Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations

Jimmy Z Liu[#1], **Suzanne van Sommeren**[#2,3], **Hailiang Huang**[4], **Siew C Ng**[5], **Rudi Alberts**[2], **Atsushi Takahashi**[6], **Stephan Ripke**[4], **James C Lee**[7], **Luke Jostins**[8], **Tejas Shah**[1], **Shifteh Abedian**[9], **Jae Hee Cheon**[10], **Judy Cho**[11], **Naser E Dayani**[12], **Lude Franke**[3], **Yuta Fuyuno**[13], **Ailsa Hart**[14], **Ramesh C Juyal**[15], **Garima Juyal**[16], **Won Ho Kim**[10], **Andrew P Morris**[17], **Hossein Poustchi**[9], **William G Newman**[18], **Vandana Midha**[19], **Timothy R Orchard**[20], **Homayon Vahedi**[9], **Ajit Sood**[19], **Joseph Y Sung**[5], **Reza Malekzadeh**[9], **Harm-Jan Westra**[3], **Keiko Yamazaki**[13], **Suk-Kyun Yang**[21], **The International Multiple Sclerosis Genetics Consortium**[22], **The International IBD Genetics Consortium**[22], **Jeffrey C Barrett**[1], **Behrooz Z Alizadeh**[23], **Miles Parkes**[7], **Thelma BK**[16], **Mark J Daly**[4], **Michiaki Kubo**[13,25], **Carl A Anderson**[1,25], and **Rinse K Weersma**[2,25]

[1]Wellcome Trust Sanger Institute, Hinxton, UK [2]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, The Netherlands [3]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands [4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA [5]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong [6]Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, Riken, Yokohama, Japan [7]Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK [8]Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK [9]Digestive Disease Research Institute, Shariati Hospital, Tehran, Iran [10]Yonsei University College of Medicine, Seoul, Korea [11]Icahn School of Medicine, Mount Sinai New York, New York, USA [12]Department of Gastroenterology, Emam Hospital, Tehran, Iran [13]Laboratory for Genotyping Development, Center for Integrative Medical Sciences, Riken, Yokohama, Japan [14]IBD Unit, St Mark's Hospital, Harrow, Middlesex, UK [15]National Institute of Immunology, Aruna Asaf Ali Road, New Delhi, India [16]Department of Genetics, University of Delhi South Campus, New Delhi, India [17]Welcome Trust Center for Human Genetics, Oxford U.K. and Department of Biostatistics, University of Liverpool, Liverpool, UK [18]Manchester

Centre for Genomic Medicine, University of Manchester and Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK [19]Department of Medicine, Dayanand Medical College and Hospital, Ludhiana, India [20]St Mary's Hospital, London, UK [21]Asan Medical Center, University of Ulsan College Medicine, Seoul, Korea [22]A list of members and affiliations appears in the Supplementary Note [23]Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands

[#] These authors contributed equally to this work.

Abstract

Ulcerative colitis and Crohn's disease are the two main forms of inflammatory bowel disease (IBD). Here, we report the first trans-ethnic association study of IBD, with genome-wide or Immunochip genotype data from an extended cohort of 86,640 European individuals and Immunochip data from 9,846 individuals of East-Asian, Indian or Iranian descent. We implicate 38 loci in IBD risk for the first time. For the majority of IBD risk loci, the direction and magnitude of effect is consistent in European and non-European cohorts. Nevertheless, we observe genetic heterogeneity between divergent populations at several established risk loci driven by a combination of differences in allele frequencies (*NOD2*), effect sizes (*TNFSF15*, *ATG16L1*) or a combination of both (*IL23R*, *IRGM*). Our results provide biological insights into the pathogenesis of IBD, and demonstrate the utility of trans-ethnic association studies for mapping complex disease loci and understanding genetic architecture across diverse populations.

Inflammatory bowel diseases (IBD) are chronic, relapsing intestinal inflammatory diseases affecting more than 2.5 million people in Europe, with increasing prevalence in Asia and developing countries[1,2]. IBD is thought to arise from an inappropriate activation of the intestinal mucosal immune system in response to commensal bacteria in a genetically susceptible host.

To date, 163 genetic loci have been associated with IBD via large-scale genome-wide association studies (GWAS) in cohorts of European descent. Smaller GWAS performed in populations from Japan, India and Korea have reported six novel genome-wide significant associations outside of the HLA region. Three of these loci (13q12, *FCGR2A* and *SLC26A3*) subsequently achieved genome-wide significant evidence of association in European cohorts. The remaining three loci demonstrated consistent direction of effect and nominally significant evidence of association ($P<1\times10^{-4}$) in previous European GWAS studies[3,4,5,6]. A number of loci initially associated with IBD in European cohorts have now also been shown to underlie risk in non-Europeans, including *JAK2*, *IL23R* and *NKX2-3*. The evidence of shared IBD risk loci across diverse populations suggests that combining genotype data across cohorts of different ethnicities will enable the detection of additional IBD associated loci. Such trans-ethnic association studies have successfully identified loci for other complex diseases, including Type-2 Diabetes and Rheumatoid Arthritis [7, 8].

In this study we aggregate genome-wide or Immunochip genotype data from 96,486 individuals. Compared to our previously published GWAS meta-analysis, this study includes

an extra 11,535 individuals of European ancestry and 9,846 individuals of non-European ancestry. Using these data we aim to 1) identify novel IBD risk loci and 2) compare the genetic architecture of IBD susceptibility across ancestrally divergent populations.

Results

Study design

Following quality-control (QC) and 1000 genomes imputation (Phase I - March2012), 5,956 Crohn's disease cases, 6,968 ulcerative colitis cases and 21,770 population controls of European descent were used to perform genome-wide association studies of Crohn's disease, ulcerative colitis or IBD (Crohn's disease and ulcerative colitis) (Online Methods). Replication was undertaken using an additional 16,619 Crohn's disease cases, 13,449 ulcerative colitis cases and 31,766 population controls genotyped on the Immunochip. The replication cohort included 2,025 Crohn's disease cases, 2,770 ulcerative colitis cases and 5,051 population controls of non-European ancestry (Table 1, Supplementary Figure 1 and 2), so principal component analysis was used to assign individuals to one of four ancestral groups (European, Iranian, Indian or East Asian) (Supplementary Figure 3). Case-control association tests were performed within each ancestry group using a linear mixed model (MMM)9 (Online Methods). A fixed-effect meta-analysis was undertaken to combine summary statistics from our European-only GWAS meta-analysis with those from the European replication cohort. We next performed a Bayesian trans-ethnic meta-analysis, as implemented in Mantra, to enable heterogeneity in effect size to be correlated with genetic distance between populations, as estimated by the mean F_{st} across all SNPs10 (Online Methods). For the trans-ethnic meta-analysis, the 6,392 cases and 7,262 population controls of European ancestry that were present in both the GWAS and replication cohorts were excluded from the Immunochip replication study (Supplementary Figure 2). To maximise power for our solely Immunochip-based comparisons across ancestral groups, the mixed model association analysis was repeated after reinstating these individuals to the Immunochip cohort.

Trans-ethnic meta-analysis identifies 38 new IBD loci

In total, 38 new disease associated loci were identified at genome-wide significance in either the association analysis of individual ancestry groups ($P<5\times10^{-8}$) or in the transethnic meta-analysis that included all ancestries (logBF>6) for ulcerative colitis, Crohn's disease or IBD (Table 2, Supplementary Table 2, Supplementary Figures 4-7). To reduce false-positive associations we required all loci only implicated in disease risk via the transethnic meta-analysis (i.e. logBF>6 but $P>5\times10^{-8}$ in each individual ancestry cohort) to show no significant evidence of heterogeneity across all four ancestry groups ($I^2> 85.7$) (Online Methods).

Twenty-five of the 38 newly associated loci overlap with those previously reported for other traits, including immune-mediated diseases, while 13 have not previously been associated to any disease or trait (Supplementary Table 3, Online Methods). A likelihood modeling approach showed that 27 of the 38 novel loci are associated with both Crohn's disease and ulcerative colitis (designated here as IBD loci), with seven of these demonstrating evidence

of heterogeneity of effect between the two diseases. Of the remaining 11 loci, seven were classified as Crohn's disease-specific and four as ulcerative colitis-specific (Table 2, Supplementary Table 2).

As a result of our updated sample QC, seventeen of the 194 independent SNPs reported at genome-wide significance in our previous European-only GWAS meta-analysis6 failed to reach this threshold in the present study. Sixteen of these loci still demonstrated strong suggestive evidence of association in the current European cohort ($5\times10^{-8}<P<8.7\times10^{-6}$, representing a False Discovery Rate of ~0.001) (Supplementary Table 2). SNP rs2226628 on chromosome 11 failed to achieve even suggestive evidence of association in our current European association analysis (P=0.0024). Our previous European-only meta-analysis incorporated a number of principal components as covariates in a logistic regression test of association and, interestingly, if we adopt the approach taken in Jostins & Ripke et al (2012), we observe a more significant P-value of 7.38×10^{-6}. This observation, plus divergent allele frequencies at this SNP across European populations (1000 Genomes release 14: GBR=0.20, CEU=0.28, IBS=0.39, FIN=0.47), suggests the previously reported signal of association may have been driven, at least in part, by population stratification (which is now being better accounted for in the linear mixed model analysis)6. In summary, we now consider 231 independent SNPs within 200 loci to be associated with IBD risk (Supplementary Table 2.

