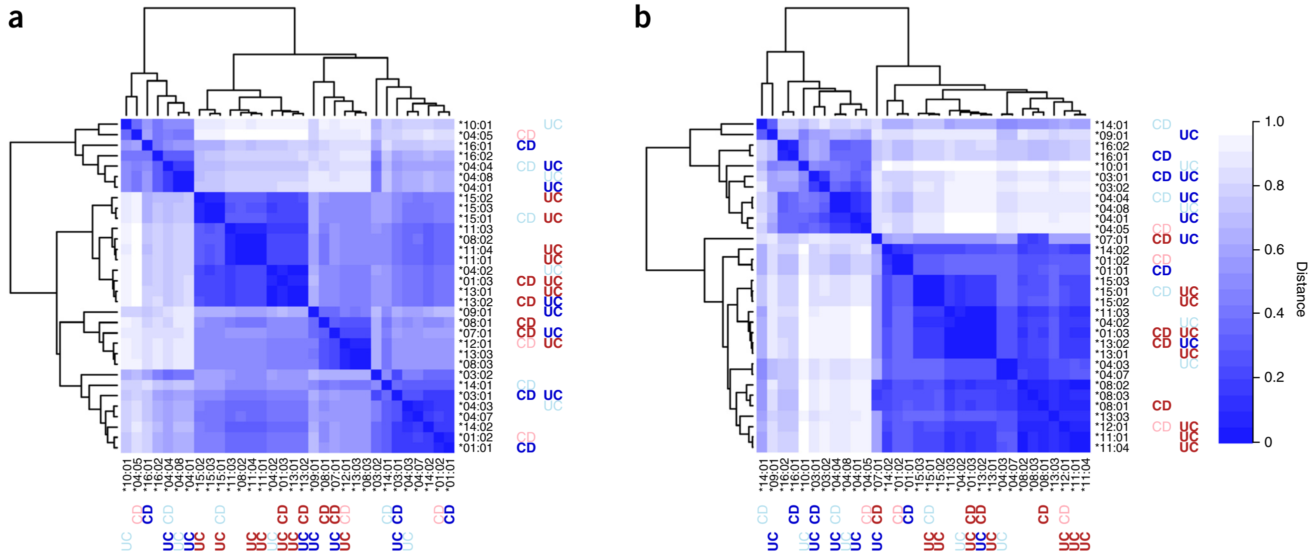


Figure 4 HLA-DR peptide-binding groove electrostatic properties and risk of IBD. The electrostatic potential of all HLA-DR molecules associated with ulcerative colitis (UC) or Crohn's disease (CD) and of all common HLA-DR molecules (frequency >1%) was calculated. HLA-DR molecules associated with increased or decreased risk of IBD at a study-wide significance level ($P < 5 \times 10^{-6}$) are shown in dark red or dark blue, respectively. Respective risk associations at a suggestive level ($1 \times 10^{-4} < P < 5 \times 10^{-6}$) are shown in pale red and pale blue. Comparisons of electrostatic potential among HLA-DR molecules were performed in a pairwise, all-versus-all fashion (Online Methods) to produce distance matrices that are displayed as symmetrical heat maps (scale ranges from 0 (identical) to 1 (maximum difference)). (**a**) The electrostatic potential in seven regions within the peptide-binding groove (Online Methods and **Supplementary Fig. 12**), which interact with the presented peptide, were compared among the HLA-DR molecules and pooled onto a single Euclidian distance matrix. Distance-based clustering identified four clusters, with an enrichment of risk alleles in two of these. Comparison of the electrostatic potential at individual peptide-binding groove regions is shown in **Supplementary Figure 13**. (**b**) Heat map representing electrostatic potential differences among the HLA-DR molecules at a spherical region that encompasses amino acids 67, 70 and 71 of the HLA-DRβ chain (associated with risk for ulcerative colitis and Crohn's disease; **Supplementary Table 13**). Distance-based clustering identified two clusters that correlate with the directionality of effect in IBD.

associated with colonic Crohn's disease and that HLA-DRB1*07:01 was associated with the absence of colon involvement (**Supplementary Fig. 3**), in line with previous suggestions[10]. This explains the shared associations for Crohn's disease and ulcerative colitis at HLA-DRB1*01:03 and strongly suggests that this allele is critically involved in determining the colonic immune response to local flora.

