Genetic susceptibility for inflammatory bowel disease across ethnicities and diseases
van Sommeren, Suzanne

DOI:
10.33612/diss.100597247

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
CHAPTER 3

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


Nature Genetics 2015;47(9):979–86

*These authors contributed equally
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.
INTRODUCTION

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. 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 in previous European GWAS studies. 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.

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.

MATERIALS AND 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). 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 1,000 cases and 1,000 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).
Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.

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 $r^2 < 0.2$, 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) were 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 MMM9. 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 (λ). There was minimal inflation in the observed test statistics (λ < 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 I² > 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 calculator45 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 MANTRA10. 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**

Evoker47 was used to manually inspect signal intensity plots of all non-HLA loci with association P-value < 10⁻⁷ (for MMM) or log₁₀ BF > 6 (for MANTRA) in any of the three phenotypes. At each locus (defined here as a ±150 kb 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 r² > 0.6 from the lead SNP in the region with a per-population association P < 5 × 10⁻⁸ or log₁₀ BF > 6. The log₁₀ 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 modeling**

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. Four multinomial logistic regression models with parameters

\[ \beta_{\text{Crohn's disease}} \] and \[ \beta_{\text{ulcerative colitis}} \] were fitted with the following constraints:
1. Crohn's disease-specific model: \[ \beta_{\text{ulcerative colitis}} = 0 \] (1 d.f.)
2. ulcerative colitis-specific model: \[ \beta_{\text{Crohn's disease}} = 0 \] (1 d.f.)
3. IBD unsaturated model:
   \[ \beta_{\text{Crohn's disease}} = \beta_{\text{ulcerative colitis}} = \beta_{\text{IBD}} \] (1 d.f.)

A fourth unconstrained model with 2 d.f. was also estimated with \[ \beta_{\text{Crohn's disease}} \] and \[ \beta_{\text{ulcerative colitis}} \] 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 (±250 kb either side of the reported SNP) for another phenotype were reported. Only SNPs with association P < 5 \times 10^{-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). Gene expression normalization was performed as described previously correcting for up to 40 principal components. 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.
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 (Fst) 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 (rg) 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 rg is significantly different from 0 (or 1), rg 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.

RESULTS

Study design
Following quality-control (QC) and 1000 genomes imputation (Phase I – March 2012), 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
Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.

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 ($P < 5 \times 10^{-8}$) or in the transethnic meta-analysis that included all ancestries ($\log\text{BF} > 6$) 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. $\log\text{BF} > 6$ but $P > 5 \times 10^{-8}$ 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 ($5 \times 10^{-8} < P < 8.7 \times 10^{-6}$, 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 $7.38 \times 10^{-6}$. 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

---

Table 1. Cohort sample sizes.