Forty-one of the 163 IBD SNPs originally associated in our previous European-only GWAS meta-analysis replicated in at least one non-European cohort if we consider a one-tailed Bonferroni corrected significance threshold of $P<6.1\times10^{-4}$ (0.05/163) (Supplementary Table 2. Nine of the fourteen non-HLA loci (10 Crohn's disease and 4 ulcerative colitis) that have been identified at genome-wide significant levels in previous non-European GWAS cohorts from Japan, India and South Korea 3, 4, 12, 13, 14 were associated to either Crohn's disease or ulcerative colitis in the East-Asian, Indian and/or Iranian cohorts with a $P<1.0\times10^{-5}$ (Supplementary Table 6). Four of the five remaining SNPs (or reliable proxy SNPs) were not present on the Immunochip. The previously reported association at rs2108225 (*SLC26A3*) on chromosome 7 showed an association signal of $P=2.64\times10^{-3}$ in the current East Asian cohort but is strongly associated to European IBD ($P=1.04\times10^{-18}$).

We next performed a series of analyses to prioritize genes within newly-associated loci for causality. *cis*-eQTL analysis from two datasets totalling peripheral blood samples of 1240 individuals revealed that 12 of the 38 newly-associated SNPs have *cis*-eQTL effects (False Discovery Rate <0.05) (Online Methods - Supplementary Table 7). Two SNPs showed *trans*-eQTL effects; SNP rs653178 in a locus harbouring *SH2B3* and *ATXN2* is associated to multiple other immune-mediated diseases, including celiac disease and rheumatoid arthritis. It has a *trans*-eQTL effects on 14 genes, including genes within IBD associated loci (*TAGAP*, *STAT1*). rs616597 has a *cis*-eQTL effect on *NFKBIZ* and has *trans*-eQTL effects on *FXL13*, *ALPL*, *HSQP1L* and *PDHX* (Supplementary Table 7)15. Both SNPs reside in known DNase1 hypersensitivity and histone modification sites in multiple cell lines (Supplementary Table 8). In contrast to the high number of SNPs tagging eQTLs, only three of the 38 SNPs were in high linkage disequilibrium (LD, $r^2>0.8$) with known missense coding variants (Supplementary Table 9).

To enable a meaningful comparison with our previously published results, we re-created the GRAIL connectivity network including all loci that now acheive genome-wide signficant evidence of association (Supplementary Figure 8). Twelve genes in the previous GRAIL network were removed in this new network. We found these genes had significantly larger GRAIL p-values (Wilcoxon P-value = 6×10^{-4}) and fewer interaction partners (11.2 vs. 16.0) than genes remaining in the network. Sixty two genes were connected into the GRAIL network for the first time, only 36 of which are located within the newly associated loci (including *NFKBIZ*, *CD28* and *OSMR*). Thus, 26 genes from previously established IBD loci are brought into the network for the first time, 12 of which are the only GRAIL gene reported for their loci, including *TAGAP* and *IKZF1*. Genes within the 16 previously associated loci that failed to reach genome-wide significance in our current study have similar average connectivities as other genes in the network (17.8 vs 16.4 respectively, Wilcoxon p-value=0.94), thus further supporting their likely involvement in IBD risk. 37 out of 56 DAPPLE candidate genes were identified as candidates in the GRAIL analysis (Supplementary Table 10).

Biological implications of newly associated IBD loci

Previous GWAS studies have highlighted components in several key pathways underlying IBD susceptibility, many involved in innate immunity, T cell signaling and epithelial barrier function. Accepting the need for fine mapping to pinpoint causal variants within the newly identified loci, the current study expands the range of pathways implicated.

The process of autophagy, which is an intracellular process during which cytoplasmic content is engulfed by double-membrane autophagosomes and delivered to the vacuole or lysosome for degradation and recycling, has been implicated in Crohn's disease pathogenesis since the identification of *ATG16L1* and *IRGM* as Crohn's disease susceptibility genes. The newly identified Crohn's disease gene *ATG4B* is a cysteine protease with a central role in this process, reinforcing the importance of autophagy in Crohn's disease pathogenesis. Likewise, the importance of epithelial barrier function in IBD pathogenesis (previously highlighted by associations with *LAMB1* and *HNF4a*16) is underscored by the new association at *OSMR,* which modulates a barrier-protective host response in intestinal inflammation.

Many of the newly identified candidate genes including *LY75, CD28*, *CCL20*, *NFKBIZ*, *AHR*, and *NFATC1*, modulate specific aspects of the T cell response. Thus, beyond the involvement of Th17 cells (previously identified through associations with e.g. *IL23R)*, our results now implicate all three components of T cell activation (TCR ligation, co-stimulation, and IL-2 signalling). Importantly, these processes are critical for memory development and are common to both CD4[+] and CD8[+] T-cells.

The function of leading new positional candidate genes is discussed in Box 1. (Box 1 – candidate genes within associated loci).

Comparing non-European IBD versus European IBD

Recent large-scale trans-ethnic genetic studies of complex diseases have shown that the majority of risk loci are shared across divergent populations 8, 17, 18. The true extent of

sharing is difficult to characterize because the sizes of non-European cohorts are often much smaller than their European counterparts, limiting power to detect associated loci. Despite our study including a large cohort of 9,846 non-European samples and being the largest non-European study of IBD, this number is still small in comparison with the European cohort of 86,640 individuals. As such, we expect that the majority of known risk loci will not be associated in the non-European populations at genome-wide significance. Nevertheless, we observed a striking positive correlation in direction of effect when comparing the 231 independently associated SNPs in European and East Asian cohorts, ($P < 1.0 \times 10^{-22}$ for Crohn's disease and $P < 1.0 \times 10^{-31}$ for ulcerative colitis) (Figure 1). Furthermore, of 3,900 suggestively associated SNPs (5×10^{-5} $P < 5 \times 10^{-8}$) from the European-only IBD association analysis, 2,566 have the same direction of effect in the East Asian analysis ($P = 5.92 \times 10^{-88}$). Consistent with the concordant direction of effect at associated SNPs, there was high genetic correlation (r_G) between the European and East Asian cohort when considering the additive effects of all SNPs genotyped on the Immunochip19 (Crohn's disease $r_G = 0.76$, ulcerative colitis $r_G = 0.79$) (Supplementary Table 11). Given that rare SNPs (minor allele frequency (MAF) < 1%) are more likely to be population-specific, these high r_G values also support the notion that the majority of causal variants are common (MAF>5%). Although the Indian and Iranian cohort sizes are small compared to the East Asian cohort we observed similar trends for homogeneity of ORs at associated loci (Supplementary Figure 9 and 10) and high genetic correlation Immunochip-wide (Supplementary Table 11). Together with the strong effect size correlations at known risk loci, these results indicate that the majority IBD risk loci are shared across ancestral populations. Therefore, ancestry matched groups of IBD cases and controls can be combined across divergent populations to amass the large sample sizes needed to detect further disease associated loci.

Not all IBD risk loci are shared across populations, as evidence by r_G being significantly less than 1 ($P < 8.2 \times 10^{-4}$) for all pairwise population comparisons. In most cases, apparent differences in genetic risk are explained by different allele frequencies across populations. For instance, consistent with previous genetic studies of Crohn's disease in East Asians 2, the three coding variants in *NOD2* (nucleotide-binding oligomerisation domain-containing protein 2) that have a large effect on IBD risk in Europeans (ORs = 2.13 to 3.03) have a risk-allele frequency (RAF) of zero in East Asians. Beyond these three coding variants, there is also evidence of at least four additional low-frequency independent *NOD2* variants on the Immunochip that are associated with Crohn's disease in Europeans (HH, personal communication). In the East Asian cohort, two of these had a RAF of zero, while we were not powered to detect association at the other two because we observed less than four copies of the risk allele (MAF < 0.0004). Furthermore, no SNP within NOD2 achieved even suggestive evidence of association in the East Asian cohort (all $P > 7.18 \times 10^{-4}$). Larger sample sizes and a more complete ascertainment of variants (particularly in non-European cohorts) will be required to better assess the genetic architecture of *NOD2* across divergent populations. Similarly, at the *IL23R* (interleukin 23 receptor) gene, previous studies have shown that there is substantial genetic heterogeneity between European and East Asian individuals in IBD risk 2. In line with these observations, the *IL23R* SNP with the largest effect in European Crohn's disease and ulcerative colitis (rs80174646) has a RAF of one in East Asians, while secondary *IL23R* variants observed in Europeans were also not