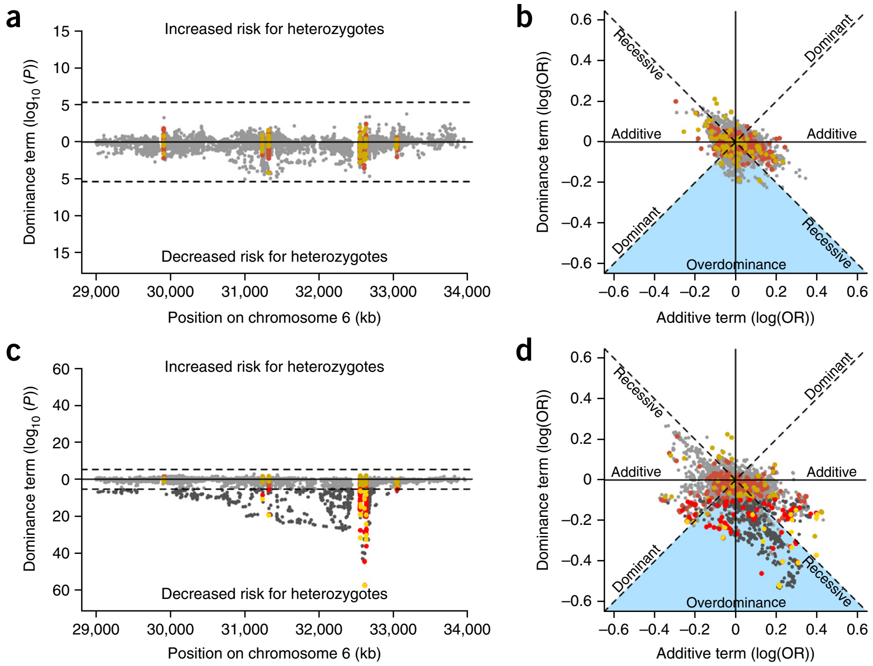


Figure 5 Non-additive effect models in Crohn's disease and ulcerative colitis. (**a**–**d**) Evidence for non-additive effects of common variants (frequency >5%) across the MHC region tested under a general model of additive and dominance effects (Online Methods) in Crohn's disease and ulcerative colitis. (**a**,**c**) The P values and directionality for departure from additive effect (dominance term) are represented on the y axis for Crohn's disease (**a**) and ulcerative colitis (**c**). HLA alleles and amino acid variants are in yellow and red, respectively, and SNPs are represented in dark gray. Variants with a non-significant ($P > 5 \times 10^{-6}$; dashed line) dominance term are plotted in less pronounced colors. A clear enrichment for lower risk in heterozygotes is observed in ulcerative colitis, as suggested by the large number of significant negative dominance terms (lower part of the plot). This effect is absent in Crohn's disease or much less important. (**b**,**d**) The dominance term OR is illustrated (y axis) versus the additive term (x axis) for Crohn's disease (**b**) and ulcerative colitis (**d**). Protective and risk-associated minor alleles are shown on the left and right sides of the plot, respectively. Strictly recessive or dominant variants are expected to fall on the diagonals, whereas strictly additive variants lay on or close to the x axis. The y axis is the expected position for variants with pure over- or underdominance. In ulcerative colitis, many alleles fall into the region of the plot for protective dominant, risk recessive or overdominance (blue triangle) (see **Supplementary Table 9** for pairwise comparison of *HLA-DRB1* alleles). These non-additive effects are observed for many variants in ulcerative colitis (**c**,**d**) (for example, HLA-DRB1*03:01 and HLA-DQB1*02:01) but are mostly absent in Crohn's disease (**a**,**b**), with one notable exception being the HLA-B*08 allele (**Supplementary Fig. 6**).

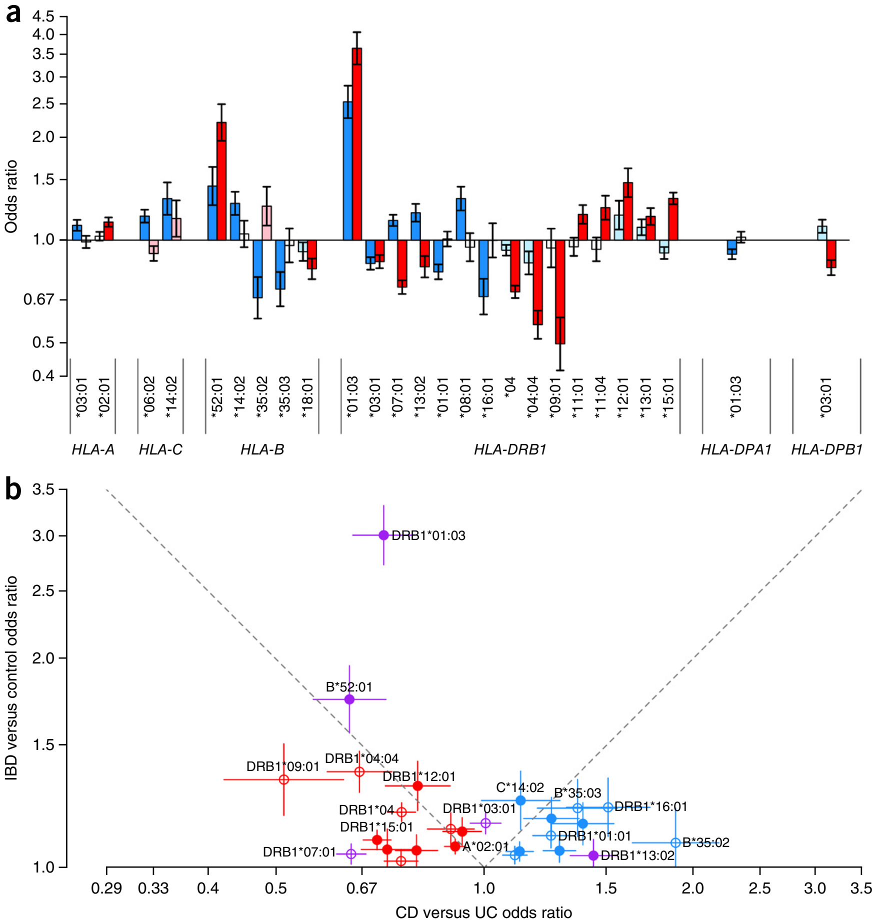


Figure 6 Comparison of OR values in Crohn's disease and ulcerative colitis for HLA alleles identified from HLA-focused models. (**a**) OR values from the primary univariate association analyses in Crohn's disease and ulcerative colitis for all alleles identified in the HLA-focused models of Crohn's disease and/or ulcerative colitis are presented with 95% confidence intervals. OR values for Crohn's disease and ulcerative colitis are in blue and red, respectively; darker colors indicate a study-wide significant effect ($P < 5 \times 10^{-6}$), lighter colors indicate a nominal significance level ($0.05 > P \geq 5 \times 10^{-6}$) and white indicates non-significance ($P \geq 0.05$) (for specific effect and significance values, refer to **Fig. 3** and **Supplementary Tables 2** and **3**). Allele HLA-B*52:01 is indicated for ulcerative colitis in place of the equivalent HLA-C*12:02 allele to simplify the display of this shared signal. (**b**) For the same HLA alleles, OR values (with 95% confidence intervals) for an IBD analysis are plotted against the OR values for the Crohn's disease versus ulcerative colitis analysis, with the IBD risk allele as the reference. Empty circles represent variants where the absence of the allele is the reference. Alleles identified as significant in Crohn's disease or ulcerative colitis only are plotted in blue and red, respectively; variants identified as significant in both are shown in purple. Note that HLA-DRB1*07:01 and HLA-DRB1*13:02 have opposite directions of effect in Crohn's disease and ulcerative colitis. Shared association signals are expected to fall into the upper triangle of the plot. Most variants fall outside of this region, highlighting the difference between Crohn's disease and ulcerative colitis in the MHC region.

Given that classical HLA alleles consist of combinations of specific amino acids at multiple positions, we tested whether the association with disease could be better explained by single amino acid positions. Indeed, we observed very strong association signals at many single-residue variants in Crohn's disease (for example, five amino acids of HLA-DRβ at positions 67, 70 and 71) and in ulcerative colitis (for example, four amino acid variants in HLA-DQα at positions 50, 53 and 215 and four amino acid variants in HLA-DRβ at positions 98 and 104) and also performed per-position omnibus analyses that confirmed the predominant association with HLA-DRβ position 11 in ulcerative colitis, as previously reported[5], and with HLA-DRβ position 70 in Crohn's disease (**Supplementary Fig. 4** and **Supplementary Tables 4** and **5**). Although the hypothesis of a positional effect is appealing, the interpretation of these position-based tests is not straightforward in the context of multiple alleles that are likely causal (**Supplementary Fig. 5**, **Supplementary Table 6** and **Supplementary Note**). Furthermore, in this study, amino acid–based models did not capture the association at *HLA-DRB1* in a more parsimonious way than the HLA allele–based models (**Supplementary Note**). To further explore the basis for the observed HLA associations, we performed three-dimensional protein structure modeling followed by analysis of the electrostatic properties of the binding groove of associated ($P < 1 \times 10^{-4}$) and common (frequency > 1%) HLA-DRB1 molecules. These analyses suggested that HLA-DR molecules associated with increased risk of ulcerative colitis and Crohn's disease share structural and electrostatic properties within or near the peptide-binding groove that are largely distinct from those of HLA-DR molecules associated with decreased risk of ulcerative colitis and Crohn's disease (**Fig. 4**).

