Supplementary Note

Contents

1. Sample ascertainment

1.1 Inflammatory bowel disease cases

Following ethical approval by Cambridge MREC (reference: 03/5/012), individuals with inflammatory bowel disease (IBD) were consented into the study and donated blood or saliva for DNA extraction at IBD clinics in and around clinical centres that contribute samples to the United Kingdon Inflammatory Bowel Disease Genetics Consortium (UKIBDGC) (Cambridge, Dundee, Edinburgh, Exeter, London, Manchester, Newcastle, Norwich, Nottingham, Oxford, Sheffield, Torbay and the Scottish early onset IBD project). Ascertainment was based on a confirmed diagnosis of Crohn's disease (CD) or ulcerative colitis (UC) using conventional endoscopic, radiological and histopathological criteria. We included all subtypes of CD and UC and the collection was not specifically enriched for family history or early age of onset.

1.2 Population controls

To maximise the number of cases we could sequence within our budget, and negate the need to ascertain population controls as part of this experiment, we obtained whole-genome sequence data from 3,910 UK population controls ascertained and sequenced by the UK10K consortium. A full description of this cohort is provided in the UK10K manuscript [1]. Briefly, this cohort consists of 6,557 samples from the Avon Longitudinal Study of Parents and Children (http://www.bristol.ac.uk/alspac/) and 2,575 from the Twins UK cohort (http://www.twinsuk.co.uk).

2. Whole-genome sequencing and data processing

2.1 Sequence data generation

Low read-depth whole-genome sequencing (WGS) of 1,817 UC cases, 2,697 CD cases and 3,910 controls was performed at the Wellcome Trust Sanger Institute (WTSI), while 2,354 controls were sequenced by the Beijing Genomics Institute [1]. DNA (1-3 µg) extracted from the blood or saliva of IBD cases, or lymphoblastoid cell lines (ALSPAC) or PBMCs (TwinsUK) from controls, was sheared to 100–1000 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation. Following size selection (300-500 bp insert size), DNA libraries were sequenced using the Illumina HiSeq platform as paired-end 100 base reads according to manufacturer's protocol.

2.2 Read mapping

Sequence data was aligned to the human reference by the sequence centre. Due to changes in the informatics pipeline over the course of the sequencing, two different versions of the GRCh37 human reference were used:

R.1 The reference used in Phase I of 1,000 Genomes [2] – the GRCh37 primary assembly.

R.2 The reference used in Phase II of 1,000 Genomes Project [3] – the new assembly integrates reference sequences from **R.1**, human herpesvirus and the concatenated decoy sequences.

`BWA` (v0.5.9-r16) [4] was used for sequencing reads alignments. For each *fasta* file

(<*seq.fasta*>) produced from per-lane level sequencing, the following steps were employed for *BAM* file generation:

a) Align pair-end reads to target reference

```
bwa aln -q 15 -b1 <reference.fa> <seq.fasta> > seq.1.sai
bwa aln -q 15 -b2 <reference.fa> <seq.fasta> > seq.2.sai
```

b) Create *SAM* files

```
bwa sampe <reference.fa> <seq.1.sai> <seq.2.sai> <seq.fasta> <seq.fasta>
```

c) Create correct read pairing information using `samtools-0.1.16` (r963:234) [5] to resolve unusual flag information on *SAM* records

```
samtools fixmate <seq.sam> <seq_fixmate.bam>
```

d) Create coordinate sorted *BAM* files from name sort *BAM*

```
samtools sort <seq_fixmate.bam> <seq_sorted.bam>
```

The *BAM* files produced from the pipeline above were submitted to the European Genome-phenome Archive (EGA):
https://www.ebi.ac.uk/ega/datasets/EGAD00001000409
https://www.ebi.ac.uk/ega/datasets/EGAD00001000401

2.2.1 Unifying BAM files to the same reference genome: BridgeBuilder

The computational cost for realignment of all sequence reads to the same reference genome is high. Thus we developed the software `BridgeBuilder` (github.com/wtsi-hgi/bridgebuilder) to efficiently realign all BAM files to the **R.2** reference. This method avoids the need to perform a computationally expensive realignment of all reads to the new reference, and instead only requires alignment of reads to a

subset defined by the differences between the two reference sequences. Any reads that align to the differential reference are remapped.

`BridgeBuilder` has three components, executed in the following order:

`baker`

Generation of the "reference bridge", mapping the old reference to the new reference. The result is metaphorically a collection of bridges representing regions of the former reference and their new destination in the latter.

`binnie`

The alignment of every read against the reference bridge produced by baker to determine whether remapping is required. For every input file, binnie populates each original aligned read in to one of three bins:

a) Unchanged reads

Reads that do not align to the reference bridge and do not require remapping.

b) Bridged reads

Reads which align to the reference bridge with a superior mapping score and require remapping.

c) Newly mapped reads

Reads that did not have an alignment previously, but now align to the reference bridge and thus can be mapped to the newer reference.

`brunel`

Takes the sorted binnie bins as "blueprints", interleaving reads to maintain co-ordinate sort order and generate the final new alignment.

2.2.2 BAM quality control

Automatic quality control of the *BAM* files was performed using pipelines developed at the WTSI. For each sample a subset of metrics was compared to hard-coded thresholds (that have typically been determined empirically from previous datasets) to raise either a warning or generate a complete failure for that sample. The metrics used during this autoQC process and thresholds are described in Supplementary Table 1.

`bamcheckR` [6] was also used to generate *BAM* statistics supplementary to those output from `samtools stats` and evaluate overall sample quality.

2.3 Variant calling

Next, we converted *BAM* files into genomic positions. For Single Nucleotide Variants (SNVs) and small INsertions and DELetions (INDELs), we used `samtools` and `bcftools` to first produce a *BCF* file that contained genomic locations, and then used this information to call genotypes. We used `GenomeSTRiP` [7] for Copy Number Variant (CNV) discovery. In the following sections, we briefly explain how different types of genetic variants were called.

2.3.1 SNVs and INDELs

SNVs and INDELs were called using `samtools-0.19` and `bcftools-0.19` (version: 0.1.19-58-g3d123cd) [8] by pooling the alignments from 8,354 sample-level low read-depth BAM files. Genotype likelihood files (bcf) for all-samples and all-sites were created with the `samtools mpileup` command

```
samtools mpileup -EDVS -C50 -pm3 -F0.2 -d 10000 -g -f hs37d5.fa
```

Variants were then called using the `bcftools` command to produce a VCF file

```
bcftools view -Ngvm0.99 <in.bcf>
```

Male samples were called as diploid in the *Pseudo-Autosomal Region* (PAR) on chromosome X, and haploid otherwise. The non-PAR regions were defined as:

X: 1-60000

X: 2699521-154931043

The pipeline (`run-mpileup`) used to created the calls is available from:

`https://github.com/VertebrateResequencing/vr-codebase/blob/develop/scripts/run-mpileup`

2.3.2 Copy number variants

CNVs were called using `GenomeSTRiP 2.0`, which was designed to discover and genotype shared deletions, duplications and multiallelic copy number variants (mCNVs) across whole-genome sequences from multiple individuals. As this study uses low coverage sequences, power to detect variation is limited to larger CNVs. Thus `GenomeSTRiP 1.0`, which is more sensitive to smaller deletions and therefore usually recommended as a complementary CNV analysis, was not used for this project.

