doi:10.1038/nature11582

Supplementary methods:

Other supplementary material:

1. Generation of IBD locus list

1a. GWAS data and analysis

i. Cohorts and samples

Seven Crohn's disease collections and eight ulcerative colitis collections with genome-wide SNP genotype data were used in this analysis (Supplementary Table 1, tab 1). The CD cohorts contained a total of 6299 cases and 15148 controls, and the UC cohorts contained a total of 7211 cases and 20783 controls (the control sets contain largely overlapping samples). Four different chips were used: two produced by Affymetrix (the GeneChip Human Mapping 500K Array and the Genome-Wide Human SNP Array 6.0) and two produced by Illumina (the HumanHap300 BeadChip and the HumanHap550 BeadChip). The majority of these samples were incorporated into our previous meta-analyses using summary statistics only[1-2].

ii. QC and imputation

Technical quality control was performed on genotypes generated by various GWAS platforms, with quality control conducted on each dataset separately using a common approach. In addition to previously reported QC on each dataset, the following quality control parameters were applied: (i) missing rate per SNP < 0.05 (before sample removal below), (ii) missing rate per individual < 0.02, (iii) heterozygosity per individual ± 0.2, (iv) missing rate per SNP < 0.02 (after sample removal above), (iv) missing rate per SNP difference in cases and controls < 0.02, (vi) Hardy-Weinberg equilibrium (in controls) $P < 10^{-6}$, (vii) Hardy-Weinberg equilibrium (in cases) $P < 10^{-10}$. Study sample sizes varied between 270 and 8,000 individuals (Supplementary Table 1, tab 1). The number of SNPs per study after quality control varied between 290,000 and 780,000. On average, the quality control processes excluded 11 individuals per study (with a range of 0–57 individuals) and 20,000 SNPs per study (with a range of 2,000–180,000 SNPs). These exclusions suggest that while previous QC was sufficient in the sample dimension, tighter QC should have been applied in the SNP dimension (details below).

After quality control, the GWAS datasets together comprised 49,441 individuals and, for the next steps of the 'genetic quality control' analysis, a set of 21,681 SNPs common to all platforms and successfully genotyped in each GWAS sample was extracted. These SNPs were then further pruned to remove LD (leaving no pairs with $r^2 > 0.05$) and lower frequency SNPs (minor allele frequency < 0.05), leaving 17,385 SNPs suitable for robust relatedness testing and population structure analysis (see below).

Imputation of untyped SNPs was performed within each study in batches of 300 individuals. These batches were randomly drawn in order to keep the same case-control ratio as in the total sample from that study. We used Beagle 3.1[3]. Imputation was performed with CEU+TSI HapMap phase 3 data (UCSC hg18/NCBI 36) using a chunk size of 10Mb with 410 phased haplotypes comprising 1,252,901 SNPs, using default parameters. λ was carefully monitored before and after imputation.

Genetic quality control included relatedness testing and principal components analyses based on 17,385 LD independent SNPs, present on all platforms in this study. Relatedness testing was done with PLINK[4], reporting pairs with genome identity (pi-hat) > 0.9 as 'identical samples' and with pi-hat > 0.2 as being closely related. After random shuffling, one individual from each related pair was excluded from downstream analysis. From groups with multiple related pairs (for example, a family), only one individual was kept.

iii. Principal component analysis

Principal component estimation was done with the same collection of SNPs on the non-related subset of individuals. We estimated the first 20 principal components and tested each of them for phenotype association (using logistic regression with study indicator variables included as covariates) and evaluated their impact on the genome-wide test statistics using λ (the genomic control inflation factor based on the median χ^2) after genome-wide association of the specified principal component. Based on this we decided which principal components to include (e.g. 1,2,3,4,5,6,7,8,9,12,13,14,18,19 and 20 for IBD) for downstream analysis as associated covariates (Supplementary Figure 1).

iv. Association analysis

A genome-wide association analyses was carried out for Crohn's disease (CD), ulcerative colitis (UC) and all inflammatory bowel disease (IBD). The CD and UC scans used only the CD and UC cohorts, and the IBD scan used all cohorts with duplicates across CD and UC cohorts removed (as described above). The CD scan had a total of 5,956 QC+ cases and 14,927 QC+ controls, the UC scan had 6,968 cases and 20,464 controls, and the IBD scan had 12,882 cases and 21,770 controls.

Association testing was carried out in PLINK, using the dosage data from the imputation and using 10, 7, 15 principal components for CD, UC, IBD respectively as covariates, chosen as described above from the first 20 principal components. The CD, UC and IBD scans had genomic inflation (λ_{GC}) values of 1.137, 1.129, and 1.169 respectively (Supplementary Figure 2).

1b. Immunochip data and analysis

i. Description of the Immunochip

The Immunochip is a custom Illumina Infinium chip comprising 196,524 SNPs and small indels selected primarily based on GWAS analysis of 12 autoimmune and inflammatory diseases. The chip has two purposes: fine mapping of 289 established associations corresponding to 187 distinct loci, and deep replication of suggestive, but not yet proven, associations. Fine-mapping regions were defined as 0.2cM centered on GWAS hit SNPs, and all SNPs and short indels in these regions from the 1000 Genomes Project low coverage pilot CEU samples[5], as well as variants discovered in resequencing experiments conducted by groups collaborating in the chip design were selected for inclusion[6]. Replication of autoimmune and inflammatory GWAS (including Crohn's disease and ulcerative colitis) contributed the bulk of the remaining SNP lists. Approximately 25,000 SNPs were included as replication of

unrelated diseases as part of the WTCCC2 project, which serve as useful null SNPs for these analyses. In total, approximately 240,000 SNPs were selected for inclusion, with an assay design success rate of ~80%.

ii. Cohorts, samples and genotyping

Sample collections from 15 countries were genotyped using the Immunochip in 11 different genotyping centers (Supplementary Table 1, tab 2). Genotyping was performed in 20 batches, with each center performing between one and three batches. A total of 60,828 samples were genotyped on the Immunochip, including 20,076 CD cases, 15,307 UC cases and 25,445 controls. These numbers include many samples that were also present in the GWAS cohorts, which are to be used for fine mapping and not for locus discovery. Samples with a mean intensity outside a 95% confidence interval were removed.

iii. Genotype calling and QC

Because many of the variants on Immunochip do not meet the manufacturer's quality standards set for GWAS products, rigorous QC was essential. Furthermore, because samples with poor quality DNA or with other genome-wide problems can adversely affect the genotype calls at high quality samples, a first stage of "coarse" QC was performed on genotypes called using Illumina's GenomeStudio program. Samples with >5% missing data, genome-wide heterozygosity outside a 95% confidence interval in each batch (Bonferroni corrected for sample size), samples of non-European ancestry (see below, Supplementary Figure 3) or with abnormal mean intensity values were removed from further analysis.

Normalized intensities for all remaining samples in all batches were then centrally called using the optiCall clustering program[7] version 0.3.0 with HWE blanking disabled and no-call cutoff set to 0.7. Duplicate and related samples (PI_HAT > 0.1) were identified using the pairwise IBD calculator in PLINK applied to a set of SNPs in linkage equilibrium (also used for PCA, see below for details). The duplicate or relative sample with more missing data was removed. A set of 692 SNPs present on both the Immunochip and all four GWAS chips were also used to remove Immunochip samples that were also present in the GWAS. Samples without a phenotype definition of Crohn's disease, ulcerative colitis or healthy control were removed. Finally, samples with > 2% missing data in this improved dataset were removed.

SNP QC was performed on the sample-clean dataset described above, and SNPs with >2% missing data or HWE p-value < 10^{-10} in controls were removed. To further ensure the quality of genotype calls in our analysis we performed 3-fold manual inspection[8] of 3,356 variants, including those with meta analysis p<10^{-5} which fulfilled at least one of the following criteria: (a) Cochran heterogeneity p < 0.01 between GWAS and Immunochip (N=871), (b) lie outside fine-mapping regions known to be associated with immune-mediated disease (N=797), (c) are one of the 3 most significantly associated SNPs in a fine-mapping region (N=851), (d) any SNP with p < 5×10^{-8} which did not fit those criteria (N=195), (e) random SNPs as a comparator (N=642). 1015 SNPs were removed after failing manual QC, and 29 had genotypes manually adjusted (blind to phenotype and association statistics) to correct "recoverable errors".

iv. Principal component analysis

As with the GWAS data, principal component analysis was used to identify ethnic outliers, and to generate covariates to control for population stratification.

To identify outliers on the continental scale we used a reference set consisting of 662 HapMap founder samples genotyped on the Illumina Human1M, the Affymetrix Human SNP Array 6.0, and the Illumina Omni2.5 for the HapMap3 and 1000 Genomes Projects. This reference set has a total of 3,268,731 SNPs, of which 83,689 are present on the Immunochip. PLINK was used to LD prune the data such that no pair of SNPs had $r^2 > 0.2$, and to remove potentially problematic GC/AT SNPs, SNPs within known high LD regions[9] and SNPs with MAF < 5%. We projected the Immunochip samples on the principal component axes generated using these 17,891 SNPs from the 662 reference samples using the R package snpMatrix[10]. All samples that did not cluster with the European samples were excluded (Supplementary Figure 3).

To resolve within-Europe relationships, we performed PCA within the remaining Immunochip samples. LD pruning was performed within European controls (this was performed three times, to properly break up the LD in fine-mapping regions), and SNPs present in high LD regions or with MAF < 5% were removed, leaving a total of 19,111 SNPs. Principal component axes were generated within the controls, and projected onto the cases to generate principal components for all samples. The first four principal component axes seemed to capture significant population structure (Supplementary Figure 3), and addition of components beyond the fourth as association covariates in a subset of the Immunochip data did not further reduce the genomic inflation factor.

v. Association analysis

Immunochip-wide association analyses were carried out for Crohn's Disease (CD), Ulcerative Colitis (UC) and all inflammatory bowel disease (IBD). The CD, UC and IBD scans all used the entire control dataset. The CD scan had a total of 14,763 QC+ cases, the UC scan had 10,920 cases, the IBD scan had 25,683 cases, and all scans used the 15,977 QC+ controls.

Association tests were performed using additive logistic regression in PLINK conditional on the first four principal components (see above). Test statistic inflation was computed from a set of 3,120 SNPs chosen based on GWAS of schizophrenia, psychosis and reading/mathematics ability. Genomic inflation factors were relatively low, given the large sample size and presence of polygenic risk: $\lambda_{GC\_CD} = 1.353$, $\lambda_{GC\_UC} = 1.154$, $\lambda_{GC\_IBD} = 1.234$ (Supplementary Figure 4). This residual inflation may reflect additional polygenic risk showing weak association in our large sample size, as indicated by very low values of λ_{1000}, the equivalent inflation factor for a study with 1000 cases and 1000 controls: CD=1.023, UC=1.012, IBD=1.012.

For comparison, we also performed an association test on all IBD samples using the Cochran-Mantel-Haenszel method stratified by country of origin of the samples. This is one of the methods used to analyze standard GWAS replication, where PCs are usually not available. The genomic inflation value for the IBD all analysis was

λ_{GC_IBD} = 2.00, showing that without the genome-wide SNP data on the Immunochip this replication analysis would have shown severe inflation.

1c. GWAS/Immunochip replication meta-analysis

A combined analysis was performed using both the GWAS and the Immunochip association results comprising 20,700 Crohn's disease cases, 17,865 ulcerative colitis cases and 37,747 healthy controls.

i. Analysis of primary signals

All SNPs that showed $p < 0.01$ in the CD, UC or IBD GWAS scans were selected for replication in the Immunochip dataset. A fixed-effect meta-analysis was performed using odds ratios and standard errors from the GWAS hit and the Immunochip tag SNP with the highest r^2 to the GWAS hit. If no Immunochip tag SNP with $r^2 > 0.4$ was available we did carry the signal forward for replication. The Cochran heterogeneity p-value was also calculated, and SNPs with low p-values in this test were manually inspected (see above).

SNPs with $p < 5 \times 10^{-8}$ in any of the three phenotypes in this analysis were combined into clumps if they had $r^2 > 0.1$. SNPs within these clumps were tested for evidence of association independent of the strongest signal in the clump by calculating an approximate conditional Z-score

$$Z_i' = Z_i - r_{i/hit} Z_{hit}$$

Where Z_i is the Z score of the SNP being tested, Z_{hit} is the Z score of the strongest signal in the clump, and $r_{i/hit}$ is the correlation coefficient between the strongest signal and the SNP being tested. If $P(Z_i' > 0) < 5 \times 10^{-8}$ then this clump was considered to have a secondary signal, and the SNP with the Z_i' largest in magnitude was recorded as a secondary signal in this clump. All other SNPs in the clump were then tested for a tertiary signal independent of the first two, using

$$Z_i' = Z_i - r_{i/hit} Z_{hit} - r_{i/2nd} Z_{2nd}$$

We did not test for additional signals after the third. Theoretically, this could be extended to an arbitrary number of signals, but the approximation will become less accurate as additional signals are tested for.