Although we performed the primary analyses on the basis of a dose effect model, our sample size allowed us to investigate the effects further by testing for non-additive effects. In fact, we found significant departure from additive effects in ulcerative colitis but not in Crohn's disease (**Fig. 5a–c**). Specifically, we found evidence of decreased heterozygosity in individuals with ulcerative colitis for genotyped and imputed variants across the MHC region and at HLA genes, mostly in class II genes (**Supplementary Tables 7** and **8**). This heterozygote advantage could be explained by an enrichment of dominant protective and recessive risk alleles[11] that were absent or had much less of an effect in Crohn's disease (**Fig. 5** and **Supplementary Fig. 6**). Notably, we also detected multiple overdominant effects in ulcerative colitis, the strongest of which was captured by HLA-DRB1*03:01 (**Fig. 5**, **Supplementary Figs. 6** and **7**, and **Supplementary Table 9**). This allele is mostly found on the ancestral haplotype 8.1, a relatively common (~5–10%) haplotype that is conserved in European populations and is implicated in other immune diseases[12–14]. The overdominance effect of this haplotype in ulcerative colitis is possibly due to the presence of both dominant protective and recessive risk alleles, which would be consistent with the reported recessive risk of this haplotype in the ulcerative colitis–related biliary disease primary sclerosing cholangitis (PSC; **Supplementary Figs. 8** and **9**)[15,16]. Analogous with an infectious paradigm[11], these data may suggest that decreased HLA class II heterozygosity might impair the ability to appropriately control colonic microbiota in ulcerative colitis.

Although there is a substantial challenge in defining the causal alleles for Crohn's disease and ulcerative colitis in the MHC, given the linkage disequilibrium (LD) structure in the region, a number of conclusions can be drawn, regardless of the models tested. First, the high-density mapping of this region in a large cohort demonstrated the significant contribution of the MHC region to disease risk, a contribution that is not apparent in the previous GWAS. Second, for both Crohn's disease and ulcerative colitis, it would appear that variation within HLA genes as opposed to variation in other genes within the MHC region has a predominant role in disease susceptibility. Third, whereas the contribution of class I and class II HLA variants to disease risk is relatively equivalent in Crohn's disease, HLA class II variation has a more important role in ulcerative colitis. Fourth, in contrast to the majority of non-MHC susceptibility loci being shared by Crohn's disease and ulcerative colitis, most associated HLA alleles have a predominant role in either Crohn's disease or ulcerative colitis, with very few conferring shared IBD risk (**Fig. 6**). Finally, the decreased heterozygosity in ulcerative colitis suggests that the ability to recognize a broader set of antigens, potentially of colonic microbial origin, is important to mount protective immunity.

URLs. Additional data on the heterogeneity of effects for associated alleles, http://www.medgeni.org/goyette_nature_gen_2015; European Bioinformatics Institute (EBI) sequence database, ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/; International Inflammatory Bowel Disease Genetics Consortium, http://www.ibdgenetics.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank the International PSC study group (http://www.ipscsg.org/) for sharing data. We are grateful to B.A. Lie and K. Holm for helpful discussions. J.D.R. holds a Canada Research Chair, and this work was supported by a US National Institute of Diabetes and Digestive and Kidney Diseases grant (NIDDK; R01 DK064869 and U01 DK062432). The laboratory of A.F. is supported by the German Ministry of Education and Research (BMBF) grant program e:Med (sysINFLAME). A.F. receives infrastructure support from the Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence 'Inflammation at Interfaces' and holds an endowment professorship (Peter Hans Hofschneider Professorship) of the Foundation for Experimental Biomedicine (Zurich, Switzerland). Grant support for T.H.K. and A.F. was received from the European Union Seventh Framework Programme (FP7/2007-2013, grant number 262055, ESGI). M.N.C. is supported by the Intramural Research Program of the US National Institutes of Health (NIH), Frederick National Laboratory, Center for Cancer Research. This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US government. J.C.B. was supported by a Wellcome Trust grant (WT098051). D.M. and V.K. are supported by the NIHR Cambridge Biomedical Research Centre. L.P.S. is supported by an NIDDK grant (U01 DK062429-14). J.A.T. is supported by the UK Medical Research Council. D.P.B.M. is supported by the Leona M. and Harry B. Helmsley Charitable Trust, the European Union (305479) and by grants from the NIDDK (U01 DK062413, P01 DK046763-19, U54 DE023789-01), the National Institute of Allergy and Infectious Diseases (NIAID; U01 AI067068) and the Agency for Healthcare Research and Quality (AHRQ; HS021747). R.H.D. holds the Inflammatory Bowel Disease Genetic Research endowed chair at the University of Pittsburgh and was supported by an NIDDK grant (U01 DK062420) and a US National Cancer Institute grant (CA141743). S.L.H. and J.R.O. would like to also acknowledge the support of the US NIH (R01 NS049477 and 1U19 A1067152) and the National Multiple Sclerosis Society (RG 2899-D11). S.L. wishes to acknowledge support from the Australian National Health and Medical Research Council (R.D. Wright Career Development Fellowship, APP1053756).