significantly associated with disease (rs6588248, P = 0.65; rs7517847, P = 0.04). These two secondary variants are common in East Asians (rs6588248, MAF = 0.39; rs7517847, MAF = 0.42) and, assuming the effect sizes observed in Europeans, we have 100% power to detect association to rs7517847 at $P < 5 \times 10^{-8}$ but only 84% power to detect association to rs6588248 at P< 0.05. Therefore, we cannot rule out the possibility that rs6588248 is involved in Crohn's disease susceptibility in East Asia. Both variants show significant heterogeneity of effect between the European and East Asian Crohn's disease cohorts (P < 2.44×10^{-4}). However, *IL23R* clearly plays a role in East Asian IBD, evidenced by the association at rs76418789 with both Crohn's disease and ulcerative colitis in East Asians (IBD P = 1.83×10^{-13}). The same variant was previously implicated in a GWAS of Crohn's disease in Koreans (Supplementary Table 6)[4]. This variant, which has a much lower allele frequency in Europeans (MAF=0.004) than East Asians (MAF=0.07), demonstrates suggestive evidence of association in European IBD (P = 3.99×10^{-6}, OR = 0.66), and becomes genome-wide significant (P = 2.31×10^{-10}, OR = 0.53) after conditioning on the three known European risk variants (rs11209026, rs6588248 and rs7517847).

We were well powered to detect genetic heterogeneity between our East Asian and European cohorts at several alleles of large effect in Europeans. (Figure 2 – Supplementary Figure 10). For example, at *ATG16L1* the reported Crohn's disease risk variant in Europeans (rs12994997) has a RAF of 0.53 and OR of 1.27. The variant shows no evidence of association in East Asians (P = 0.21), driven at least in part by a significant difference in allele frequency (RAF = 0.24, F_{st} = 0.15). However, assuming the effect size at this SNP in the East Asian cohort was equal to that seen in the European cohort, we would still have more than 80% power to detect suggestive evidence of association (P<5×10^{-5}). In addition to differences in allele frequency we also observe evidence of heterogeneity of odds at this SNP (OR_{EA} = 1.06; P = 8.45×10^{-4}). The previously reported lead SNP at the *IRGM* locus in Europeans also shows only nominally significant evidence of association in East Asian Crohn's disease (rs11741861, European P = 5.89×10^{-44}, East Asian P = 2.62×10^{-3}) as well as evidence of heterogeneity of effect (European OR = 1.33 vs. East Asian OR = 1.13; heterogeneity P = 1.20×10^{-3}). However, not all loci demonstrating significant heterogeneity of odds have lower effect in the non-European cohort; Two of the three independent signals at *TNFSF15/TNFSF8* have much larger effect on East Asian IBD risk (rs4246905: OR = 1.15/1.75; rs13300483: OR = 1.14/1.70) despite similar allele frequencies. The third European risk variant was not significantly associated in East Asians (rs11554257, P = 0.21), though this may reflect a lack of power (76% power to detect this variant at P < 0.05 assuming identical ORs).

Although the incidence of IBD is rising in developing countries, comparable data on clinical phenotype of disease in European and non-European populations is limited. We collected sub-phenotype data on 4,686 IBD patients from East Asia, India and Iran and compared this with available clinical phenotypes across 35,128 Europeans. Given the fact that this is the largest cohort available for clinical comparisons between European and no-European IBD we performed basic comparative statistical analyses. Overall our data showed some demographic differences between the European and non-European populations with a male predominance in Crohn's disease (67% of non-European Crohn's disease patients are male compared to 45% in Europeans, P=7.09×10^{-78}). Furthermore we observed more stricturing

behaviour (P=2.02 × 10⁻³³) and perinatal disease (P=5.36 × 10⁻³³) and less inflammatory Crohn's disease (p=4.28 × 10⁻³²) in the non-European population. In ulcerative colitis there was a lower rate of extensive colitis reported in the non-European population (p=1.52 × 10⁻³⁴) which was also reflected in a lower rate of colectomy (p=1.23 × 10⁻⁶⁹) (Supplementary table 12). Although these data have been collected retrospectively the current findings are in line with previously reported prospectively collected clinical findings in incident cases in non-European IBD[2].

Discussion

We identified 38 additional IBD susceptibility loci by adding an extra 11,535 individuals of European descent and 9,846 individuals of non- European descent to our previously reported European-only cohort of 75,105 samples. Given trans-ethnic association studies principally identify risk loci shared across populations, we would expect to identify a similar number of associated loci had all the individuals in this study been of the same ancestry. Our analyses suggest that significant differences in effect size are minimal at all but a handful of associated loci, further indicating that trans-ethnic association studies represent a powerful means of identifying new loci in complex diseases like IBD. Furthermore, the near complete sharing of genetic risk among individuals of diverse ancestry has significant consequences for association studies and disease risk prediction in non-European populations. Firstly, a significant association in one population makes the locus in question a very strong candidate for involvement in IBD risk worldwide. Secondly, our data suggest that ORs estimated from a very large association study are likely to better represent the effect size of the associated variants in a second, ancestrally diverse population, than those estimated from a significantly smaller study in the second population itself (because of the larger sampling variance in the second study). Finally, because rare alleles are more likely than common variants to be population specific, the significant number of IBD risk loci shared across ancestral populations implies that the underlying causal variants at these loci are common. This adds further weight to the growing number of arguments against the 'synthetic association' model explaining a large proportion of GWAS loci [20, 21, 22].

While the majority of risk loci are shared across populations, we were able to detect a handful of loci demonstrating heterogeneity of effect between populations. Major European risk variants in *NOD2* and *IL23R* are not present in individuals of East Asian ancestry. The relatively small sample size of the non-European cohorts, and the fact that Immunochip SNP selection was only based on resquencing data from individuals of European ancestry, hinders our ability to identify association to sites that are monomorphic in Europeans but polymorphic in non-Europeans. Targetted resequencing efforts in large numbers of non-European IBD cases and controls, similar to those undertaken in European cohorts , may identify such associations and thus provide further insight into the genetic architecture of IBD. The much smaller number of individuals in the non-European cohorts also reduces power to detect heterogeneity of effect versus the European cohort and therefore we may be overestimating the degree of sharing between the various ancestry groups.

In addition to allele frequencies differing between ancestral populations, patterns of linkage disequilibrium can also vary greatly; such differences further complicate comparisons of

complex disease genetic architecture across diverse populations. For example, we observed significant heterogeneity of odds at the *TNFSF15*/*TNFSF8* and *ATG16L1* loci, potentially suggesting that gene-environment interactions increase the variance explained by these associations in either European (*ATG16L1*) or non-European (*TNFSF15*/*TNFSF8*) populations. Though this hypothesis is attractive, the heterogeneity of effect size could also be underpinned by differential tagging of untyped causal variants at these loci in one or both populations. Although Immunochip provides dense coverage of 186 previously associated loci, SNP selection was based on low-coverage sequence data from a pilot release of the 1000 genomes project. Approximately 240,000 SNPs were selected for inclusion, with an assay design success rate of approximately 80%. Therefore it is possible that causal variants could remain untyped, even within the dense 'fine-mapping' regions of Immunochip, and the chances of this occuring are greater still in populations of non-European ancestry. Until the causal variants that underlie these associated loci have been identified (or all SNPs within these loci are included in our association tests) we cannot rule out the possibility that differential tagging of untyped causal variants is driving the observed heterogeneity of effect.

In summary, we have performed the first trans-ethnic association study of IBD and identified 38 risk loci, raising the number of known IBD risk loci to 200. Together, these loci explain 13.1% and 8.2% of variance in disease liability in Crohn's disease and ulcerative colitis respectively. The majority of these loci are shared across diverse ancestry groups, with only a handful of demonstrating population specific effects driven by heterogeneity in risk allele frequency (e.g. *NOD2*) or effect size (e.g. *TNFSF15*/*TNFSF8*). Concordance in direction of effect is significantly enriched among SNPs demonstrating only suggestive evidence of association, indicating that larger transethnic association studies represent a powerful means of identifying more IBD risk loci. By leveraging imputation based on tens of thousand of reference haplotypes, or directly sequencing large numbers of cases and controls, these studies will more thoroughly survey causal variants and thus have increased ability to model the genetic architecture of IBD across diverse ancestral populations.

URLs

NHGRI GWAS Catalog http://www.genome.gov/admin/gwascatalog.txt

functionGVS (http://snp.gs.washington.edu/SeattleSeqAnnotation134/

Variant Explorer http://molgenis70.target.rug.nl/index.htm

Online Methods

Ethical approval

The recruitment of study subjects was approved by the ethics committees or institutional review boards of all individual participating centers or countries. Written informed consent was obtained from all study participants.