<table>
<thead>
<tr>
<th>Population</th>
<th>CD</th>
<th>CD controls</th>
<th>UC</th>
<th>UC controls</th>
<th>IBD</th>
<th>IBD controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>European GWAS</td>
<td>5,956</td>
<td>14,927</td>
<td>6,968</td>
<td>20,464</td>
<td>12,882</td>
<td>21,770</td>
</tr>
<tr>
<td>European Immunochip</td>
<td>14,594</td>
<td>26,715</td>
<td>10,679</td>
<td>26,715</td>
<td>25,273</td>
<td>26,715</td>
</tr>
<tr>
<td>Non-European Immunochip</td>
<td>2,025</td>
<td>5,051</td>
<td>2,770</td>
<td>5,051</td>
<td>4,795</td>
<td>5,051</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>22,575</td>
<td>46,693</td>
<td>20,417</td>
<td>52,230</td>
<td>42,950</td>
<td>53,536</td>
</tr>
</tbody>
</table>
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 $P > 5 \times 10^{-8}$ 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.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>SNP</th>
<th>BP position</th>
<th>Reference allele</th>
<th>Best phenotype</th>
<th>LR phenotype</th>
<th>Log10 BF</th>
<th>Het I2</th>
<th>Eur. OR</th>
<th>Eur. P</th>
<th>Candidate Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs1748195</td>
<td>63049593 G</td>
<td>CD</td>
<td>CD</td>
<td>6.08</td>
<td>0</td>
<td>1.07</td>
<td>(1.04–1.1)</td>
<td>7.13 × 10^{-8}</td>
<td>USP1</td>
</tr>
<tr>
<td>1</td>
<td>rs34856868</td>
<td>92554283 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.16</td>
<td>0</td>
<td>0.82</td>
<td>(0.77–0.88)</td>
<td>9.80 × 10^{-9}</td>
<td>BTBD8</td>
</tr>
<tr>
<td>1</td>
<td>rs11583043</td>
<td>101466054 A</td>
<td>UC</td>
<td>IBD_U</td>
<td>8.34</td>
<td>66.5</td>
<td>1.08</td>
<td>(1.05–1.11)</td>
<td>6.05 × 10^{-9}</td>
<td>SLC30A, EDG1</td>
</tr>
<tr>
<td>1</td>
<td>rs6025</td>
<td>169519049 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.43</td>
<td>0</td>
<td>0.84</td>
<td>(0.79–0.89)</td>
<td>2.51 × 10^{-8}</td>
<td>SELP, SELE, SELL</td>
</tr>
<tr>
<td>1</td>
<td>rs10798069</td>
<td>186875459 A</td>
<td>CD</td>
<td>IBD_S</td>
<td>7.24</td>
<td>0</td>
<td>0.93</td>
<td>(0.91–0.95)</td>
<td>4.25 × 10^{-9}</td>
<td>PTGS2, PLA2G4A</td>
</tr>
<tr>
<td>1</td>
<td>rs7555082</td>
<td>198598663 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>7.97</td>
<td>0</td>
<td>1.13</td>
<td>(1.09–1.17)</td>
<td>1.47 × 10^{-10}</td>
<td>PTPRC</td>
</tr>
<tr>
<td>2</td>
<td>rs11681525</td>
<td>145492382 C</td>
<td>CD</td>
<td>CD</td>
<td>8.8</td>
<td>59.3</td>
<td>0.86</td>
<td>(0.82–0.90)</td>
<td>4.08 × 10^{-11}</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>rs4664304</td>
<td>160794008 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.34</td>
<td>0</td>
<td>1.06</td>
<td>(1.04–1.08)</td>
<td>2.61 × 10^{-9}</td>
<td>MARCH7, LY75, PLA2R1</td>
</tr>
<tr>
<td>2</td>
<td>rs3116494</td>
<td>204592021 G</td>
<td>UC</td>
<td>IBD_S</td>
<td>7.03</td>
<td>0</td>
<td>1.08</td>
<td>(1.05–1.11)</td>
<td>1.30 × 10^{-7}</td>
<td>ICOS, CD28, CTLA4</td>
</tr>
<tr>
<td>2</td>
<td>rs111781203</td>
<td>228660112 G</td>