AUTHOR CONTRIBUTIONS

J.D.R., M.J.D., K.V.S., V.K., T.H.K. and A.F. jointly supervised research. J.D.R., P.G., G.B., R.H.D., J.C.B., D.P.B.M., J.A.T., M.N.C., V.K. and A.F. conceived and designed the experiments. P.G., G.B., D.M. and L.J. performed statistical analysis. P.G., G.B., L.J. and E.S.G. analyzed the data. V.A., S.L.H., J.R.O., I.T., S.L. and L.P.S. contributed reagents, materials or analysis tools. P.G., G.B., E.E., H.H. and S.R. performed data quality control and imputation. J.D.R., P.G., G.B., D.M., D.P.B.M., V.K., T.H.K. and A.F. wrote the manuscript. All authors read and approved the final manuscript before submission.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
2. Rioux, J.D. *et al.* Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc. Natl. Acad. Sci. USA* **106**, 18680–18685 (2009).
3. 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).
4. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
5. Achkar, J.P. *et al.* Amino acid position 11 of HLA-DRβ1 is a major determinant of chromosome 6p association with ulcerative colitis. *Genes Immun.* **13**, 245–252 (2012).
6. Jones, D.C. *et al.* Killer Ig-like receptor (KIR) genotype and HLA ligand combinations in ulcerative colitis susceptibility. *Genes Immun.* **7**, 576–582 (2006).
7. Kulkarni, S. *et al.* Genetic interplay between *HLA-C* and *MIR148A* in HIV control and Crohn disease. *Proc. Natl. Acad. Sci. USA* **110**, 20705–20710 (2013).
8. Satsangi, J. *et al.* Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* **347**, 1212–1217 (1996).
9. Oksenberg, J.R. *et al.* Mapping multiple sclerosis susceptibility to the *HLA-DR* locus in African Americans. *Am. J. Hum. Genet.* **74**, 160–167 (2004).
10. Newman, B. *et al. CARD15* and *HLA-DRB1* alleles influence susceptibility and disease localization in Crohn's disease. *Am. J. Gastroenterol.* **99**, 306–315 (2004).
11. Lipsitch, M., Bergstrom, C.T. & Antia, R. Effect of human leukocyte antigen heterozygosity on infectious disease outcome: the need for allele-specific measures. *BMC Med. Genet.* **4**, 2 (2003).
12. Alper, C.A., Fleischnick, E., Awdeh, Z., Katz, A.J. & Yunis, E.J. Extended major histocompatibility complex haplotypes in patients with gluten-sensitive enteropathy. *J. Clin. Invest.* **79**, 251–256 (1987).
13. Aly, T.A. *et al.* Multi-SNP analysis of MHC region: remarkable conservation of HLA-A1-B8-DR3 haplotype. *Diabetes* **55**, 1265–1269 (2006).
14. Wiencke, K., Spurkland, A., Schrumpf, E. & Boberg, K.M. Primary sclerosing cholangitis is associated to an extended B8-DR3 haplotype including particular *MICA* and *MICB* alleles. *Hepatology* **34**, 625–630 (2001).
15. Donaldson, P.T. & Norris, S. Evaluation of the role of MHC class II alleles, haplotypes and selected amino acid sequences in primary sclerosing cholangitis. *Autoimmunity* **35**, 555–564 (2002).
16. Liu, J.Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* **45**, 670–675 (2013).

Members of the International Inflammatory Bowel Disease Genetics Consortium:

Clara Abraham[32], Jean-Paul Achkar[33,34], Tariq Ahmad[35], Leila Amininejad[36,37], Ashwin N Ananthakrishnan[38,39], Vibeke Andersen[40,41], Carl A Anderson[19], Jane M Andrews[42], Vito Annese[11,12], Guy Aumais[29,43], Leonard Baidoo[17], Robert N Baldassano[44], Tobias Balschun[4], Peter A Bampton[45], Murray Barclay[46], Jeffrey C Barrett[19], Theodore M Bayless[47], Johannes Bethge[48], Joshua C Bis[49], Alain Bitton[50], Gabrielle Boucher[1], Stephan Brand[51], Steven R Brant[47], Carsten Büning[52], Angela Chew[53,54], Judy H Cho[55], Isabelle Cleynen[56], Ariella Cohain[57], Anthony Croft[58], Mark J Daly[7,8], Mauro D′Amato[59], Silvio Danese[60], Dirk De Jong[61], Martine De Vos[62], Goda Denapiene[63], Lee A Denson[64], Kathy L Devaney[38], Olivier Dewit[65], Renata D′Inca[66], Marla Dubinsky[67], Richard H Duerr[17,18], Cathryn Edwards[68], David Ellinghaus[4], Jonah Essers[69,70], Lynnette R Ferguson[71], Eleonora A Festen[72], Philip Fleshner[20], Tim Florin[73], Denis Franchimont[36,37], Andre Franke[4], Karin Fransen[74], Richard Gearry[46,75], Michel Georges[76,77], Christian Gieger[78], Jürgen Glas[50], Philippe Goyette[1], Todd Green[8,69], Anne M Griffiths[79], Stephen L Guthery[80], Hakon Hakonarson[44], Jonas Halfvarson[81,82], Katherine Hanigan[58], Talin Haritunians[20], Ailsa Hart[83], Chris Hawkey[84], Nicholas K Hayward[85], Matija Hedl[32], Paul Henderson[86,87], Xinli Hu[88], Hailiang Huang[7,8], Ken Y Hui[55], Marcin Imielinski[44], Andrew Ippoliti[20], Laimas Jonaitis[89], Luke Jostins[5,6], Tom H Karlsen[26–28], Nicholas A Kennedy[90], Mohammed Azam Khan[91,92], Gediminas Kiudelis[89], Subra Kugathasan[93], Limas Kupcinskas[94], Anna Latiano[11], Debby Laukens[62], Ian C Lawrance[54], James C Lee[95], Charlie W Lees[90], Marcis Leja[96], Johan Van Limbergen[79], Paolo Lionetti[97], Jimmy Z Liu[19], Edouard Louis[98], Gillian Mahy[99], John Mansfield[100], Dunecan Massey[95], Christopher G Mathew[101,102], Dermot P B McGovern[20], Raquel Milgrom[103], Mitja Mitrovic[74,104], Grant W Montgomery[85], Craig Mowat[105], William Newman[91,92], Aylwin Ng[38,106], Siew C Ng[107], Sok Meng Evelyn Ng[32], Susanna Nikolaus[48], Kaida Ning[32], Markus Nöthen[108], Ioannis Oikonomou[32], Orazio Palmieri[11], Miles Parkes[95], Anne Phillips[105], Cyriel Y Ponsioen[109], Urõs Potocnik[104,110], Natalie J Prescott[101,102], Deborah D Proctor[32], Graham Radford-Smith[58,111], Jean-Francois Rahier[112], Soumya Raychaudhuri[88], Miguel Regueiro[17], Florian Rieder[33], John D Rioux[1,29], Stephan Ripke[7,8], Rebecca Roberts[46], Richard K Russell[86], Jeremy D Sanderson[113], Miquel Sans[114], Jack Satsangi[90], Eric E Schadt[57], Stefan Schreiber[4,48], L Philip Schumm[21], Regan Scott[17], Mark Seielstad[115,116], Yashoda Sharma[32], Mark S Silverberg[103], Lisa A Simms[58], Jurgita Skieceviciene[89], Sarah L Spain[102], A Hillary Steinhart[103], Joanne M Stempak[103], Laura Stronati[117], Jurgita Sventoraityte[94], Stephan R Targan[20], Kirstin M Taylor[113], Anje ter Velde[109], Emilie Theatre[76,77], Leif Torkvist[118], Mark Tremelling[119], Andrea van der Meulen[120], Suzanne van Sommeren[72], Eric Vasiliauskas[20], Severine Vermeire[56,121], Hein W Verspaget[120], Thomas Walters[79,122], Kai Wang[44], Ming-Hsi Wang[33,47], Rinse K Weersma[72], Zhi Wei[123], David Whiteman[85], Cisca Wijmenga[74], David C Wilson[86,87], Juliane Winkelmann[124,125], Ramnik J Xavier[8,38], Sebastian Zeissig[48], Bin Zhang[57], Clarence K Zhang[126], Hu Zhang[127,128], Wei Zhang[32], Hongyu Zhao[126], Zhen Z Zhao[85], Australia and New Zealand IBDGC[129], Belgium IBD Genetics Consortium[129], Italian Group for IBD Genetic Consortium[129], NIDDK Inflammatory Bowel Disease Genetics Consortium[129], United Kingdom IBDGC[129], Wellcome Trust Case Control Consortium[129] & Quebec IBD Genetics Consortium[129]

