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


Ulcerative colitis and Crohn’s disease are the two main forms of inflammatory bowel disease (IBD). Here we report the first trans-ancestry 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. Nevertheless, the direction and magnitude of effect are consistent in European and non-European cohorts. Moreover, we observe genetic heterogeneity between divergent populations at several established risk loci driven by differences in allele frequency (NOD2) or effect size (TNFSF15 and ATG16L1) of a combination of these factors (IL23R and IRGM). Our results provide biological insights into the pathogenesis of IBD and demonstrate the usefulness of trans-ancestry association studies for mapping loci associated with complex diseases and understanding genetic architecture across diverse populations.

IBD is composed of chronic, relapsing intestinal inflammatory diseases affecting more than 2.5 million people in Europe, with increasing prevalence in Asia and developing countries1,2. IBD is thought to arise from inappropriate activation of the intestinal mucosal immune system in response to commensal bacteria in a genetically susceptible host.

Thus far, 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 new genome-wide significant associations outside of the human leukocyte antigen (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 a consistent direction of effect and nominally significant evidence of association (P < 1 x 10^(-4)) in previous European GWAS analyses3-6. 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 NKKX2-3. The evidence of shared IBD risk loci across diverse populations

---

1Wellcome Trust Sanger Institute, Hinxton, UK. 2Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. 3Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. 4Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. 5Department of Medicine and Therapeutics, Institute of Digestive Disease, LKS Institute of Health Science, State Key Laboratory of Digestive Disease, Chinese University of Hong Kong, Hong Kong. 6Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan. 7Inflammatory Bowel Disease Research Group, Addenbrooke’s Hospital, Cambridge, UK. 8Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK. 9Digestive Disease Research Institute, Shariati Hospital, Tehran, Iran. 10Icahn School of Medicine, Mount Sinai Hospital, New York, New York, USA. 11Department of Gastroenterology, Emam Hospital, Tehran, Iran. 12Laboratory for Genotyping Development, Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan. 13Inflammatory Bowel Disease Unit, St Mark’s Hospital, Harrow, UK. 14National Institute of Immunology, New Delhi, India. 15Department of Genetics, University of Delhi South Campus, New Delhi, India. 16Department of Biostatistics, University of Liverpool, Liverpool, UK. 17Manchester Centre for Genomic Medicine, University of Manchester and Central Manchester University Hospitals National Health Service (NHS) Foundation Trust, Manchester, UK. 18Department of Medicine, Dayanand Medical College and Hospital, Ludhiana, India. 19Department of Gastroenterology and Hepatology, St. Mary’s Hospital, London, UK. 20Department of Gastroenterology and Hepatology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea. 21Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. 22Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. 23These authors contributed equally to this work. 24These authors jointly supervised this work. Correspondence should be addressed to R.K.W. (r.k.weersma@umcg.nl) or C.A.A. (carl.anderson@sanger.ac.uk).

Received 30 October 2014; accepted 24 June 2015; published online 20 July 2015; doi:10.1038/ng.3359
suggests that combining genotype data from cohorts of different ancestry will enable the detection of additional IBD-associated loci. Such trans-ancestry association studies have successfully identified susceptibility 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. In comparison to our previously published GWAS meta-analysis, this study includes an additional 11,535 individuals of European ancestry and 9,846 individuals of non-European ancestry. Using these data, we aim to identify new IBD risk loci and compare the genetic architecture of IBD susceptibility across ancestrally divergent populations.

RESULTS
Study design
After quality control and 1000 Genomes Project imputation (Phase I–August 2012), we used 5,956 Crohn’s disease cases, 6,968 ulcerative colitis cases and 21,770 population controls of European descent to perform GWAS of Crohn’s disease, ulcerative colitis and IBD (Crohn’s disease and ulcerative colitis together) (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 and Supplementary Figs. 1 and 2), so principal-component analysis was used to assign individuals to 1 of 4 ancestral groups (European, Iranian, Indian or East Asian) (Supplementary Fig. 3). Case-control association tests were performed within each ancestry group using a linear mixed model (MMM) (Online Methods). A fixed-effects meta-analysis was undertaken to combine the summary statistics from our European-only GWAS meta-analysis with those from the European replication cohort. We next performed a Bayesian trans-ancestry meta-analysis, as implemented in MANTRA, to enable heterogeneity in effect sizes to be correlated with the genetic distance between populations, as estimated by the mean fixation index ($F_{ST}$) across all SNPs (Online Methods). For the trans-ancestry 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 Fig. 2). To maximize power for our solely Immunochip-based comparisons across ancestral groups, the mixed-model association analysis was repeated after reinstating these individuals in the Immunochip cohort.