GWAS cohort, quality control and analysis

Cohorts and quality control—The GWAS cohorts and QC are described in detail in Jostins & Ripke et al. (2012). Briefly, seven Crohn's disease and eight ulcerative colitis collections with genome-wide SNP data were combined. Samples were genotyped on a combination of Affymetrix GeneChip Human Mapping 500K, Affymetrix Genome-Wide Human SNP Array 6.0, Illumina HumanHap300 BeadChip and Illumina HumanHap550 BeadChip arrays. After SNP and sample QC, the Crohn's disease data consisted of 5,956 cases and 14,927 controls, the ulcerative colitis data consisted of 6,968 cases and 20,464 controls, and Crohn's disease+ulcerative colitis combined (IBD) data consisted of 12,882 cases and 21,770 controls. The number of SNPs per collection varied between 290,000 and 780,000.

Imputation—Genotype imputation was performed using the pre-phasing/imputation stepwise approach implemented in IMPUTE2 / SHAPEIT (chunk size of 3 Mb and default parameters 37, 38. The imputation reference set consisted of 2,186 phased haplotypes from the full 1000 Genomes Project dataset (August 2012, 30,069,288 variants, release "v3.macGT1").

Association Analysis—A genome-wide association analyses was carried out for Crohn's Disease, Ulcerative Colitis and inflammatory bowel disease (IBD – Crohn's disease and ulcerative colitis cases combined). After applying MAF > 1% and INFO score > 0.6 filters to all imputed variants, around 9 million variants were found suitable for association analysis. Association tests was carried out in PLINK, using the post-imputation genotype dosage data and using 10, 7 or 15 principal components for Crohn's disease, ulcerative colitis or IBD respectively as covariates, chosen from the first 20 principal components. The Crohn's disease, ulcerative colitis and IBD scans had genomic inflation (LambdaGC) values of 1.129, 1.114, and 1.160 respectively. Accounting for inflation due to sample size and polygenic effects, these Crohn's disease, ulcerative colitis and IBD LambdaGC values are equivalent to LambdaGC1000 (the inflations factor from a sample size of 1000 cases and 1000 controls) 39 values of 1.015, 1.011 and 1.010 respectively.

Immunochip cohort, QC and analysis

Description of Immunochip—The Immunochip is an Illumina Infinium microarray comprising 196,524 SNPs and small indel markers selected based on results from genome-wide association studies of 12 different immune-mediated diseases. The Immunochip enables 1) replication of all nominally associated SNPs (P < 0.001) from the index GWAS scans and 2) fine-mapping of 186 loci associated at genome-wide significance with at least one of the 12 index immune-mediated diseases. Within fine-mapping regions, SNPs from the 1000 genomes project pilot phase 1 (European cohorts), plus selected autoimmune disease resequencing efforts, were selected for inclusion (with a design success rate of around 80%). The chip also contains around 3,000 SNPs added as part of the WTCCC2 project replication phase. These SNPs are useful for QC purposes because they have not previously been associated with immune-mediated diseases ("null" SNPs).

European ancestry cohorts—Recruitment of patients and matched controls genotyped with the Immunochip was performed in 15 countries in Europe, North America and Oceania (Table 1). Diagnosis of IBD was based on accepted radiologic, endoscopic, and histopathologic evaluation. All included cases fulfill clinical criteria for IBD. Genotyping was performed across 36 batches, and included a total of 19,802 Crohn's disease cases, 14,864 ulcerative colitis cases and 34,872 population controls. The Immunochip cohort includes 3,424 Crohn's disease cases, 3,189 ulcerative colitis cases and 7,379 population controls present in the GWAS cohort. The overlapping Immunochip samples were excluded for the trans-ethnic association analysis but included in the modelling of European vs non-European IBD because this was based solely on Immunochip data.

East Asian, Indian and Iranian ancestry cohorts—East Asian IBD patients and controls were recruited from the following countries: Japan (Institute of Medical Science, University of Tokyo, RIKEN Yokohama Institute and Japan Biobank), Korea (Yonsei University College of Medicine and Asan Medical Centre, Seoul), Hong Kong (Chinese University of Hong Kong). Indian IBD cases and controls were recruited from Dayanand Medical College and Hospital, Ludhiana and University of Delhi South Campus. Iranian cases and controls were recruited from the Tehran University of Medical Sciences. Samples recruited as part of a European cohort but who cluster with a non-European cohort in PCA (see below) were reassigned to the non-European cohort. In total, 6,598 East Asian, 3,088 Indian and 1,393 Iranian individuals were genotyped on the Immunochip (Table 1, Supplementary Table 1, Supplementary Figure 1, Supplementary Figure 2).

Phenotype data—Detailed phenotype data (including gender, ethnicity, age of disease onset, smoking status, family history, extraintestinal manifestations and surgery) were available for 47,799 European IBD cases and 3,986 non-European IBD cases (Supplementary Table 12). Disease location and behaviour were assessed with the Montreal classification. Clinical demographics and disease phenotype in the European and Non-european cohorts were compared using chi-square analysis (SPSS 20)

Genotyping and calling—The Immunochip samples were genotyped in 36 batches. Normalized intensities for all samples were centrally called using the optiCall clustering program 40 with Hardy-Weinberg equilibrium blanking disabled and the no-call cutoff set to 0.7. Before calling all data, we first established the optimal composition of sample sets. Calling per genotyping batch turned out to give the most reliable genotype clustering (compared to calling individual ancestral populations separately within each genotyping batch, calling all individuals per ancestry group together or calling all available data together).

Quality Control—Quality control (QC) was performed separately in each population (East Asian, Iranian, Indian and European) using PLINK 41. Individuals were assigned to populations based on principal component analysis (PCA). PCA was performed using EIGENSTRAT 42 on a set of 15,552 Immunochip SNPs that had a pairwise r2 < 0.2, MAF > 0.05, and were present in 1000 Genomes Phase 2 data. The first two principal components were estimated in the 1000 Genomes individuals and projected onto all Immunochip cases

and controls. As expected, a clear separation between the different populations was seen (Supplementary Figure 3). Samples were assigned to the population that they clustered with, and those that did not cluster with any of the reported populations were removed.

Marker QC—SNPs meeting the following criteria were removed: (i) not on autosomes,(ii) call rate lower than 98% across all genotyping batches in the population and/or lower than 90% in one of the genotyping batches, (iii) not present in 1000 Genomes Phase 1, (iv) fail Hardy Weinberg equilibrium (FDR $< 10^{-5}$ across all samples or within each genotyping batch), (v) heterogeneous allele frequencies between the different genotyping batches within one population (FDR $< 10^{-5}$; in genotyping batches with more than 100 samples), (vi) different missing genotyping rate between cases and controls (P $< 10^{-5}$), (vii) monomorphic in the population. Following marker QC 125,141 SNPs remained in the East Asian dataset, 145,857 SNPs in the Indian dataset, 152,232 in the Iranian dataset and 144,245 in the Caucasian dataset.

Sample QC—Samples with a low call rate (<98%) and samples with outlying heterozygosity rate (FDR <0.01) where removed. Identity by descent was calculated using an LD-pruned set of SNPs with MAF > 0.05. Sample pairs with an identity by descent of > 0.8 were considered duplicates, and pairs with an identity by descent of > 0.4 and < 0.8 where considered related. For all duplicated and related pairs, the sample with the lowest genotype call rate was removed. After sample QC 6,543 (2,824 cases, 3,719 controls) East Asian samples, 2,413 (1,423 cases, 990 controls) Indian samples, 890 (548 cases, 342 controls) Iranian samples and 65,642 (31,664 cases, 33,977 controls) European samples remained.

Per-population association analysis—Case-control association tests for Crohn's disease, ulcerative colitis and IBD were performed in each ancestry group (European, East Asian, Indian and Iranian) using a linear mixed model as implemented in MMM 9. A covariance genetic relatedness matrix, R, was included as a random effects component in the model to account for population stratification. To avoid biases in the estimation of R due to the design of the Immunochip, SNPs were first pruned for LD (pairwise $r^2 < 0.2$). Of the remaining SNPs, we then removed those that lie in the HLA region or had a MAF < 10%. SNPs that showed modest association ($P < 0.005$) with IBD in a linear regression model fitting the first 10 principal components as covariates were also excluded. A total of ~14,000 SNPs were used to estimate R (varies between cohorts).

Genomic inflation factor—The Immunochip contains 3,120 SNPs that were part of a bipolar disease replication effort and other non-immune-related studies. After QC, 2,544 of these were used as null markers to estimate the overall inflation of the distribution of association test statistics (lambda). There was minimal inflation in the observed test statistics (lambda<1.06) from each cohort (Supplementary Figure 4).