[32]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA. [33]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA. [34]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA. [35]Peninsula College of Medicine and Dentistry, Exeter, UK. [36]Department of Gastroenterology, Erasmus Hospital, Brussels, Belgium. [37]Department of Gastroenterology, Free University of Brussels, Brussels, Belgium. [38]Gastroenterology Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [39]Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [40]Medical Department, Viborg Regional Hospital, Viborg, Denmark. [41]Organ Center, Hospital of Southern Jutland Aabenraa, Aabenraa, Denmark. [42]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, South Australia, Australia. [43]Department of Gastroenterology, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada. [44]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [45]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, South Australia, Australia. [46]Department of Medicine, University of Otago, Christchurch, New Zealand. [47]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [48]Department for General Internal Medicine, Christian Albrechts University, Kiel, Germany. [49]Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA. [50]Division of Gastroenterology, Royal Victoria Hospital, Montreal, Quebec, Canada. [51]Department of Medicine II, Ludwig Maximilians University Hospital Munich-Grosshadern, Munich, Germany. [52]Department of Gastroenterology, Campus Charité Mitte, Universitatsmedizin Berlin, Berlin, Germany. [53]IBD Unit, Fremantle Hospital, Fremantle, Western Australia, Australia. [54]School of Medicine and Pharmacology, University of Western Australia, Fremantle, Western Australia, Australia. [55]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA. [56]Department of Clinical and Experimental Medicine, Translational Research in Gastrointestinal Disorders (TARGID), Katholieke Universiteit (KU) Leuven, Leuven, Belgium. [57]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [58]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [59]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. [60]IBD Center, Department of Gastroenterology, Istituto Clinico Humanitas, Milan, Italy. [61]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands. [62]Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium. [63]Center of Hepatology, Gastroenterology and Dietetics, Vilnius University, Vilnius, Lithuania. [64]Pediatric Gastroenterology, Cincinnati Children's Hospital Medical Center,

Cincinnati, Ohio, USA. [65]Department of Gastroenterology, Université Catholique de Louvain (UCL) Cliniques Universitaires Saint-Luc, Brussels, Belgium. [66]Division of Gastroenterology, University Hospital Padua, Padua, Italy. [67]Department of Pediatrics, Cedars-Sinai Medical Center, Los Angeles, California, USA. [68]Department of Gastroenterology, Torbay Hospital, Torbay, UK. [69]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [70]Pediatrics, Harvard Medical School, Boston, Massachusetts, USA. [71]Faculty of Medical and Health Sciences, School of Medical Sciences, The University of Auckland, Auckland, New Zealand. [72]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, the Netherlands. [73]Department of Gastroenterology, Mater Health Services, Brisbane, Queensland, Australia. [74]Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands. [75]Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand. [76]Unit of Animal Genomics, GIGA-R (Groupe Interdisciplinaire de Génoprotéomique Appliquée) Research Center, University of Liege, Liege, Belgium. [77]Faculty of Veterinary Medicine, University of Liege, Liege, Belgium. [78]Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [79]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada. [80]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah, USA. [81]Department of Medicine, Örebro University Hospital, Örebro, Sweden. [82]School of Health and Medical Sciences, Örebro University, Örebro, Sweden. [83]Department of Medicine, St Mark's Hospital, Harrow, UK. [84]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK. [85]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [86]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK. [87]Child Life and Health, University of Edinburgh, Edinburgh, UK. [88]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [89]Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania. [90]Gastrointestinal Unit, Western General Hospital, University of Edinburgh, Edinburgh, UK. [91]Genetic Medicine, Manchester Academic Health Science Centre, Manchester, UK. [92]Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK. [93]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA. [94]Department of Gastroenterology, Kaunas University of Medicine, Kaunas, Lithuania. [95]Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK. [96]Faculty of Medicine, University of Latvia, Riga, Latvia. [97]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, Florence, Italy. [98]Division of Gastroenterology, Centre Hospitalier Universitaire (CHU) de Liège, Liege, Belgium. [99]Department of Gastroenterology, Townsville Hospital, Townsville, Queensland, Australia. [100]Institute of Human Genetics, Newcastle University, Newcastle-upon-Tyne, UK. [101]Department of Medical and Molecular Genetics, Guy's Hospital, London, UK. [102]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK. [103]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada. [104]Center for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia. [105]Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK. [106]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [107]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong. [108]Department of Genomics, Life & Brain Center, University Hospital Bonn, Bonn, Germany. [109]Department of Gastroenterology, Academic Medical Center, Amsterdam, the Netherlands. [110]Faculty for Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia. [111]Department of Gastroenterology, Royal Brisbane and Womens Hospital, Brisbane, Queensland, Australia. [112]Department of Gastroenterology, Université Catholique de Louvain (UCL), Centre Hospitalier Universitaire (CHU) Mont-Godinne, Mont-Godinne, Belgium. [113]Department of Gastroenterology, Guy's and St Thomas' National Health Service (NHS) Foundation Trust, St Thomas' Hospital, London, UK. [114]Department of Digestive Diseases, Hospital Quirón Teknon, Barcelona, Spain. [115]Human Genetics, Genome Institute of Singapore, Singapore. [116]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. [117]Department of Biology of Radiations and Human Health, Agenzia Nazionale per le Nuove Tecnologie l'Energia e lo Sviluppo Economico Sostenibile (ENEA), Rome, Italy. [118]Department of Clinical Science Intervention and Technology, Karolinska Institutet, Stockholm, Sweden. [119]Gastroenterology and General Medicine, Norfolk and Norwich University Hospital, Norwich, UK. [120]Department of Gastroenterology, Leiden University Medical Center, Leiden, the Netherlands. [121]Division of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium. [122]Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [123]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA. [124]Institute of Human Genetics, Technische Universität München, Munich, Germany. [125]Department of Neurology, Technische Universität München, Munich, Germany. [126]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, USA. [127]Department of Gastroenterology, West China Hospital, Chengdu, China. [128]State Key Laboratory of Biotherapy, Sichuan University West China University of Medical Sciences (WCUMS), Chengdu, China. [129]A list of members and affiliations appears in the **Supplementary Note**.