This approach yielded 193 genome-wide significant independent signals of association. None of these signals had significant heterogeneity of effect size, and all had their Immunochip intensity cluster plots manually inspected to ensure that they were well clustered.

ii. Combination of independent signals into loci

The large number of independent signals (193) makes categorizing them into functionally separate loci problematic. We conventionally define signals as coming from the same locus if their lead SNPs lie within a certain physical or genetic distance of each other. However if this physical distance parameter is too large functionally

independent signals that are adjacent by chance may be incorrectly combined. Conversely, selecting too small a distance parameter could cause variants that act relatively proximately on the same gene to be split into independent loci.

To test the effect of this distance parameter on classification of signals into loci, we performed a null simulation. We selected randomly from the PCA SNPs to simulate null signals, and examined what proportion of signals were incorrectly merged together for a given distance parameter value. Based on this, we defined loci as 500kb units: 250kb on either side of the hit SNP. This was predicted to result in between 95% and 99% of null loci being correctly separated (Supplementary Figure 5).

We calculated an associated region for each of the 193 independent signals, defined as 250kb on either side of the hit SNP, or the extent of LD (defined as the positions of the furthest up-and-downstream variants with $r^2 > 0.5$ to the hit SNP). Overlapping regions were merged together, providing that they were associated to compatible phenotypes under the likelihood analysis (see below); i.e. regions were not merged if one was uniquely associated with CD, and the other uniquely associated with UC. The final merged regions were defined as loci, with their total extents being the maximum extent of their component signals. A total of 163 independent loci were thus defined (Supplementary Table 2, tab 1).

1d. Crohn's disease/Ulcerative colitis likelihood modeling

We used a likelihood modeling approach to classify signals into four categories according to their relative strength of association to CD and UC. We used a multinomial logistic regression model with additive log-odds ratio parameters β_{CD} and β_{UC}. The model was fitted to the Immunochip genotypes using the mlogit package in R.

We fit this model with four sets of parameter constraints:

1. CD-specific model: $\beta_{UC} = 0$, β_{CD} fitted by maximum likelihood
2. UC-specific model: $\beta_{CD} = 0$, β_{UC} fitted by maximum likelihood
3. IBD unsaturated (same-effect size) model: $\beta_{CD} = \beta_{UC} = \beta_{IBD,}$ β_{IBD} fitted by maximum likelihood
4. IBD saturated (different effect sizes) model: β_{CD} and β_{UC} both fitted by maximum likelihood

Note that models 1-3 are all constrained versions (with 1 degree of freedom) of model 4 (with 2 d.f.).

We calculated likelihoods for each model, and performed a likelihood ratio test of each of models 1-3 against model 4. If the likelihood ratio test had $p < 0.05$ for all 3 models (the 2 d.f. model is nominally significantly a better fit than any of the 1 d.f. models), we classified the signal as "saturated" (i.e. associated to both CD and UC, but with evidence of different effect sizes). Otherwise, we classified the signal according to which of the first three models had the largest likelihood (Supplementary Table 2, tab 2). Note that being classified as IBD unsaturated should be interpreted as "associated to both CD and UC, without significant evidence of different effect sizes".

In Table 1, the "IBD" section contains all loci where the main signal was classified as IBD unsaturated or IBD saturated. An exception was made for the CD associations at *PTPN22* and *NOD2*, where the correct model was "IBD saturated", as there were significant UC associations that went in the opposite direction to the CD effect.

Even within these classifications there is a significant variation in the balance of CD and UC effect sizes (see main text Figure 1b). To capture this variation, we also used polar-transformed log odds ratios as a continuous measure of CD vs UC effect size balance. This is defined as $\theta = \text{atan2}(\log(OR_{CD}), \log(OR_{UC}))$. Large values of θ correspond to associations with a stronger UC component, smaller values correspond to a stronger CD component.

1e. Comparison of this locus list to previous CD and UC lists

Because this study has access to raw genotype data from both CD and UC for the first time, it has allowed us to clarify several aspects of the 99 previously reported associations:

- While previously suspected, we have confirmed that the associations in the MHC are distinct for CD & UC (Supplementary Table 2, tab 2), and therefore should be split into two phenotype specific associations, rather than a single IBD locus.
- Conversely our improved imputation has re-localized the CD association previously reported as *VAMP3* to be the same effect as the adjacent previous UC association to *TNFRSF9*, making this a single IBD locus.
- Two previously independent associations on chromosome 2 near 102Mb (one CD, one UC) have both been shown to be associated to both CD and UC, and accordingly have been merged into independent effects in a single IBD locus. Similarly, a previous CD SNP (chromosome 2 near 198Mb), which is now associated to UC as well, was incorporated into a new nearby UC locus.
- Five previous associations (Chr2@198Mb, Chr5@36Mb, Chr6@3Mb, Chr6@44Mb, Chr13@42Mb) are no longer genome-wide significant. In four cases, our improved PCA-corrected analysis is >2 orders of magnitude less significant than the previous country-stratified analysis, suggesting that these associations may have been driven in part by uncorrected population structure. In the final instance the key SNP failed Immunochip design.

Thus, from 99 previously reported loci, one was split, three were merged and five were lost, leaving 92 established and 71 novel loci. This highlights both the overall robustness of our previous analyses as well as potential pitfalls in small-scale replication genotyping, for which correction for population stratification is difficult.

We also compared the total phenotypic variance of CD and UC explained by our loci compared to previously published estimates. In ulcerative colitis we improved from 4.1% of phenotypic variance explained by known loci to 7.5% explained by our 193 signals. For Crohn's disease we improved from 8.2% to 13.6%. Two additional comments are necessary: first, we have decided here to report phenotypic variance explained, rather than heritability, because of recent publications suggesting that it is challenging to accurately estimate narrow sense heritability in a way compared to our variance explained calculations. Second, the odds ratios estimated in this study are

smaller than previous estimates for several key loci in CD, including *NOD2*, *IL23R* and *ATG16L1*. This difference was not explained by stratification or differential ancestry, but because our new odds ratios are estimated in replication samples in this project, they may reflect less severe disease than the samples previously collected for GWAS.

1f. Epistasis analysis in IBD, UC and CD datasets

In order to search for statistical interactions between the most strongly associated SNPs from each of the identified genome-wide significant loci, we used logistic regression on an allele dosage model in R. For each pair of SNPs, the likelihood ratio test was employed to calculate a P value of the interaction term. This analysis was performed in the Immunochip dataset (for cases subdivided into IBD, UC and CD, and the complete control dataset) and the first four principal components were added as covariates. The analyses of the CD and UC subsets were inconclusive. The results for the analysis with IBD showed one suggestive result between SNPs near SLC7A10 (rs17694108) and IL2RA (rs12722515) with P value=3.26×10^{-5}. However, a physical interaction was not supported by the protein interaction analysis described in this paper. However, the proteins do seem to act downstream of one another (connected through physical interactions with ICAM1 and SLC3A2).

2. IBD genetics in the context of autoimmunity and infection

2a. Annotation of associations to other phenotypes

The IBD locus list was annotated with the NHGRI GWAS catalogue[11]. All associations with $p < 5x10^{-8}$ to any disease or primary phenotype were included. For each IBD locus we annotated all phenotypes that had at least one associated SNP within the IBD region. We also checked whether the hit SNP in the NHGRI catalogue was the same as, or in high LD with ($r^2 > 0.8$) the IBD hit SNP, and when this was the case the direction of association was checked to highlight cases where IBD went in to opposite direction to the other association (Supplementary Table 2, tab 3).

2b. Analysis of primary immunodeficiency (PID), complex autoimmune and immune-mediated disease (IMD) and IBD

To place the IBD loci in the context of other immune-related diseases, we generated lists of associations with other immune-related diseases. We included complex autoimmune and immune-mediated and diseases (IMD), and autosomal dominant or recessive primary immunodeficiencies (PID).

Autosomal dominant and recessive genes identified as causing PID were taken from Notarangelo *et al*[12]. Genes that lay within 250kb of each other were merged together into regions, giving 135 genes across 121 independent regions.

The associated regions for the IMD list were taken from the NHGRI GWAS catalogue, and included the following diseases: Primary sclerosing cholangitis, primary biliary cirrhosis, rheumatoid arthritis, type 1 diabetes, multiple sclerosis, celiac disease, atopic dermatitis, psoriasis, ankylosing spondylitis, asthma and systemic lupus erythematosus. All SNPs in the catalogue with $p < 5 \times 10^{-8}$ were included. Each SNP was given a region on 250kb on either side, and overlapping regions were merged together into loci (the same method used to define IBD loci in Section 1c). This generated a total of 156 independent IMD loci.

We assessed the overlap (Figure 2C, Supplementary Table 4) in regions in the three lists (IBD, PID and IMD). We calculated the statistical significance of the enrichment in overlaps using the method described below (Section 4a iv).

GO terms and pathways that were enriched or depleted in the IBD-unique set relative to the PID-unique and IMD-unique sets were detected using the method described below in section 4a.

3. Prioritizing causal genes within IBD loci

We performed a number of analyses designed to identify candidate genes within IBD loci (Supplementary Table 1, tab 4).

3a. GRAIL

GRAIL (Gene Relationships Across Implicated Loci)[13] is a network connectivity tool that uses text mining to calculate a network distance between genes in different implicated loci. Each gene is measured for enrichment of connectivity to genes in other associated loci measured, and a p-value is calculated.

We used the online GRAIL web tool to perform this analysis, using associated region definitions from Table 1. To reduce noise we removed associations within the HLA, and replaced regions around 4 well-established genes (*NOD2*, *IL23R*, *ATG16L1* and *PTPN22*) with the gene itself. We searched only among PubMed articles pre-2006 in order to avoid bias from the (now very large) literature that discusses or follows up the results of previous genome-wide association studies in IBD. We selected all genes with $p < 0.05$ as GRAIL implicated loci.

3b. DAPPLE

DAPPLE (Disease Association Protein-Protein Link Evaluator) is a network connectivity tool that uses protein-protein interactions (PPIs)[14]. Each gene is measured for enrichment in either direct or indirect (i.e. via other proteins) interactions with genes in other loci, and an empirical p-value is calculated by permutation.

We used the DAPPLE web tool to perform this analysis using association region definitions from Table 1. As with GRAIL, to reduce noise we removed associations within the HLA, and fixed the same 4 well-established genes as causal. We selected all genes with $p < 0.05$ as DAPPLE implicated loci.

3c. eQTLs

We identified genes whose expression showed evidence of being correlated with our associated hit SNPs, i.e. for which the IBD hit SNP is an expression quantitative trait locus (eQTL) for the gene.

We looked at three different sources of eQTLs.

1. The University of Chicago eQTL database (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl), containing eQTLs collected from a range of studies. We excluded eQTLs from studies of non-European individuals, where the patterns of LD may differ from our study. Data track kindly provided by Jacob Degner.
2. The Dixon *et al* eQTL dataset (http://www.sph.umich.edu/csg/liang/asthma/), containing eQTLs inferred from 400 lymphoblastoid cell lines of asthmatic children[15]. A p-value cut-off of $p < 10^{-5}$ was applied to the dataset.

3. The Merck Research Laboratories eQTL dataset, containing eQTLs of four tissues from 1000 morbidly obese patients[16]. A p-value cut-off of $p < 10^{-5}$ was applied to the dataset. Data kindly provided by Cong Li.

In order to take a broad view of the relationship between IBD risk and gene expression, we included data from both immune and non-immune tissue datasets. However, we recorded the tissue of discovery for each eQTL (Supplementary Table 2). The majority of eQTLs (25/37) were found in immune-related tissues (lymphoblastoid or monocyte).

We identified cases where the eQTL was the same as the IBD hit SNP, or where the eQTL was in LD with the IBD hit SNP ($r^2 > 0.8$). We selected all genes regulated by these eQTLs as eQTL implicated loci.

3d. Coding SNPs

We performed functional annotation to identify genes for which an IBD hit SNP was correlated with an amino-acid changing variant.

Functional annotation was performed using functionGVS (http://snp.gs.washington.edu/SeattleSeqAnnotation134/), using 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 ($r^2 > 0.8$) with a SNP with such a classification. The genes in which these missense variants lie were included as cSNP implicated genes.

3e. Co-expression network analysis

Genes in IBD loci implicated by the inflammatory adipose network described in section 5 were included as co-expression network implicated genes.

4. Functional analysis across IBD loci

4a. GO term and canonical pathway analysis

We performed various analyses looking at enrichment or depletion of functional terms (including GO terms and canonical pathways) in IBD loci as a whole, within subsets, or compared to other quantitative variables.

i. A note on bias within the Immunochip

The Immunochip was constructed using variant lists submitted by immune-related disease association consortia. We may therefore expect there to be a bias towards discovering loci that are associated to both IBD and other immune-related diseases. Such a bias could cause an artificial inflation in enrichment of immune-related GO terms. To investigate this possibility, we applied our GO enrichment analysis (described below) to two non-overlapping subsets of our loci: (i) the 92 loci described in our previous meta-analyses, and (ii) the 71 newly discovered loci. If our analysis for identifying new IBD loci were biased (via the Immunochip design) toward loci shared across autoimmune diseases we would expect larger enrichment odds ratios in set (ii) compared to (i). Supplementary Figure 8 shows that in fact, the opposite is true: our previous loci are, on average, slightly more strongly enriched than our new loci ($p = 2.2x10^{-9}$). This difference might suggest that the strongest IBD loci (i.e. those already known) play a more central role in key immune functions than our new discoveries.