Trans-ancestry 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 the trans-ancestry meta-analysis that included all ancestries ($\log_{10}$(Bayes factor) > 6) for ulcerative colitis, Crohn’s disease or IBD (Table 2, Supplementary Figs. 4–7 and Supplementary Tables 1 and 2). To reduce false positive associations, we required all loci only implicated in disease risk via the trans-ancestry meta-analysis ($\log_{10}$(Bayes factor) > 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 and Supplementary Table 3).

Twenty-five of the 38 newly associated loci overlapped with loci previously reported for other traits, including immune-mediated diseases, whereas 13 had not previously been associated with any disease or trait (Online Methods and Supplementary Table 4). A likelihood-modeling approach showed that 27 of the 38 newly identified loci were associated with both Crohn’s disease and ulcerative colitis (designated here as IBD-associated loci), with 7 of these loci demonstrating evidence of heterogeneity of effect between the 2 diseases. Of the remaining 11 loci, 7 were classified as specific to Crohn’s disease and 4 were classified as specific to ulcerative colitis (Table 2 and Supplementary Table 1).

As a result of our updated sample quality control procedure, 17 of the 194 independent SNPs reported at genome-wide significance in our previous European-only GWAS meta-analysis failed to reach this significance 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 (FDR) of ~0.001) (Supplementary Table 1). 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 adopted the approach taken by Jostins et al, we observed a more significant $P$ value of $7.38 \times 10^{-6}$ for this SNP. This observation, together with the divergent allele frequencies at this SNP across European populations (1000 Genomes Project release 14: GBR (British in England and Scotland), 0.20; CEU (Utah residents of Western European ancestry), 0.28; IBS (Spanish Iberian), 0.39; FIN (Finnish), 0.47) suggests that the previously reported signal of association might have been driven, at least in part, by population stratification (which is now 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-associated SNPs originally identified in our previous European-only GWAS meta-analysis replicated in at least one non-European cohort if we considered a one-tailed Bonferroni-corrected significance threshold of $P < 6.1 \times 10^{-4}$ (0.05/163) (Supplementary Table 1). Nine of the 14 non-HLA loci (10 for Crohn’s disease and 4 for ulcerative colitis) that had been identified at genome-wide levels of significance in previous non-European GWAS cohorts from Japan, India and Korea were associated with either Crohn’s disease or ulcerative colitis in the East Asian, Indian and/or Iranian cohorts with $P < 1.0 \times 10^{-5}$ (Supplementary Table 5). 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 at $P = 2.64 \times 10^{-3}$ in the current East Asian cohort but was strongly associated with IBD in the European cohort ($P = 1.04 \times 10^{-18}$).

We next performed a series of analyses to prioritize genes within the newly associated loci for causality. cis-eQTL (expression quantitative trait locus) analysis from two data sets of peripheral blood samples from a total of 1,240 individuals showed that 12 of the 38 newly associated SNPs had cis-eQTL effects (FDR < 0.05) (Online Methods and Supplementary Table 6). Two SNPs showed trans-eQTL effects.
SNP rs653178 in a locus harboring SH2B3 and ATXN2 is associated with multiple other immune-mediated diseases, including celiac disease and rheumatoid arthritis. It had trans-eQTL effects on 14 genes, including genes within IBD-associated loci (TAGAP and STAT1). rs616597 had a cis-eQTL effect on NFKBIZ and had trans-eQTL effects on FLXB13 (Supplementary Table 6) (ref. 14). Both SNPs reside in known DNase I hypersensitivity and histone modification sites in multiple cell lines (Supplementary Table 7). In contrast to the high number of SNPs tagging eQTLs, only 3 of the 38 SNPs were in high linkage disequilibrium (LD; r² > 0.8) with known missense coding variants (Supplementary Table 8).