ONLINE METHODS

Genotype data set. The cohorts used in the current study were collected from 15 countries across Europe, North America and Australia and have previously been described[4]. In total, 19,802 Crohn's disease cases, 14,864 ulcerative colitis cases and 34,872 controls of European ancestries from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) were included in the study.

Genotyping of the IIBDGC cohorts was performed in 34 different batches across 11 different genotyping centers, and additional genotyping data for 5,815 controls were obtained from the International Multiple Sclerosis Genetics Consortium (IMSGC)[17]. All participating centers received approval from their local and national institutional review boards, and informed consent was obtained from all participants in the study. All DNA samples included in the study were genotyped using the Immunochip custom genotyping array (Illumina)[4,18].

Genotype calling and quality control. After initial genotype calling using Illumina BeadStudio software and a first stage of quality control, all data were centrally recalled using optiCall (v.0.6.2)[19]. OptiCall clustering was performed for each batch separately, with a Hardy-Weinberg equilibrium P-value threshold of 1×10^{-15}, Hardy-Weinberg equilibrium blanking disabled and a genotype call threshold of 0.7. Hardy-Weinberg equilibrium was calculated with conditioning on predicted ancestry, and related individuals were removed from this calculation.

After recall, a single unified quality control procedure was performed across all genotyping batches, including the IMSGC controls. Variants that failed the Hardy-Weinberg equilibrium test in unaffected individuals, had different missing genotype rates in affected and unaffected individuals, or had significantly different allele frequencies across the batches with a false discovery rate (FDR) threshold of 1×10^{-5} for each test were removed. We also removed variants that had a missing genotype rate of >2% across the entire collection or >10% in any single batch. Variants that only failed one quality control criterion in a single batch were set to missing in the failed batch. Individuals were removed if they showed a missing genotype rate of >2%, had a significantly higher or lower inbred coefficient (F) (PLINK)[20] at FDR < 0.01 or showed a high level of relation (PI_HAT ≥ 0.4) calculated on the basis of the identity-by-state (IBS) distance between all individuals. Coefficients of inbreeding and inter-sample relationships were calculated on an LD-pruned data set of independent variants.

To control for population stratification while avoiding possible bias introduced by the enrichment of associated alleles in the data set, principal components were computed on control samples, on the basis of a set of 18,123 independent (LD-pruned) SNPs across the Immunochip, and then applied to the affected samples. To generate the LD-pruned SNPs, we removed variants in long-range LD and pruned the common variants (minor allele frequency (MAF) > 0.05) three times (PLINK)[20]. The genomic inflation factor (λ) was estimated from a set of 3,120 'null' SNPs (chosen on the basis of GWAS of schizophrenia, psychosis, and reading and mathematics ability), using a different subset of principal components. On the basis of these SNPs and investigation of the contribution of individual SNPs to the components (loadings), we chose to use the first five principal components to control for population stratification (**Supplementary Fig. 10**).

For the purpose of this study, the chromosome 6 region at 25–34 Mb, encompassing the MHC region, was extracted from the post–quality control Immunochip data set. In total, after quality control, 18,405 Crohn's disease cases, 14,308 ulcerative colitis cases and 34,241 healthy controls were successfully genotyped for 8,001 SNPs within the MHC region.

Imputation of missing genotype data. Missing genotype data, from failed genotype calls or failed quality control in single batches, were imputed using the Beagle SNP imputation package (v.3.0.4); imputation was performed using only the high-density information contained within the data set, and no external reference data set was used[21].

Imputation of HLA alleles. To avoid cohort-specific asymmetry in the data set, variants that failed quality control in only one genotyping batch were removed before the imputation of HLA alleles. Imputation of HLA alleles was performed using two independent HLA imputation pipelines, HLA*IMP2 (ref. 22) and SNP2HLA (v2)[23]; imputation of polymorphic amino acid positions and SNP variants was performed using SNP2HLA (v2). A set of additional SNP variants was also included using version 1 of SNP2HLA, which used a different reference data set and imputed SNP variants not found in the SNP2HLA (v2) reference data set.

Imputation accuracy. To benchmark the HLA imputation results generated by HLA*IMP2 and SNP2HLA, we used two cohorts from the current study (Italian and Norwegian) for which classical HLA typing was available. The Italian data set consisted of 450 ulcerative colitis cases and 280 controls for which 4-digit HLA types at *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1* were generated by sequence-based typing (SBT)[24]. The Norwegian data set contained 244 ulcerative colitis cases and 254 controls with 2-digit HLA types and a subset of 92 cases and 250 controls with 4-digit HLA types at *HLA-DRB1* generated at the Oslo University Hospital. We only considered individuals for whom both HLA alleles were successfully typed at two- and four-digit resolution at the locus under validation. For validation of HLA allele imputations, we compared the imputation results from HLA*IMP2 and SNP2HLA to the laboratory-derived types in a locus-specific and allele-specific manner. We calculated per-locus concordance (**Supplementary Table 10**) and sensitivity, specificity, positive predictive value, negative predictive value and accuracy for each allele (**Supplementary Table 11**).

Let us denote true positives, true negatives, false positives and false negatives by TP, TN, FP and FN, respectively. Given that our analyses were performed using the expected allele doses from posterior probabilities, we used the expected doses for computation:

$$\mathrm{TP}_i = \min(x_i, y_i)$$

$$\mathrm{FP}_i = y_i - \mathrm{TP}_i$$

$$\mathrm{TN}_i = \min(2 - x_i, 2 - y_i)$$

$$\mathrm{FN}_i = (2 - y_i) - \mathrm{TN}_i$$

where x_i and y_i are, respectively, the typed and imputed doses for individual i and the sum over all individuals gives the TP, FP, TN and FN values (**Supplementary Table 11**). For a given HLA allele, we calculated sensitivity, specificity, positive predictive value, negative predictive value and accuracy using the usual definitions:

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

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

$$\text{positive predictive value (PPV)} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$

$$\text{negative predictive value (NPV)} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FN})$$

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

For a given locus, we calculated the concordance as $\mathrm{TP}_{\mathrm{all}}$/total number of chromosomes (**Supplementary Table 10**).