<td>IBD</td>
<td>IBD_S</td>
<td>10.04</td>
<td>0</td>
<td>0.94</td>
<td>(0.92–0.96)</td>
<td>2.16 × 10^{-10}</td>
<td>CCL20</td>
</tr>
<tr>
<td>2</td>
<td>rs35320439</td>
<td>242737341 G</td>
<td>CD</td>
<td>IBD_S</td>
<td>7.71</td>
<td>0</td>
<td>1.09</td>
<td>(1.06–1.12)</td>
<td>9.89 × 10^{-10}</td>
<td>PDCD1, ATG4B</td>
</tr>
<tr>
<td>3</td>
<td>rs113010081</td>
<td>46457412 G</td>
<td>UC</td>
<td>IBD_U</td>
<td>7.45</td>
<td>0</td>
<td>1.14</td>
<td>(1.09–1.19)</td>
<td>9.02 × 10^{-10}</td>
<td>FLJ78302, LTF, CCR1, CCR2, CCR3, CCR5</td>
</tr>
<tr>
<td>3</td>
<td>rs616597</td>
<td>101569726 A</td>
<td>UC</td>
<td>UC</td>
<td>6.68</td>
<td>54.7</td>
<td>0.93</td>
<td>(0.90–0.96)</td>
<td>9.34 × 10^{-6}</td>
<td>NFKBIZ</td>
</tr>
<tr>
<td>3</td>
<td>rs724016</td>
<td>141105570 G</td>
<td>CD</td>
<td>CD</td>
<td>7.41</td>
<td>79.9</td>
<td>1.06</td>
<td>(1.04–1.09)</td>
<td>3.36 × 10^{-6}</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>rs2073505</td>
<td>3444003 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.87</td>
<td>0</td>
<td>1.1</td>
<td>(1.06–1.14)</td>
<td>1.46 × 10^{-7}</td>
<td>HGFAC</td>
</tr>
<tr>
<td>4</td>
<td>rs4692386</td>
<td>26132361 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.47</td>
<td>0</td>
<td>0.94</td>
<td>(0.92–0.96)</td>
<td>1.21 × 10^{-8}</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>rs6856616</td>
<td>38320536 G</td>
<td>IBD</td>
<td>IBD_U</td>
<td>9.78</td>
<td>61.6</td>
<td>1.1</td>
<td>(1.06–1.14)</td>
<td>9.72 × 10^{-7}</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>rs2189234</td>
<td>106075498 A</td>
<td>UC</td>
<td>UC</td>
<td>8.85</td>
<td>0</td>
<td>1.08</td>
<td>(1.05–1.11)</td>
<td>1.95 × 10^{-10}</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>rs395157</td>
<td>38867732 A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>19.5</td>
<td>0</td>
<td>1.1</td>
<td>(1.08–1.12)</td>
<td>2.22 × 10^{-20}</td>
<td>OSMR, FYB, LIFR</td>
</tr>
</tbody>
</table>
Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP</th>
<th>BP position</th>
<th>Reference allele</th>
<th>Best phenotype</th>
<th>LR phenotype</th>
<th>Log10 BF</th>
<th>Het I2</th>
<th>Eur. OR</th>
<th>Eur. P</th>
<th>Candidate Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>rs4703855</td>
<td>71693899</td>
<td>A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.83</td>
<td>70.3</td>
<td>0.93 [0.91–0.95]</td>
<td>7.16×10⁻¹¹</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>rs564349</td>
<td>172324978</td>
<td>G</td>
<td>IBD</td>
<td>IBD_U</td>
<td>8.12</td>
<td>37.5</td>
<td>1.06 [1.04–1.08]</td>
<td>1.54×10⁻⁷</td>
<td>C5orf4, DUSP1</td>
</tr>
<tr>
<td>6</td>
<td>rs7773324</td>
<td>382559</td>
<td>G</td>
<td>CD</td>
<td>IBD_U</td>
<td>7.67</td>
<td>0</td>
<td>0.92 [0.90–0.94]</td>
<td>1.06×10⁻⁹</td>
<td>IRF4, DUSP22</td>
</tr>
<tr>
<td>6</td>
<td>rs13204048</td>
<td>3420406</td>
<td>G</td>
<td>CD</td>
<td>IBD_S</td>
<td>7.23</td>
<td>53.5</td>
<td>0.93 [0.91–0.95]</td>