Heterogeneity of effect—We tested the heterogeneity of associations across the four ancestry groups using the Cochran's Q test. The analysis was performed in R with the metafor package, using the odds ratios and standard errors estimated from each ancestry group. The I2 statistic from the Q test quantifies heterogeneity and ranges from 0% to 100%

43, with a value of 75 or above typically taken to indicate a high degree of heterogeneity [44]. We Bonferroni corrected this threshold for the 234 independently associated SNPs and consider $I^2 > 85.7$ (Q = 27.94 with 4 degrees of freedom) to indicate significant evidence of heterogeneity.

Power calculations—All power calculations were performed using the genetic power calculator [45] assuming a disease prevalence of 0.005 and log-additive risk.

Variance explained—The proportion of variance in disease liability explained by the associated variants were estimated assuming a disease prevalence of 0.005 and log-additive risk [46]. Due to ORs likely to be more accurately estimated in the much larger European cohort, only European ORs and allele frequencies were used.

Trans-ethnic association analysis

MANTRA meta-analysis—The European, East Asian, Indian and Iranian per-population association summary statistics were combined into a trans-ethnic meta-analysis using MANTRA [10]. This method allows for differences in allelic effects arising from differences in LD between distant populations. MANTRA first assigns each population into clusters using a Bayesian partition model of relatedness defined by the mean pairwise allele frequency differences between populations (Fst) calculated using all SNPs on the Immunochip (Supplementary Figure 11). As more closely related populations are more similar to each other with respect to allele frequencies and LD with the causal variant, we would expect greater homogeneity in effect sizes. Conversely, more distant populations may exhibit greater heterogeneity in effect sizes. For each SNP, if there is no evidence for heterogeneity, all studies are placed in the same cluster and the method is equivalent to a fixed-effects meta-analysis. Where the data is consistent with heterogeneity, the studies will be assigned to different clusters, with greater weight given to clusters that match the similarity in the ancestry from the prior model of relatedness. The strength of association is measured by a Bayes Factor (BF).

Manual inspection of associated SNPs—Evoker [47] was used to manually inspect signal intensity plots of all non-HLA loci with association P-value $< 10^{-7}$ (for MMM) or $\log_{10}$ BF > 6 (for MANTRA) in any of the three phenotypes. At each locus (defined here as a +/- 150kb window spanning the most strongly associated SNP), the top 10 P-value ranked SNPs were selected for inspection. Every SNP was inspected by two different researchers. SNPs that that were passed by both researchers were taken forward.

Locus definition—Genome-wide significant loci were defined by an LD window of $r^2 >$ 0.6 from the lead SNP in the region with a per-population association $P < 5 \times 10^{-8}$ or $\log_{10}$ BF > 6. The $\log_{10}$BF > 6 threshold has been suggested to be a conservative threshold for declaring genome-wide significance [48]. Regions less than 250 kb apart were merged into a single associated locus. All LD calculations were performed using the control samples within each population.

Crohn's disease/ulcerative colitis/IBD likelihood modelling—Associated loci were classified according to their strength of association with Crohn's disease, ulcerative colitis or both using a multinomial logistic regression likelihood modelling approach within the Europeans only[6]. Four multinomial logistic regression models with parameters $\beta_{Crohn's disease}$ and $\beta_{ulcerative colitis}$ were fitted with the following constraints:

1. Crohn's disease-specific model: $\beta_{ulcerative colitis} = 0$ (1 d.f.)

2. ulcerative colitis-specific model: $\beta_{Crohn's disease} = 0$ (1 d.f.)

3. IBD unsaturated model: $\beta_{Crohn's disease} = \beta_{ulcerative colitis} = \beta_{IBD}$ (1 d.f.)

A fourth unconstrained model with 2 d.f. was also estimated with $\beta_{Crohn's disease}$ and $\beta_{ulcerative colitis}$ both fitted by maximum likelihood. Log-likelihoods were calculated for each model, and three likelihood-ratio tests were performed comparing models 1-3 against the unconstrained model. If the P-values of all three tests were less than 0.05, the SNP was classified as associated with both Crohn's disease and ulcerative colitis but with evidence of different effect sizes. Otherwise, of the three constrained models, the SNP was classified according to the model with the largest likelihood. If 'IBD unsaturated' is the best fitting model the locus can be interpreted as associated with both Crohn's disease and ulcerative colitis but with no evidence for different effect sizes.

Locus annotations and candidate gene prioritization

Associations with other phenotypes—IBD risk loci were annotated with the NHGRI GWAS Catalog accessed on August 15th 2014. Newly identified IBD loci that overlap with a GWAS locus (±250kb either side of the reported SNP) for another phenotype were reported. Only SNPs with association $P < 5\times10^{-8}$ in the GWAS catalog were considered.

Non-synonymous SNPs—Functional annotation was performed using functionGVS (dbSNP build 134). A variant was annotated as a coding SNP if it was classified as "missense" or "nonsense", or if it was in LD of $r^2 > 0.8$ (in Europeans or East Asians) with a SNP with such a classification. The genes in which these missense variants lie were included as cSNP implicated genes.

Expression quantitative trait loci—We tested whether each of the IBD associated variants showed an effect on gene expression levels of genes (*cis*-eQTLs) in whole blood. For this analysis we used gene expression and genotype data from the Fehrmann study (N=1,240) and the EGCUT study (N=891) [49, 50]. Gene expression normalization was performed as described previously correcting for up to 40 principal components[15]. eQTL effects were determined using Spearman's rank correlation and subsequently meta-analysed using a sample-weighted Z-score method. SNPs (MAF > 5%, Hardy-weinberg P-value > 0.001) were tested against probes within 250 kb of the SNP. Multiple testing correction was performed by controlling the FDR at 5%, using a null. For each significant IBD eQTL probe, we determined the variant having the largest eQTL effect size (within 250 kb of the probe). We then removed the effect of this top-associated variant using linear regression and repeated the analysis on the IBD variant. This allowed us to determine whether the eQTL

effect of the IBD variant either is the top eQTL effect in a locus or whether the IBD variant has an eQTL effect independent of the top effect within the locus.

GRAIL network analysis—GRAIL evaluates the degree of functional connectivity of a gene based on the textual relationships among genes. To avoid publication biases from large scale GWAS, we used all PubMed text before December 2006. We used the GRAIL web tool to perform this analysis and took the list of loci from Supplementary Table 10. As in the previous study, we removed associations in the MHC region, and replaced regions with the 4 well-established genes (*IL23R*, *ATG16L1*, *PTPN22* and *NOD2*) to reduce noise. Only genes with GRAIL P-value <0.05 and edges with a score > 0.5 were used in the connectivity map 51.

Protein-Protein Interaction networks (DAPPLE)—DAPPLE uses the protein-protein physical interactions to evaluate the disease association of genes. Each gene is assigned an empirical P-value based on its enrichment in interactions with other genes in the list. We used the DAPPLE web tool to perform this analysis and took the list of loci from Supplementary Table 10. As in the GRAIL analysis, we removed associations in the MHC region, and used the 4 established genes instead of their regions. Genes with Dapple P-value < 0.05 were reported 52.

ENCODE regulatory features—The following regulatory features from the Encyclopedia of DNA Elements (ENCODE) 53 were used to annotate IBD risk loci: DNaseI hypersensitivity sites, transcription factor binding sites, histone modification and DNA-polymerase sites. The cell types in which they occur are also reported. Regulatory elements were extracted using the Variant Explorer tool.

Modelling European vs. Non European IBD risk

Effect size and frequency comparisons—For each associated SNP for a given phenotype as defined from the likelihood modelling, we estimated correlation between logORs in European and non-European populations using a weighted linear regression with the inverse variance of the non-European logOR as weights. For an associated SNP, differences in the effect size between two populations were tested using t-tests for a significant difference in log odds ratios (ORs). Fixation index (F_{st}) values for a SNP between two populations were calculated using the Weir and Cockerham method on allele frequencies in control samples only 54. The proportion of variance explained by each associated locus per population was calculated using a liability threshold model53 assuming a disease prevalence of 500 per 100,000 and log-additive disease risk.