Final data set. As an additional quality control step, we performed manual cluster inspection for any genotyped SNPs in the region tagging ($r^2 > 0.8$) imputed SNPs, amino acid variants and HLA alleles reported in our proposed models for Crohn's disease and ulcerative colitis.

To not duplicate HLA alleles in our data set and given the mostly equivalent imputation quality of the two pipelines used, we opted to keep HLA imputations for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* derived from the SNP2HLA pipeline because information on amino acid variants and additional SNPs was also obtained from this pipeline, whereas the HLA allele information for the *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* genes was obtained from the HLA*IMP2 pipeline.

Genotyped SNP variants, as well as imputed SNPs, HLA alleles and amino acid variants at polymorphic amino acid positions were combined into a single data set for analysis. Variants with a MAF of less than 0.05% in controls and variants showing an imputation quality (INFO) score of <0.5 were removed

doi:10.1038/ng.3176

from the analyses. The final data set contained 8,939 SNP variants, 138 HLA alleles at 4-digit resolution, 90 HLA alleles at 2-digit resolution and 741 single–amino acid variants.

Heterogeneity. To evaluate the heterogeneity of effects and allele frequencies between subgroups of different European ancestry, we clustered the individuals into relatively large population subgroups. These clusters were determined using *k*-means clustering on the principal components. On the basis of the decline in the within-cluster sum of squared distances, we determined that the optimal choice for the number of clusters (*k*) was in a range of 6–10. For each value of *k* in this range, we compared the clustering obtained to self-reported country of origin, when available. The objective was to be able to identify known population structures, while keeping a homogeneous group in a single cluster. On the basis of these criteria, we stratified the data set into nine clusters (**Supplementary Fig. 11**). For every reported variant, we evaluated and illustrated the heterogeneity of effect sizes between these nine clusters using forest plots (data on heterogeneity of effects are available; see URLs).

Association testing and conditional analyses. Unless otherwise stated, all analyses were corrected for five principal components and performed in R (v 2.15.2) on expected allele counts (additive dose from posterior probability) (see **Supplementary Tables 4**, **12** and **13** for primary univariate association results for single–amino acid variants, SNPs and HLA alleles, respectively). Four-digit HLA alleles were prioritized over two-digit alleles in the final selection of the association signals included in our models. We calculated the threshold for statistical significance in our study as $P < 5 \times 10^{-6}$ for a study-wide type 1 error rate of 5% with 9,852 independent tests.

Given the imputed nature of the HLA allele data within our data set, the complexity of signals and the burden of dimensionality, a standard forward conditional logistic regression approach was avoided. We opted instead to identify all HLA alleles showing study-wide significant association across the MHC region, in the primary univariate association data set, and evaluate the independent effects of these alleles through single pairwise reciprocal conditional logistic regression (**Supplementary Tables 2** and **3**). This approach allowed the identification of independent association signals, composed either of single HLA alleles or groups of equivalent alleles (**Fig. 3**).

Gene-based analyses (omnibus). We tested the association with phenotypes at each HLA gene using logistic regression under an additive model of effect, including all four-digit alleles with a frequency greater than 0.5%. Evidence of association at the gene was given as the *P* value of the likelihood test for this regression model versus the null model (including only principal components). Evidence of association at a given gene, conditional on other HLA genes, was given as the *P* value of the likelihood ratio test for the full model (including all HLA alleles) versus the partial model (not including the alleles at the given gene). In a similar fashion, we also tested the remaining association at each HLA gene, conditional on our final set of associated alleles (final model). We note that the different HLA genes had a different number of distinct alleles; thus, there are variable levels of complexity (degrees of freedom) for the model and the interpretation of these gene-based tests are not straightforward in the context of the likely presence of multiple causal alleles (**Supplementary Note**).

Variance explained (pseudo R^2). To represent the importance of genetic variation in the MHC region for Crohn's disease and ulcerative colitis, we computed an estimate of the variance explained for different models. Variance explained is not well defined for binary outcomes, and many different metrics exist to represent it[25]. Given the correlation between the variants in the MHC region, we computed McKelvey and Zavoina's pseudo R^2 on the logit scale[26]. This metric can be computed using correlated variables and is independent of disease prevalence. Let β be the vector of fitted coefficients from logistic regression and *S* be the estimated covariance matrix of the predictor, then the McKelvey and Zavoina's pseudo R^2 is given by:

$$R^2_{\mathrm{MZ}} = \frac{\beta' S \beta}{\beta' S \beta + \pi^{2/3}}$$

We note that this estimation of variance explained should not be directly compared to values given by other metrics or to heritability estimates based on Gaussian liability.

We computed the variance explained by a regression model including all the HLA alleles with a frequency greater than 0.5%. We did the same separately for HLA alleles within the class I and class II loci. The variance explained by class I alleles, after the inclusion of class II alleles, was computed as the improvement in R^2 for the full model (all HLA alleles) in comparison to only class II alleles. This difference in R^2 estimates the specific contribution of class I alleles to the total variance explained that cannot be attributed to class II alleles. We did the same for class II alleles. To be able to compare our results to the SNPs identified by the GWAS[4], we computed R^2 for the published GWAS index SNPs in the MHC region for Crohn's disease and ulcerative colitis.

Subphenotype analyses. The IIBDGC has collected detailed subphenotype information for a subset of samples included in this study. These phenotypes include demographics, disease location and behavior in Crohn's disease, disease extent in ulcerative colitis, surgery and PSC. Quality control of this subphenotype information and genotype-phenotype association testing has been performed in the context of another project from the IIBDGC (C.W. Lees, personal communication). In the context of this fine-mapping project, we considered disease location in Crohn's disease and PSC in ulcerative colitis to evaluate the impact of disease heterogeneity. Association tests were performed within the subset of samples with a known subphenotype.

Non-additive effects, heterozygote advantage and overdominance. We tested for evidence of non-additive effects at each variant using logistic regression including terms for additive and non-additive effects, as described below. Rare variants with MAFs below 5% were excluded from this analysis, given that the very low number of homozygotes precludes evaluation of the model. For this particular analysis, we used best guess genotype data, unless otherwise stated. We computed evidence of non-additive effect as the *P* value of the Wald statistic for the dominance term. We also computed the evidence of association for an allele under the general model using the likelihood ratio test.

Suppose a genetic variant with alleles G and g. The genotypes (GG, Gg, gg) can be coded as $u = (1, 0, -1)$ and $v = (0, 1, 0)$, respectively. In this context, $u = \text{dose} - 1$ and $v = 1 - |u|$. The effect of a specific genotype is then given as $au + dv$, where *a* and *d* are, respectively, the additive and dominance effects, as estimated by logistic regression. This parameterization can be generalized to expected allele counts (additive dose from posterior probability).