To enable a meaningful comparison with our previously published results, we recreated the GRAIL connectivity network using all loci that now achieved genome-wide significant evidence of association (Supplementary Fig. 8). Twelve genes in the previous GRAIL network were removed in this new network. We found that these genes had significantly larger GRAIL P values (Wilcoxon P value = 6 × 10⁻⁴) and fewer interaction partners (11.2 versus 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 were located within the newly associated loci (including NFKBIZ, CD28 and OSMR). Thus, 26 genes from previously established IBD loci are brought into the network.
null
The concordant direction of effect at associated SNPs, there was high genetic correlation ($r_G$) between the European and East Asian cohorts when considering the additive effects of all SNPs genotyped on the Immunochip (Supplementary Table 10). Given that rare SNPs (minor allele frequency (MAF) < 1%) are more likely to be specific to a particular population, 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 were small in comparison to the East Asian cohort, we observed similar trends for homogeneity of odds ratios at associated loci (Supplementary Figs. 9 and 10) and high genetic correlations with the European cohort (Supplementary Table 10). Together with the strong effect size correlations at known risk loci, these results indicate that the majority of IBD risk loci are shared across ancestral populations. Therefore, ancestry-matched groups of IBD cases and controls can be combined from divergent populations to mass 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 were explained by different allele frequencies across populations. For instance, consistent with previous genetic studies of Crohn’s disease in East Asians (rs6588248, MAF = 0.39; rs7517847, MAF = 0.04), these two secondary variants are common in East Asians (rs6588248, MAF = 0.39; rs7517847, MAF = 0.04), and, assuming the effect sizes observed in Europeans, we had 100% power to detect association with rs7517847 at $P < 5 \times 10^{-8}$ but only 84% power to detect association with rs6588248 at $P = 0.05$. Therefore, we cannot rule out the possibility that rs6588248 is involved in Crohn’s disease susceptibility in East Asians. Both variants showed significant heterogeneity of effect between the European and East Asian Crohn’s disease cohorts ($P < 2.44 \times 10^{-4}$). However, IL23R clearly has a role in IBD in the East Asian population, as evidenced by the association at rs76418789 with both Crohn’s disease and ulcerative colitis in Koreans (Supplementary Table 5) (ref. 4). This variant, which has a much lower allele frequency in Europeans (MAF = 0.004) than East Asians (MAF = 0.07), demonstrated suggestive evidence of association with IBD in Europeans ($P = 3.99 \times 10^{-4}$; OR = 0.66) and became genome-wide significant ($P = 2.31 \times 10^{-10}$; OR = 0.53) after conditioning on the three known European risk variants (rs1209026, 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 (Fig. 2 and Supplementary Fig. 10). For example, at ATG16L1, the reported Crohn’s disease risk variant in Europeans (rs12994997) had a RA of 0.53 and an OR of 1.27. The variant showed no evidence of association in East Asians ($P = 0.21$), a finding driven at

---

**Figure 1:** Comparison of odds ratios for Crohn’s disease and ulcerative colitis risk variants in Europeans and East Asians. (a, b) For each SNP, odds ratios (on a log scale) were estimated within each population for Crohn’s disease (a) and ulcerative colitis (b). The color 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(OR) in East Asians. Significance and goodness-of-fit are shown in red.
least in part by a significant difference in allele frequency (RAF = 0.24 in East Asians; $F_{ST} = 0.15$). 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 observed evidence of heterogeneity of odds at this SNP (East Asian $OR = 1.06$; $P = 8.45 \times 10^{-4}$). The previously reported lead SNP at the $IRGM$ locus in Europeans also showed only nominally significant evidence of association with Crohn’s disease in East Asians ($rs11741861$: European $P = 5.89 \times 10^{-44}$, East Asian $P = 2.62 \times 10^{-3}$) as well as evidence of heterogeneity of effect (European $OR = 1.33$ versus East Asian $OR = 1.13$; heterogeneity $P = 1.20 \times 10^{-3}$). However, not all loci demonstrating significant heterogeneity of odds had lower effect sizes in the non-European cohort: two of the three independent signals at $TNFSF15-TNFSF8$ had much larger effects on IBD risk in East Asians ($rs4246905$: European $OR = 1.15$ and East Asian $OR = 1.75$; $rs13300483$: European $OR = 1.14$ and East Asian $OR = 1.70$), despite similar allele frequencies in the two populations.

The third European risk variant was not significantly associated in East Asians ($rs11554257$: $P = 0.21$), although this might reflect a lack of power (76% power to detect this variant at $P < 0.05$ when assuming an identical odds ratio).