Genetic correlations—The proportion of genetic variation tagged by Immunochip SNPs that is shared between European and each non-European cohort (r_G) was estimated using the bivariate linear mixed-effects model implemented in GCTA 55.The method was applied across Immunochip individuals for each European vs. non-European pairwise comparison for Crohn's disease and ulcerative colitis, with 20 PCs as covariates and assuming a disease prevalence of 0.005. To test whether r_G is significantly different from 0 (or 1), r_G was fixed at 0 (or 1) and a likelihood ratio test comparing this constrained model with the

unconstrained model was applied. An rG of 0 means that no genetic variants are shared between the two populations, while a value of 1 means that all the genetic variance tagged in one population is shared with the other. In the European cohort, only 10,000 cases and 10,000 controls (selected at random) were included due to computation limitations, while all non-European samples were included.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Molodecky NA, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 2012; 142:46–54. [PubMed: 22001864]

2. NG SC, et al. Incidence and phenotype of inflammatory bowel disease based on results from the Asia-pacific Crohn's and colitis epidemiology study. Gastroenterology. 2013; 145:158–165. [PubMed: 23583432]

3. Asano K, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat Genet. 2009; 41:1325–1329. [PubMed: 19915573]

4. Yang SK, et al. Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut. 2014; 63:80–87. [PubMed: 23850713]

5. Juyal G, et al. Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. Gut. 2014 doi: 10.1136.

6. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]

7. Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014; 46:234–244. [PubMed: 24509480]

8. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376–381. [PubMed: 24390342]

9. Pirinen M, Donnelly P, Spencer C. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann Appl Stat. 2013; 7:369–390.

10. Morris AP. Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 2011; 35:809–822. [PubMed: 22125221]

11. Rioux JD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 2007; 39:596–604. [PubMed: 17435756]

12. Yamazaki K, et al. A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population. Gastroenterology. 2013; 144:781–788. [PubMed: 23266558]

13. Okada Y, et al. HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. Gastroenterology. 2011; 141:864–871. [PubMed: 21699788]

14. Juyal G, et al. An investigation of genome-wide studies reported susceptibility loci for ulcerative colitis shows limited replication in north Indians. PLoS One. 2011; 6:e16565. [PubMed: 21304977]

15. Westra HJ, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]

16. Beigel F, et al. Oncostatin M mediates STAT3-dependent intestinal epithelial restitution via increased cell proliferation, decreased apoptosis and upregulation of SERPIN family members. PLoS One. 2014; 7:e93498. [PubMed: 24710357]

17. Dastani Z, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. 2012; 8:e1002607. [PubMed: 22479202]

18. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]

19. Lee SH, et al. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2542. [PubMed: 22843982]

20. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010; 8:e1000294. [PubMed: 20126254]

21. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. 2011; 9:e1000580. [PubMed: 21267062]

22. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 2011; 9:e1000579. [PubMed: 21267061]

23. Beaudoin M, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLoS Genet. 2013; 9:e1003723. [PubMed: 24068945]

24. Rivas MA, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011; 43:1066–73. [PubMed: 21983784]

25. Kalinski P. Regulation of immune responses by prostaglandin E2. JImmunol. 2012; 188:21–28. [PubMed: 22187483]

26. Bonifaz L, et al. Efficient targeting of protein antigen to the dendritic cell receptor DEC-205 in the steady state leads to antigen presentation on major histocompatibility complex class I products and peripheral CD8+ T cell tolerance. J Exp Med. 2002; 196:1627–638. [PubMed: 12486105]

27. Fukaya T, et al. Conditional ablation of CD205+ conventional dendritic cells impacts the regulation of T-cell immunity and homeostasis in vivo. Proc Natl Acad Sci U S A. 2012; 109:11288–11293. [PubMed: 22736794]

28. Izadpanah A, Dwinell MB, Eckmann L, Varki NM, Kagnoff MF. Regulated MIP-3alpha/CCL20 production by human intestinal epithelium: mechanism for modulating mucosal immunity. Am J Physiol Gastrointest Liver Physiol. 2001; 280:G710–719. [PubMed: 11254498]

29. Kaser A, et al. Increased expression of CCL20 in human inflammatory bowel disease. J Clin Immunol. 2004; 24:74–85. [PubMed: 14997037]

30. Varona R, Cadenas V, Flores J, Martínez AC, Márquez G. CCR6 has a non-redundant role in the development of inflammatory bowel disease. Eur J Immunol. 2003; 33:2937–46. [PubMed: 14515278]

31. Miyake T, et al. IκBζ is essential for natural killer cell activation in response to IL-12 and IL-18. Proc Natl Acad Sci U S A. 2010; 107:17680–17685. [PubMed: 20876105]

32. Hildebrand DG, et al. IκBζ is a transcriptional key regulator of CCL2/MCP-1. J Immunol. 2013; 190:4812–4820. [PubMed: 23547114]

33. Okamoto K, et al. IkappaBzeta regulates T(H)17 development by cooperating with ROR nuclear receptors. Nature. 2010; 464:1381–1385. [PubMed: 20383124]

34. Duarte JH, Di Meglio P, Hirota K, Ahlfors H, Stockinger B. Differential influences of the aryl hydrocarbon receptor on Th17 mediated responses in vitro and in vivo. PLoS One. 2013; 8:e79819. [PubMed: 24244565]

35. Li Y, et al. Exogenous stimuli maintain intraepithelial lymphocytes via aryl hydrocarbon receptor activation. Cell. 2011; 147:629–640. [PubMed: 21999944]

36. Serfling E, et al. NFATc1/αA: the other face of NFAT factors in lymphocytes. Cell Commun Signal. 2012; 10:16. [PubMed: 22764736]

37. Matute JD, et al. A new genetic subgroup of chronic granulomatous disease with autosomal recessive mutations in p40 phox and selective defects in neutrophil NADPH oxidase activity. Blood. 2009; 114:3309–3315. [PubMed: 19692703]

38. Conway KL, et al. p40phox expression regulates neutrophil recruitment and function during the resolution phase of intestinal inflammation. J Immunol. 2012; 189:3631–40. [PubMed: 22914050]

37. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011; 9:179–181. [PubMed: 22138821]

38. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda). 2011; 1:457–470. [PubMed: 22384356]

39. Freedman ML, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36:388–393. [PubMed: 15052270]

40. Shah TS, et al. OptiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics. 2012; 28:1598–1603. [PubMed: 22500001]

41. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

42. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

43. Higgins JP, Thompson SP. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21:1539–1558. [PubMed: 12111919]

44. Higgins JP, et al. Measuring inconsistency in meta-analyses. BMJ. 2003; 327:557–560. [PubMed: 12958120]

45. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003; 19:149–150. [PubMed: 12499305]

46. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol. 2011; 35:310–317. [PubMed: 21374718]

47. Morris JA, Randall JC, Maller JB, Barrett JC. Evoker: a visualization tool for genotype intensity data. Bioinformatics. 2010; 26:1786–1787. [PubMed: 20507892]

48. Distani Z, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. 2012; 8:e1002607. [PubMed: 22479202]

49. Schramm K, et al. Mapping the genetic architecture of gene regulation in whole blood. PLoS One. 2014; 9:e93844. [PubMed: 24740359]

50. Fehrmann RS. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011; 7:e1002197. [PubMed: 21829388]

51. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5:e1000534. [PubMed: 19557189]

52. Rossin EJ, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011; 7:e1001273. [PubMed: 21249183]

53. Myers RM, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9:e1001046. [PubMed: 21526222]

54. Cockerham CC, Weir BS. Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics. 1984; 40:157–164. [PubMed: 6733226]

55. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2542. [PubMed: 22843982]

> **Box: Selected candidate genes in the newly-associated IBD susceptibility loci**
>
> *PTGS2:* encodes COX-2, an enzyme that converts arachidonic acid into prostaglandins and which is the pharmacological target of non-steroidal anti-inflammatory drugs. Prostaglandins were once thought to be exclusively pro-inflammatory (hence the anti-inflammatory moniker of NSAIDs) although there is now increasing evidence that some may play important anti-inflammatory roles by inhibiting T cell activation and promoting regulatory T cell development [25]. Consistent with this, NSAIDs are generally avoided in IBD as they are known to precipitate disease flares.
>
> *LY75:* encodes DEC-205 (also known as CD205), a cell surface receptor that is highly expressed on dendritic cells and is involved in the endocytosis of extracellular antigens and their presentation on MHC class I molecules [26]. This receptor has been shown to play an important role in T cell function and homeostasis [27].
>
> *CD28:* a key co-stimulatory molecule that plays an important role in T cell activation. This locus also contains other genes that are also involved in T cell co-stimulation, including *ICOS* and *CTLA4*. If T cells are stimulated in the absence of co-stimulatory signal, this typically leads to anergy - one of the three main processes that can bring about tolerance; an important means of preventing aberrant immunological responses to intestinal antigens.
>
> *CCL20:* a chemokine that is produced by the intestinal epithelium [28] and which binds and activates CCR6. This interaction is important in regulating the migration of T cells (especially regulatory T cells) and dendritic cells to the gut, with increased production of CCL20 being detectable during inflammation [29]. Consistent with this, murine models of IBD are modulated if mice lack CCR6 [30]. The *CCR6* locus is itself associated with IBD.
>
> *NFKBIZ:* encodes NF Kappa B inhibitor zeta, an inducible regulator of NFKB. This gene has been shown to have several functions, including roles in natural killer cell activation [31] and monocyte recruitment [32]. Recently, however, NFKBIZ has also been shown to be a critical regulator of Th17 development through its interaction with ROR nuclear receptors [33]. Accordingly, this association thus further underlines the importance of Th17 cells in IBD pathogenesis.
>
> *OSMR:* encodes the Oncostatin M receptor, a cytokine receptor component which heterodimerises with other proteins to form both the oncostatin M receptor and the IL-31 receptor. Levels of oncostatin M are elevated in biopsies from patients with active IBD and are thought to promote intestinal epithelial cell proliferation and wound healing - thereby augmenting the barrier function of the intestinal epithelium in intestinal inflammation [16].
>
> *AHR:* encodes the aryl hydrocarbon receptor, a ligand-activated transcription factor that can bind a range of aromatic hydrocarbons - including several compounds derived from dietary components. This receptor is highly expressed on Th17 cells and its ligation leads to their expansion and enhanced production of cytokines, including IL-22 [34]. Moreover, deficiency of this receptor (or its ligands) also disrupts intraepithelial lymphocyte

homeostasis, leading to failure to control intestinal microbial load and composition, and aberrant immune activation resulting in epithelial damage [35]. Accordingly, this association further highlights the importance of the interaction between genes and the environment in IBD pathogenesis.