Under this parameterization, the effects of the genotypes GG, Gg and gg are given by *a*, *d* and $-a$, respectively. A strictly additive model would have $d = 0$ and $a \neq 0$, whereas a dominant or recessive model would have $d = a$ or $d = -a$. If the dominance term has a higher protective effect than the additive term, that is, if $d < -|a|$, the model is one of overdominance, where being a heterozygote provides protection in comparison to both homozygotes.

For each HLA gene, we also tested for evidence of heterozygote advantage. We coded each individual as a homozygote or heterozygote at the gene, as determined from the imputed two-digit and four-digit alleles. Association of phenotype and zygosity at each gene was tested using logistic regression. Evidence of association was given as the *P* value of the likelihood test.

Pairwise comparisons of common two-digit alleles at *HLA-DRB1* were conducted to better understand non-additive effects (**Supplementary Table 9**). For each pair of alleles, analysis was constrained to the subset of individuals carrying only these two alleles as homozygotes or heterozygotes. Genotype effects were evaluated within each subset. Overdominance was tested as the effect observed in heterozygotes versus the homozygotes with the lowest risk. Heterozygote advantage was tested as the effect observed in heterozygotes versus the pooled homozygotes.

Comparative structural modeling of HLA-DR molecules. Comparative structure models for all HLA-DR molecules associated with ulcerative colitis or Crohn's disease and for all common HLA-DR molecules (population frequency > 1%) were generated using the program MODELLER[27]. Templates for comparative structure modeling were identified by querying the RCSB Protein Database[28] using the sequence of HLA-DRB1*01:01 and the

DELTA-BLAST algorithm for humans (Taxonomy ID 9606, *E*-value threshold of 0.05). Of 36 templates identified, 10 crystallographically resolved HLA-DRβ1 structures (Protein Data Bank (PDB): 4MD5, 1FV1, 1KLU, 3PDO, 1D5M, 1JH8, 4MDI, 3L6F, 4I5B and 2IPK) were retained, which had high resolution (<2.5 Å) and favorable markers of structural quality (Ramachandran plot, DOPE, Verfiy3D and WHAT_CHECK scores)[29–32]. The sequences for the extracellular domains of target HLA molecules were retrieved from the European Bioinformatics Institute (EBI) sequence database (see URLs), aligned using ClustalW2 with the BLOSUM matrix and Neighbor-Joining clustering algorithm and manually adjusted as indicated[33]. The modeled peptide was standardized to an alanine 12-mer, which was removed before calculating protein electrostatics to allow comparison of the electrostatic potentials within the peptide-binding groove.

Electrostatic potential calculations. Atom charges and radii were assigned and side chains were protonated for pH 7.4 using the PARSE force field in PDB2PQR[34]. Protein electrostatic potential was calculated by solving the linearized Poisson-Boltzmann equation in APBS for a cubic grid of 353 points at a spacing of 0.33 Å (ref. 35). Other parameters were set as follows: ionic solution of 0.15 M univalent positive and negative ions; protein dielectric of 2; solvent dielectric of 78; temperature of 310 K; and a probe radius of 1.4 Å. HLA class II molecules typically make contact with nine amino acids of the presented peptide; of these, seven peptide residues (at peptide positions 1, 2, 3, 4, 6, 7 and 9) make contact within the peptide-binding groove[36]. Peptide residues at positions 5 and 8 are elevated away from the peptide-binding groove. For comparison of the electrostatic potentials of the peptide-binding groove, radii were chosen so that the space around each coordinate encompassed all side-chain atoms of the relevant peptide residue; the electrostatic potential within 1 Å of the molecular surface of the HLA molecule was not examined.

The peptides of the template HLA structures were used to define the geometric average coordinates of the positions of the side-chain atoms of peptide residues 1, 2, 3, 4, 6, 7 and 9. The electrostatic potential within a 3.5- to 5-Å radius for each coordinate was considered for comparison of the electrostatic properties of the peptide-binding groove among HLA-DR molecules (**Supplementary Fig. 12**). For comparison of the electrostatic potential around amino acid positions 67, 70 and 71 (associated with both ulcerative colitis and Crohn's disease in different analyses; **Supplementary Tables 4–6**), the geometric average coordinates of the positions of the side-chain atoms of these amino acids were calculated and the electrostatic potential within a 5-Å radius of these coordinates was considered for comparisons among different molecules. Comparisons of electrostatic potential were performed using the Hodgkin's index as described previously[27,28] in a pairwise, all-versus-all fashion to produce a distance matrix[37,38]. Distance matrices were displayed as a symmetrical heat map with reordering such that electrostatically similar molecules were clustered together, according to the dendrogram (**Supplementary Fig. 13**). A pooled heat map was created using the Euclidian distances between the individual distance matrices from the seven peptide-binding groove regions (**Fig. 4a**). Heat maps were created for electrostatic potential comparisons at individual regions of the HLA-DR molecule (seven regions in the peptide-binding groove and one region defined by residues 67, 70 and 71, as detailed above) (**Fig. 4b**).

17. International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
18. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
19. Shah, T.S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28**, 1598–1603 (2012).
20. 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).
21. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
22. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput. Biol.* **9**, e1002877 (2013).
23. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8**, e64683 (2013).
24. Gourraud, P.A. *et al.* HLA diversity in the 1000 Genomes dataset. *PLoS ONE* **9**, e97282 (2014).
25. Veall, M.R. & Zimmermann, K.F. Pseudo-R^2 measures for some common limited dependent variable models. *J. Econ. Surv.* **10**, 241–259 (1996).
26. McKelvey, R.D. & Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **4**, 103–120 (1975).
27. Eswar, N. *et al.* Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* Chapter 2, Unit 2.9 (2007).
28. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
29. Hooft, R.W., Vriend, G., Sander, C. & Abola, E.E. Errors in protein structures. *Nature* **381**, 272 (1996).
30. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
31. Lüthy, R., Bowie, J.U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
32. Shen, M.Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
33. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
34. Dolinsky, T.J. *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522–W525 (2007).
35. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. & McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA* **98**, 10037–10041 (2001).
36. Jones, E.Y., Fugger, L., Strominger, J.L. & Siebold, C. MHC class II proteins and disease: a structural perspective. *Nat. Rev. Immunol.* **6**, 271–282 (2006).
37. Richter, S., Wenzel, A., Stein, M., Gabdoulline, R.R. & Wade, R.C. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res.* **36**, W276–W280 (2008).
38. Wade, R.C., Gabdoulline, R.R. & De Rienzo, F. Protein interaction property similarity analysis. *Int. J. Quantum Chem.* **83**, 122–127 (2001).

doi:10.1038/ng.3176