Default `GenomeSTRiP` configurations were used, as per the example config files provided within the software releases. Window sizing parameters, which define the size of CNVs that can be detected, matched those used for the 1,000 Genomes Project's low coverage (6-8x) dataset:

```
tilingWindowSize 5000
tilingWindowOverlap 2500
```

`maximumReferenceGapLength 2500`

`boundaryPrecision 200`

`minimumRefinedLength 2500`

Because reads realigned from **R.1** to **R.2** using `BridgeBuilder` did not contain appropriate metadata information for use by `GenomeSTRiP 2.0`, these reads were excluded from discovery and genotyping.

2.4 Variant filtering

Following variant calling, a number of machine learning methods were used to assess qualities of each called variant. We used this quality information to filter the raw call set to produce a set of high quality variant sites.

2.4.1 SNVs

Support Vector Machines (SVMs) were used to identify poor quality SNP calls in the sequence data. A SVM is a supervised learning model that trains on highly confident known sites to determine the probability that sites outside of the training set are true, based on various quality metrics generated with `samtools-0.19` [8], including:

- DP: Raw read depth
- MQ: Root-mean-square mapping quality of covering reads
- AN: Total number of alleles in called genotypes
- MDV: Maximum number of high-quality non-Ref reads in samples
- EDB: End Distance Bias

- RPB: Read Position Bias

Five independent SVMs were run in parallel and only sites that passed at least two out five SVM runs were considered high quality. The training set for each SVM consisted of 1,000 'good sites' that overlapped with HapMap3 [9] and 1,000 'bad sites' with quality score ($QUAL$) < 10 in the raw vcf file. To ensure we had a balanced number of variants selected across the full minor allele frequency (MAF) spectrum in both good and bad training sets (i.e. not all bad sites were rare and not all good sites were common), we preserved the original MAF proportion in each of the SVM training sets. The following MAF bins were used for preserving the MAF range:

- $0 \leq$ MAF $< 0.5\%$
- $0.5\% \leq$MAF$< 5\%$
- MAF$\geq 5\%$

The pipeline (`run-filter`) used to filter SNVs is available from:

`https://github.com/VertebrateResequencing/vr-codebase/blob/develop/scripts/run-filter`

2.4.2 INDELs

Variant Quality Score Recalibration (VQSR) was used to filter INDELs. For short INDELs called with `samtools-0.19`, the `GATK UnifiedGenotyper` [10] (version 2.1-5-gf3daab0) was used for recall in order to generate the annotations needed for recalibration. The `GATK VariantRecalibrator` was then used for INDEL filtering, followed by `GATK ApplyRecalibration` to assign VQSLOD (variant quality score log odds ratio) values to each INDEL. For INDEL filtering VQSR considers the following annotations generated using UnifiedGenotyper:

- DP: Approximate read depth (reads with MQ= 255 or with bad mates are removed)
- FS: Phred-scaled p-value using Fisher's exact test to detect strand bias
- ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
- MQRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read mapping qualities

The Mills-Devine dataset [11], an INDEL call set that has been validated to a high degree of confidence, and is recommended by the GATK workflow, was used as the truth set for the VQSR model training. A truth sensitivity threshold of 97%, which corresponded to a minimum VQSLOD score of 1.0659 was chosen for INDEL filtering.

2.4.3 Copy number variants

Initial CNV filtering was performed in accordance with the default thresholds set in the `GenomeSTRiP 2.0 CNVDiscoveryPipeline` workflow. These thresholds are generous, and many poor-quality sites are expected to remain: nevertheless, this process removed 86,379 variants (out of 179,774) variants from the discovery set, and made manual quality control more manageable. The filters applied at this step include:

Deletion or mixed CNV length $> 1,000$. Given the search windows used, this still allows variants slightly smaller than those we expect to confidently detect to be included.

Duplication length $> 2,000$. This follows the recommendations of Handsaker et al, who note that small duplications appear to have a higher false discovery rate

than equivalently sized deletions or mixed CNVs [12].

Call rate > 0.9, to remove those variants with excessive missingness.

Density > 0.5. Density is calculated by dividing GSELENGTH (the effective CNV length) by GCLENGTH (the denominator of GC content).

Cluster separation > 5. This measure checks that appropriate cluster separation was achieved by the Gaussian mixture model used in read depth genotyping.

GSVDJFRACTION > 0. Remove variants with any evidence of V(D)J recombination, based on the vdjregions.bed file provided with the `GenomeSTRiP` metadata.

We then applied the following additional filters based on:

Missing sample data. We removed 1,103 copy number variants that were driven by 95 control samples with a large stretch of missing data on chromosome 6.

GSELENGTH > 60,000. For shorter copy number variants, we observed considerable differences in sensitivity across different mean coverage depths (Supplementary Figure 4).

Biallelic sites. We only kept biallelic sites, for simplicity when association testing. However, because `GenomeStrip 2.0` is capable of calling multiallelic CNVs, we noted an abundance of common sites where a small fraction of alt individuals contain a CNV in the opposite direction to the majority call, possibly due in part to our particularly low coverage. At sites where this fraction of inconsistent directions is less than 10% of the alt calls made, we retain the site as biallelic.

2.5 Genotype refinement

Post SNV and INDEL quality control (Section 2.4.1 and 2.4.2), genotypes at all passing sites were refined via imputation, as is standard in low-coverage sequencing studies. To increase computational efficiency, imputation was performed in

batches of 3,000 sites, with a buffer region of 500 sites up- and down-stream, using default parameters in BEAGLE [13] (version: v4.r1196):

```
java -jar b4.r1196.jar gl=<in.vcf.gz> out=<out.vcf.gz>
```

The pipeline (`run-beagle`) used for genotype refinement is available from:

`https://github.com/VertebrateResequencing/vr-codebase/blob/develop/scripts/run-beagle`

2.5.1 Sample quality control

The following sample quality control criteria were applied based on refined genotypes (Section 2.5):

Excessive heterozygosity rate ±3.5 standard deviations from the mean. Heterozygosity rate was calculated using `PLINK2 --ibc` (version: 1.9) [14], which computes the method-of-moments F coefficient:

$$\text{heterozygosity rate } (F) = \frac{[\text{observed homozygosity count}] - [\text{expected count}]}{[\text{total observations}] - [\text{expected count}]} \quad (2.1)$$

Duplicated or related individuals with $\hat{\pi} > 0.25$ (second-degree relatives or closer). To identify duplicate and related individuals, SNVs were first pruned such that no two sites within 5,000kb had an $r^2 > 0.2$. Identity-By-State (IBS) was then calculated for each pair of individuals using only variants with MAF > 1%. The degree of recent shared ancestry for each pair of individuals (Identity-By-Descent, $\hat{\pi}$) was then estimated using the following `PLINK2` commands:

```
plink --bfile <plinkfile> --indep-pairwise 5000 1000 0.2
plink --bfile <plinkfile> --maf 0.01 --extract <file.prune.in> --genome
```

One individuals from each pair with $\hat{\pi} > 0.25$ was then removed from this particular analysis.

Individuals of non-European ancestry were identified and removed based on a principal component model built on genotype data from 11 different HapMap3 populations (Supplementary Figure 10). In total $1,343,150$ sites were present in both the HapMap3 data and our sequenced samples. These sites were then pruned such so that no pair of SNPs had $r^2 > 0.2$, and known regions of high LD were excluded. Principal components were generated based on the HapMap3 samples and the factor loading used to project the principal components for our sequenced samples. All individuals with a second principal component score less than 0.08 were excluded. The following PLINK2 commands were used to identify individuals of divergent ancestry:

```
plink --bfile <ibd-hm3> --exclude range high-LD-regions.txt
      --indep-pairwise 5000 1000 0.2
plink --bfile <ibd-hm3> --extract <ibd-hm3.prune.in> --maf 0.05 --pca
```

2.5.2 Variant quality control

In order to improve the genotype refinement quality and reduce the false-positive rate in the association study,the following variants were removed after initial BEAGLE genotype refinement:

Hardy Weinberg exact test P-value in controls $< 10^{-7}$;

Removal of sequencing centre batch effects in controls. The control data were sequenced at two different centres (WTSI and BGI) (Section 1.). To investigate the presence of batch effects, we fitted a logistic regression model to assess differences in allele frequencies between two centres for each variant. Variants with P-value $\leq 10^{-3}$ were removed from subsequent analysis;

Variants with $> 10\%$ **missing genotypes following genotype refinement**,

where the minimum posterior probability required to call a genotype was 0.9;

SnpGap (3) filters SNPs within 3 base pairs of an indel;

IndelGap (2) filters clusters of INDELs separted by 2 or fewer base pairs allowing only one to pass.