PTK2B: encodes Protein tyrosine kinase 2 beta (also known as Pyk2), an important intracellular kinase for diverse signalling pathways, including MAP kinase and JNK. Functions include roles in monocyte migration and neutrophil degranulation.

NFATC1: encodes Nuclear factor of activated T-cells, cytoplasmic 1 - an NFAT transcription factor that is specifically expressed upon activation of T and B cells following ligation of their respective receptors. This expression supports lymphocyte proliferation and inhibits activation-induced cell death leading to enhanced immune responses [36]. NFAT transcription factors are the main molecular targets of calcineurin inhibitors, such as cyclosporine, which are used in the treatment of IBD.

A – Crohn's disease

B – Ulcerative colitis
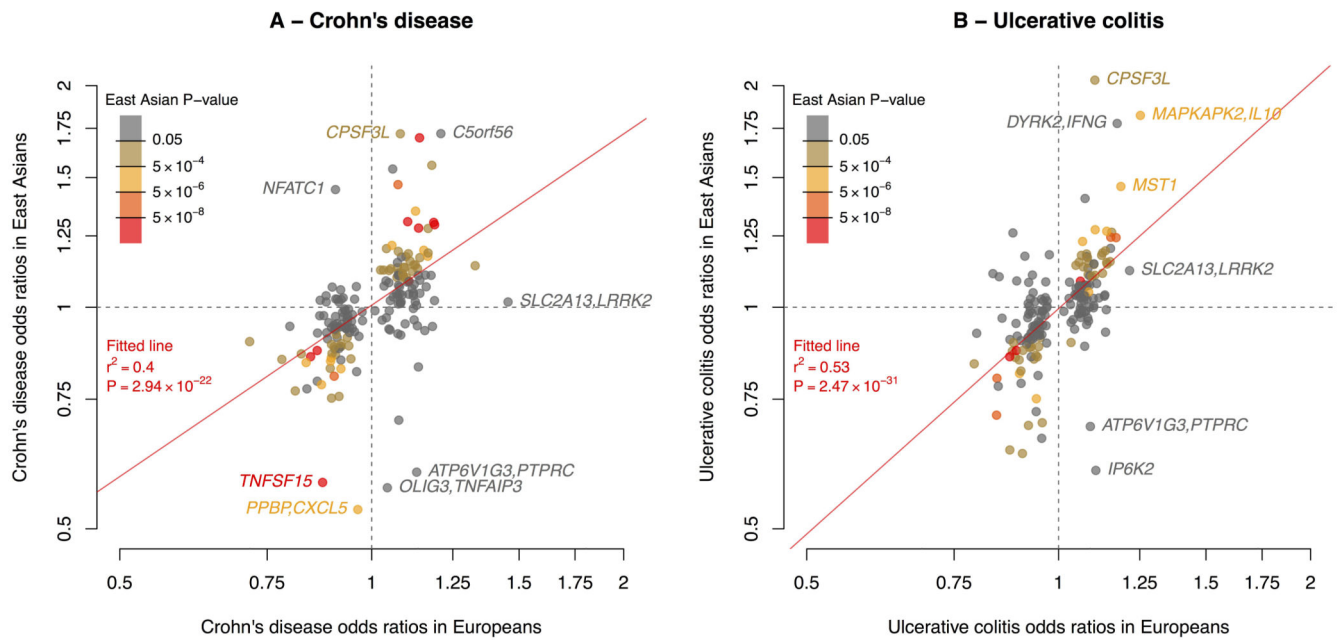
Figure 1. Comparison of Crohn's disease and ulcerative colitis risk variant odds ratios in Europeans and East Asians.

For each SNP, ORs (on the log-scale) were estimated within each population for Crohn's disease (A) and ulcerative colitis (B). The colour of each point denotes the association P-value for that phenotype in East Asians. The red line denotes the best fitting least squares regression line, weighted by the inverse of the variance of the log ORs in East Asians. The significance and goodness of fit are shown in red.
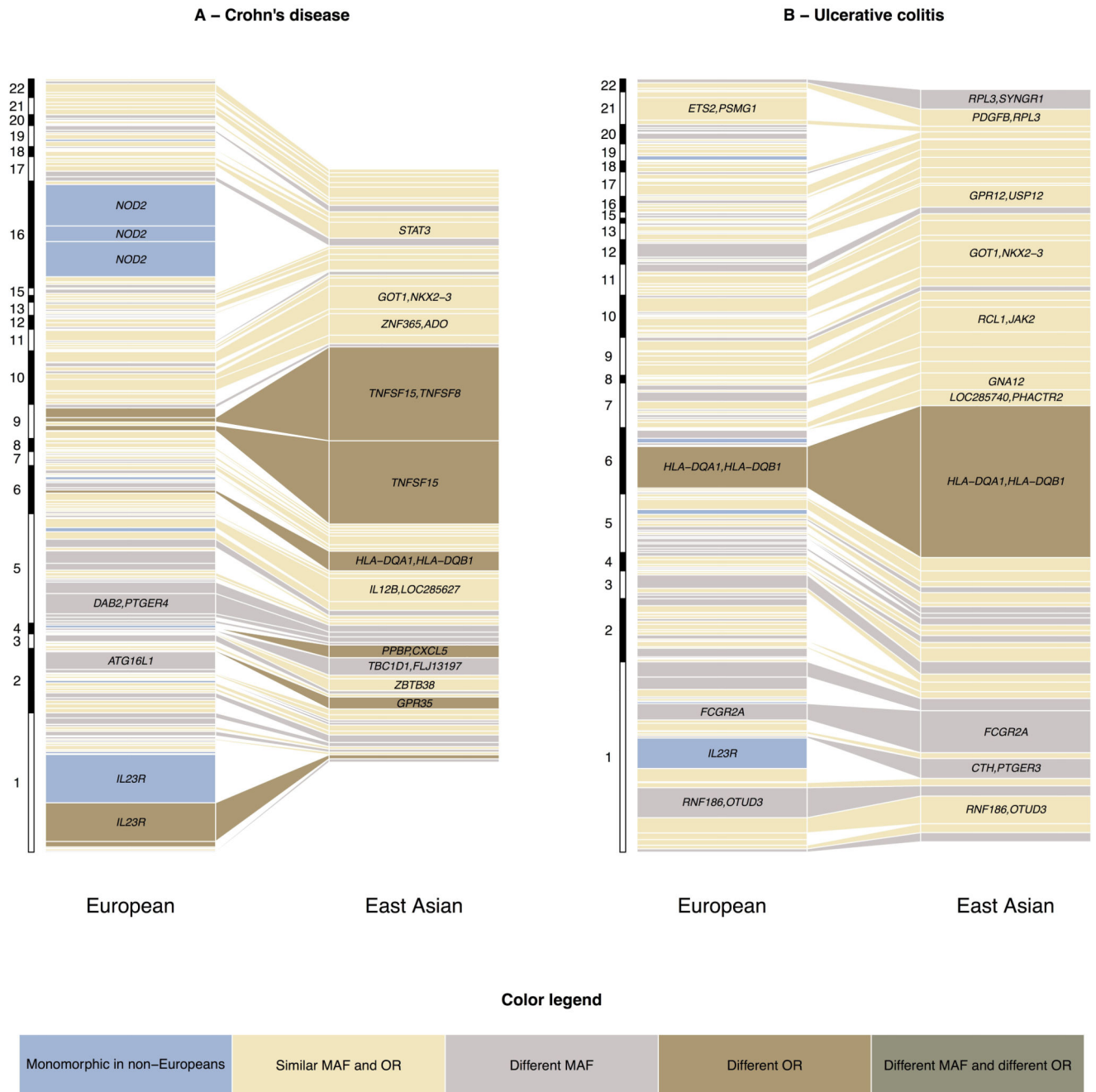
Figure 2. Comparison of Variance explained per risk variant between East Asian and European Crohn's Disease and ulcerative colitis.