<td>2.89×10⁻⁸</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>rs7758080</td>
<td>149577097</td>
<td>G</td>
<td>CD</td>
<td>IBD_S</td>
<td>7.88</td>
<td>0</td>
<td>1.08 [1.05–1.11]</td>
<td>7.27×10⁻⁹</td>
<td>MAP3K7IP2</td>
</tr>
<tr>
<td>7</td>
<td>rs1077773</td>
<td>17442679</td>
<td>G</td>
<td>UC</td>
<td>UC</td>
<td>5.86</td>
<td>76.7</td>
<td>0.93 [0.91–0.95]</td>
<td>5.96×10⁻⁹</td>
<td>AHR</td>
</tr>
<tr>
<td>7</td>
<td>rs2538470</td>
<td>148220448</td>
<td>A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>10.93</td>
<td>54.6</td>
<td>1.07 [1.05–1.09]</td>
<td>3.00×10⁻¹¹</td>
<td>CNTNAP2</td>
</tr>
<tr>
<td>7</td>
<td>rs17057051</td>
<td>27227554</td>
<td>G</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.74</td>
<td>15.9</td>
<td>0.94 [0.92–0.96]</td>
<td>5.50×10⁻⁸</td>
<td>PTK2B, TRIM 35, EPHX2</td>
</tr>
<tr>
<td>8</td>
<td>rs1715107</td>
<td>49129242</td>
<td>A</td>
<td>UC</td>
<td>IBD_U</td>
<td>7.49</td>
<td>39.3</td>
<td>0.9 [0.87–0.93]</td>
<td>6.40×10⁻⁸</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>rs3740415</td>
<td>104232716</td>
<td>G</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.26</td>
<td>0</td>
<td>0.95 [0.93–0.97]</td>
<td>1.03×10⁻⁷</td>
<td>NFKB2, TRIM8, TMEM180</td>
</tr>
<tr>
<td>12</td>
<td>rs7954567</td>
<td>6491125</td>
<td>A</td>
<td>CD</td>
<td>CD</td>
<td>8.25</td>
<td>0</td>
<td>1.09 [1.06–1.12]</td>
<td>1.30×10⁻⁹</td>
<td>CD27, TNFRSF 1A, LTBR</td>
</tr>
<tr>
<td>12</td>
<td>rs653178</td>
<td>112007756</td>
<td>A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>6.57</td>
<td>49.7</td>
<td>1.06 [1.04–1.08]</td>
<td>1.11×10⁻⁹</td>
<td>SH2B3, ALDH2, ATXN2</td>
</tr>
<tr>
<td>12</td>
<td>rs11064881</td>
<td>120146925</td>
<td>A</td>
<td>IBD</td>
<td>IBD_U</td>
<td>7.02</td>
<td>31.7</td>
<td>1.1 [1.06–1.14]</td>
<td>5.95×10⁻⁸</td>
<td>PRKAB1</td>
</tr>
<tr>
<td>13</td>
<td>rs9525625</td>
<td>43018030</td>
<td>A</td>
<td>CD</td>
<td>CD</td>
<td>8.55</td>
<td>37.3</td>
<td>1.08 [1.05–1.11]</td>
<td>1.41×10⁻⁹</td>
<td>AKAP1, TFSF11</td>
</tr>
<tr>
<td>17</td>
<td>rs3853824</td>
<td>54880993</td>
<td>A</td>
<td>CD</td>
<td>IBD_S</td>
<td>8.46</td>
<td>50.4</td>
<td>0.92 [0.90–0.94]</td>
<td>1.17×10⁻¹⁰</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>rs17736589</td>
<td>76737118</td>
<td>G</td>
<td>UC</td>
<td>UC</td>
<td>6.53</td>
<td>53.4</td>
<td>1.09 [1.06–1.12]</td>
<td>4.34×10⁻⁸</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>rs9319943</td>
<td>56879827</td>
<td>G</td>
<td>CD</td>
<td>CD</td>
<td>6.33</td>
<td>33.4</td>
<td>1.08 [1.05–1.11]</td>
<td>9.05×10⁻⁷</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>rs7236492</td>
<td>77220616</td>
<td>A</td>
<td>CD</td>
<td>IBD_S</td>
<td>6.6</td>
<td>0</td>
<td>0.91 [0.88–0.94]</td>
<td>9.09×10⁻⁹</td>
<td>NFATC1, TST</td>
</tr>
<tr>
<td>22</td>
<td>rs727563</td>
<td>41867377</td>
<td>G</td>
<td>CD</td>
<td>CD</td>
<td>7.1</td>
<td>76</td>
<td>1.1 [1.07–1.13]</td>
<td>1.88×10⁻¹⁰</td>
<td>TEF, NHP2, PMM1, L3MBTL2, CHADL</td>
</tr>
</tbody>
</table>