This lack of observable bias, while initially surprising, can largely be explained by our experimental design, and the specifics of the SNP selection process for the Immunochip. As part of that design we included the top 2000 most associated SNPs each from the earlier CD and UC GWAS meta-analyses regardless of function or association with other phenotype (corresponding to $p < 0.0009$ for CD and $p < 0.0004$ for UC). This subset of SNPs therefore represents a functionally unbiased, genome-wide replication set that includes 147 (55 new, 92 known) of our 163 reported loci. Therefore the non-IBD immune disease focused part of the Immunochip contributed to only 16 of our loci – a number too small to bias our analyses as shown above.

ii. Description of enrichment methodology

We wish to assess the enrichment of a particular functional term (e.g. a GO term) in causal IBD genes. Given a list of causal genes, we could easily calculate an enrichment odds ratio λ_i of a functional term i in IBD genes relative to the genome as a whole, and perform a statistical test of $\lambda_i = 1$ vs $\lambda_i > 1$. However, we do not know the causal variant for most IBD regions, and most IBD regions contain multiple genes. To compensate for this, we use an extension of the standard odds ratio method that takes into account the presence of non-causal genes.

Assume that we have M loci, designated by $j = (1, \ldots, M)$ each of which contains N_j genes. For each associated locus j we set an indicator variable δ_{ij} to 1 if the functional

term *i* is present in locus *j*, and 0 otherwise. We also calculate a genome-wide frequency for term f_i that is equal to the proportion of all genes that contain the term *i*.

We calculate g_i, the frequency of term *i* in causal genes, given an enrichment odds ratio λ_i

$$g_i = \left(1 + \frac{1 - f_i}{\lambda_i f_i}\right)^{-1}$$

We then assume that all other genes have a frequency of the term f_i. Assuming that there is exactly one causal gene in the region, the log likelihood L_i is given by

$$L_i = \sum_j \delta_{ij} log\left(1 - (1 - f_i)^{N_j}(1 - g_i)\right) + \sum_j (1 - \delta_{ij}) log\left((1 - f_i)^{N_j}(1 - g_i)\right)$$

We fit the parameter λ_i by maximum likelihood using the Nelder-Mead, implemented in the statistical package R. We assess the significance of the parameter λ_i by performing a likelihood ratio test of of $\lambda_i = 1$ vs $\lambda_i \neq 1$.

iii. Extension to arbitrary predictors

We can extend the method above to include arbitrary predictors $X = \{x_{jk}\}$, by extending the definition of g_i to a generalized logistic model

$$g_i = \left(1 + \frac{1 - f_i}{f_i} exp(-\beta_0 - \beta X)\right)^{-1}$$

We keep the enrichment odds ratio (in this case as $\lambda_i = \exp(\beta_0)$) but also include terms for other predictors β. The predictors *X* can be discrete (e.g. x = 0 for UC, x = 1 for CD), or continuous (e.g. the polar-transformed odds ratio θ described above). The model is fitted by maximum likelihood in the same way as the simple enrichment model, and likelihood ratio tests can be used to assess the significances of the parameters.

iv. Extension to enrichment of interval overlap

We can extend the above methodology to assess the enrichment in overlap between sets of genomic intervals. Assume that our loci have lengths l_j, our genomic intervals have lengths l_k, and the total length of the genome is l_g We can use the equations in sections 4a.ii and 4a.iii above by setting

$$f_i = \frac{1}{l_g} \sum_k (l_k + l_i)$$

This extension enables us to evaluate the significance of the overlap between our IBD loci and GWAS associations (Figure 2A).

v. Terms, pathways and examined

We examined 15,526 human GO terms dated 28/02/2012. We also used canonical pathways (186 taken from KEGG, 430 taken from Reactome, and 217 from Biocarta).

When testing for enrichment we considered testing all genes in the 163 loci, or just those genes prioritized by at least one of the methods described in Section 2. Supplementary Figure 10 shows that the enrichments were substantially stronger in the prioritized list, which demonstrates the validity of our prioritization strategies and highlights the consistency between orthogonal functional annotations. Supplementary Figure 11 shows that the estimated odds ratios are slightly higher using this approach than using the entire gene list, suggesting this using these prioritized genes introduces a slight bias towards the detection of well-studied pathways, though this bias is relatively small. We therefore used only the prioritized list for enrichments discussed in the text.

4b. Immune cell enrichment analysis

To assess whether genes near risk alleles are specifically expressed in individual immune cell types, we used a separately published approach[17]. Here we present a summary of the approach.

We used two high quality gene expression datasets, a mouse dataset curated by the Immunological Genome Projects (ImmGen) and a human dataset curated by the Genomics Institute of the Novartis Research Foundation (GNF). We applied standard quality control and quantile-normalization to both datasets[18]. The ImmGen dataset consists of 223 mouse immune cell types from different lineages at multiple developmental stages, sorted by FACS and assayed in at least triplicate[19]. We mapped the mouse genes to 14,624 human homologous genes within the hg18/build36 of the human reference genome. The GNF dataset consists of 17,581 genes from 79 diverse human tissue types, including peripheral blood cells, neurological tissues, and tissues from visceral organs; each cell type was measured in duplicate[20].

We use SNPs from genome-wide association studies and cell-specific expression profiles to identify candidate pathogenic cell types in three major steps:

1. In order to assess cell-specific expression, we first divide the absolute expression of each gene in each cell type by the Euclidean norm of the vector of the gene's expression values across all cell types. To make this specificity score non-parametric, we rank the specificity of all genes in each cell type, and then convert each rank to a cell-specificity percentile score between 0 (most specific) and 1 (least specific).
2. We calculate a cell-specificity score for each SNP that is associated with a given disease. To do so, we first identify all genes that are implicated by a given SNP by defining a region containing the disease-associated SNP and all SNPs in LD using the same approach as GRAIL and DAPPLE (see above). All genes that overlap with this region are considered implicated by the SNP. In each cell type, we score each SNP based on the percentile of the most specifically expressed gene near that SNP. As the number of genes in LD with each SNP is variable, this locus score is adjusted for multiple hypothesis testing. Under the null, these "locus p scores" should be uniformly distributed between 0 and 1.

3. Finally, to score each cell type we take the log average of the locus p scores of all the disease-associated SNPs. To assess significance of this score, we match each set of disease-associated SNPs with sets of random SNPs from the genome-wide catalog that are 1) not known to be associated with diseases, 2) matched for the total number of SNPs, and 3) matched for the number of genes in LD with each SNP. We report an empirical p-value that equals the proportion of simulated p-values achieving higher significance than the analytical p-value.

Data used in this analysis are available from ImmGen (http://www.immgen.org/suggestions/dataRequest.do) and Novartis GNF (http://biogps.org/downloads/).

4c. Selection analysis

To test selection on IBD loci, we used data, provided by Joe Pickrell, generated using the TreeMix method developed by Pickrell and Pritchard[21] They constructed population trees from the Human Genetic Diversity Panel data, and produced a per-variant score that measures the extent to which population allele frequencies at that site are over-dispersed relative to this tree. The most over-dispersed sites are likely to have been subjected to directional (positive or negative) selection, whereas those that match the tree most closely are likely to have been subjected to balancing selection.

We picked the best tag SNP for each of our associated variants using LD information taken from HapMap (picking only the UC associated variant from the HLA), and extracted the scores for these variants. Because the score is confounded with allele frequency, we calculated empirical p-values for each variant as follows: pick all variants with an allele frequency within 1 percentage point of the hit variant's allele frequency, and measure the proportion of variants with a score greater than the score of the hit variant. We calculated p-values for directional selection (the proportion of variants with a scores higher than the hit variant), and p-values for balancing selection (the proportion with scores lower than the hit variant), as well as two-tailed p-values.

For set-based tests of selection, we used a Fisher's method combination of the empirical p-values to generate a set-wise empirical p-value.

In order to assess whether extent or direction of selection was correlated with specific functions, we used the GO term enrichment method described in section 4a. We converted our selection p-values to Z scores using an inverse normal transformation, and tested for association between these scores and GO terms.

5. Identifying causal directions of regulation among IBD loci: Co-expression module and causal network analyses

5a. Gene expression datasets

We explored expression in several tissues from a cohort which contained 950 patients who underwent Roux-en-Y gastric bypass surgery at Massachusetts General Hospital (MGH)[16]. Liver, subcutaneous adipose and omental adipose tissues were collected from each participant. Genomic DNA was isolated from liver tissues, and total RNA was extracted from each of the tissue types. Each RNA sample was profiled on a custom Agilent array with 39,280 oligonucleotide probes targeting transcripts representing 34,266 known and predicted genes, including high-confidence noncoding RNA sequences. Each DNA sample was genotyped on the Illumina 650Y BeadChip array. Phasing was performed using BEAGLE with default parameters, and imputation was performed using minimac with default parameters on all markers in the February 2012 genotypes release from the 1,000 Genomes Project. Other publicly available expression datasets that were used only for co-expression network construction included blood and adipose tissue from 1,675 Icelandic individuals[22], and liver tissue from 427 samples[23].

5b. eSNP Analysis

The top IBD-associated markers, as reported in Table 1, were each tested for *cis* and *trans* eSNP association in the MGH liver, subcutaneous adipose, and omental adipose datasets. Significant eSNPs were identified using a method previously described[23]. The *cis* eQTL for a given marker was defined as the gene with associated expression levels whose transcription start or stop site was located within 1 megabase (Mb) of the genetic polymorphism. All other associations were considered *trans*. SNP associations were identified using linear regression and the Kruskal-Wallis test. Based on the number of potential *cis* SNP-gene pairs that met the proximity criteria, Bonferroni correction for multiple testing was applied to the association P-values. Within the 163 IBD-associated loci, *cis* eSNPs with a corrected P-value less than 0.05 were reported as significant. All identified *cis* eSNPs genome-wide were then tested for *trans* eQTL associations at 10% FDR for a significance value threshold of $P \leq 10^{-5}$. Where SNP associations to the same trait were identified in high LD with each other, the SNP with the most significant p-value was reported.

5c. Omental Adipose Bayesian Network Construction

Using gene expression data from the omental adipose tissue set, we applied a method similar to one previously described for Bayesian network construction[24]. Given a set of nodes defined by the genes present in the dataset, 1000 independent simulations were employed to identify a range of plausible network structures, which were then combined to obtain a consensus network with confidence values on each directed edge. Each simulation started with a different randomly generated Bayesian network seed. Three types of prior information were used in the edge seeding of each

simulation. First, protein-protein interaction (PPI) information was retrieved from public (BIND, BioGRID, HPRD, MINT, Reactome, DIP, and IntAct) and commercial (Ingenuity, Proteome, MetaBase, and NetPro) databases, providing scale-free structural priors. Second, transcription factor (TF) binding data was also included in the network seeding. In addition to directly supporting causal relationships between TF's and their target genes, these data were used to identify PPI subnetworks in which at least half of the genes' expression was modulated by a given transcription factor. All the genes in such a subnetwork were then considered to be child nodes of the corresponding TF. Third, eQTL signatures identified in the omental adipose data were used in several ways in seeding the simulations[24]. For a given genetic marker with both a cis-acting and a trans-acting eQTL, the cis-acting eQTL gene was defined as a parent of trans-acting eQTL gene with a prior probability of 1. Genes were then tested for eQTL pleiotropy to identify additional causal relationships that were not marginally significant[25]. Finally, for gene pairs not already addressed by the above information sources, cis-acting and trans-acting eQTLs at a relaxed significance threshold were used to assign fractional prior probabilities.

For model fitting, gene expression was discretized into one of three possible states (downregulated, upregulated, or no change) guided by modified k-means clustering[24]. In each simulation, up to N^2 iterations of MCMC were run until the network's Bayesian Information Criterion (BIC) score was maximized, which typically occurred at roughly N iterations, where N is the number of nodes in the network. In each iteration, a randomly chosen edge was added, removed, or flipped; if the change improved the network's fit to the data, then it was kept. After completing the simulations, we then determined the consensus network by retaining only those edges represented in at least 30% of the 1,000 reconstructed networks. Cycles were eliminated by removing the minimal number of edges with the lowest simulation support in order to satisfy the acyclic property of Bayesian networks. Lastly, genes that were parents of a large number of downstream nodes, and whose simulated changes in expression level modulated expression in many other nodes (as measured against background variation across all genes in the network) were labeled as causal regulators[24] (and Zhang B et al., 2012, Gene Network Remodeling in Alzheimer's Disease, under review).