Although the incidence of IBD is rising in developing countries, comparable data on the clinical phenotype of disease in European and non-European populations are limited. We collected subphenotype data on 4,686 patients with IBD from East Asia, India and Iran and compared these data with available clinical phenotypes for 35,128 Europeans. Given that the current cohort is the largest available for clinical comparisons of IBD in Europeans and non-Europeans, we performed basic comparative statistical analyses. Overall, our data showed some demographic differences between the European and non-European populations, with male predominance in Crohn’s disease (67% of non-European patients with Crohn’s disease were male in comparison to 45% of European patients; $P = 7.09 \times 10^{-78}$). Furthermore, we observed more stricturing behavior ($P = 2.02 \times 10^{-33}$) and perianal disease ($P = 5.36 \times 10^{-33}$) and less inflammatory Crohn’s disease.

Figure 2 Comparison of variance explained per risk variant for Crohn’s disease and ulcerative colitis between East Asians and Europeans. (a, b) 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 association $P < 0.01$ 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 frequency ($F_{ST} > 0.1$/monomorphic in East Asians), significant heterogeneity of odds ratios ($P < 2.5 \times 10^{-4}$) or both.
findings are in line with previously reported prospectively collected data. Although these data have been collected retrospectively, the current findings are in line with previously reported prospectively collected clinical findings in incident cases of IBD in non-Europeans².

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 that trans-ancestry 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-ancestry association studies represent a powerful means of identifying new loci in complex diseases such as IBD. Furthermore, the nearly complete sharing of genetic risk among individuals of diverse ancestry has important consequences for association studies and disease risk prediction in non-European populations. First, a significant association in one population makes the locus in question a very strong candidate for involvement in IBD risk worldwide. Second, our data suggest that odds ratios 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 substantially 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 specific to a particular population, the substantial 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 loci²⁰–²².

Although the majority of risk-associated 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 hinder our ability to identify association with sites that are monomorphic in Europeans but polymorphic in non-Europeans. Targeted 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 IBD²³,²⁴. The much smaller number of individuals in the non-European cohorts also reduces power to detect heterogeneity of effect versus the European cohort, and we therefore may be overestimating the degree of sharing between the various ancestry groups.

In addition to allele frequencies differing between ancestral populations, patterns of LD can also vary greatly; such differences further complicate comparisons of genetic architecture for complex disease 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. Although this hypothesis is attractive, the heterogeneity in effect sizes could also be underpinned by differential tagging of untyped causal variants at these loci in one or both populations. Although the 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 the 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 the association tests), we cannot rule out the possibility that differential tagging of untyped causal variants is driving the observed heterogeneity in effects.

In summary, we have performed the first trans-ancestry association study of IBD and identified 38 risk loci, increasing the number of known IBD risk loci to 200. Together, these loci explain 13.1% and 8.2% of the variance in disease liability for Crohn’s disease and ulcerative colitis, respectively. The majority of these loci are shared across diverse ancestry groups, with only a handful demonstrating population-specific effects driven by heterogeneity in RAF (for example, NOD2) or effect size (for example, TNFSF15-TNFSF8). Concordance in direction of effect is significantly enriched among SNPs demonstrating only suggestive evidence of association, indicating that larger trans-ancestry association studies may represent a powerful means of identifying more risk loci for IBD. By leveraging imputation based on tens of thousands 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.


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

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

ACKNOWLEDGMENTS
R.K.W. is supported by a VIDI grant (016.136.308) from the Netherlands Organization for Scientific Research (NWO) and the Broad Medical Research Program of the Broad Foundation (IBD-0318). L.F. is supported by the Netherlands Organization for Scientific Research (NWO), through NWO VENI grant 916.10.135 and NWO VIDI grant 917.14.374. The research leading to these results has received funding from the European Community’s Health Seventh Framework Programme (FP7/2007–2013) under grant agreement 239867. T.B.K. is supported by Centre of Excellence grant BT/01/COE/07/UDSC/2008 from the Department of Biotechnology of the government of India (New Delhi, India). The collection of Iranian samples has been supported by the Iranian Medical University of Sciences, Iran. UK case collections were supported by the National Association for Colitis and Crohn’s Disease, the Wellcome Trust, the Medical Research Council UK and the Peninsular College of Medicine and Dentistry, Exeter. We also acknowledge the Wellcome Trust for Health Research (NIHR) Biomedical Research Centre awards to Guy’s and St Thomas’ NHS Trust/King’s College London and to Addenbrooke’s Hospital/University of Cambridge School of Clinical Medicine. A.P.M. is supported by the Wellcome Trust under award WT098017. J.Z.L., T.S., J.C.B. and C.A.A. are supported by the Wellcome Trust (098051).