Each box represents an independently associated SNP for Crohn's disease (A) and ulcerative colitis (B). The size of each box is proportional to the amount of variance in disease liability explained by that variant. Only SNPs with an association P-value < 0.01 are included in the East Asian panel. The color of each box denotes whether any difference in variance explained is due to differences in allele frequencies ($F_{st} > 0.1$/monomorphic in East Asians), significant heterogeneity of odds ratios ($P < 2.5 \times 10^{-4}$) or both.

Table 1

Cohort sample sizes.

GWAS and Immunochip trans-ethnic meta-analysis					
Population	**CD**	**CD controls**	**UC**	**UC controls**	**IBD**	**IBD controls**
European GWAS	5956	14927	6968	20464	12882	21770
European Immunochip	14594	26715	10679	26715	25273	26715
Non-European Immunochip	2025	5051	2770	5051	4795	5051
Total	22575	46693	20417	52230	42950	53536

Table 2

Newly associated IBD risk loci.

The IBD, ulcerative colitis or Crohn's disease loci are identified through a trans-ethnic analysis of genome-wide and Immunochip genotype data from a cohort of 86,682 European individuals and 9,846 individuals of non-European descent. Loci achieving genome-wide significance ($P < 5 \times 10^{-8}$) in one of the individual cohorts of European (Eur), East Asian, Indian or Iranian descent, or a log10 Bayes Factor > 6.0 in the combined trans-ethnic association analysis, are considered significantly associated loci. Loci having a logBF>6 but P>5×10⁻⁸ in each individual ancestral cohort were required to show no significant evidence of heterogeneity across all four ancestry groups (I^2> 85.7). Association P-values and ORs of non-European cohorts are given in Supplementary Table 2.

Ch r.	SNP	BP positio n	[a]Reference allele	[b]Best phenotype	[c]LR phenotype	[d]Log1 0 BF	[e]Het I^2	Eur. OR	Eur. P	Candidate Genes
1	rs1748195	63049593	G	CD	CD	6.08	0	1.07 (1.04-1.1)	7.13×10^{-8}	*USP1*
1	rs34856868	92554283	A	IBD	IBD_U	6.16	0	0.82 (0.77-0.88)	9.80×10^{-9}	*BTBD8*
1	rs11583043	101466054	A	UC	IBD_U	8.34	66.5	1.08 (1.05-1.11)	6.05×10^{-8}	*SLC30A, EDG1*
1	rs6025	169519049	A	IBD	IBD_U	6.43	0	0.84 (0.79-0.89)	2.51×10^{-8}	*SELI,SELE,SE LL*
1	rs10798 069	186875459	A	CD	IBD_S	7.24	0	0.93 (0.91-0.95)	4.25×10^{-9}	*PTGS2,PLA2G4A*
1	rs7555082	198598663	A	CD	IBD_U	7.97	0	1.13 (1.09-1.17)	1.47×10^{-10}	*PTPRC*
2	rs11681 525	145492382	C	CD	CD	8.8	59.3	0.86 (0.82-0.90)	4.08×10^{-11}	-
2	rs4664304	160794008	A	IBD	IBD_U	6.34	0	1.06 (1.04-1.08)	2.61×10^{-8}	*MARCH7,LY75,PLA2R1*
2	rs31164 94	204592021	G	UC	IBD_S	7.03	0	1.08 (1.05-1.11)	1.30×10^{-7}	*ICOS,CD28,CT LA4*
2	rs11178 1203	228660112	G	IBD	IBD_U	10.04	0	0.94 (0.92-0.96)	2.16×10^{-10}	*CCL20*
2	rs35320 439	242737341	G	CD	IBD_S	7.71	0	1.09 (1.06-1.12)	9.89×10^{-10}	*PDCD1,ATG4B*
3	rs11301 0081	46457412	G	UC	IBD_U	7.45	0	1.14 (1.09-1.19)	9.02×10^{-10}	*FLJ78302,LTF, CCR1,CCR2, CCR3,CCR5*
3	rs616597	101569726	A	UC	UC	6.68	54.7	0.93 (0.90-0.96)	9.34×10^{-6}	***NFKBIZ***
3	rs724016	141105570	G	CD	CD	7.41	70.9	1.06 (1.04-1.09)	3.36×10^{-6}	-
4	rs2073505	3444503	A	IBD	IBD_U	6.87	0	1.1 (1.06-1.14)	1.46×10^{-7}	*HGFAC*
4	rs4692386	26132361	A	IBD	IBD_U	6.47	0	0.94 (0.92-0.96)	1.21×10^{-8}	-
4	rs6856616	38325036	G	IBD	IBD_U	9.78	61.6	1.1 (1.06-1.14)	9.72×10^{-7}	-
4	rs2189234	106075498	A	UC	UC	8.85	0	1.08 (1.05-1.11)	1.95×10^{-10}	-
5	rs395157	38867732	A	IBD	IBD_U	19.5	0	1.1 (1.08-1.12)	2.22×10^{-20}	***OSMR,FYB, LIFR***
5	rs4703855	71693899	A	IBD	IBD_U	6.83	70.3	0.93 (0.91-0.95)	7.16×10^{-11}	-
5	rs564349	172324978	G	IBD	IBD_U	8.12	37.5	1.06 (1.04-1.08)	1.54×10^{-7}	*C5orf4, DUSP1*

Ch r.	SNP	BP position n	[a]Reference allele	[b]Best phenotype	[c]LR phenotype	[d]Log1 0 BF	[e]Het I²	Eur. OR	Eur. P	Candidate Genes
6	rs7773324	382559	G	CD	IBD_U	7.67	0	0.92 (0.90-0.94)	1.06×10^{-9}	IRF4,DUSP22
6	rs13204048	3420406	G	CD	IBD_S	7.23	53.5	0.93 (0.91-0.95)	2.89×10^{-8}	-
6	rs7758080	149577079	G	CD	IBD_S	7.88	0	1.08 (1.05-1.11)	7.27×10^{-9}	MAP3K7IP2
7	rs1077773	17442679	G	UC	UC	5.86	76.7	0.93 (0.91-0.95)	5.96×10^{-9}	AHR
7	rs2538470	148220448	A	IBD	IBD_U	10.93	54.6	1.07 (1.05-1.09)	3.00×10^{-11}	CNTNAP2
8	rs17057051	27227554	G	IBD	IBD_U	6.74	15.9	0.94 (0.92-0.96)	5.50×10^{-8}	PTK2B,TRIM 35,EPHX2,
8	rs7011507	49129242	A	UC	IBD_U	7.49	39.3	0.9 (0.87-0.93)	6.40×10^{-8}	-
10	rs3740415	104232716	G	IBD	IBD_U	6.26	0	0.95 (0.93-0.97)	1.03×10^{-7}	NFKB2, TRIM8, TMEM180
12	rs7954567	6491125	A	CD	CD	8.25	0	1.09 (1.06-1.12)	1.30×10^{-9}	CD27,TNFRSF1A,LTBR
12	rs653178	112007756	G	IBD	IBD_U	6.57	49.7	1.06 (1.04-1.08)	1.11×10^{-8}	SH2B3, ALDH2,ATXN 2
12	rs11064881	120146925	A	IBD	IBD_U	7.02	31.7	1.1 (1.06-1.14)	5.95×10^{-8}	PRKAB1
13	rs9525625	43018030	A	CD	CD	8.55	37.3	1.08 (1.05-1.11)	1.41×10^{-9}	AKAP1, TFSF11
17	rs3853824	54880993	A	CD	IBD_S	8.46	50.4	0.92 (0.90-0.94)	1.17×10^{-10}	-
17	rs17736589	76737118	G	UC	UC	6.53	53.4	1.09 (1.06-1.12)	4.34×10^{-8}	-
18	rs9319943	56879827	G	CD	CD	6.33	33.4	1.08 (1.05-1.11)	9.05×10^{-7}	-
18	rs7236492	77220616	A	CD	IBD_S	6.6	0	0.91 (0.88-0.94)	9.09×10^{-9}	NFATC1, TST
22	rs727563	41867377	G	CD	CD	7.1	76	1.1 (1.07-1.13)	1.88×10^{-10}	TEF, NHP2L1, PMM1, L3MBTL2, CHADL

[a]The minor allele in the European cohort was chosen to be the reference allele.

[b]Phenotype with the largest MANTRA Bayes factor

[c]The preferred phenotype (ulcerative colitis, Crohn's disease or IBD (i.e. both)) from our likelihood modeling approach to classify loci according to their relative strength of association. IBD_S and IBD_U refer to the IBD saturated and IBD unsaturated models, respectively (see main text and Online Methods).

[d]MANTRA log10 Bayes Factor.

[e]Heterogeneity I² percentage. Candidate genes are identified by one of the gene prioritization methods we performed (eQTL, GRAIL, DAPPLE and cSNP annotation - see main text and Online Methods). Genes in bold are prioritized by > 2 gene prioritization strategies. UC, Ulcerative Colitis; CD, Crohn's Disease; IBD, Inflammatory Bowel Disease; BP, Base Position; Chr, chromosome; OR, odds ratio.