5d. Enrichment in Co-expression Network Modules

A previously described algorithm was employed to construct weighted gene co-expression analysis (WCGNA) networks on 15 expression datasets[26]. We first constructed a matrix of Pearson correlations between all gene expression pairs. This was then converted into a weighted adjacency matrix using the power function $f(x) = x^{\beta}$, where parameter β was minimized such that the weighted adjacency matrix was approximately scale-free. We used a model fitting index proposed by Zhang *et al*[26] to determine how well a network had a scale-free topology. The maximum value of this index, which describes a perfectly scale-free network, is 1, and 0.8 was the minimum fit required for our final co-expression networks. To identify modules of highly co-regulated genes within these networks, we used average linkage hierarchical clustering to group genes based on the topological overlap of their connectivity, followed by a dynamic cut-tree algorithm to cut clustering dendrogram branches into non-overlapping gene modules[27]. This, in effect, defines groups of genes with high intra-connectivity, relative to their background connectivity

to other genes in the network. In the MGH omental adipose co-expression network, we identified 16 modules (Supp table).

We then screened all 211 modules from co-expression networks for enrichment of IBD-associated genes. Fold-enrichment was calculated as ((A/B)/(C/D)), where A is the number of IBD-associated genes in the module of interest, B is the number of IBD-associated genes, C is the total number of genes in the module of interest, and D is the total number of genes in the full co-expression network. Hypergeometric tests were used to generate exact estimates of the statistical significance of each module's enrichment (Supp table). The module that was most significantly enriched for IBD-associated genes was generated from the omental adipose co-expression network and is most correlated with macrophage gene expression. The portion of the Bayesian network defined by the genes contained in this module and their connections was labeled as the IBD subnetwork.

To investigate the relationship between IBD pathogenesis and response to Mycobacterium tuberculosis (M.Tb) infection, we constructed a weighted co-expression network from gene expression data generated on dendritic cells that had been isolated from 70 individuals and subsequently infected with M.Tb[28]. Of the 12 modules identified in the M.Tb network (Supplementary Figure 9C), 5 significantly overlapped with the IBD subnetwork and were enriched for the GO pathways, such as chemotaxis ($p = 1.6x10^{-8}$) highlighted in (Supplementary Figure 9E).

5e. Integrative Network Reconstruction

Previous publications have described the integrative network reconstruction approach used to construct the human omental adipose network, and include details of the algorithm and implementation used for the constructions[29-32], simulations demonstrating the robustness and the increase in accuracy achieved by integrating DNA variation and RNA expression (compared to RNA expression alone) [24], and applications that have demonstrated the utility of the approach with respect to leading to novel causal genes of disease and the biological context in which they operate[23,29,31-38].

To reconstruct the Bayesian network presented in the manuscript we inputted gene expression data and eQTL data (in the form of structure priors as previously described[31-32]) into our standard Bayesian network reconstruction process. The omental gene expression data and cis eQTL data used to construct the networks has been previously fully described[16] and is freely available from the Sage Bionetworks Synapse tool (https://synapse.sagebase.org/), from the GEO database under the super series accession number GSE24335, and from the Massachusetts General Hospital at http://www.samscore.org. Given the underdetermined nature of our system (i.e., there are many more unknowns than data we have to estimate the unknowns uniquely), we protect against overfitting and ensure robustness by generating thousands of network structures from a Monte Carlo Markov Chain (MCMC) process using different random seed numbers (thousands of random seed numbers are generated by a master process, then each slave process starts an MCMC process using one of the generated seed numbers). Once the thousands of network structures have been generated (typically 1,000 to 10,000 structures are generated; in the present case 1,000 network structures were generated), common features are extracted to derive a consensus

network. That is, edges that are consistent across 30% or greater network structures from the thousands of structures generated are used to derive the final network structure. We have demonstrated previously that this type of consensus network is robust and highly consistent (i.e., multiple repeats of this process generally lead to the same network structure)[24,32]. Because the consensus network may contain loops after this consensus process, which is prohibited in Bayesian networks, we ensure the final network is a directed acyclic graph by removing edges if and only if 1) the edge was involved in a loop, and 2) the edge was the most weakly supported of all edges making up the loop.

The software, RIMBANet, for constructing the Bayesian networks is freely available at: http://www.mssm.edu/research/institutes/genomics-institute/rimbanet and comes complete with instructions on how to run the software and specific examples with step-by-step constructions on reproducing previously published results with this software.

Supplementary Tables

Note: All supplementary tables available as additional supplementary files.

Supplementary Table 1: GWAS & Immunochip samples used in this study. The first tab shows details of all GWAS samples included in our meta-analysis, including the number of unique controls, since several CD and UC datasets from the same group used overlapping controls. The second tab shows details of the Immunochip samples used, broken down by 11 genotyping centers, and then by nationality within those groups. "Non-overlap" refers to Immunochip samples that do not overlap with GWAS samples from tab 1.

Supplementary Table 2: Complete details of 163 IBD loci. The first tab is a more complete summary of the loci shown in main text table 1, containing all prioritized genes, and gives the overview of our 163 IBD loci. The second tab contains detailed association statistics, including disease-specific and platform-specific odds ratios and p-values. The third tab shows details about overlaps with other disease phenotypes (described in Section 2 above)[39-51].The fourth tab shows details about our gene prioritization approaches (Section 3 above).

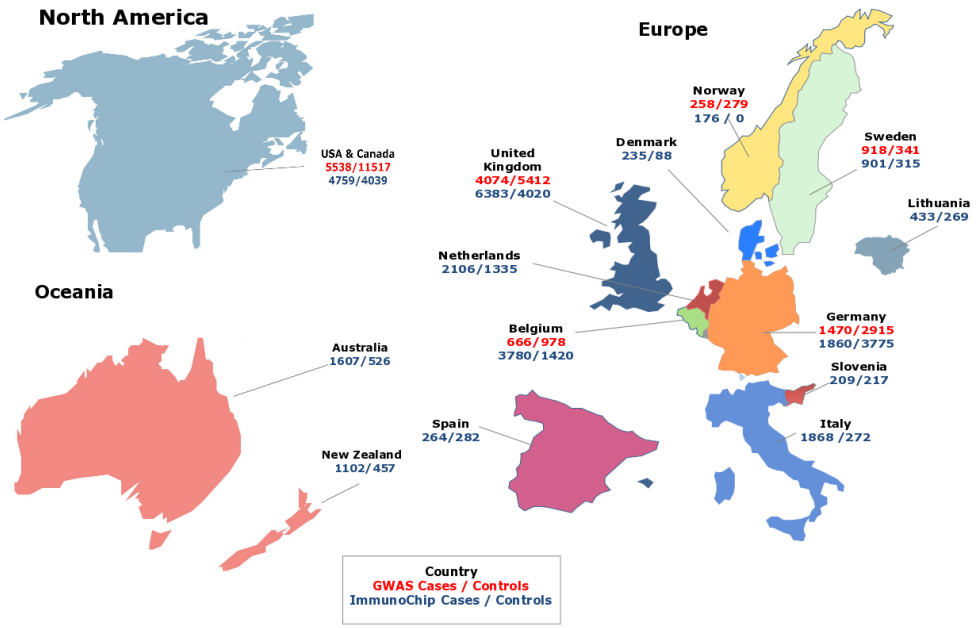
Supplementary Table 3: Disease overlaps. Details of disease overlaps with IMD, PID, and MSMD described in section 2 of the methods. The first tab gives overlap and enrichment statistics broken down by IMD phenotype. The second tab lists all PIDs with genes that overlap IBD loci. The third tab lists the 8 MSMD loci, the associated alleles in MSMD and IBD, and their functional effects[12,52-57]. The fourth tab gives details of the overlapping loci between IBD and leprosy, along with their inferred directions of effect.

Supplementary Table 4: GO term and pathway enrichment. Detailed enrichment statistics for all GO terms (tab 1) and canonical pathways (tab 2), using the technique described in section 4a of the methods. All terms with $p < 10^{-4}$ in the IBD enrichment analysis are included. Fields of the form NhitsA and p_A show the number of loci associated with phenotype A that are annotated with this term, and the statistical significance of this enrichment. betaAB and p_AB give the β statistic (defined in Section 4aii) and enrichment p-value of the term in phenotype A compared to phenotye B. betaAxisA and p_axisA give the β statistic and enrichment p-value of the term in phenotype A relative to the other two phenotypes. thetabeta and p_theta are the β statistic and significance of the correlation between θ (the CD-UC balance defined in Section 1d) and the functional term.

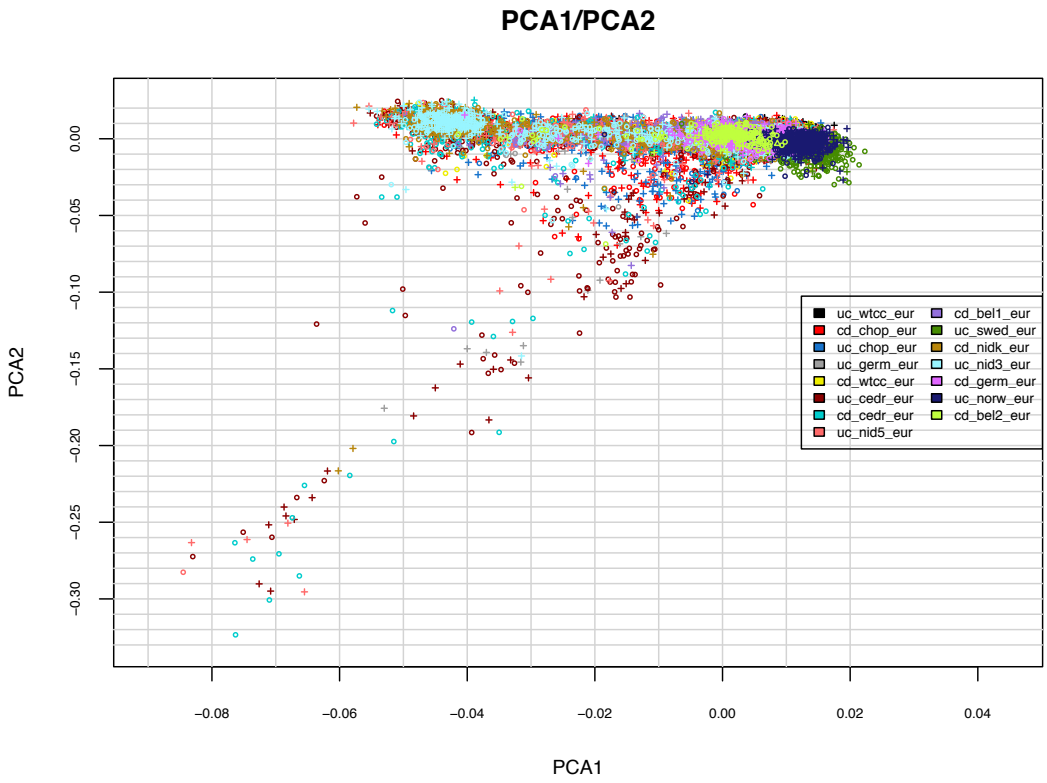
Supplementary Table 5: Signals of selection at IBD loci. Selection statistics for individual SNPs (tab 2) and sets of SNPs (tab 1) calculated from TreeMix, as well as GO enrichment for selection described in Section 4c above (tab 3).

Supplementary Table 6: Enrichment scores for genes in IBD loci within co-expression modules. Fold-enrichment was calculated as ((A/B)/(C/D)), where A is the number of IBD-associated genes in the module of interest, B is the number of IBD-associated genes, C is the total number of genes in the module of interest, and D is the total number of genes in the full co-expression network. Hypergeometric tests were used to generate exact estimates of the statistical significance of each module's enrichment