AUTHOR CONTRIBUTIONS
COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

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

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 quality control are described in detail in Jostins et al.8. Briefly, seven Crohn’s disease and eight ulcerative colitis collections with genome-wide SNP data were combined. Samples were genotyped on a combination of the 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 quality control, 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 the data for Crohn’s disease and ulcerative colitis combined (IBD) 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 prephasing/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 data set (August 2012; 30,069,288 variants, release v3.macGT1).

Association analysis. Genome-wide association analyses were carried out for Crohn’s disease, ulcerative colitis and IBD (the Crohn’s disease and ulcerative colitis cases combined). After applying filters requiring MAF > 1% and imputation INFO score > 0.6 to all imputed variants, around 9 million variants were found to be suitable for association analysis. Association tests were carried out in PLINK, using the post-imputation genotype dosage data and using 10, 7 and 15 principal components for Crohn’s disease, ulcerative colitis and IBD, respectively, as covariates, chosen from the first 20 principal components. The Crohn’s disease, ulcerative colitis and IBD scans had genomic inflation (λ_{GC}) 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 λ_{GC} values are equivalent to λ_{GC,1,000} (the inflation 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, quality control and analysis. Description of the Immunochip. The Immunochip is an Illumina Infinium microarray comprising 196,524 SNPs and small indel markers selected on the basis of results from GWASs of 12 different immune-mediated diseases. The Immunochip enables replication of all nominally associated SNPs (P < 0.001) from the index GWAS scans and fine mapping of 186 loci associated at genome-wide significance with at least 1 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 Wellcome Trust Case Control Consortium 2 (WTCCC2) project replication phase. These SNPs are useful for quality control purposes because they have not previously been associated with immune-mediated diseases (‘null’ SNPs).

Cohorts of European ancestry. 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 radiological, endoscopic and histopathological evaluation. All included cases fulfilled 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 included 3,424 Crohn’s disease cases, 3,189 ulcerative colitis cases and 7,379 population controls present in the GWAS cohort. The overlapping Immunochip markers were excluded from the trans-ancestry association analysis but included in the modeling of European versus non-European IBD because this was based solely on Immunochip data.

Cohorts of East Asian, Indian and Iranian ancestry. East Asian patients with IBD 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) and Hong Kong (Chinese University of Hong Kong). Indian IBD cases and controls were recruited from Dayanand Medical College and Hospital, Ludhiana, and the 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 that clustered with a non-European cohort in principal-component analysis 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 Figs 1 and 2, and Supplementary Table 12).

Phenotype data. Detailed phenotype data (including sex, ancestry, 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 11). Disease location and behavior were assessed with the Montreal classification. Clinical demographics and disease phenotype were compared in the European and non-European cohorts using χ² 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 program40 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 (in comparison 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 was performed separately in each population (East Asian, Iranian, Indian and European) using PLINK41. Individuals were assigned to populations on the basis of principal-component analysis. This analysis was performed using EIGENSTRAT42 on a set of 15,552 Immunochip SNPs that had pairwise r² < 0.2 and MAF > 0.05 and were present in 1000 Genomes Project Phase 2 data. The first two principal components were estimated for the 1000 Genomes Project individuals and projected onto all Immunochip cases and controls. As expected, a clear separation between the different populations was seen (Supplementary Fig. 3). Samples were assigned to the population with which they clustered, and those that did not cluster with any of the reported populations were removed.

Marker quality control. SNPs were removed if they (i) were not on autosomes; (ii) had a call rate lower than 98% across all genotyping batches in the population and/or lower than 90% in one of the genotyping batches; (iii) were not present in 1000 Genomes Project Phase 1 data; (iv) failed Hardy-Weinberg equilibrium (FDR < 1 × 10⁻⁵) across all samples or within each genotyping batch; (v) had heterogeneous allele frequencies between the different genotyping batches within one population (FDR < 1 × 10⁻⁵); in genotyping batches with more than 100 samples; (vi) had different missing genotype rates for cases and controls (P < 1 × 10⁻³); and (vii) were monomorphic in the population. After marker quality control, 125,141 SNPs remained in the East Asian data set, 145,857 SNPs remained in the Indian data set, 152,232 SNPs remained in the Iranian data set and 144,245 SNPs remained in the European data set.

Sample quality control. Samples with a low call rate (<89%) and samples with an 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 identity by descent of >0.8 were considered duplicates, and pairs with identity by descent of >0.4 and <0.8 were considered related. For all duplicate and related pairs, the sample with the lowest genotype call rate was removed. After sample quality control, 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 (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 r² < 0.2). Of the remaining SNPs, we then removed those that lay in the HLA region or had