Following these exclusions, a second round of genotype refinement was undertaken using `BEAGLE`. Supplementary Table 3 summarises the results from the above variant quality control steps.

2.5.3 Data quality evaluation

We evaluated our data quality by comparing the variant overlap with the 1000 Genomes Project Phase 3 European data to assess the sensitivity and specificity of our call set. We then evaluated the genotypic quality of our sequencing data by means of genotypic concordance rate (r^2) comparing to five genotyped datasets with partially overlapping samples.

Sensitivity and specificity compared to 1000 Genome Project Phase 3 European panel

To assess how well our data represents the variation in the European population, we compared the biallelic SNVs in autosomal regions identified in our project to that in the 1,000GP Phase 3 European panel (503 individuals). The left panel of Supplementary Figure 3 shows the percentage of SNVs identified at different QC stages in the IBD sequencing project that are also present in the 1000GP set. As the QC criteria becomes more stringent, the sensitivity of our call set increases. 98% of SNVs with MAF $\geq$ 1% overlap with 1000GP after the genotype refinement stage, and this percentage increases to more than 99% for variants which are retained

for association testing. 55 million variants at the post genotype-refinement stage were not previously seen before, the majority of which were singletons (~ 53M), doubletons (~ 0.5M) or rare (MAF ≤ 1%) variants (~ 10M) in our data. Details of the number of sites are listed in Supplementary Table 5a

Overall, our data covers the majority of low frequency (91.0%) and common variants (99.1%) discovered in 1000GP Phase 3 European panel (Supplementary Figure 3 left panel). This indicates that our variant filtering strategies have limited the number of false-positive sites and provided good sensitivity when compared to 1000GP dataset.

Genotypic accuracy compared against GWAS and Immunochip datasets To evaluate the sequencing accuracy after genotype refinement, we compared the probability dosage yield from our sequencing data to existing genotype datasets on the overlapping samples - including that from an IBD Immunochip project [15], the Wellcome Trust Case Control Consortium (WTCCC) 1 Crohn's disease GWAS project [16] (Section 4.1.1) and the WTCCC2 UC GWAS project [17] (Section 4.1.2). Summary statistics of the overlapping samples and variants between the sequencing cohort and the comparison cohorts are listed in Supplementary Table 4. Across individuals present in a given microarray dataset and our sequenced cohort, we calculated a Dosage r^2 at each site present in both datasets. For the majority of variants with MAF ≥ 1%, the sequencing genotypes were > 90% concordant with other genotype data (Supplementary Figure 2). The Immunochip data has lower mean r^2 because it contains fewer shared low frequency variants, and it therefore has a larger confidence interval. Overall, we conclude that our sequencing data is comparable to genotyped data for common variations.

3. Whole-genome sequence association studies

3.1 Single-variant association study

Single variant logistic regression association tests were performed using `SNPTEST v2.5` [18] based on the post refinement genotype likelihoods.

$$\log \frac{p_i}{1-p_i} = \alpha + \beta G_{ij} \tag{3.2}$$

where G_{ij} denotes the genotype of the ith individual at the jth variant.
Three independent genome-wide single-locus based association studies were performed conditional on the first 10 principal components for 2,513 CD cases, 1,767 UC cases and 4,280 IBD cases versus the same 3,652 controls post QC samples (Section 2.5.1). In total, ~ 12.7M variants with MAF $\geq 0.1\%$ were tested for association. Genomic inflation factors (λ_{1000}) for an equivalent study of 1000 cases and 1000 controls are $\lambda_{CD} = 1.04$, $\lambda_{UC} = 1.05$ and $\lambda_{IBD} = 1.06$ (Supplementary Figure 5).

3.1.1 Additional variant quality control

Additional variant filtering was applied post single variant association testing, in addition to that described in Section 2.5.2.

minSVM Score < 0.1. As described in Section 2.4.1, five SVM scores were available for each site. We removed those that had a SVM score less than 0.1 in any of the 5 runs.

Imputation $r^2 \geq 0.4$. Variants with an imputation quality score less than 0.4 in SNPTEST2 were removed.

Hardy Weinberg equilibrium exact test P-value in controls $< 10^{-6}$.

3.2 Rare variant burden association study

Rare variant burden tests were performed using the Robust Variance Score (RVS) statistic developed by Derkach et al (2014) [19], as shown in Equation 3.3. This method adjusts for differences in read depth between cases and controls by calculating the variance of the score separately for each group, as described in Equation 3.4.

$$S_j = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) E\left(G_{ij}|D_{ij}\right) \tag{3.3}$$

$$Var\left(S_j\right) = \sum_{cases} \left(1-\bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) + \sum_{controls} \left(\bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) \tag{3.4}$$

The corresponding test statistic for association at a single site, $T_j = \frac{{S_j}^2}{Var(S_j)}$ is chi-squared distributed, with one degree of freedom. The test incorporates the expected value of the genotype given the data, $E(G_{ij}|D_{ij})$, which reflects the dosage of the alternate allele at the given site, and is calculated using genotype probabilities (Equation 3.5). By using a statistic based on genotype probabilities, this method accounts for uncertainty in the genotype call, helping to adjust for the poor individual genotype quality observed in low coverage data.

$$E\left(G_{ij}|D_{ij}\right) = \sum_{g=0}^{2} gP(G_{ij} = g|D_{ij}) \tag{3.5}$$

The basic statistic is then extended to perform a joint analysis of multiple rare variants. The individual variant score statistics are summed together to give an overall score, while the variance component is calculated by combining the *covariance* matrices of the cases and controls, after estimating them separately. Significance is then evaluated using bootstrap permutation.

This test was implemented as an extension to the software suite ANGSD [20]. Code is available at https://github.com/katiedelange/angsd.

3.2.1 Generating genotype probabilities

Genotype refinement via imputation produces a set of 'smoothed' genotype probabilities, making use of population-level information to remove noise and improve confidence in genotype calls made (see Section 2.5). However, when the true signal is low, such as for sites of rare variation, this refinement step tends to be overzealous, and generates poorly calibrated individual genotype probabilities (Supplementary Figure 8).

Therefore for rare variant analyses, we used genotype probabilities generated directly from the samtools Genotype Quality (GQ) field, without any genotype refinement. The GQ value represents the phred-scaled genotype probability of the most likely genotype. We assumed that, given the low MAF ($\leq 0.5\%$ in controls) of the variants being considered here, the rare homozygote is not observed and thus we defined the genotype probabilities as described below:

$$\begin{aligned} P(\text{Genotype called in VCF}) &= 1 - 10^{\frac{-\text{GQ}}{10}}, \\ P(\text{Alt}) &= 1 - P(\text{Genotype called in VCF}), \end{aligned} \tag{3.6}$$

where the possible (Call,Alt) pairs are (RR,RA), (RA,RR), and (AA,RA)

3.2.2 Additional variant quality control

Additional site filtering was used, as rare sites are more susceptible to differences in read depth between cases and controls (Supplementary Figure 11). As well as the QC procedures described in Sections 2.4 and 2.5, the following filters were used:

Missingness calculated from GQ-generated genotype probabilities ≤ 0.1, as this rate differs slightly from that produced following genotype refinement.