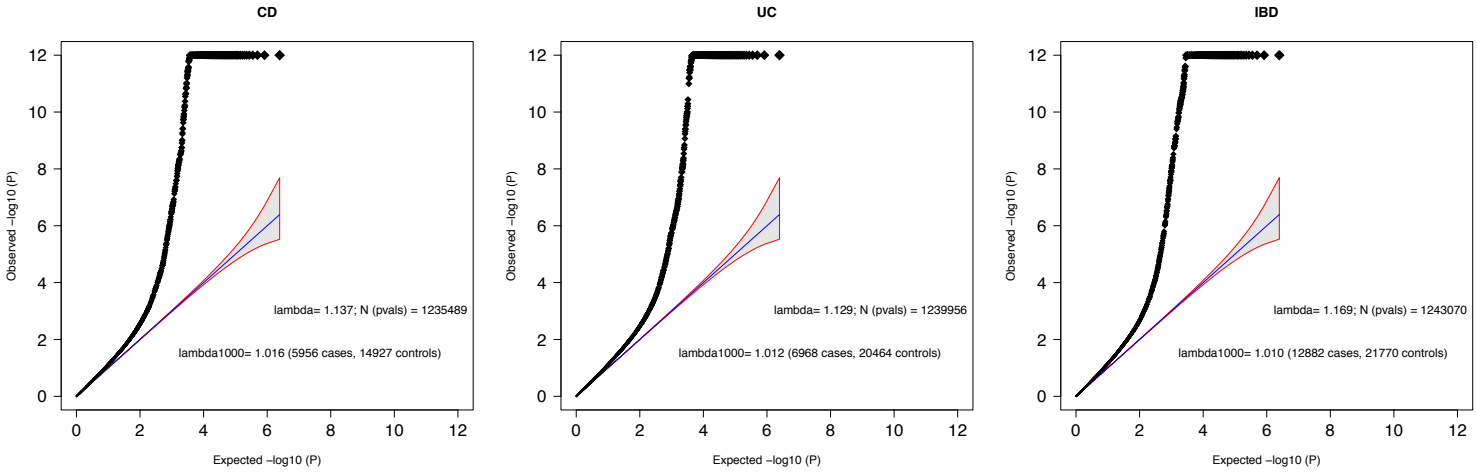
Supplementary Figures



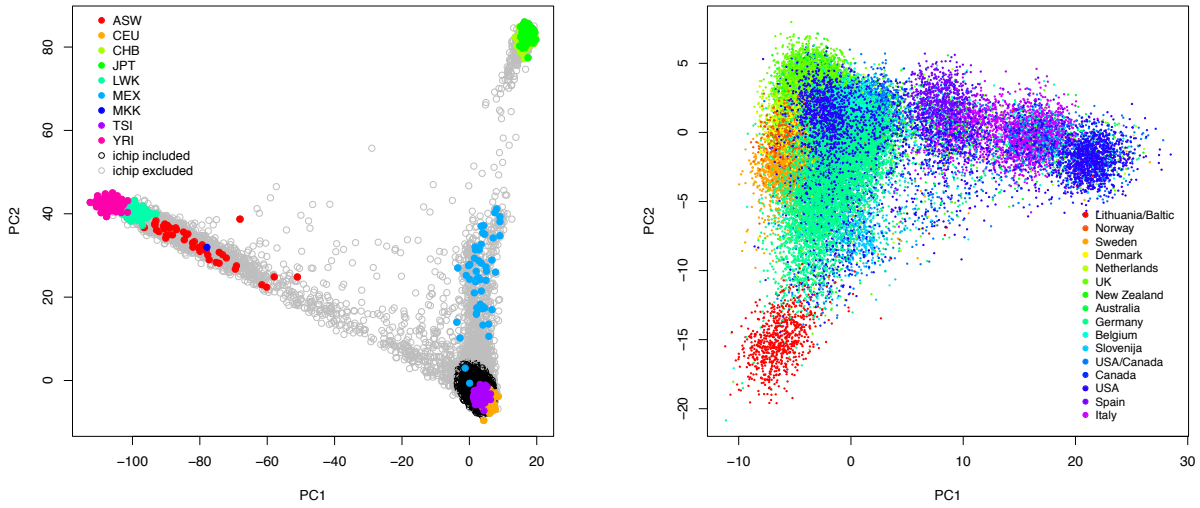
Supplementary Figure 1: Global distribution of IBD GWAS and Immunochip samples. Numbers of quality control passing IBD and control samples from each country participating in this study. The numbers for the Immunochip samples (numbers in blue) only include samples that are not also present in the GWAS (numbers in red).



Supplementary Figure 2: PCA of GWAS data. All GWAS samples plotted on the first two principal components, colored by study. Circles are cases, crosses controls.



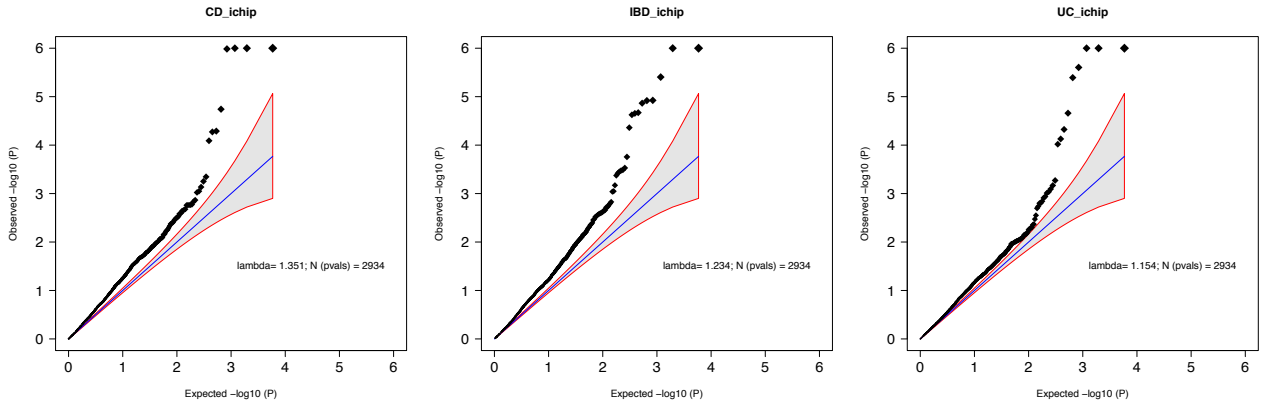
Supplementary Figure 3: QQ plots for GWAS. QQ plots, lambda and lambda_1000 values for the CD, UC and IBD GWAS analyses. Grey shapes show 95% confidence interval under the null.



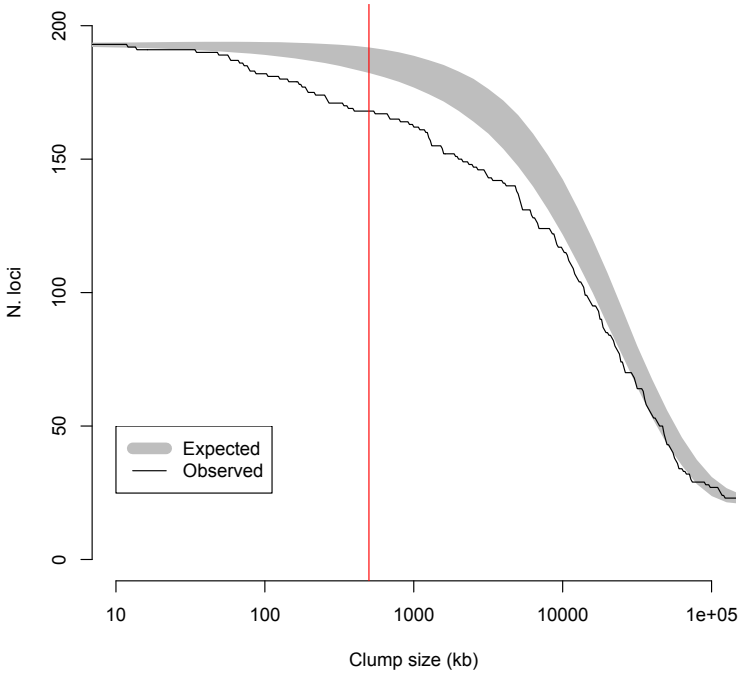
(a) Continental PCA with ethnic outliers

(b) Within Europe PCA by country

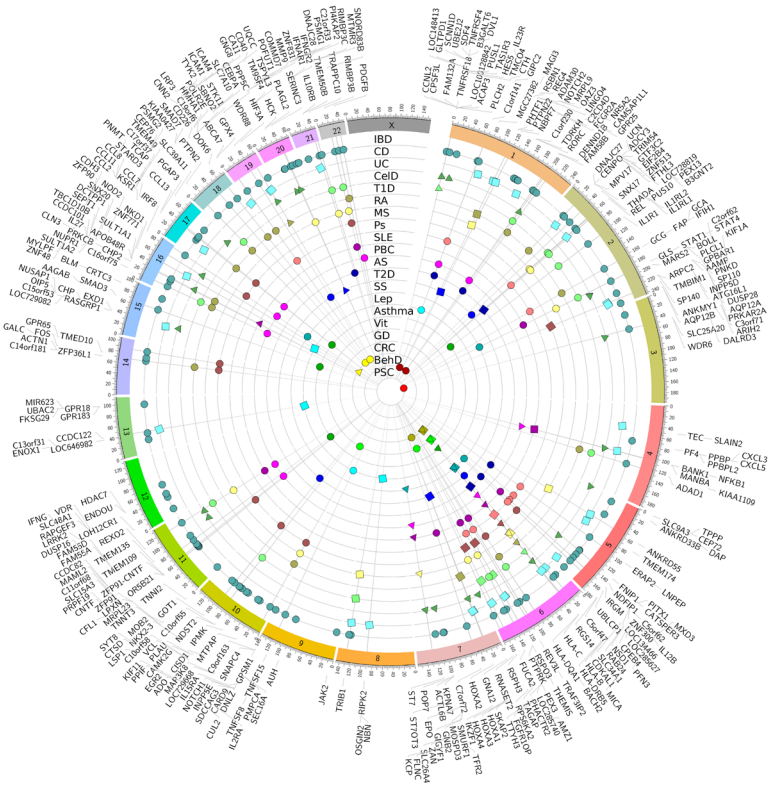
Supplementary Figure 4: PCA of Immunochip data. A) PCA projection of Immunochip samples onto world-wide axes. Our samples are anchored near the CEU and TSI HapMap populations, but with substantial spread. Samples excluded as ethnic outliers are marked in grey. B) PCA calculated in Immunochip controls projected onto cases & controls, colored by country of origin. We accurately capture both North-South and East-West variation within Europe.



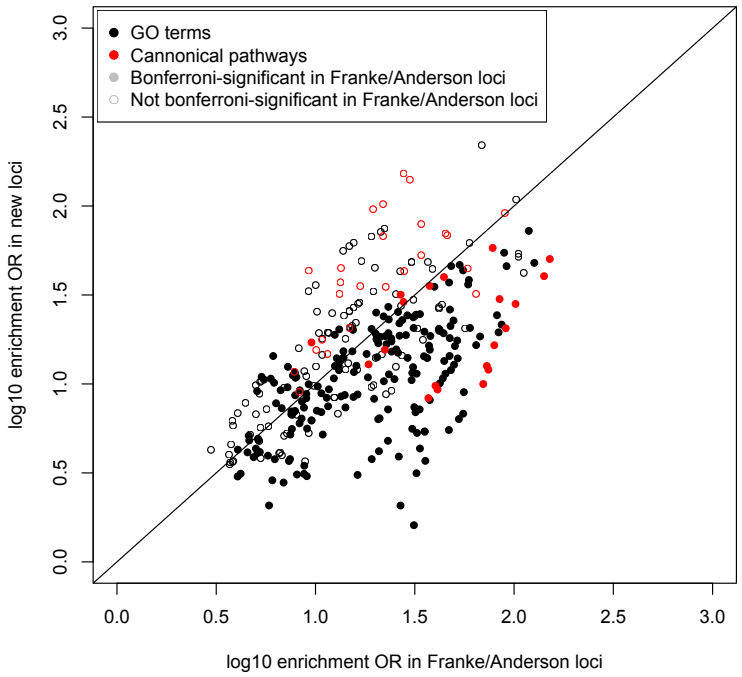
Supplementary Figure 5: QQ plots for Immunochip. QQ plots and lambda values for the CD, UC and IBD Immunochip analyses. Grey shapes show 95% confidence interval under the null.



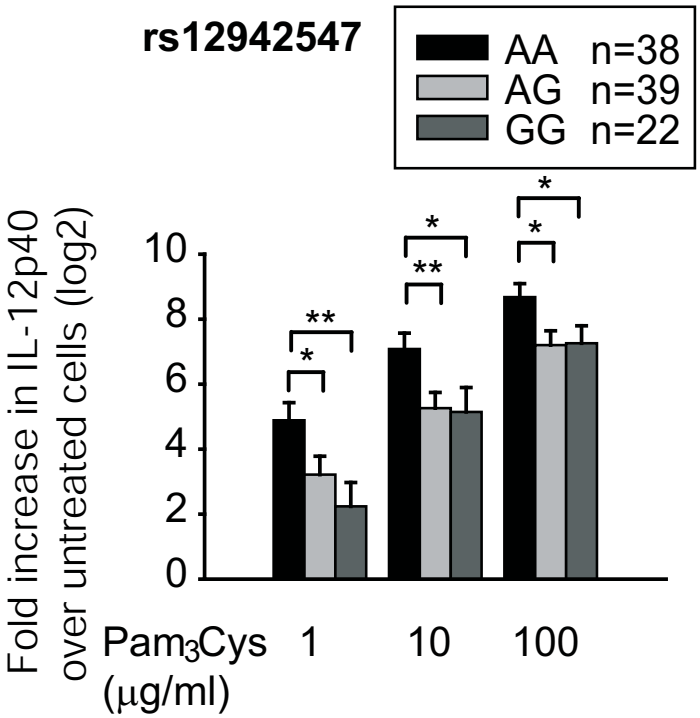
Supplementary Figure 6: From association signals to loci. The x-axis shows varying thresholds of proximity for two statistically independent signals to be considered in the same locus. The y-axis shows the number of loci for a particular threshold, from 193 (the total number of independent signals) at the left when no signals are combined to fewer than 50 when even extremely distant signals 100Mb apart are combined. The 95% confidence interval from simulations of 193 randomly placed signals (grey shaded area) demonstrates that even choosing random regions yields many within 1Mb of each other, and half within 10Mb. Our regions (black line) are often closer than expected by chance, possibly caused by the physical proximity of functionally related genes associated to IBD. The red vertical line shows our chosen threshold of 500kb for merging.