The minor allele in the European cohort was chosen to be the reference allele. Phenotype with the largest MANTRA Bayes factor. Phenotype with the largest MANTRA Bayes factor 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). MANTRA log10 Bayes Factor. 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.
least in part, by population stratification (which is now being better accounted for in the linear mixed model analysis). 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 \( P < 6.1 \times 10^{-4} \) (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 \( P < 1.0 \times 10^{-5} \) (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 \( P = 2.64 \times 10^{-3} \) in the current East Asian cohort but is strongly associated to European IBD (\( P = 1.04 \times 10^{-18} \)).

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 1,240 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 \( NFKBIZ \), \( CD28 \) and \( OSMR \) (supplementary table 7).

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 achieve genome-wide significant 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 = \( 6 \times 10^{-4} \)) 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, 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 HNF4α) 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, NFKBI, 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 \( (5 \times 10^{-5} \leq P < 5 \times 10^{-8}) \) 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 × 10^{-4}). Larger sample sizes and a more

**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. 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. This receptor has been shown to play an important role in T cell function and homeostasis.

**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 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.

**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. NFAT transcription factors are the main molecular targets of calcineurin inhibitors, such as cyclosporine, which are used in the treatment of IBD.
Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.

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.
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 \( \text{IL23R} \) (interleukin 23 receptor) gene, previous studies have shown that there is substantial genetic heterogeneity between European and East Asian individuals in IBD risk. In line with these observations, the \( \text{IL23R} \) SNP with the largest effect in European Crohn’s disease and ulcerative colitis (rs80174646) has a RAF of one in East Asians, while secondary \( \text{IL23R} \) 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,
Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.

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 (\(P < 2.44 \times 10^{-8}\)). 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 \(P = 1.83 \times 10^{-13}\)). The same variant was previously implicated in a GWAS of Crohn’s disease in Koreans (supplementary table 6). 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 (\(P = 3.99 \times 10^{-6}\), OR = 0.66), and becomes genome-wide significant (\(P = 2.31 \times 10^{-10}\), 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 (\(P < 5 \times 10^{-5}\)). In addition to differences in allele frequency we also observe evidence of heterogeneity of odds at this SNP \(\text{OR} = 1.37\; (P = 8.45 \times 10^{-46})\). 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 \(P = 5.89 \times 10^{-44}\), East Asian \(P = 2.62 \times 10^{-3}\) as well as evidence of heterogeneity of effect (European OR = 1.33 vs. East Asian OR = 1.13; heterogeneity \(P = 1.20 \times 10^{-3}\)). 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, \(P = 7.09 \times 10^{-78}\)). Furthermore we observed more stricturing behaviour (\(P = 2.02 \times 10^{-33}\)) and perianal disease (\(P = 5.36 \times 10^{-33}\)) and less inflammatory Crohn’s disease (\(P = 4.28 \times 10^{-32}\)) in the non-European population. In ulcerative colitis there was a lower rate of extensive colitis reported in the non-European population (\(P = 1.52 \times 10^{-34}\)) which was also reflected in a lower rate of colectomy (\(P = 1.23 \times 10^{-69}\)) (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.

**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.  

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 resequencing 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 \( \text{TNFSF15/TNFSF8} \) and \( \text{ATG16L1} \) loci, potentially suggesting that gene-environment interactions increase the variance explained by these associations in either European (\( \text{ATG16L1} \)) or non-European (\( \text{TNFSF15/TNFSF8} \)) 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 occurring 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.

**SUPPLEMENTARY DATA**

Supplementary data are available online: https://www.nature.com/articles/ng.3359#supplementary-information

**REFERENCES**


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

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.