High confidence observations $> 99\%$ **of non-missing data**, where a high confidence observation is that with a genotype probability ≥ 0.9 for the most likely genotype.

High confidence alternate allele observations ≥ 2 in the complete dataset. This excluded singletons from the analyses, as they contained too many false positives, particularly amongst the very low coverage ulcerative colitis samples (Supplementary Figure 11a).

INFO score ≥ 0.6, calculated separately for all appropriate association cohorts (CD, UC, IBD, controls). For association tests in IBD, variants had to pass this filter within the CD and UC cohorts individually, as well as across the entire IBD subset. The INFO score α (Equation 3.7) is the same as that implemented in `SNPTEST` and `IMPUTE2` [18], and can be interpreted as describing the amount of 'missing' information, such that the observed data in a sample of size N is equivalent to a set of perfectly observed genotypes in a sample of size αN.

$$\alpha = \frac{\frac{2N_{case/control}}{\hat{\theta}(1-\hat{\theta})} - \frac{\sum_{i=1}^{N_{case/control}} E(G_{ij}|D_{ij}) - E(G_{ij}^2|D_{ij})}{\hat{\theta}^2(1-\hat{\theta})^2}}{\frac{2N_{case/control}}{\hat{\theta}(1-\hat{\theta})}}, \text{ where } \hat{\theta} = \frac{\sum_{i=1}^{N_{total}} E(G_{ij}|D_{ij})}{2N_{total}} \tag{3.7}$$

3.2.3 Coding variation in genes

Burden tests were performed across sites with a MAF $\leq 0.5\%$ in controls and falling within a given gene as defined by annotation with an Ensembl ID. For each gene, two sets of burden tests were performed to include all functional coding variants, and all predicted damaging functional coding variants. The particular `Variant Effect Predictor` [21] annotations used to define these variant groups are detailed in Supplementary Table 8. `Combined Annotation Dependent Depletion` (CADD) scores [22] were used to further subset annotated sites into those that were predicted to have damaging consequences (CADD score ≥ 21).

Every test was repeated to independently check for association with CD, UC and IBD at every gene containing one or more relevant variants. This resulted in a total of 100,335 tests, with an average of 5.84 variants contributing to each test (Supplementary Table 9). To reduce computational load, adaptive permutation was used, whereby the significance of the test would be evaluated every 10^x permutations (starting from $x = 5$). Only tests with fewer than 100 permutations more significant than the unpermuted sample were continued. Results from these tests are summarised in Supplementary Figure 9, and Supplementary Table 10.

For *NOD2*, the only gene for which we observed a significant signal, we evaluated the independence of this signal from the known common coding variants rs2066844, rs2066845, and rs2066847. Individuals with a minor allele at any of these sites were assigned to one group, and those with reference genotypes to another. Burden testing for this new phenotype in both variant sets that contained a significant CD vs controls signal produced $P_{functional} = 0.0117$ and $P_{damaging} = 0.7311$. On average, contributing rare variants were at an elevated frequency in non-NOD2 canonical mutation carriers, compared to those individuals with a minor allele at any of these three sites.

3.2.4 Gene set tests

To increase power to detect rare variant associations across coding regions, individual gene results were combined into gene sets as defined in Supplementary Table 11. The gene sets were analysed using a meta-analysis approach, rather than performing a complete burden test on all constituent variants, to overcome any differences in the direction of effect of rare variants in the genes included in the set. The absolute scores for each gene in the set were summed, as were the variances, across 100,000 permutations. Thus, while covariance was included for intra-gene variant relationships, the inter-gene covariance was not accounted for, although we expect this to be of minimal consequence. Individual set statistics were then evaluated against the statistics from the set of *all* genes, in an approach based on Purcell et al's SMP method [23], to account for residual case-control coverage bias.

Given the relative strength of the *NOD2* signal, each gene set test was performed both with and without *NOD2* (where appropriate). Results from these tests can be found in Supplementary Table 12.

3.2.5 Non-coding variation in enhancers

Using the same approach outlined above for individual genes, burden tests were performed across enhancer regions as defined by the FANTOM5 project [24]. Within each robustly defined enhancer, we tested all observed rare variation, as well as the subset predicted to disrupt or create a transcription factor binding motif. Disruption or creation of a transcription factor binding motif was determined using the same approach employed by Huang et al [25], thus we considered all ENCODE transcription factor ChIP-seq motifs [26] with an overall information content (IC)

≥ 14 bits (equivalent to 7 perfectly conserved positions) and checked if a given variant created or disrupted that motif at a high-information site (IC≥ 1.8).

We again repeated each test to independently check for association to UC, CD and IBD at every enhancer with one or more relevant variants, resulting in 121,848 tests, with an average of 2.27 variants contributing to each test (Supplementary Table 13).

3.2.6 Enhancer set tests

Individual enhancers were combined into enhancer sets based on cell and tissue-specific expression. Using pre-defined tracks (http://enhancer.binf.ku.dk/presets/) as described by Andersson et al, we tested all enhancers that were positively differentially expressed in each of 69 cell types and 41 tissues (Supplementary Table 17) [24]. Note that positive differential expression is not the same as exclusive expression in a given cell/tissue.

Using the same SMP-based approach that was used to analyse gene sets, we tested the cell and tissue enhancer sets against the background of all robustly defined FANTOM5 enhancers, both for all observed rare variation and that predicted to disrupt or create a transcription factor binding motif. Results from these tests are summarised in Supplementary Table 14.

4. GWAS cohort and imputation

4.1 GWAS cohort description

We collected a large GWAS cohort that consisted of three distinct studies: the Wellcome Trust Case Control Consortium (WTCCC) 1 Crohn's disease GWAS [16], the WTCCC2 ulcerative colitis GWAS [17], and a new IBD GWAS collected and genotyped at the Wellcome Trust Sanger Institute between 2014 and 2015. Cumulatively these studies contain over 12,000 IBD cases and 15,000 controls, genotyped on a combination of different chips.

4.1.1 WTCCC1

Post-QC, the WTCCC1 study contains 1,748 CD cases and 2,936 controls, genotyped on the Affymetrix 500K chip. As the genotypes were originally aligned to reference build 35, the UCSC software tool `liftOver` [27] was used to update the data to reference build 37. Successful conversion was achieved for a total of 458,817 sites.

4.1.2 WTCCC2

Similarly, post-QC the WTCCC2 study included 2,361 UC cases and 5,417 controls (some of which overlapped with the WTCCC1 study), genotyped on the Affymetrix 6.0 array. The reference was updated to build 37 from build 36 using liftOver. As strand alignment had not been performed on this dataset, misaligned SNPs were detected using `SHAPEIT -check` (version: v2.r790) [28]. Ambiguous SNPs with a MAF > 0.4 were removed, and a final pass to flip misaligned SNPs was performed by comparing sample allele frequencies to the European allele frequencies in

the 1000 Genomes Project. After lift over and strand alignment, 735,782 sites remained.

4.1.3 Novel GWAS cohort

A novel GWAS cohort (GWAS3) was collected, consisting of 5,695 CD cases, 5,299 UC cases, 764 indeterminate IBD cases, and 10,484 controls. Both cases and controls were genotyped at the Wellcome Trust Sanger Institute; controls on the Human Core Exome v12.0 chip, and cases on the Human Core Exome v12.1 chip. Genotypes were called using optiCall [29], and then strand aligned using files provided by William Rayner (http://www.well.ox.ac.uk/ wrayner/strand/). Sites not included on both versions of the chip were removed, leaving a total of 535,434 genotyped sites. Prior to sample quality control, these sites were then pruned further to remove those with an **excessive missingness rate** > 5%. Per SNP genotype missingness rate was calculated using `PLINK2 -missing` (version: 1.9) [14].

Samples were filtered using the following quality control thresholds:
Excessive heterozygosity rate ±3 standard deviations from the mean. Heterozygosity rate was computed using `PLINK2 -het` (version: 1.9) [14], which calculates the method-of-moments F coefficient (see Equation 2.1).
Excessive missingness rate > 1%. Per sample missingness rate was calculated using `PLINK2 -missing` (version: 1.9) [14].
Mismatching gender between that recorded at patient recruitment and that determined genetically (unless a valid explanation for the mismatch was available). Genetic genders were obtained using `PLINK2 -check-sex` (version: 1.9) [14], which imputes the inbreeding coefficient F (Equation 2.1) for the X chromosome. Under Hardy-Weinberg Equilibrium, females should have an X-chromosome F

coefficient close to 0, while for males it should be close to 1.

Duplicated or related individuals with kinship coefficient > 0.177 (first-degree relatives or closer). Kinship coefficients were calculated for samples passing the heterozygosity and missingness checks, using markers with a MAF > 0.05 and the software `KING` [30]. The sample with the lowest call rate (or mismatching gender, if applicable) of each related pair was removed.

Non-European samples as determined by Principal Component Analysis (PCA). Principal components were calculated together with samples from the HapMap3 project [31], using `SMARTPCA.perl` [32]. Individuals with a PC2 score less that 0.067 were defined as non-European and removed from further analysis.

A final set of quality control filters were then used to remove markers still performing poorly on the high-quality samples:

Significant difference in call rate between cases and controls. Significance was evaluated using `PLINK2 -test-missing` (version: 1.9) [14], and those sites with $p < 1e^{-5}$ were removed.

Hardy-Weinberg equilibrium (HWE) exact test P-value in controls $< 1e^{-5}$. Tests for HWE were performed with `PLINK2 -hwe` (version: 1.9) [14], using the mid-p modifier.

Genotyping batch effect, affecting 429 markers. These sites were identified by computing within-sample principal components (PCs) using common variants (MAF > 1%), which highlighted a clear outlier group of case samples all belonging to one genotyping batch. PC1 was then used to split cases into outliers and non-outliers, and an association test between these groups was used to identify significant sites ($p < 1e^{-5}$). Once these sites were removed, the within-sample PCs no longer produced any outlier groups.

This left a high-quality dataset consisting of 510,520 genotyped sites in 9,239

cases (4,474 CD, 4,173 UC, 592 indeterminate IBD), and 9,500 controls. Before imputation, these sites were further pruned to those with a MAF $> 0.1\%$, leaving a total of 296,203 markers.

4.2 Imputation

Whole genome sequences were imputed for the genotyped samples using a reference panel containing the IBD-affected and control sequence data described in Section 1., together with the 1000 Genomes Project Phase 3 whole genomes. Given the size of the resulting reference and genotype panels, imputation was performed using `PBWT` [33] so results could be obtained in a tractable amount of time.

4.2.1 Novel sequencing reference panel

Re-phasing of IBD sequencing samples

Following the second round of genotype refinement on the sequencing data, `SHAPEIT2` (version: v2.r790) [28] was used to increase the accuracy of the estimated haplotypes. To maintain computational efficiency, batches of 100,000 sites were phased, with 5,000 sites in buffer regions either side of these. The maximum chromosome length was set to 249,250,621 base pairs. `SHAPEIT2` was run with the following parameters:

```
--input-map <1000GP_phase1interim_jun2011_genetic_map.txt>
--thread 16 --window 0.5 --states 200
```

`bcftools convert` (version: 1.1-82-g4f3a265) was used to combine the original VCF with the new phase information. Batches were merged using `bcftools concat` (version: 1.1-82-g4f3a265) and phase determined by matching overlapping heterozygous sites. The pipeline (`run-shapeit`) used to perform haplotype estima-

tion is available from:

https://github.com/VertebrateResequencing/vr-codebase/blob/develop/scripts/run-shapeit

Creation of a new IBD, 1000G Phase 3 and UK10K imputation reference

The haplotypes from 4,686 IBD samples (retaining those excluded from association analyses due to non-European ancestry) are then combined with 3,781 UK10K and 2,504 1000 Genomes Phase 3 control sequences, to create a new reference panel enriched with low frequency and rare variants detected from our IBD whole-genome sequences.

4.2.2 Preparation of GWAS data for imputation

Three separate imputation panels were created for input into PBWT:

A: All WTCCC1 cases and controls

B: All WTCCC2 cases, and controls not already included in panel A

C: All GWAS3 cases and controls not already included in panels A or B

Prior to imputation, we also removed any genotyped samples already included in the sequencing study (as these would be present in the reference panel). After imputation we had whole genome sequences for 11,987 cases and 15,191 controls (Supplementary Table 6).

4.3 Variance explained

4.3.1 Heritability estimation

Using the imputed whole genome sequences, we applied the Restricted Maximum Likelihood (REML) method implemented in GCTA [34, 35, 36] to estimate the vari-

ance explained by all the autosomal SNVs. Individuals in three GWAS cohorts and variants that passed quality controls post imputation (Section 4.2) were used to estimate the genetic heritability (h^2_g) explained for UC and CD, respectively.
Since heritability estimation represents the sum of association across all variants, even small spurious associations due to imperfect quality control could accumulate to greatly inflate estimates of h^2_g. We thus applied a set of additional filtering to eliminate spurious associations, only including variants with MAF$\geq$ 0.1%, imputation $r^2 \geq 0.6$, missing rate $\leq$ 1% and Hardy-Weinberg equilibrium P-value $\geq 10^{-7}$ in controls for each GWAS cohort. After merging GWAS cohorts, we next filtered samples such that no pair of samples had an IBD ≥ 0.025 using the `"-grm-cutoff"` option in `GCTA`. Reassurringly, the heritability explained was consistent regardless of whether or not an additional relatedness filter was used (e.g for CD, $h^2_{g\_0.025} = 0.284$ (SE=0.016) and $h^2_g = 0.272$ (SE=0.013)).

To transform the h^2_g estimate on the observed case-control risk scale to the liability scale, as described in Yang *et al* (2011) [34], we used a population prevalence of 0.005 and 0.0025 for CD and UC respectively. This workflow is documented in detail in Supplementary Figure 7.

We checked the reliability and robustness of our h^2_g estimates by estimating each genetic heritability in four ways:

i) Univariate estimation: using constrained REML in `GCTA` to estimate h^2_g for all SNPs with MAF $\geq$ 0.1% and individuals with relatedness < 0.025.

ii) Chromosome-partitioning: joint variant analysis across autosomes. GRM was constructed for each autosome and genetic variance for each chromosome was estimated in an analysis in which all chromosomal GRMs were fitted jointly as described in Lee *et al* (2012) [37].

iii) MAF-partitioning: similar to (ii), we estimated multiple genetic variance components by grouping SNVs into three MAF bins: $\geq 5\%$, $0.5\% - 5\%$ and $0.1\% - 0.5\%$.

iv) LD-adjusted GRMs were computed using `LDAK` [38].

The total SNP-heritabilities estimated based on univariate analysis, chromosome-partitioning analyses, MAF-partioning and with LD-adjusted approaches were consistent and similar to those from previously published studies, suggesting that our estimates are robust and reliable (Supplementary Table 16.

4.3.2 Data generation and quality control

We tested each cohort separately for association to UC, CD and IBD using a missing data likelihood score test as implemented in `SNPTEST v2.5` [18], conditioning on the first ten principal components as computed for each cohort when excluding the MHC region (chromosome 6:28-34Mb). We filtered all output to sites with MAF $\geq 0.1\%$, and INFO ≥ 0.4, and then used METAL to perform a standard error weighted meta-analysis of all three GWAS cohorts with our sequencing cohort (which was also pre-filtered to MAF $\geq 0.1\%$ and INFO ≥ 0.4).

The output of the fixed-effects meta-analysis was further filtered, and sites with high evidence for heterogeneity (I2> 0.90) or strong evidence for deviations from HWE in controls ($P_{HWE} < 1 \times 10^{-7}$) were discarded. Only sites at which all cohorts passed our quality control filters were included in our analysis. In addition, we discarded all variants for which the meta-analysis p-value was not lower than any of the cohort-specific p-values. Any sites which were included in the Immunochip or the IIBDGC datasets and were not at least nominally significantly associated with IBD in these datasets were also excluded from our analyses. Finally, and

in order to minimise the false positive associations due to bad imputation, sites which did not have an info score of 0.8 or more in at least three of the four datasets included in our meta-analysis were filtered out (two of the three for CD and UC, as we only have data from three cohorts for each of these).

4.3.3 Locus definition

An LD window was calculated for every genome-wide significant variant in any of the three traits (CD, UC, IBD), defined by the left-most and right-most variants that had an r^2 of 0.6 or more with the most associated SNP. LD was calculated in 1000 Genomes phase 3, release v5 (based on 20130502 sequence freeze and alignments), and only individuals of GBR and CEU ancestry were included in the calculation. Overlapping LD windows were subsequently merged, as well as windows with a distance of 500Kb or less between the lead variants of each locus, and the variant with the strongest evidence of association was kept as the lead variant for that respective locus. This process was conducted separately for each trait. A locus was annotated as known when there was at least one variant in it that was previously reported to be of genome-wide significance (irrespective of the LD between that variant and the most associated variants in the locus in our study). Otherwise, a locus was annotated as putatively novel. The PMIDs of the previous studies we included in our search for known IBD associations are described in Supplementary Table 15.

References

[1] Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

[2] The 1000 Genomes Project Consortium. 1000 genomes project phase I (2010). URL `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz`.

[3] The 1000 Genomes Project Consortium. 1000 genomes project phase II (2011). URL `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz`.

[4] Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

[5] Li, H. *et al.* The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[6] wtsi-hgi. wtsi-hgi/seq_autoqc. `https://github.com/wtsi-hgi/seq_autoqc`. Accessed: 2016-6-7.

[7] Handsaker, R. E., Korn, J. M., nemesh, J. & McCarroll, S. A. .

[8] SAM tools - browse /samtools/0.1.19 at SourceForge.net. `https://sourceforge.net/projects/samtools/files/samtools/0.1.19/`. Accessed: 2016-6-7.

[9] The International HapMap project. the international hapmap project phase 3 (2009). URL `ftp://ftp.ncbi.nlm.nih.gov/hapmap/phase_3/`.

[10] McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303

(2010).

[11] Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research* **21**, 830–839 (2011).

[12] Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature Genetics* 1–10 (2015).

[13] Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).

[14] Purcell, S. & Chang, C. Plink2 (2014). URL `https://www.cog-genomics.org/plink2`.

[15] Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

[16] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[17] Barrett, J. C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330–1334 (2009).

[18] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499–511 (2010).

[19] Derkach, A. *et al.* Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics (Oxford, England)* **30**, 2179–2188 (2014).

[20] Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

[21] McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26**, 2069–70 (2010).

[22] Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310–5 (2014).

[23] Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–90 (2014).

[24] Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).

[25] Huang, H. *et al.* Association mapping of inflammatory bowel disease loci to single variant resolution. *bioRxiv* 028688 (2015).

[26] Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research* **42**, 2976–2987 (2014).

[27] Hinrichs, A. S. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

[28] Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *American journal of human genetics* **93**, 687–96 (2013).

[29] Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics (Oxford, England)* **28**, 1598–603 (2012).

[30] Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**, 2867–73 (2010).

[31] The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–320 (2005).

[32] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

[33] Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

[34] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76–82 (2011).

[35] Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics* **88**, 294–305 (2011).

[36] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta (2011). URL `http://www.complextraitgenomics.com/software/gcta/index.html`.

[37] Hong Lee, S. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).

[38] Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).

Supplementary Table 1: autoQC Quality Control Classification Metrics

Metric	Warning	Failure
Duplicate Reads	⩾ 10.0%	⩾ 20.0%
Error Rate	⩾ 1.0%	⩾ 2.0%
Mapped Bases	⩽ 95.0%	⩽ 90.0%
Insertion:Deletion Ratio%	⩽ 67.5%, ⩾ 82.5%	⩽ 45.0%, ⩾ 110.5%
Overlapping Bases	⩾ 4.0%	⩾ 8.0%
Indel vs Read Cycle Deviation	⩾ 2.0%	⩾ 10.0%
Max Contiguous Quality High IQR	⩾ 30.0%	⩾ 50.0%
Max Base Content Total Baseline Deviation	⩾ 0.5	⩾ 1.5
Max Base Content Max Baseline Deviation	⩾ 5.0	⩾ 15.0
Insert peaks	-	<80% of reads within 25bp of the peak
Contiguous drop off cycles	Low value < 25 at a rate = 20%	Low value < 25 at a rate ⩾ 30%
Multiple max peaks	-	6 or more warnings from: - Indel vs Read Cycle Deviation - Max Base Content Total Baseline Deviation - Max Base Content Max Baseline Deviation
Multiple indel peaks	-	3 or more warnings from: - Indel vs Read Cycle Deviation

Supplementary Table 2: Sequencing sample quality control summary

Criteria	UC	CD	UK10K	Total
Initial sample size	1,817	2,697	3,910	8,354
Average coverage	2.05x	3.84x	7x	4.39x
BAM QC	-12	-107	-244	-363
heterozygosity rate (±3.5 s.d.)	-2	-16	-13	-31
Relatedness ($\hat{\pi} > 0.25$)	-33	-50	-7	-90
Ancestry Outliers	-3	-11	-1	-15
Post-QC sample size	1,767	2,513	3,652	7,932

Supplementary Table 3: Sequencing autosomal variants quality control summary

	SNPs	INDELs	Total
Raw variant calling	87,456,881	7,683,401	95,140,282
Post SVM/VQSR filtering	72,166,448	4,522,487	76,688,935
Post genotype refinement filtering	70,344,218	3,205,131	73,549,349
Post UC-association filtering	56,430,118	1,776,521	58,206,639
Post CD-association filtering	59,985,208	1,922,610	61,907,818
Post IBD-association filtering	67,201,374	1,938,154	69,139,528

Supplementary Table 4: Summary of the datasets used to evaluate the genotype accuracy of our sequencing data

	UK10K GWAS	**IBD-ichip**	**CD-GWAS**	**UC-GWAS**
Total samples	3,777	53,279	1,748	2,361
Total SNPs	919,415	177,367	458,858	757,728
Common samples	3,666	1,476	96	332
Common SNPs	895,421	152,158	446,140	739,631

Supplementary Table 5: IBD sequencing autosomal variants compared to 1000 Genomes Projects Phase 3 European panel.

(a) SNVs kept at different stages of our IBD sequencing project, and their overlap with 1000GP Phase 3 European panel (503 samples). The left-hand-side panel of Supplementary Figure 3 is a graphical representation of these values.

		Singleton	Doubleton	Tripleton-.5%	.5-1%	1-5%	>5%
Variant discovery	in 1000GP	4,802,093	1,205,334	1,576,257	1,112,007	2,460,806	5,546,446
	Total	71,555,627	1,800,559	2,189,422	2,030,079	3,772,684	6,108,510
SVM filtering	in 1000GP	4,541,020	1,183,336	1,537,694	1,079,949	2,376,196	5,357,949
	Total	58,820,369	1,731,931	1,891,848	1,472,594	2,729,738	5,519,968
Genotype refinement	in 1000GP	4,484,552	1,166,126	1,509,623	1,056,948	2,313,001	5,173,738
	Total	58,066,170	1,697,928	1,748,091	1,220,697	2,390,652	5,220,680
UC association	in 1000GP	3,990,628	1,126,957	1,456,086	1,020,611	2,237,109	4,998,181
	Total	44,828,935	1,627,078	1,628,445	1,046,432	2,266,573	5,032,655
CD association	in 1000GP	4,197,863	1,127,572	1,456,346	1,020,716	2,237,151	4,998,181
	Total	48,369,564	1,628,484	1,630,659	1,054,469	2,269,376	5,032,656
IBD association	in 1000GP	4,295,649	1,126,036	1,454,470	1,019,630	2,235,317	4,993,282
	Total	55,604,960	1,625,925	1,627,348	1,048,880	2,266,540	5,027,721

(b) Number of 1000GP SNVs overlap with different stages of IBD sequencing cohort. These are the numerical values represented in the right-hand-side panel of Supplementary Fig. 3

	Singleton	Doubleton	Tripleton-.5%	.5-1%	1-5%	>5%
Variant Discovery	3,987,923	1,448,809	1,699,915	1,131,041	2,421,432	5,876,796
SVM filtering	3,780,362	1,403,372	1,657,460	1,102,798	2,357,992	5,641,270
Genotype refinement	3,736,078	1,385,178	1,632,905	1,083,811	2,304,311	5,430,787
UC association	3,353,624	1,302,278	1,565,404	1,044,464	2,226,128	5,231,381
CD association	3,517,836	1,327,946	1,576,068	1,046,919	2,226,927	5,231,601
IBD association	3,597,609	1,337,451	1,578,194	1,046,634	2,225,203	5,226,318
Total	8,716,645	1,798,814	1,871,612	1,195,609	2,507,644	6,042,565

Supplementary Table 6: Imputed GWAS cohort summary

Cohort	Case	Control	Total
WTCCC1	1,206	2,918	4,124
WTCCC2	1,921	2,776	4,697
GWAS3_CD	4,264	9,495	13,759
GWAS3_UC	4,072	9,495	13,567
GWAS3_IBD	8,860	9,495	18,355
Total	11,987	15,189	27,176

Supplementary Table 7: Association statistics for rs78534766 (chr16:50335074, ADCY7 Asp439Glu) across cohorts. Missingness in cases and controls is zero for the sequenced data due to the genotype refinement step (see section 2.5 in supplement) and there is also zero missingness in the imputed data.

Cohort	Cases	Controls	OR [95% CI]	Pvalue	MAF (ctrls)	Method	Info	Missingness (cases/controls)	Phet
WTCCC2	1,921	2,918	2.62 [1.63-4.22]	7.03E-05	0.0061	imputed	0.82	N/A	
GWAS3	4,072	9,495	2.05 [1.53-2.75]	1.43E-06	0.0065	directly genotyped	N/A	0.00025/0.0024	
Low coverage sequencing	1,767	3,652	2.14 [1.27-3.60]	0.0042	0.0060	sequenced	0.88	N/A	
All discovery	7,760	16,065	2.19 [1.75-2.74]	9.20E-12		(meta-analysis)			0.69
UK Biobank	982	136,464	1.70 [1.18-2.44]	0.0189	0.0061	directly genotyped	N/A	0.0000/0.0004	
Replication genotyping	450	3,905	4.10 [1.76-9.51]	0.0009	0.0069	directly genotyped	N/A	0.0000/0.0044	
All directly genotyped	5,504	149,864	2.06 [1.63-2.60]	1.62E-09		(meta-analysis)			0.19
All	13,264	165,929	2.16[1.77-2.62]	1.17E-14		(meta-analysis)			0.39

Supplementary Table 8: Variant annotations included in each of the gene-based burden test subsets.

Annotation	Functional coding	Predicted damaging
frameshift_variant	✓	✓
stop_gained	✓	CADD≥21
initiator_codon_variant	✓	CADD≥21
splice_donor_variant	✓	CADD≥21
splice_acceptor_variant	✓	CADD≥21
missense_variant	✓	CADD≥21
stop_lost	✓	CADD≥21
inframe_deletion	✓	X
inframe_insertion	✓	X

Supplementary Table 9: The number of gene-based burden tests performed for each combination of annotation set and phenotype, with the average number of variants contributing to each of those tests given in parentheses.

Test	Functional coding	Predicted damaging	Total
UC	18,149 (6.82500)	14,850 (4.24795)	32,999 (5.66529)
CD	18,670 (7.42341)	15,406 (4.56283)	34,076 (6.13012)
IBD	18,293 (6.88088)	14,967 (4.25991)	33,260 (5.70144)

Supplementary Table 10: Genes with $P < 5e^{-4}$ in the gene-based burden tests. For each gene exceeding this threshold, the bam files for the three variants with the largest contribution to the overall gene signal were inspected, and any with questionable variant calls were excluded from this table.

Gene Name	Ensembl ID	P value	Phenotype	Annotation set	Effect
NOD2	ENSG00000167207	0.0000001	CD	Functional coding	Risk
NOD2	ENSG00000167207	0.0000004	CD	Predicted damaging	Risk
NOD2	ENSG00000167207	0.000001	IBD	Predicted damaging	Risk
NOD2	ENSG00000167207	0.000003	IBD	Functional coding	Risk
IGKC	ENSG00000211592	0.000037	CD	Functional coding	Risk
WWP1	ENSG00000123124	0.000065	IBD	Functional coding	Protective
VWA5A	ENSG00000110002	0.00007	CD	Functional coding	Risk
CTB-78H18.1	ENSG00000253110	0.000081	IBD	Functional coding	Risk
KRT16	ENSG00000186832	0.000129	IBD	Functional coding	Protective
DCTD	ENSG00000129187	0.000175	IBD	Functional coding	Protective
CADM4	ENSG00000105767	0.000183	IBD	Functional coding	Risk
UGT1A3	ENSG00000243135	0.000239	IBD	Predicted damaging	Risk
LRRC55	ENSG00000183908	0.00025	CD	Functional coding	Risk
LRRC55	ENSG00000183908	0.00025	CD	Predicted damaging	Risk

Continued on next page

Table 10 – *Continued from previous page*

Gene Name	Ensembl ID	P value	Phenotype	Annotation set	Effect
MYO19	ENSG00000141140	0.000314	CD	Functional coding	Protective
DOCK8	ENSG00000107099	0.000353	CD	Functional coding	Risk
ERBB3	ENSG00000065361	0.000388	UC	Predicted damaging	Protective
SOAT2	ENSG00000167780	0.000448	IBD	Functional coding	Protective
ARHGAP19-SLIT1	ENSG00000269891	0.000453	IBD	Predicted damaging	Risk
IL23R	ENSG00000162594	0.000492	CD	Predicted damaging	Protective

Supplementary Table 11: Genes used in the main gene-set burden tests: implicated by a coding variant in the fine-mapping credible sets recently defined by Huang et al [25], eQTL mapping, or by implication of causal coding variants in the literature.

Gene ID	Name	Dis.	Gene ID	Name	Dis.
ENSG00000085978	*ATG16L1*	CD	ENSG00000164308	*ERAP2*	CD
ENSG00000187796	*CARD9*	IBD	ENSG00000136634	*IL10*	IBD
ENSG00000013725	*CD6*	CD	ENSG00000115607	*IL18RAP*	IBD
ENSG00000143226	*FCGR2A*	IBD	ENSG00000134460	*IL2RA*	CD
ENSG00000176920	*FUT2*	CD	ENSG00000005844	*ITGAL*	UC
ENSG00000115267	*IFIH1*	UC	ENSG00000095110	*NXPE1*	UC
ENSG00000162594	*IL23R*	IBD	ENSG00000079263	*SP140*	CD
ENSG00000173531	*MST1*	IBD	ENSG00000106952	*TNFSF8*	IBD
ENSG00000167207	*NOD2*	CD			
ENSG00000134242	*PTPN22*	CD			
ENSG00000166949	*SMAD3*	IBD			
ENSG00000105397	*TYK2*	IBD			

Supplementary Table 12: P values for burden tests performed on gene set described above. Results of the burden test excluding *NOD2* are shown in parentheses.

	Functional coding	Predicted damaging	Loss of function
UC	0.13 (0.128)	0.114 (0.246)	0.457 (0.457)
CD	0 (0.201)	0 (0.004)	0.222 (0.222)
IBD	0 (0.021)	0 (0.017)	0.24 (0.24)

Supplementary Table 13: The number of enhancer-based burden tests performed for each combination of annotation set and phenotype, with the average number of variants contributing to each of those tests given in parentheses.

Test	All variants	Variants affecting a TFBM	Total
UC	28,292 (2.64099)	11,532 (1.29067)	39,824 (2.24997)
CD	29,628 (2.74679)	12,403 (1.30912)	42,031 (2.32255)
IBD	28,453 (2.62155)	11,540 (1.28631)	39,993 (2.23627)

Supplementary Table 14: Enhancer set-based tests with $P < 0.01$. 'TFBM' refers to set tests performed only using rare variants predicted to create or disrupt a transcription factor binding motif, while 'All' includes all rare variants within the relevant enhancer region. No set test reaches significance after multiple correction testing for the 660 tests performed.

Cell/tissue type	P-value	Phenotype	Annotation set	Number of enhancers	Number of variants
skeletal muscle tissue	0.00058	CD	All	67	222
skeletal muscle tissue	0.00068	IBD	All	61	188
skeletal muscle cell	0.00253	IBD	TFBM	293	397
melanocyte	0.0039	CD	All	379	1,241
stromal cell	0.00398	IBD	TFBM	272	401
cardiac fibroblast	0.00425	UC	TFBM	192	278
osteoblast	0.00565	UC	TFBM	168	234
cardiac fibroblast	0.00674	IBD	TFBM	196	280
melanocyte	0.0096	IBD	All	358	1,104
thyroid gland	0.00993	CD	All	109	352

Supplementary Table 15: Publications used to determine known IBD loci.

Pubmed ID	Citation
17554261	Parkes et al. 2007. Nature Genetics 39 (7): 830–32.
19915572	Barrett et al. 2009. Nature Genetics 41 (12): 1330–34.
20228798	Franke et al. 2010. Nature Genetics 42 (4): 292–94.
20228799	McGovern et al. 2010. Nature Genetics 42 (4): 332–37.
21102463	Franke et al. 2010. Nature Genetics 42 (12). 1118–25.
21297633	Anderson et al. 2011. Nature Genetics 43 (3): 246–52.
22412388	Kenny et al. 2012. PLoS Genetics 8 (3).
23128233	Jostins et al. 2012. Nature 491 (7422). 119–24.
23266558	Yamazaki et al. 2013. Gastroenterology 144 (4): 781–88.
23850713	Yang et al. 2014. Gut 63 (1): 80–87.
25082827	Julià et al. 2014. Human Molecular Genetics 23 (25): 6927–34.
26192919	Liu et al. 2015. Nature Genetics 47 (9): 979–89.
26974007	Ellinghaus et al. 2016. Nature Genetics 48 (5): 510-8.
NA	Huang et al. 2015. bioRxiv. doi:10.1101/028688.

Supplementary Table 16: Estimate of heritability using four different approaches. All analyses were carried out after excluding one individual from every pair with relatedness > 0.025, estimated from markers with MAF ≥ 0.1%. The prevalences of CD and UC were assumed to be 0.005 and 0.0025, respectively.

Method	$h^2_{g_{CD}}$ (SE)	$h^2_{g_{UC}}$ (SE)
Univariant	0.284 (0.016)	0.211 (0.012)
Chr-partitioning	0.270 (0.012)	0.233 (0.056)
MAF partitioning	0.293 (0.014)	0.226 (0.020)
LD-adjusted	0.268 (0.013)	0.215 (0.012)

Supplementary Table 17: Cell and tissue types for which FANTOM5 defines preferentially expressed enhancer sets.

Cell types	Tissue types
neuronal stem cell	lymph node
myoblast	large intestine
osteoblast	blood
ciliated epithelial cell	throat
blood vessel endothelial cell	testis
mesothelial cell	stomach
T cell	heart
mast cell	brain
sensory epithelial cell	eye
astrocyte	penis
mesenchymal cell	female gonad
fat cell	uterus
chondrocyte	vagina
melanocyte	adipose tissue
hepatocyte	esophagus
skeletal muscle cell	salivary gland
macrophage	skeletal muscle tissue
keratinocyte	smooth muscle tissue
vascular associated smooth muscle cell	urinary bladder
tendon cell	pancreas
dendritic cell	tongue
stromal cell	submandibular gland

Continued on next page

Table 17 – *Continued from previous page*

Cell types	Tissue types
neuron	parotid gland
reticulocyte	blood vessel
corneal epithelial cell	placenta
monocyte	thyroid gland
acinar cell	lung
natural killer cell	skin of body
hepatic stellate cell	spleen
pericyte cell	liver
urothelial cell	small intestine
cardiac myocyte	gallbladder
basophil	kidney
neutrophil	spinal cord
lymphocyte of B lineage	umbilical cord
endothelial cell of lymphatic vessel	meninx
epithelial cell of Malassez	prostate gland
lens epithelial cell	thymus
epithelial cell of prostate	tonsil
epithelial cell of esophagus	olfactory region
mammary epithelial cell	internal male genitalia
preadipocyte	
keratocyte	
trabecular meshwork cell	
respiratory epithelial cell	
enteric smooth muscle cell	

Continued on next page

Table 17 – *Continued from previous page*

Cell types	Tissue types
kidney epithelial cell	
amniotic epithelial cell	
cardiac fibroblast	
fibroblast of choroid plexus	
fibroblast of the conjuctiva	
fibroblast of gingiva	
fibroblast of lymphatic vessel	
fibroblast of periodontium	
fibroblast of pulmonary artery	
hair follicle cell	
intestinal epithelial cell	
iris pigment epithelial cell	
placental epithelial cell	
retinal pigment epithelial cell	
bronchial smooth muscle cell	
smooth muscle cell of the esophagus	
smooth muscle cell of trachea	
uterine smooth muscle cell	
skin fibroblast	
gingival epithelial cell	
fibroblast of tunica adventitia of artery	
endothelial cell of hepatic sinusoid	
smooth muscle cell of prostate	