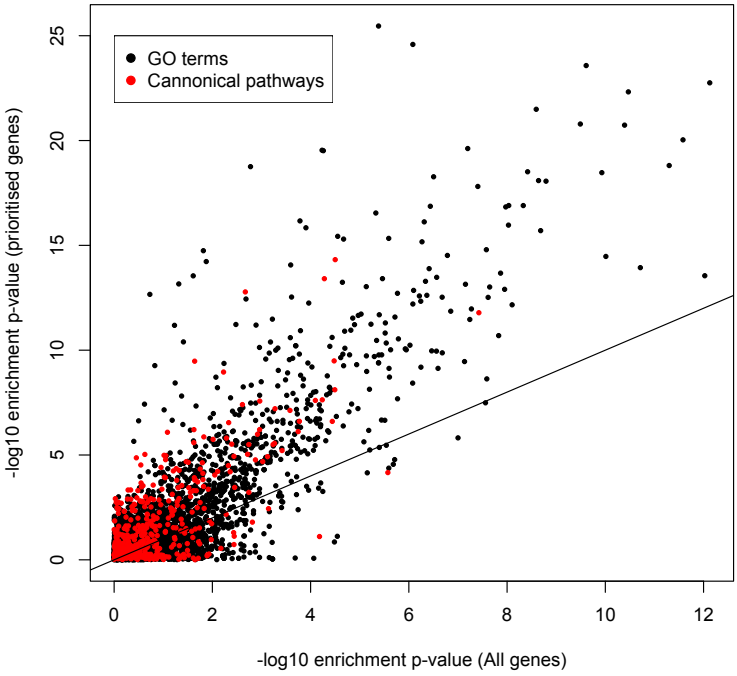
Supplementary Figure 7: 'Circus' plot of associations to other IMD diseases.
Each radial line represents an IBD locus, ordered by genomic position and labeled around the rim, and each circular line represents a phenotype, with all points on a line colored according to the phenotype key given. Points sit at the intersection of a radial and a circular lines, and represent sharing of an IBD locus with a given phenotype. The location in Table 1 of each IBD locus is shown by shapes: triangles for UC-specific, square for CD-specific and circle for shared.



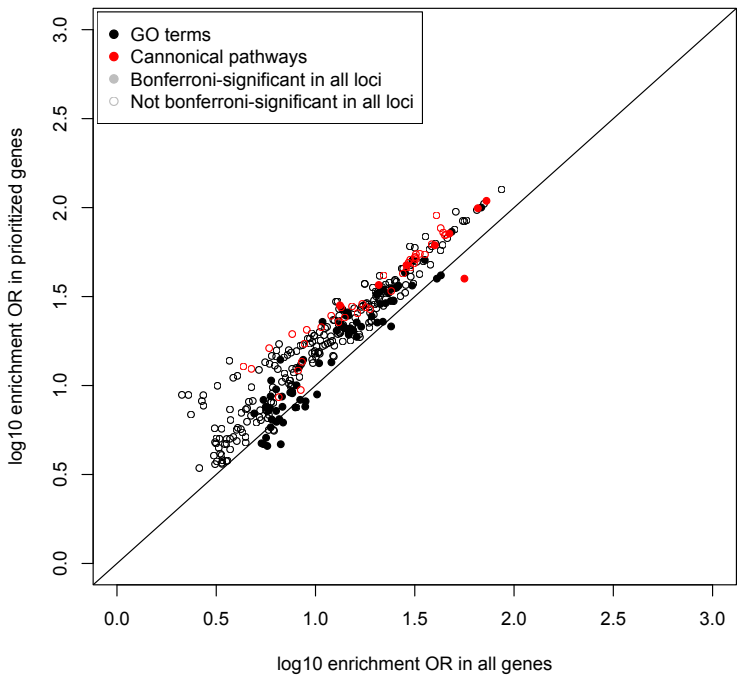
Supplementary Figure 8: GO enrichment in known vs. new loci. The enrichment odds ratios for GO terms (Supplementary Methods 4a) are plotted in previously known loci (x-axis) and new loci identified here (y-axis). In contrast to the possibility that our immune related GO enrichment is driven by Immunochip biased discovery, the odds ratios are, on average, slightly but significantly higher when estimated in known loci only. This could suggest that loci related to core biological processes tend to have large effect sizes, and are therefore more likely to have been previously discovered. Note that while the effect sizes are similar in both set of loci, our newly discovered set allow us to detect *significant* enrichment for the first time in many more GO terms (open circles) than was possible using previous loci (filled circles).



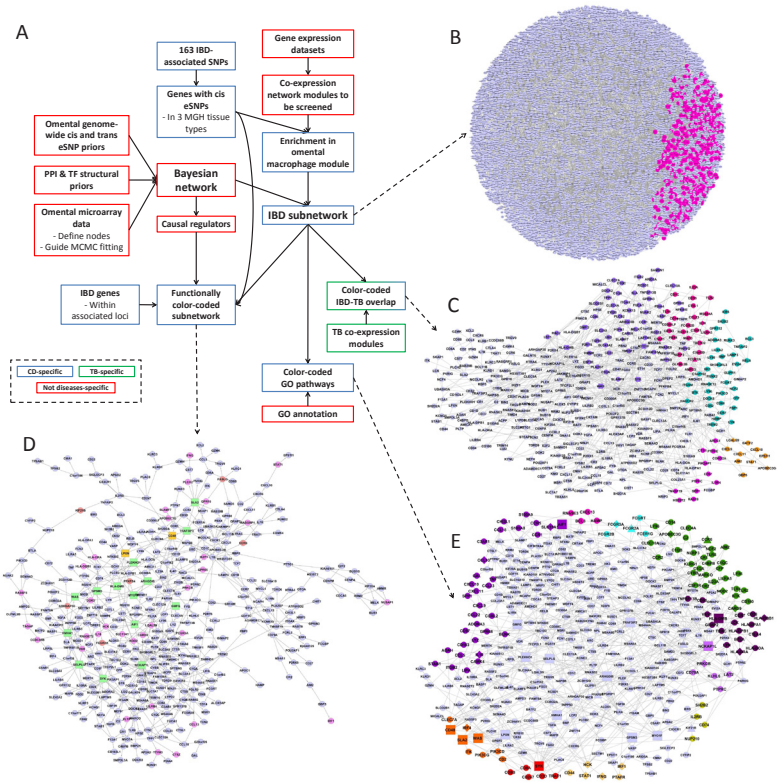
Supplementary Figure 9: Functional impact of IBD-associated risk alleles.
STAT3 risk carriers at rs12942547 demonstrate increased IL12p40 secretion upon pattern-recognition receptor initiated stimulation. Peripheral blood monocyte-derived macrophages from healthy controls were stimulated with Pam3Cys (activates toll-like receptor 2) (Calbiochem, La Jolla, CA) at the doses listed. After 24 hours, IL12p40 concentrations in the supernatant were measured by ELISA (R&D Systems Inc. Minneapolis, MI, USA). The IBD risk allele is the A allele, which is associated, likely via autocrine-mediated cytokine effects, with increased IL12p40 secretion with PRR-initiated stimulation compared to the G allele. *, p-value < 0.05; **, p-value < 0.01.



Supplementary Figure 10: Relative enrichment of all and prioritized genes in IBD loci. Enrichment p-values for GO terms (black dots) and canonical pathways (red dots) calculated on all 1,438 genes in IBD loci (x-axis) and just the 300 prioritized as described in Supplementary Methods Section 3 (y-axis). Nearly all significantly enriched terms and pathways are more significant in our prioritized genes alone (above the dashed x=y line), demonstrating that the prioritization procedure has successfully enriched for specific pathways and functions.



Supplementary Figure 11: Comparison of odds ratio estimates for enrichment using all and prioritized genes in IBD loci. Odds ratios for GO terms (black dots) and canonical pathways (red) calculated on all 1,438 genes in IBD loci (x-axis) and just the 300 prioritized as described in Supplementary Methods Section 3 (y-axis). The slight bias towards higher estimated ORs in prioritized genes suggests we are more likely to prioritize genes from well-studied pathways. While this suggests our current list of prioritized genes is biased away from relevant, but unknown biology, it does not imply that our currently highlighted genes are not actually involved in IBD pathogenesis.



Supplementary Figure 12: IBD network analysis.

A) Workflow of the Bayesian network-based analysis of newly IBD-associated genes' functional relationships. Blue boxes represent data products guided by evidence for disease association, green boxes represent results guided by an expression dataset in tuberculosis (MTB) infection, and red boxes represent analytical and data products that did not utilize any IBD or MTB datasets. **B)** Full Bayesian network constructed from the omental adipose expression dataset. Genes that are in the IBD subnetwork are highlighted in pink. **C)** Gene overlap between the IBD subnetwork and top MTB modules. Colors other than light purple correspond to modules in the TB-infected macrophage coexpression network. Square = causal regulator. Diamond = IBD-associated gene. Triangle = gene associated with both IBD and a *cis* eSNP. **D)** Functionally color-coded IBD subnetwork, representing genes in the omental macrophage module and edges in the omental adipose Bayesian network. Nodes disconnected from the main subnetwork were removed, with the 465 remaining genes displayed here. Pink circle = IBD-associated gene. Peach circle = gene associated with both IBD and a *cis* eSNP. Green square = causal regulator. Yellow square = gene that is both IBD-associated and also a causal regulator. E) GO-pathway color-coded IBD subnetwork. Clockwise, from left: Dark purple = inflammatory response. Pink = defense response to bacterium. Turquoise= IgG binding. Green = innate immune response. Plum = T cell costimulation. Light purple = B cell receptor signaling. Lime = cytokine-mediated signaling. Yellow = interferon gamma-mediated signaling. Red = T cell receptor complex. Orange = T cell activation. Square = causal regulator. Diamond = IBD-associated gene. Triangle = gene associated with both IBD and a *cis* eSNP. All network illustrations are available for download in Cytoscape format as Supplementary Data from http://www.nature.com/nature (file name All_network.cys).

Discussion of other pathways highlighted by the 163 loci and subsequent analyses

Ubiquitination and NFκB

Ubiquitination is a mechanism of post-translational modification that can lead to protein degradation and is implicated in innate immunity, adaptive immunity, autophagy and NFκB activation. IBD loci are enriched for this process, with 15 containing prioritized 'ubiquitination-related' genes ($p = 8 \times 10^{-3}$). These include the ubiquitin-specific proteases, *USP4* and *UBE2L3*, which are involved in the second step of ubiquitination that transfers ubiquitin to the active site of a ubiquitin-conjugating enzyme. *TNFAIP3* is UC-predominant and encodes the ubiquitin editing protein A20 implicated in multiple immune-mediated diseases[58]. Importantly, ubiquitination is regarded as a tractable inflammation-related therapeutic target. NFκB is a master transcriptional regulator of the inflammatory response, and controller of epithelial integrity and mucosal immune homeostasis in the presence of gut microflora[59]. Ubiquitin has multiple roles in the NFκB pathway, including processing of NFκB precursors (p105) and IkB kinase activation and degradation. NFκB dysregulation has been linked to several autoimmune and inflammatory conditions, including IBD. In the current study, new associations were identified at loci encoding several key constituents of NFκB, including *REL* on chromosome 2 (IBD), *RELA* on chromosome 11 (CD) and *NFKB1* itself on chromosome 4 (UC).

Autophagy

Among the autophagy genes associated with IBD, *ATG16L1* remains specific to CD. *IRGM* and *LRRK2* are now also associated with UC, but with significantly smaller effect sizes than those seen in CD. We also observed a novel UC-specific association with *SMURF1*, a ubiquitin ligase that was recently identified, through an image based genome-wide siRNA screen, as an important mediator of selective viral autophagy and mitophagy[60]. These findings, together with the recently described UC-association with a negative regulator of autophagy (*DAP*)[61], extend the role of autophagy to UC.

Th17-cell differentiation: RORC

Genetic studies have implicated genes involved in T-cell differentiation, specifically in the differentiation of Th1 and Th17 cells, in multiple immune-mediated diseases[62]. We identify a locus on chromosome 1q21 harboring RORC (or RORγt), a nuclear receptor and the master transcriptional regulator of the differentiation of naïve CD4+ T cells into IL17-producing Th17 cells[63]. RORγt has a ligand-binding pocket, so it is an excellent candidate for pharmacological intervention. Recently, a high-affinity synthetic ligand specific for both RORγt and RORα was shown to inhibit the development and function of Th17 cells[64].

Transforming Growth Factor beta (TGFβ) signaling

TGFβ limits immune responses and is a potent pro-fibrogenic agent, inducing collagen synthesis in the GI tract[65]. In this study the TGFβ pathway showed the

greatest enrichment in IBD loci relative to both other immune-mediated diseases ($p=6.7 \times 10^{-5}$) and primary immunodeficiencies ($p = 8.5 \times 10^{-5}$). Twelve loci contained a gene involved in this pathway, and many of these were involved in production or degradation of TGFβ signaling components. In addition to confirming the *SMAD3* (TGFβ signaling) association, we report two new associations at *SMAD7* and *SMURF1*, known promoters of type I TGF-beta receptor ubiquitination. We also report a novel association near the *FURIN* gene, a protein responsible for cleaving and activating the TGFβ complex precursor.

Overlap with colorectal cancer

An important clinical complication in UC is colorectal cancer (CRC). Previous GWAS studies identified *CDH1* and *RHPN2* as risk loci shared between UC and CRC[66]. *SMAD7* is an intracellular antagonist of TGFβ signaling and is a known susceptibility gene for CRC[67]. The newly observed association between *SMAD7* and UC increases the genetic contribution to this shared risk.

Supplementary References

1. Franke, A. *et al*. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
2. Anderson, C.A. *et al*. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43**, 246-52 (2011).
3. Browning, B.L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85**, 847-61 (2009).
4. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
5. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
6. Trynka, G. *et al*. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
7. Shah, T.S. *et al*. optiCall: A robust genotype-calling algorithm for rare, low frequency and common variants. *Bioinformatics* **28**, 1598-603 (2012).
8. Morris, J.A., Randall, J.C., Maller, J.B. & Barrett, J.C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26**, 1786-7 (2010).
9. Price, A.L. et al. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* **83**, 132-5; author reply 135-9 (2008).
10. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum Hered* **64**, 45-51 (2007).
11. Hindorff, L.A. *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**, 9362-7 (2009).
12. Notarangelo, L.D. *et al*. Primary immunodeficiencies: 2009 update. *J Allergy Clin Immunol* **124**, 1161-78 (2009).
13. Raychaudhuri, S. *et al*. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).
14. Rossin, E.J. *et al*. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
15. Dixon, A.L. *et al*. A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-7 (2007).
16. Greenawalt, D.M. *et al*. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Re*s **21**, 1008-16 (2011).
17. Hu, X. *et al*. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am J Hum Genet* **89**, 496-506 (2011).
18. Irizarry, R.A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).

19. Hyatt, G. *et al*. Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol* **7**, 686-91 (2006).
20. Su, A.I. *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101**, 6062-7 (2004).
21. Pickrell, J. & Pritchard, J. Inference of population splits and mixtures from genome-wide allele frequency data. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2012.6956.1>. *Nature Precedings* (2012).
22. Emilsson, V. *et al*. Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
23. Schadt, E.E. *et al*. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).
24. Zhu, J. *et al*. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol* **3**, e69 (2007).
25. Jiang, C. & Zeng, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111-27 (1995).
26. Frazer, K.A. *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
27. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* **1**, 54 (2007).
28. Barreiro, L.B. *et al*. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci USA* **109**, 1204-9 (2012).
29. Schadt, E.E. *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-7 (2005).
30. Zhu, J. et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet *Genome Res* **105**, 363-74 (2004).
31. Zhu, J. *et al*. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol* **10**, e1001301 (2012).
32. Zhu, J. *et al*. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* **40**, 854-61 (2008).
33. Chen, Y. *et al*. Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429-35 (2008).
34. Dobrin, R. *et al*. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol* **10**, R55 (2009).
35. Emilsson, V. *et al*. Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
36. Yang, X. *et al*. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet* **41**, 415-23 (2009).
37. Zhong, H. *et al*. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* **6**, e1000932 (2010).
38. Zhong, H., Yang, X., Kaplan, L.M., Molony, C. & Schadt, E.E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86**, 581-91 (2010).
39. Ogura, Y. *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6 (2001).

40. Bonen, D.K. *et al*. Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* **124**, 140-6 (2003).
41. Graham RR, *et a*l. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat Genet* **38**, 550-5 (2006).
42. Kuballa, P., Huett, A., Rioux, J.D., Daly, M.J. & Xavier, R.J. Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. *PLoS One* **3**, e3391 (2008).
43. Fiorillo, E. *et al*. Autoimmune-associated PTPN22 R620W variation reduces phosphorylation of lymphoid phosphatase on an inhibitory tyrosine residue. *J Biol Chem* **285**, 26506-18 (2010).
44. Brest, P. *et al*. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* **43**, 242-5 (2011).
45. Di Meglio, P. *et al*. The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans. *PLoS One* **6**, e17160 (2011).
46. Plantinga TS, *et al*. Crohn's disease-associated ATG16L1 polymorphism modulates pro-inflammatory cytokine responses selectively upon activation of NOD2. *Gut* **60**, 1229-35 (2011).
47. Menard, L. *et al*. The PTPN22 allele encoding an R620W variant interferes with the removal of developing autoreactive B cells in humans. *J Clin Invest* **121**, 3635-44 (2011).
48. Pidasheva, S. *et al*. Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. *PLoS One* **6**, e25038 (2011).
49. Sarin, R., Wu, X. & Abraham, C. Inflammatory disease protective R381Q IL23 receptor polymorphism results in decreased primary CD4+ and CD8+ human T-cell functional responses. *Proc Natl Acad Sci USA* **108**, 9560-5 (2011).
50. Zhang, J. et al. The autoimmune disease-associated PTPN22 variant promotes calpain-mediated Lyp/Pep degradation associated with lymphocyte and dendritic cell hyperresponsiveness. *Nat Genet* **43**, 902-7 (2011).
51. Hedl, M. & Abraham, C. IRF5 risk polymorphisms contribute to interindividual variance in pattern recognition receptor-mediated cytokine secretion in human monocyte-derived cells. *J Immunol* **188**, 5348-56 (2012).
52. Picard, C. *et al*. Inherited interleukin-12 deficiency: IL12B genotype and clinical phenotype of 13 patients from six kindreds. *Am J Hum Genet* **70**, 336-48 (2002).
53. Bustamante, J., Picard, C., Boisson-Dupuis, S., Abel, L. & Casanova, J.L. Genetic lessons learned from X-linked Mendelian susceptibility to mycobacterial diseases. *Ann N Y Acad Sci* **1246**, 92-101 (2011).
54. Patel, S.Y., Doffinger, R., Barcenas-Morales, G. & Kumararatne, D.S. Genetically determined susceptibility to mycobacterial infection. *J Clin Pathol* **61**, 1006-12 (2008).
55. Hambleton, S. et al. IRF8 mutations and human dendritic-cell immunodeficiency. *N Engl J Med* **365**, 127-38 (2011).

56. Minegishi, Y. *et al*. Human tyrosine kinase 2 deficiency reveals its requisite roles in multiple cytokine signals involved in innate and acquired immunity. *Immunity* **25**, 745-55 (2006).
57. Minegishi, Y. *et al*. Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome. *Nature* **448**, 1058-62 (2007).
58. Martin, F. & Dixit, V.M. A20 edits ubiquitin and autoimmune paradigms. *Nat Genet* **43**, 822-3 (2011).
59. Nenci, A. *et al*. Epithelial NEMO links innate immunity to chronic intestinal inflammation. *Nature* **446**, 557-61 (2007).
60. Orvedahl, A. *et al*. Image-based genome-wide siRNA screen identifies selective autophagy factors. *Nature* **480**, 113-7 (2011).
61. Koren, I., Reem, E. & Kimchi, A. DAP1, a novel substrate of mTOR, negatively regulates autophagy. *Curr Biol* **20**, 1093-8 (2010).
62. Cho, J.H. & Gregersen, P.K. Genomics and the multifactorial nature of human autoimmune disease. *N Engl J Med* **365**, 1612-23 (2011).
63. Ivanov, II *et al*. The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells. *Cell* **126**, 1121-33 (2006).
64. Solt, L.A. *et al*. Suppression of TH17 differentiation and autoimmunity by a synthetic ROR ligand. *Nature* **472**, 491-4 (2011).
65. Di Sabatino, A. *et al*. Transforming growth factor beta signalling and matrix metalloproteinases in the mucosa overlying Crohn's disease strictures. *Gut* **58**, 777-89 (2009).
66. Barrett, J.C. *et al*. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* **41**, 1330-4 (2009).
67. Broderick, P. *et al*. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39**, 1315-7 (2007).

International IBD Genetics Consortium Contributing Members:

Murray Barclay[1], Laurent Peyrin-Biroulet[2], Mathias Chamaillard[3], Jean-Frederick Colombel[4], Mario Cottone[5], Anthony Croft[6], Renata D'Incà[7], Jonas Halfvarson[8,9], Katherine Hanigan[6], Paul Henderson[10,11], Jean-Pierre Hugot[12,13], Amir Karban[14], Nicholas A Kennedy[15], Mohammed Azam Khan[16], Marc Lémann[17], Arie Levine[18], Dunecan Massey[19], Monica Milla[20], Grant W Montgomery[21], Sok Meng Evelyn Ng[22], Ioannis Oikonomou[22], Harald Peeters[23], Deborah D. Proctor[22], Jean-Francois Rahier[24], Rebecca Roberts[2], Paul Rutgeerts[25], Frank Seibold[26], Laura Stronati[27], Kirstin M Taylor[28], Leif Törkvist[29], Kullak Ublick[30], Johan Van Limbergen[31], Andre Van Gossum[32], Morten H. Vatn[33], Hu Zhang[20], Wei Zhang[22], Australia and New Zealand IBDGC*, Belgium Genetic Consortium†, Initiative on Crohn and Colitis, NIDDK IBDGC‡, United Kingdom IBDGC, Wellcome Trust Case Control Consortium§

[1]Department of Medicine, University of Otago, Christchurch, New Zealand. [2]Gastroenterology Unit, INSERM U954, Nancy University and Hospital, France. [3]INSERM, U1019, Lille, France. [4]Univ Lille Nord de France, CHU Lille and Lille-2 University,Gastroenterology Unit, France. [5]Division of Internal Medicine, Villa Sofia-V. Cervello Hospital, University of Palermo, Palermo, Italy. [6]Inflammatory Bowel Diseases, Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia. [7]Department of Surgical and Gastroenterological Sciences, University of Padua, Padua, Italy. [8]Department of Medicine, Örebro University Hospital, Örebro, Sweden. [9]School of Health and Medical Sciences, Örebro University, Örebro, Sweden. [10]Royal Hospital for Sick Children, Paediatric Gastroenterology and Nutrition, Edinburgh, UK. [11]Child Life and Health, University of Edinburgh, Edinburgh, UK. [12]INSERM U843, Paris, France. [13]Univ-Paris Diderot Sorbonne Paris-Cité, Paris France. [14]Department of Gastroenterology, Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel. [15]Gastrointestinal Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [16]Genetic Medicine, MAHSC, University of Manchester, Manchester, UK. [17]Université Paris Diderot, GETAID group, Paris, France. [18]Pediatric Gastroenterology Unit, Wolfson Medical Center and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [19]Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK. [20]Azienda Ospedaliero Universitaria (AOU) Careggi, Unit of Gastroenterology SOD2, Florence, Italy. [21] Molecular Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia. [22]Department of Internal Medicine, Section of Digestive Diseases, Yale School of Medicine, New Haven, Connecticut, USA. [23]Dept Gastroenterology - University hospital Gent - De Pintelaan - 9000 Gent Belgium. [24]Dept Gastroenterology - UCL Mont Godinne Belgium. [25]Division of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium. [26]University

of Bern, Division of Gastroenterology, Inselspital, Bern, Switzerland. [27]Department of Radiobiology and Human Health, Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Rome, Italy. [28]Dept Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St Thomas' Hospital, London, UK. [29]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden. [30]Division of Clinical Pharmacology and Toxicology, University Hospital Zurich, Zurich, Switzerland. [31]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada. [32]Dept Gastroenterology - 3University Brussels. [33] Department of Transplantation Medicine, Division of Cancer medicine, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway.

***Australia and New Zealand IBDGC**

Jane M. Andrews[1], Peter A. Bampton[2], Murray Barclay[3], Timothy H. Florin[4], Richard Gearry[3], Krupa Krishnaprasad[5], Ian C. Lawrance[6], Gillian Mahy[7], Grant W. Montgomery[8], Graham Radford-Smith[5,9], Rebecca L. Roberts[10], Lisa A. Simms[5].

[1]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, and School of Medicine, University of Adelaide, Adelaide, Australia. [2]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, Australia. [3]Department of Gastroenterology, Christchurch Hospital and Department of Medicine, University of Otago, Christchurch, New Zealand. [4]Department of Gastroenterology, Mater Health Services, Brisbane, Australia, and School of Medicine, University of Queensland, Brisbane, Australia. [5]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [6]Centre for Inflammatory Bowel Diseases, Fremantle Hospital and School of Medicine and Pharmacology, The University of Western Australia, Fremantle, Australia. [7]Department of Gastroenterology, The Townsville Hospital and James Cook University School of Medicine, Townsville, Australia. [8]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [9]Department of Gastroenterology, Royal Brisbane and Womens Hospital, and School of Medicine, University of Queensland, Brisbane, Australia. [10]University of Otago, Department of Medicine, Christchurch, New Zealand.

†Belgium Genetic Consortium

Leila Amininijad[1], Isabelle Cleynen[2], Olivier Dewit[3], Denis Franchimont[1], Michel Georges[4], Debby Laukens[5], Harald Peeters[5], Jean-Francois Rahier[3], Paul Rutgeerts[2], Emilie Theatre[4], [6], André Van Gossum[1], Severine Vermeire[7].

[1]Erasmus Hospital, Free University of Brussels, Department of Gastroenterology, Brussels, Belgium. [2]Department of Pathophysiology, Gastroenterology section, KU Leuven, Leuven, Belgium. [3]Department of Gastroenterology, Clinique Universitaire

St-Luc, Brussels, Belgium. [4]Unit of Animal Genomics, Groupe Interdisciplinaire de Gnoprotomique Applique (GIGA-R) and Faculty of Veterinary Medicine, University of Lige, Lige, Belgium. [5]Ghent University Hospital, Department of Gastroenterology and Hepatology, Ghent, Belgium. [6]Division of Gastroenterology, Centre Hospitalier Universitaire, Universit de Lige, Lige, Belgium. [7]Division of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium.

‡NIDDK Inflammatory Bowel Disease Genetics Consortium‡

Guy Aumais[1], Leonard Baidoo[2], Arthur M. Barrie III[2], Karen Beck[2], Edmond-Jean Bernard[3], David G. Binion[2], Alain Bitton[4], Steve R. Brant[5], Judy H. Cho[6,7], Albert Cohen[8], Kenneth Croitoru[9], Mark J. Daly[10,11], Lisa W. Datta[5], Colette Deslandres[12], Richard H. Duerr[2,13], Debra Dutridge[14], John Ferguson[7], Joann Fultz[2], Philippe Goyette[15], Gordon R. Greenberg[9], Talin Haritunians[14], Gilles Jobin[16], Seymour Katz[17], Raymond G. Lahaie[18], Dermot P. McGovern[14,19], Linda Nelson[2], Sok Meng Ng[7], Kaida Ning[7], Ioannis Oikonomou[7], Pierre Paré[20], Deborah D. Proctor[7], Miguel D. Regueiro[2], John D. Rioux[15], Elizabeth Ruggiero[7], L. Philip Schumm[21], Marc Schwartz[2], Regan Scott[2], Yashoda Sharma[7], Mark S. Silverberg[9], Denise Spears[5], A. Hillary Steinhart[9], Joanne M. Stempak[9], Jason M. Swoger[2], Constantina Tsagarelis[4], Wei Zhang[7], Clarence Zhang[22], Hongyu Zhao[22].

[1]University of Montreal, Maissonneuve – Rosemont Hospital, Quebec Association of Gastroenterologists, Montréal, Québec, Canada. [2]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [3]Hôpital Hôtel Dieu, Montréal, Québec, Canada. [4]Division of Gastroenterology, McGill University Health Centre, Royal Victoria Hospital, Montréal, Québec, Canada. [5]Inflammatory Bowel Disease Center, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA. [7]Department of Internal Medicine, Section of Digestive Diseases, Yale School of Medicine, New Haven, Connecticut, USA. [8]Division of Gastroenterology, Hôpital Général Juif Sir Mortimer B. Davis Jewish General Hospital, Montréal, Québec, Canada. [9]Mount Sinai Hospital Inflammatory Bowel Disease Centre, University of Toronto, Toronto, Ontario, Canada. [10]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [11]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [12]Hopital Sainte Justine, Montréal, Québec, Canada. [13]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA. [14]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [15]Université de Montréal and the Montreal Heart Institute, Research Center, Montréal, Québec, Canada. [16]Pavillon Maisonneuve, Montréal, Québec, Canada. [17]Long Island Clinical Research Associates, Great Neck, New York, USA. [18]CHUM – Hopital Sainte-Luc, Montréal, Québec, Canada. [19]Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [20]Laval

University, Quebec City, Québec, Canada. [21]Department of Health Studies, University of Chicago, Chicago, Illinois, USA. [22]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, USA.

§Members of the Wellcome Trust Case Control Consortium

Jan Aerts[1], Tariq Ahmad[2], Hazel Arbury[1], Anthony Attwood[1,3,4], Adam Auton[5], Stephen G Ball[6], Anthony J Balmforth[6], Chris Barnes[1], Jeffrey C Barrett[1], Inês Barroso[1], Anne Barton[7], Amanda J Bennett[8], Sanjeev Bhaskar[1], Katarzyna Blaszczyk[9], John Bowes[7], Oliver J Brand[8,10], Peter S Braund[11], Francesca Bredin[12], Gerome Breen[13,14], Morris J Brown[15], Ian N Bruce[7], Jaswinder Bull[16], Oliver S Burren[17], John Burton[1], Jake Byrnes[18], Sian Caesar[19], Niall Cardin[5], Chris M Clee[1], Alison J Coffey[1], John MC Connell[20], Donald F Conrad[1], Jason D Cooper[17], Anna F Dominiczak[20], Kate Downes[17], Hazel E Drummond[21], Darshna Dudakia[16], Andrew Dunham[1], Bernadette Ebbs[16], Diana Eccles[22], Sarah Edkins[1], Cathryn Edwards[23], Anna Elliot[16], Paul Emery[24], David M Evans[25], Gareth Evans[26], Steve Eyre[7], Anne Farmer[14], I Nicol Ferrier[27], Edward Flynn[7], Alistair Forbes[28], Liz Forty[29], Jayne A Franklyn[10,30], Timothy M Frayling[2], Rachel M Freathy[2], Eleni Giannoulatou[5], Polly Gibbs[16], Paul Gilbert[7], Katherine Gordon-Smith[19,29], Emma Gray[1], Elaine Green[29], Chris J Groves[8], Detelina Grozeva[29], Rhian Gwilliam[1], Anita Hall[16], Naomi Hammond[1], Matt Hardy[17], Pile Harrison[31], Neelam Hassanali[8], Husam Hebaishi[1], Sarah Hines[16], Anne Hinks[7], Graham A Hitman[32], Lynne Hocking[33], Chris Holmes[5], Eleanor Howard[1], Philip Howard[34], Joanna MM Howson[17], Debbie Hughes[16], Sarah Hunt[1], John D Isaacs[35], Mahim Jain[18], Derek P Jewell[36], Toby Johnson[34], Jennifer D Jolley[3,4], Ian R Jones[29], Lisa A Jones[19], George Kirov[29], Cordelia F Langford[1], Hana Lango-Allen[2], G Mark Lathrop[37], James Lee[12], Kate L Lee[34], Charlie Lees[21], Kevin Lewis[1], Cecilia M Lindgren[8,18], Meeta Maisuria-Armer[17], Julian Maller[18], John Mansfield[38], Jonathan L Marchini[5], Paul Martin[7], Dunecan CO Massey[12], Wendy L McArdle[39], Peter McGuffin[14], Kirsten E McLay[1], Gil McVean[5,18], Alex Mentzer[40], Michael L Mimmack[1], Ann E Morgan[41], Andrew P Morris[18], Craig Mowat[42], Patricia B Munroe[34], Simon Myers[18], William Newman[26], Elaine R Nimmo[21], Michael C O'Donovan[29], Abiodun Onipinla[34], Nigel R Ovington[17], Michael J Owen[29], Kimmo Palin[1], Aarno Palotie[1], Kirstie Parnell[2], Richard Pearson[8], David Pernet[16], John RB Perry[2,18], Anne Phillips[42], Vincent Plagnol[17], Natalie J Prescott[9], Inga Prokopenko[8,18], Michael A Quail[1], Suzanne Rafelt[11], Nigel W Rayner[8,18], David M Reid[33], Anthony Renwick[16], Susan M Ring[39], Neil Robertson[8,18], Samuel Robson[1], Ellie Russell[29], David St Clair[13], Jennifer G Sambrook[3,4], Jeremy D Sanderson[40], Stephen J Sawcer[43], Helen Schuilenburg[17], Carol E Scott[1], Richard Scott[16], Sheila Seal[16], Sue Shaw-Hawkins[34], Beverley M Shields[2], Matthew J Simmonds[8,10], Debbie J Smyth[17], Elilan Somaskantharajah[1], Katarina Spanova[16], Sophia Steer[44], Jonathan Stephens[3,4], Helen E Stevens[17], Kathy Stirrups[1], Millicent A Stone[45,46], David P Strachan[47], Zhan Su[5], Deborah PM Symmons[7], John R Thompson[48], Wendy Thomson[7], Martin D Tobin[48], Mary E Travers[8], Clare Turnbull[16], Damjan Vukcevic[18], Louise V Wain[48], Mark Walker[49], Neil M Walker[17], Chris Wallace[17], Margaret Warren-Perry[16], Nicholas A Watkins[3,4], John Webster[50], Michael N Weedon[2], Anthony G Wilson[51], Matthew Woodburn[17], B Paul Wordsworth[52], Chris Yau[5], Allan H Young[27,53], Eleftheria Zeggini[1], Matthew A Brown[52,54], Paul R Burton[48], Mark J Caulfield[34], Alastair Compston[43], Martin Farrall[55], Stephen CL Gough[8,10,30], Alistair S Hall[6], Andrew T Hattersley[2,56], Adrian VS Hill[18], Christopher G Mathew[9], Marcus Pembrey[57], Jack Satsangi[21], Michael R Stratton[1,16], Jane Worthington[7], Matthew E Hurles[1], Audrey

Duncanson[58], Willem H Ouwehand[1,3,4], Miles Parkes[12], Nazneen Rahman[16], John A Todd[17], Nilesh J Samani[11,59], Dominic P Kwiatkowski[1,18], Mark I McCarthy[8,18,60], Nick Craddock[29], Panos Deloukas[1], Peter Donnelly[5,18], Jenefer M Blackwell[61, 62], Elvira Bramon[63], Juan P Casas[64] , Aiden Corvin[65], Janusz Jankowski[66], Hugh S Markus[67], Colin NA Palmer[68], Robert Plomin[14], Anna Rautanen[18], Richard C Trembath[9], Ananth C Viswanathan[69], Nicholas W Wood[70], Chris C A Spencer[18], Gavin Band[18], Céline Bellenguez[18], Colin Freeman[18], Garrett Hellenthal[18], Eleni Giannoulatou[18], Matti Pirinen[18], Richard Pearson[18], Amy Strange[18], Hannah Blackburn[1], Suzannah J Bumpstead[1], Serge Dronov[1], Matthew Gillman[1], Alagurevathi Jayakumar[1], Owen T McCann[1], Jennifer Liddle[1], Simon C Potter[1], Radhi Ravindrarajah[1], Michelle Ricketts[1], Matthew Waller[1], Paul Weston[1], Sara Widaa[1], Pamela Whittaker[1].

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK. [2]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, EX1 2LU, UK.[3]Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 0PT, UK. [4]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 0PT, UK. [5]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK. [6]Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, LS2 9JT, UK. [7]ARC Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK. [8] Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. [9]Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London, SE1 9RT, UK. [10]Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research, University of Birmingham, Birmingham, B15 2TT, UK. [11]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. [12] IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK. [13]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [14]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. [15]Clinical Pharmacology Unit, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK. [16]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. [17]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. [18]The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [19]Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham, B15 2FG, UK.[20]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, G12 8TA, UK. [21]Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [22]Academic Unit of Genetic Medicine, University of Southampton, Southampton, UK. [23]Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK. [24]Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. [25]MRC Centre for

Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol, BS8 2BN, UK. [26]Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 0JH, UK. [27]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK. [28]Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK. [29]MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK. [30]University Hospital Birmingham NHS Foundation Trust, Birmingham, B15 2TT, UK. [31]University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, OX3 7LD, UK. [32]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB, UK. [33]Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, AB25 2ZD, UK. [34]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK. [35]Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK. [36]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK. [37]Centre National de Genotypage, 2, Rue Gaston Cremieux, Evry, Paris 91057, France. [38]Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. [39]ALSPAC Laboratory, Department of Social Medicine, University of Bristol, BS8 2BN, UK. [40]Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK. [41]NIHR-Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. [42]Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY, UK. [43]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK. [44]Clinical and Academic Rheumatology, Kings College Hosptal National Health Service Foundation Trust, Denmark Hill, London SE5 9RS, UK. [45]University of Toronto, St. Michael's Hospital, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada. [46]University of Bath, Claverdon, Norwood House, Room 5.11a Bath Somerset BA2 7AY, UK. [47]Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK. [48]Departments of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester, LE1 7RH, UK. [49]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [50]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. [51]School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, S10 2JF, UK. [52]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Windmill Road, Headington, Oxford, OX3 7LD, UK. [53]UBC Institute of Mental Health, 430-5950 University Boulevard Vancouver, British Columbia, V6T 1Z3, Canada. [54]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland, 4102, Australia. [55]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [56]Genetics of Diabetes, Peninsula College of

Medicine and Dentistry, University of Exeter, Barrack Road, Exeter, EX2 5DW, UK. [57]Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK. [58]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [59]Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, LE3 9QP, UK. [60]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, OX3 7LJ, UK. [61]Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, 100 Roberts Road, Subiaco, Western Australia 6008. [62]Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK. [63]Department of Psychosis Studies, NIHR Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London SE5 8AF, UK. [64]Department Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT and Dept Epidemiology and Public Health, University College London WC1E 6BT, UK. [65]Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Eire. [66]Department of Oncology, Old Road Campus, University of Oxford, Oxford OX3 7DQ, UK, Digestive Diseases Centre, Leicester Royal Infirmary, Leicester LE7 7HH, UK and Centre for Digestive Diseases, Queen Mary University of London, London E1 2AD, UK. [67]Clinical Neurosciences, St George's University of London, London SW17 0RE, UK.[68]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK.[69]NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London EC1V 2PD, UK. [70]Department Molecular Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK.