Chapter 3

The role of rare and low frequency variation in IBD risk

3.1 Introduction

Genome wide association studies (GWAS) have identified 215 risk loci for inflammatory bowel disease (Parkes et al., 2007; Anderson et al., 2011; Kenny et al., 2012; Yamazaki et al., 2013; Julià et al., 2014; Yang et al., 2014b; Liu et al., 2015; Ellinghaus et al., 2016), nearly all of which are driven by common variation. The high correlation between common variants in close proximity has driven the success of GWAS, but also makes it difficult to narrow these associations down to individual causal variants, or even to identify which gene is likely to be affected. In contrast, rare variants (which plausibly have larger effect sizes) can be more straightforward to interpret, but are more difficult to assess. Because they are poorly tagged by neighbouring variation, each rare variant must be directly captured in order to be tested for association.

Recent reductions in the cost of DNA sequencing means that rare variants may now be captured at scale. In order to maximise sample size, early IBD sequencing studies concentrated on genes in GWAS-implicated loci (Rivas et al., 2011; Beaudoin et al., 2013; Hunt et al., 2013; Prescott et al., 2015), which can logically be extended to study the entire exome. However, coding variation has been shown to explain at most 20% of the IBD associations uncovered using GWAS (Huang et al., 2015), with the remaining variants lying in non-coding, presumed regulatory, regions of the genome. Low coverage whole genome sequencing has therefore been suggested as a cost-effective approach to capture both coding and non-coding variation in large numbers of samples (Li et al., 2011). This approach is well suited to explore rarer variants than are accessible using GWAS (Cai et al., 2015; Danjou et al., 2015), although the low individual sequencing depth precludes the capture of extremely rare and private mutations.

3.1.1 Chapter overview

In this chapter, I investigate the role of rare, low frequency and structural variation in inflammatory bowel disease risk using low coverage whole genome sequences from 4,280 IBD cases and 3,652 controls. In order to maximise the number of IBD patients included in this study, the cases were sequenced to a lower average depth (2-4x) than the controls (7x), which were already available via managed access from the UK10K project (Walter et al., 2015). For structural variants, which are particularly challenging to call in low coverage data, even very careful filtering and joint analysis was not sufficient to overcome this bias. However, for rare and low frequency variation the use of joint calling, genotype refinement, and specially designed test statistics (Chapter 2) allows the false positive rate to be adequately controlled.

I observe a significant burden of rare, damaging missense variation in the gene *NOD2*, as well as a more general burden of such variation amongst known inflammatory bowel disease risk genes. However, I note the need to perform larger sequence-based studies in order to properly resolve the precise variation that is contributing to this observation. At current sample sizes, I do not detect any burden of rare variation within cell- and tissue-specific enhancer regions.

In collaboration, I then impute from these sequences into both new and existing GWAS cohorts in order to test for association at ∼12 million low frequency variants across 16,267 cases and 18,841 controls. We discovered a missense variant in

ADCY7 that approximately doubles the risk of ulcerative colitis (MAF=0.6%, OR=2.19). However, despite good power to detect such associations, we did not identify any other new low frequency risk variants, suggesting that such variants as a class explain very little disease heritability.

3.1.2 Contributions

This study was conceived and designed by the UK IBD Genetics Consortium (UKIBDGC), with case ascertainment, phenotyping and sample collection performed by the numerous clinics that contribute to this effort: please see Appendix A for a full list of contributors. DNA sample preparation, sequencing, read alignment, and initial quality control of the whole genome sequences used in this chapter was performed by the Wellcome Trust Sanger Institute sequencing pipeline facility and the human genetics informatics team. Calling of single nucleotide polymorphisms and insertion-deletions, genotype refinement, quality control analyses (except where indicated), and heritability analyses were performed by Yang Luo. Code for identifying variants predicted to create or disrupt a transcription factor binding motif was provided by Hailiang Huang. Imputation of GWAS datasets using an IBD-specific reference panel was performed by Shane McCarthy; quality control and conditional analysis of the resulting meta-analysis was performed by Loukas Moutsianas. Analysis of the UK BioBank replication cohort was performed by Luke Jostins. Unless stated, I carried out all other analyses.

3.2 Data preparation

3.2.1 Low coverage whole genome sequencing

Sample ascertainment

Individuals were consented into the study based on a confirmed diagnosis of Crohn's disease or ulcerative colitis using standard endoscopic, radiological and histopathological criteria. No selection was made for patients based on family history or early age of onset, and all subtypes of CD and UC were included. Blood or saliva samples were donated for DNA extraction at UK clinics involved in the UK IBD Genetics Consortium (Cambridge, Dundee, Edinburgh, Exeter, London, Manchester, Newcastle, Norwich, Nottingham, Oxford, Sheffield, Torbay and the Scottish early onset IBD project). Ethical approval was granted by the Cambridge MREC (reference: 03/5/012).

Control samples were collected by the UK10K Consortium, including individuals from both the Avon Longitudinal Study of Parents and Children (Boyd et al., 2013) and the Twins UK cohort (Moayyeri et al., 2013). Full details of selection criteria may be found in the UK10K flagship paper by Walter et al. (2015).

Sequencing and data processing

Whole genome sequencing of 1,817 ulcerative colitis cases at 2x average coverage, and 2,697 Crohn's disease cases at 4x average coverage, was performed at the Wellcome Trust Sanger Institute (WTSI). For each sample, 1-3μg of DNA was sheared to 100-1000bp using a Covaris E210 or LE220 machine, then prepared for sequencing using an Illumina paired-end DNA library preparation kit. The resulting libraries were selected for insert sizes of 300-500bp, and then sequenced on the Illumina HiSeq platform as paired-end 100bp reads (according to the manufacturer's protocol). Controls were whole genome sequenced to 7x average coverage using the same protocol, with 1,556 samples processed at the WTSI and 2,354 at the Beijing Genomics Institute (BGI).

Sequencing reads were aligned to the human reference genome by their respective sequencing centres. Case data was aligned to hs37d5, the reference genome used in Phase II of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2011), which consists of the GrCH37 primary assembly plus sequences from human herpesvirus and concatenated decoy sequences. Control data was originally aligned to the GrCH37 primary assembly that was used in Phase I of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), but was later updated to hs37d5 using the software BridgeBuilder (Luo et al., 2017). Automatic quality control of the resulting BAM files was performed by the WTSI pipelines.

3.2.2 Variant calling and imputation improvement

Generating a SNP and indel call set

Single nucleotide polymorphisms (SNPs) and small insertion-deletions (indels) were called jointly across 8,354 pooled sample-level BAM files that passed automatic quality control. First, genotype likelihoods were obtained using samtools-0.19 (Li et al., 2009) and then converted to variant calls with bcftools-0.19 (Li et al., 2013b). Before refinement of these genotypes via imputation improvement, initial quality control was applied to remove low-confidence sites.

Initial SNP filtering

A set of Support Vector Machines (SVMs) were trained to identify poor quality SNP calls. Training data consisted of $1,000$ sites that overlapped with HapMap3 (Altshuler et al., 2010), and were therefore deemed highly likely to be true sites of variation, and $1,000$ sites with a quality score QUAL < 10 in the raw VCF file. Because the composition of HapMap3 (and established variant databases in general) is heavily skewed towards common variation, training variants were selected so as to roughly preserve the expected true MAF distribution in the human population within three MAF bins ($0 \leq$MAF$< 0.5\%$, $0.5\% \leq$MAF$< 5\%$, and MAF$\geq 5\%$). The models were then trained using the following variant call features:

- DP: Raw read depth

- MQ: Root-mean-square mapping quality of reads covering the site
- AN: Total number of alleles in called genotypes
- MDV: Maximum number of high-quality non-reference reads in samples
- EDB: End distance bias
- RPB: Read position bias

Five independent SVMs were run in parallel, and only SNPs labelled as high-quality by at least two of the five SVMs were taken forward for imputation improvement.

Initial indel filtering

Indels were filtered using VQSR, or Variant Quality Score Recalibration (DePristo et al., 2011), trained on the Mills-Devine high-confidence indel call set (Mills et al., 2011). VQSR assigns each indel a variant quality score log odds ratio (VQSLOD) based on the following features:

- DP: Approximate read depth, after reads with MQ= 255 or 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 alternate vs. reference read position bias
- MQRankSum: Z-score from Wilcoxon rank sum test of alternate vs. reference read mapping qualities

A minimum VQSLOD score of 1.0659, which corresponds to a truth sensitivity threshold of 97%, was used to select high-quality indels.

Genotype refinement

Genotypes at all SNP and indel sites that passed initial filtering were refined via imputation. To increase the computational efficiency of this process, imputation improvement was performed in batches of $3,000$ sites, with a buffer region of 500 sites on either side, using BEAGLE v4.1 (Browning and Browning, 2016) with default parameters.

After an initial round of refinement, a number of poor-quality sites not identified during initial quality control became apparent. These were removed using the following filters:

- Evidence for a deviation from Hardy-Weinberg equilibrium in controls, where the p-value $< 1 \times 10^{-7}$
- Removal of sequencing centre batch effects in controls, where the p-value $< 1 \times 10^{-3}$ when testing for association with sequencing centre
- Variants with $> 10\%$ missing genotypes following genotype refinement, where the minimum posterior probability required to call a genotype was 0.9
- SNPs within 3 base pairs of an indel
- Clusters of indels separated by 2 or fewer base pairs, so that only one may pass

Following these exclusions, a second round of genotype refinement was performed using BEAGLE v4.1 to ensure that neighbouring variant calls had not been adversely affected by imputation with poor-quality sites.

Challenges of calling structural variants in a large low coverage sequencing study

Copy number variants (CNVs) are usually detected via the identification of localised changes in read depth, an individual read that spans a deletion or insertion breakpoint, or read pairs that map unexpectedly far apart. However, the low average read depth of this particular dataset means that this form of variant detection is not particularly sensitive for individual samples. I therefore called CNVs using GenomeSTRiP 2.0 (Handsaker et al., 2015), 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.

The actual discovery and genotyping process can be broken down into several modules, as summarised in Figure 3.1. To improve efficiency, I ran the pre-processing steps separately for each chromosome and cohort (CD, UC and controls). Computational resource restrictions also required the discovery and genotyping processes to be run separately across each chromosome, which led to a need for manual intervention at the sample filtering step during discovery to ensure that filtering considered all chromosomes at once.

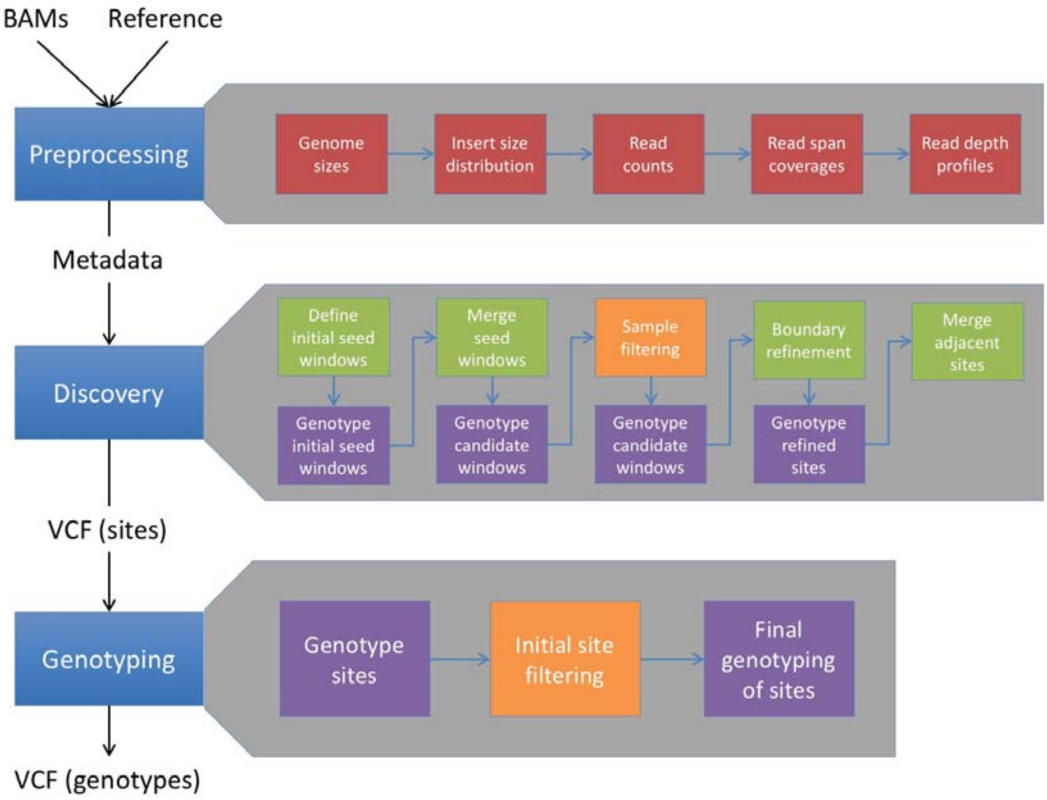

Figure 3.1: Overview of the modular structure employed by GenomeSTRiP 2.0 to discover and genotype CNVs across a number of low coverage whole genome sequences.

Default GenomeSTRiP configurations were used, as per the example configuration 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 GrCH37 to hs37d5 using BridgeBuilder did not contain appropriate metadata information for use by GenomeSTRiP 2.0, these reads were excluded from discovery and genotyping.

3.2.3 Quality control

Sample filtering

Individuals failing on one or more of the following filtering criteria (when calculated using refined genotypes) were removed from the dataset:

- Heterozygosity rate ±3.5 standard deviations from the mean.
- Duplicate or closely-related individuals with $\hat{\pi} > 0.25$ (indicating second-degree relatives or closer). To identify these individuals, SNPs were first pruned such that no two sites within 5,000kb had an $r^2 > 0.2$, and the Identity-By-State value for each pair of individuals was then calculated using only variants with MAF $> 1\%$. Only one individual from each duplicate or related pair was removed.
- Individuals of non-European ancestry, as identified using a principal component analysis projected from 11 HapMap2 populations.

Site filtering for SNPs and indels

In addition to the SNP and indel site filters applied in section 3.2.2, the following criteria were used to remove lower quality sites prior to association testing:

- Minimum score < 0.1 in any of the five independent SVM runs
- INFO score < 0.4
- Evidence for a deviation from Hardy-Weinberg equilibrium in controls, where the p-value $< 10^{-6}$

Site filtering for 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. (2015), who note that small duplications appear to have a higher false discovery rate than equivalently sized deletions or mixed CNVs.
- Call rate > 0.9, to remove those variants with excessive missingness.
- Density > 0.5, with density 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.

I then apply the following dataset-specific quality control filters:

- Remove CNVs attributable to missing sample data. Specifically, an excess of very large copy number variants with a MAF of 1-2% was observed (Figure 3.2), that I traced down to $1,103$ copy number variants that were driven by 95 control samples with a large stretch of missing data on chromosome 6.

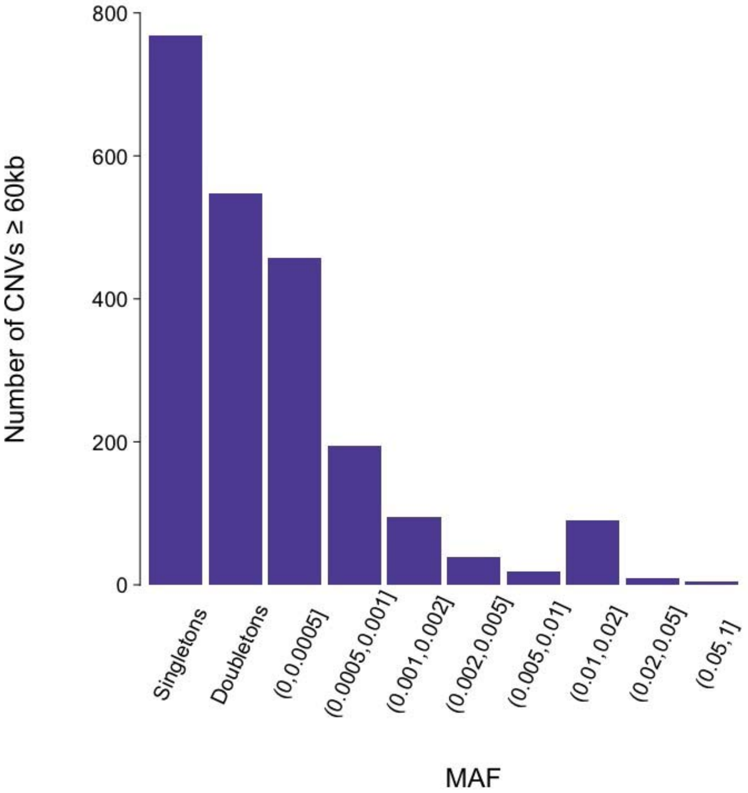

Figure 3.2: Due to a stretch of missing data on chromosome 6 for 95 control samples, there is an apparent excess of large copy number variants with a MAF of 1-2%.

- Remove CNVs with GSELENGTH ≤60,000. For shorter copy number variants, I observed considerable differences in sensitivity across different mean coverage depths (Figure 3.3).

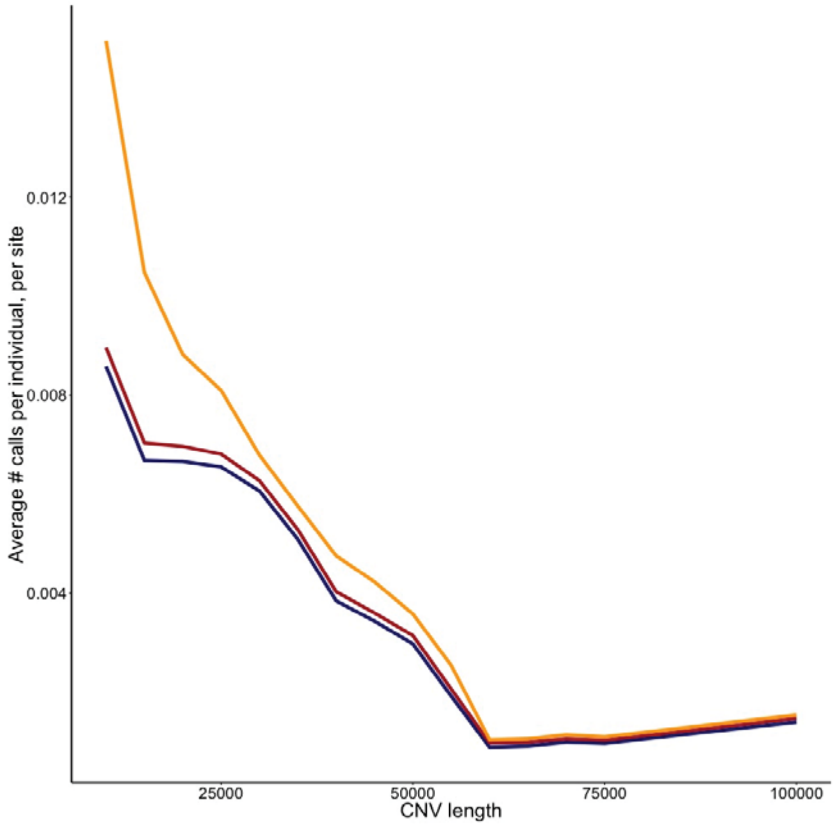

Figure 3.3: The average number of calls per individual per site, across different copy number variant (CNV) lengths. UK10K controls (7x) in yellow, Crohn's disease cases (4x) in red, and ulcerative colitis cases (2x) in blue.

- Keep only biallelic sites, for simplicity when association testing. However, because GenomeStrip 2.0 is capable of calling multi-allelic CNVs, I noted an abundance of common sites where a small fraction of non-reference individuals contain a CNV in the opposite direction to the majority call, possibly due in part to the particularly low coverage seen in this dataset. At sites where this fraction of inconsistent directions is less than 10% of the alternate calls made, I retain the site as biallelic.

3.3 Structural variation

Following quality control, I observed an approximately equal number of variants in cases and controls, but retained only 1,475 CNVs. Of these, just 59 had a MAF $> 0.1\%$ and were taken forward for single site association testing. Following association testing using a likelihood score test, as implemented in SNPTEST v2.5 (Marchini and Howie, 2010), no individual CNV was significantly associated after correction for multiple testing.

I then considered the 1,464 CNVs with a MAF $\leq 0.5\%$ in controls, performing a simple chi-squared test to compare the cumulative minor allele frequencies of these variants between cases and controls (Table 3.1). I note that there is a significant genome-wide excess of rare duplications in controls ($P = 0.0002$), suggesting that even after very stringent filtering the data remains too noisy for meaningful conclusions to be drawn. Therefore, to avoid including any bias due to sequencing depth heterogeneity between cases and controls, I tested within cases only for a burden of CNVs in known IBD regions (Liu et al., 2015) compared to regions not previously associated with IBD. However, the number of CNVs contributing to these tests were very small (Table 3.1), and no significant results were obtained.

Table 3.1: Testing for an association of structural variation with IBD.

	Variation	Number of CNVs	Cumulative MAF in A	Cumulative MAF in B	*P*-value
A) Cases vs B) Controls	Deletions	668	0.00019	0.00017	0.0499
	Duplications	796	0.00020	0.00023	0.0002
	Combined	1,464	0.00019	0.00020	0.1200
A) IBD vs B) Non-IBD Regions	Deletions	5	0.00012	0.00019	0.2967
	Duplications	11	0.00013	0.00020	0.1227
	Combined	16	0.00012	0.00019	0.0684

These results suggest that high coverage whole genome sequencing of more individuals, preferably with balanced coverage between cases and controls, will be required to evaluate the contribution of rare CNVs to IBD risk.

3.4 Rare variation

Low coverage sequencing is not generally a suitable study design with which to accurately capture very rare and private variants, particularly as joint-calling and cross-sample genotype refinement adds little information at sites where nearly all individuals are homozygous for the major allele. Nevertheless, given how difficult such variants are to impute from GWAS data (recently, McCarthy et al. (2016) showed that even a reference panel of over 32,000 individuals offers little imputation accuracy for MAF $< 0.1\%$), this sequence dataset represents the largest source of rare variation in an IBD cohort to date. Because of this, it was decided that the potential role of rare variation in IBD risk within this dataset was worth investigating.

Due to the sequencing depth heterogeneity between cases and controls, existing rare variant burden methods will give systematically inflated test statistics. I therefore performed rare variant burden testing across both genes and putative enhancers using unrefined genotype probabilities and an extension of the Robust Variance Score statistic by Derkach et al. (2014), which was developed to account for this type of bias as described in Chapter 2.

3.4.1 Additional quality control

Additional site filtering was required prior to rare variant association testing, as these types of studies are more susceptible to differences in read depth between cases and controls (as discussed in Chapter 2). This filtering consisted of removing:

- Singleton variants, observed only once in the population.
- Variants with a missingness rate >0.9, when calculated using genotype probabilities estimated from the samtools genotype quality (GQ) field
- Low confidence observations (maximum genotype probability ≤ 0.9) comprising $\geq 1\%$ of non-missing data
- Sites with INFO < 0.6 in the appropriate cohorts

I will note here that the singleton variants removed from this analysis have actually been the primary focus of other rare variant association studies in complex traits, such as schizophrenia and educational attainment (Ganna et al., 2016; Genovese et al., 2016), where they have been shown to have an important role. However, in this dataset we observe distinct differences in the specificity of variant calling between the lowest coverage group (2x) and the higher coverage groups (4x and 7x), as shown in Figure 2.6. This bias cannot be fully accounted for during association testing, and was not able to be overcome using more stringent filtering techniques. Therefore, in order to maintain a well-controlled Type I error rate, it was necessary to remove all such sites from the analysis. As with structural variants, high coverage whole genome sequencing of more individuals, preferably with balanced coverage between cases and controls, will be required to assess the contribution of ultra rare variation to IBD risk.

3.4.2 Burden testing across coding regions

Gene-based burden tests

For each of 18,670 genes, as defined by annotation with an Ensembl ID, I tested for a differential burden of rare (MAF $\leq 0.5\%$ in controls) variation between the sequenced cases and controls. Two separate burden tests were performed for each gene: one aggregating all functional coding variants and one for all predicted damaging functional coding variants, as defined in Table 3.2. Variant annotations were assigned using the Variant Effect Predictor by McLaren et al. (2010) and the Combined Annotation Dependent Depletion (CADD) score by Kircher et al. (2014). The CADD score is used to estimate the deleteriousness of a given variant in the human genome, with higher scores indicating a variant is more likely to be deleterious: the threshold of 21 used here represents the median value of all possible canonical splice sites and non-synonymous variants.

Table 3.2: Variant annotations used to define 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

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 (Table 3.3). To correct for this multiple testing, I used a Bonferroni-adjusted threshold for significance of 5×10^{-7}, reflecting an overall alpha value of 0.05. This does not take into account the correlation between the different tests (as the predicted damaging variant set is a direct subset of the functional coding set, and the CD and UC individuals are a subset of the IBD set) and therefore may be too stringent a threshold.

Table 3.3: 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 (6.83)	14,850 (4.25)	32,999 (5.67)
CD	18,670 (7.42)	15,406 (4.56)	34,076 (6.13)
IBD	18,293 (6.88)	14,967 (4.26)	33,260 (5.70)

For each gene with a final p-value $< 5 \times 10^{-4}$, I inspect the BAM files for the three variants with the largest individual contributions to the overall gene signal (as determined using single-site association testing with the RVS statistic at each site), in order to assess the quality of variant calling at that position. This manual inspection was used to identify sites where, for example, all the alternate alleles lie at the ends of reads, or predominantly on reads sequenced in one direction. I also check for regions that appear to have been generally difficult to map, or contain an excess of potential errors around the variant call (Figure 3.4). Details of the genes passing this quality control check can be found in Table 3.4, while the full tests are summarised in Figures 3.5 and 3.6.

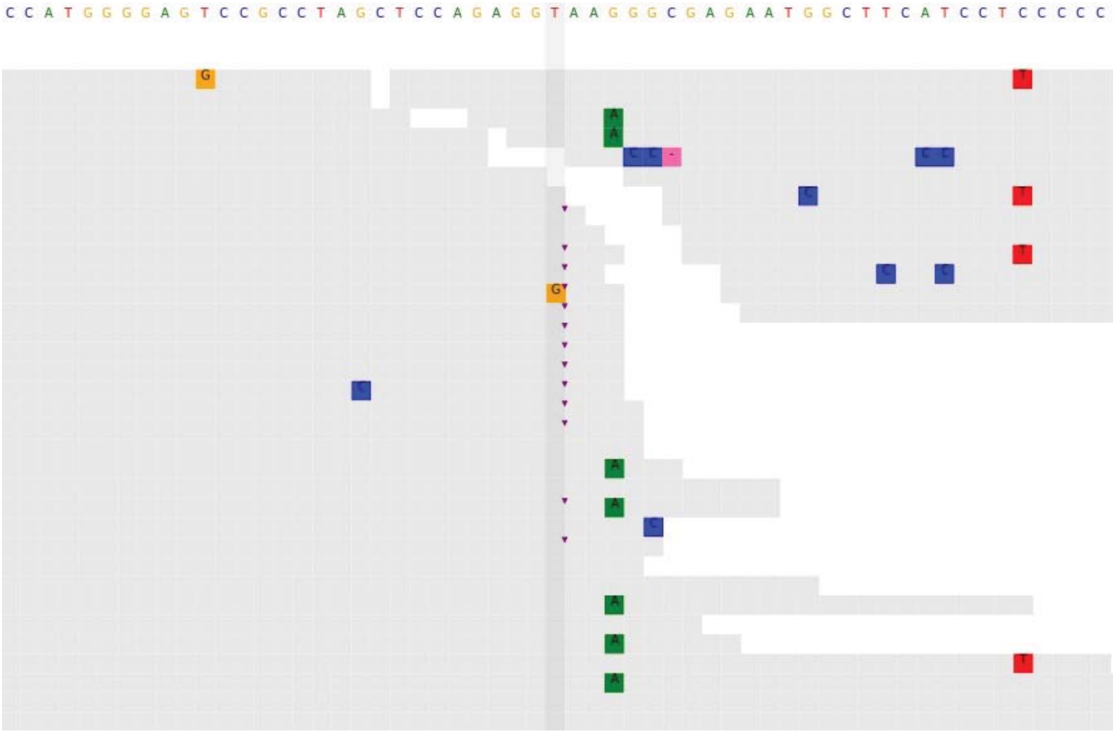

Figure 3.4: Manual inspection of variant calling at nominally associated sites, to identify low quality sites that may have passed the broad quality control thresholds.

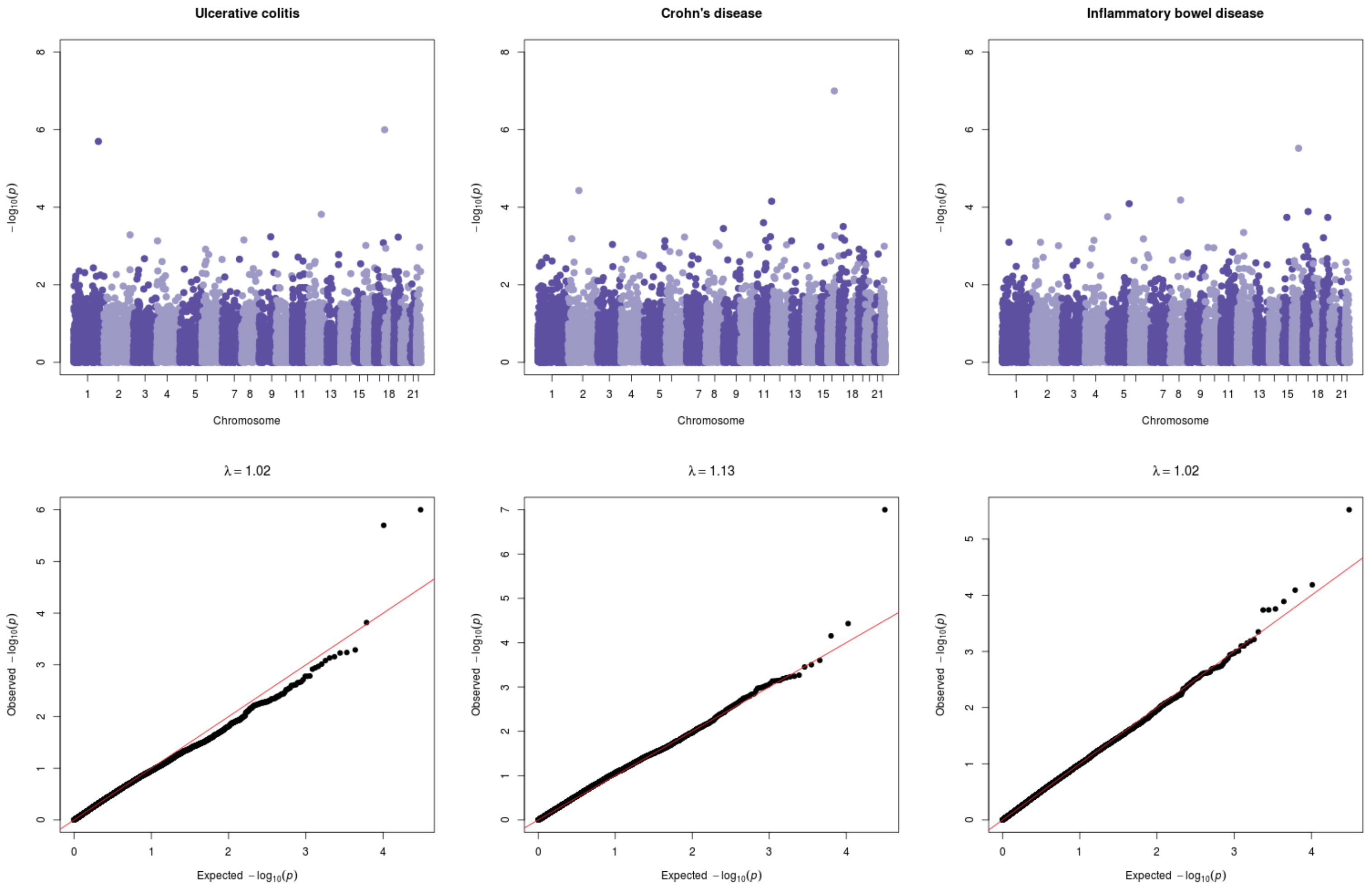

Figure 3.5: Manhattan and QQ plots showing the results of gene-based burden tests using rare, functional coding variation.

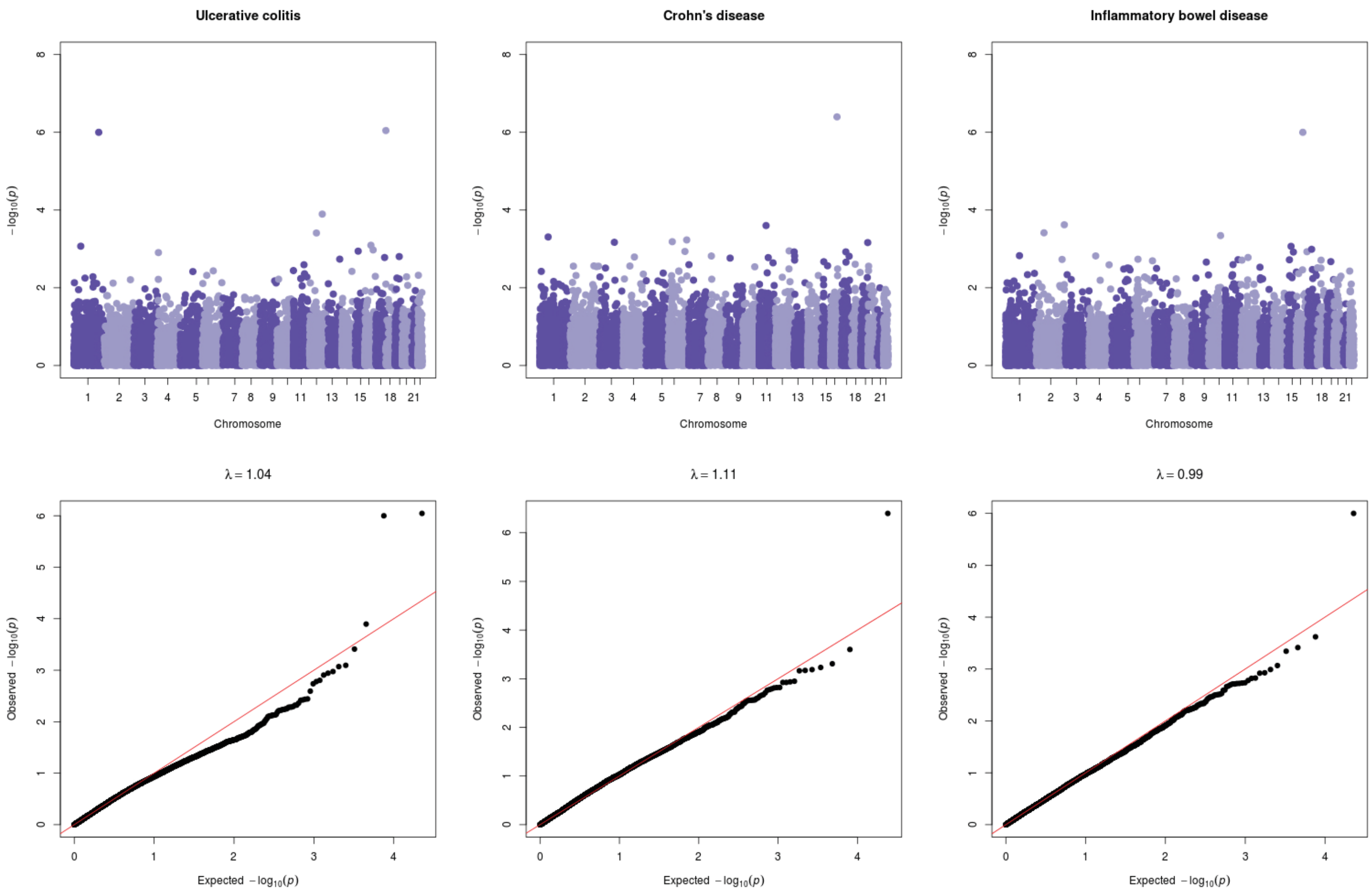

Figure 3.6: Manhattan and QQ plots showing the results of gene-based burden tests using rare, functional coding variation that is predicted to be damaging.

Table 3.4: Genes with a p-value $< 5 \times 10^{-4}$ 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
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

The only gene for which I detected a significant burden of rare variants was *NOD2* ($P_{functional} = 1 \times 10^{-7}$), the well-known Crohn's disease risk gene. To ensure this association was not due to the known low frequency *NOD2* risk variants, I evaluated the independence of the rare variant signal against the common IBD-associated 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 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.

When compared to a previous targeted sequencing study by Rivas et al. (2011), which investigated *NOD2* in 350 CD cases and 350 controls, I discover a number of additional variants (Figure 3.7). These additional variants can be seen to be contributing to the significant burden of rare variation in *NOD2*, with evidence of a signal remaining even after removal of the previously discovered rare variants ($P_{functional} = 5.4 \times 10^{-4}$, $P_{damaging} = 7.5 \times 10^{-5}$). However, cumulatively these additional variants explain just 0.13% of the variance in Crohn's disease liability, compared to 1.15% for the previously known *NOD2* variants (starred in Figure 3.7). This highlights the fact that the low frequency of very rare variants means that they cannot account for much of the overall population variability in disease risk.

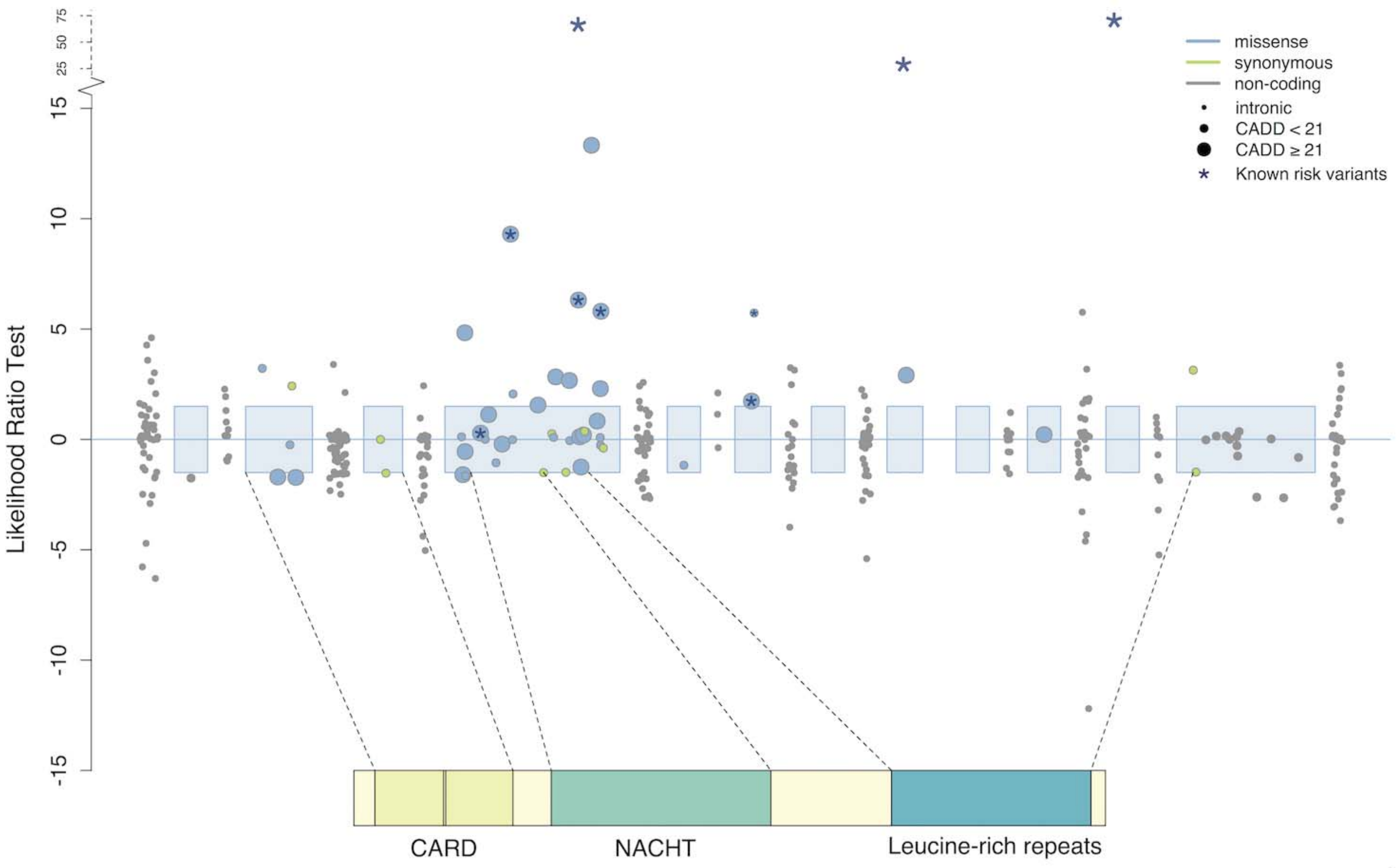

Figure 3.7: Each point represents the contribution of an individual variant to my *NOD2* burden test. Three common variants (rs2066844, rs2066845, rs2066847) are shown, and the six rare variants identified by Rivas et al. (2011) using targeted sequencing are starred. Exonic regions (not to scale) are marked in blue, with their corresponding protein domains highlighted.

Gene set tests

Some genes that have been implicated by IBD GWAS had suggestive p-values, but did not reach exome-wide significance ($P = 5 \times 10^{-7}$, Table 3.4). To test if the allelic series of associated variation observed in *NOD2* might also exist at other known IBD genes, I combined the individual gene results to perform gene set tests across IBD risk genes.

For these tests I created two separate definitions of IBD risk genes. The first, more stringent, definition included only genes that have been confidently implicated in IBD risk (Table 3.5) through fine-mapping and eQTL studies (Huang et al., 2015; Fairfax et al., 2014; Wright et al., 2014). A second, broader definition of IBD-associated genes was created to also include 63 additional genes that were implicated by two or more candidate gene approaches in Jostins et al. (2012).

Table 3.5: IBD-associated genes implicated by a coding variant in the fine-mapping credible sets recently defined by Huang et al. (2015), or with a plausible eQTL association.

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

I first tested the stringent gene set (after excluding *NOD2*, which otherwise dominates the test) using an enrichment procedure that allows for genes with opposite directions of effect to be combined, as described in Chapter 2. To account for residual bias due to sequencing depth differences between cases and controls (that is not fully accounted for using the RVS statistic with such large burden tests), I evaluate the significance of the gene set within the context of the exome-wide gene set. The test was performed to 10^5 permutations separately for CD, UC and IBD, and for each of the functional coding and predicted damaging variant definitions. The results from these tests are summarised in Table 3.6.

Table 3.6: P-values for burden tests performed on the stringently-defined set of IBD risk genes. Results for the Crohn's disease burden test excluding *NOD2* are shown in parentheses.

	Functional coding	Predicted damaging
UC	0.7330	0.4615
CD	0.0001 (0.2291)	0.0000 (0.0045)
IBD	0.2275	0.0026

I detect a burden of rare variants in the twelve confidently implicated Crohn's disease genes ($P_{damaging\_CD} = 0.0045$) and seven confidently implicated inflammatory bowel disease genes ($P_{damaging\_IBD} = 0.0026$) that contained at least one damaging missense variant. This signal is driven by a mixture of genes where rare variants are risk increasing (e.g. *NOD2*) and risk decreasing (e.g. *IL23R*), as shown in Figure 3.8. It is notable that this burden is not detected when considering all functional coding variation, highlighting the value of being able to predict the likely functional impact of a variant in order to better refine the signal to noise ratio of the burden tests. Similarly, I observe no signal in the second, less stringently defined, set of IBD-associated genes (Table 3.7). Figure 3.8 highlights how the broader gene set definition contributes a number of genes that are not associated with IBD in this dataset, causing the signal to be diluted. This observation underscores the importance of using methods such as fine-mapping and eQTL associations when causally assigning an association signal to a particular gene.

Table 3.7: The burden of rare, predicted damaging (CADD ≥21) coding variation in IBD gene sets.

Gene sets	Constituents	Phenotype	*P*-value
NOD2	*NOD2*	CD	4×10^{-7}
Other IBD genes implicated by causal coding or eQTL variants (genes in brackets had zero contributing rare variants)	*CARD9, FCGR2A, IFIH1, IL23R, MST1, (SMAD3), TYK2, (IL10), IL18RAP, (ITGAL), NXPE1, TNFSF8*	UC	0.4615
	ATG16L1, CARD9, CD6, FCGR2A, FUT2, IL23R, MST1, (NOD2), PTPN22, (SMAD3), TYK2, ERAP2, (IL10), IL18RAP, (IL2RA), (SP140), TNFSF8	CD	0.0045
	CARD9, FCGR2A, IL23R, MST1, (SMAD3), TYK2, (IL10), IL18RAP, TNFSF8	IBD	0.0026
Other IBD GWAS genes	Genes implicated by two or more candidate gene approaches in Jostins et al. (2012)	UC	0.9512
		CD	0.9438
		IBD	0.9307

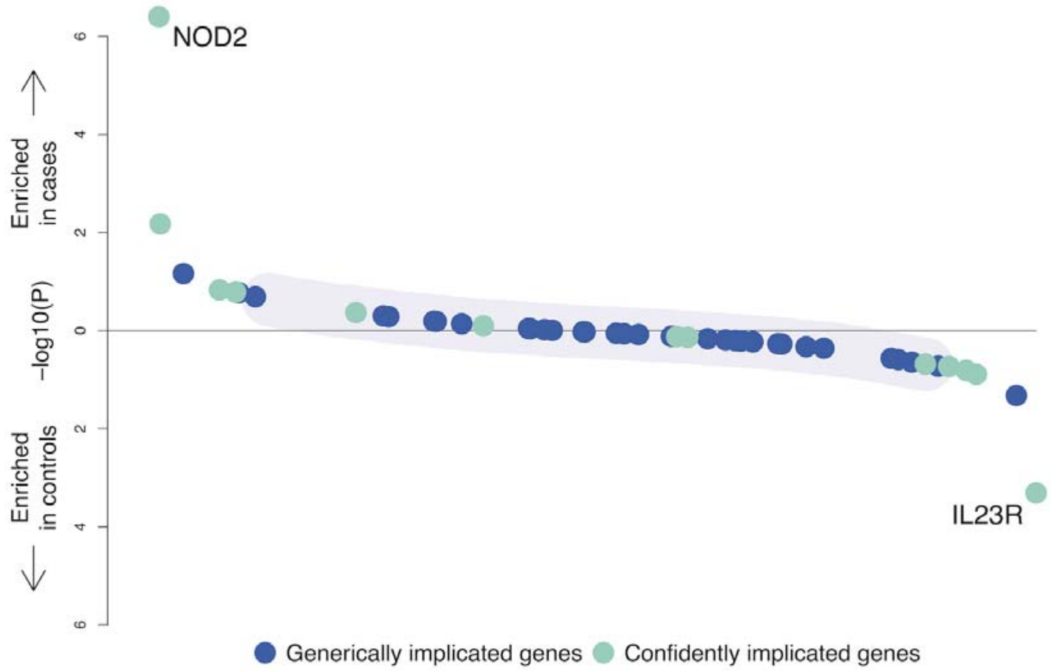

Figure 3.8: The burden of rare damaging variants in Crohn's disease. Each point represents a gene in my confidently implicated (green) or generically implicated (blue) gene sets. Genes are ranked on the x-axis from most enriched in cases to most enriched in controls, and position on the y-axis represents significance. The purple shaded region indicates where 75% of all genes tested lie. The burden signal is driven by a mixture of genes where rare variants are risk increasing (e.g. *NOD2*) and risk decreasing (*IL23R*).

3.4.3 Burden testing across non-coding regions

Enhancer-based burden tests

Using the same approach outlined above for individual genes, I evaluated the role of rare (MAF $\leq 0.5\%$ in controls) regulatory variation using burden tests across enhancer regions. I consider enhancer regions as defined by the FANTOM5 project (Andersson et al., 2014), which used cap analysis of gene expression (CAGE) to identify enhancer activity through the presence of balanced bidirectional capped transcripts. In particular, I focus my testing on those enhancers that were transcribed at a significant expression level in at least one of the 432 primary cell or 135 tissue samples tested by the FANTOM5 consortium, which are referred to as 'robust enhancers' by Andersson et al. (2014). The locations of these robust enhancers were downloaded using the `robust_enhancers.bed` track available at http://enhancer.binf.ku.dk/presets/.

As with the gene-based burden tests, I looked to restrict the tested variants to those sites predicted to have some sort of functional impact, in order to maximise power. However, estimating the likely functional impact of variation within an enhancer region is a challenging task, as understanding is generally limited to a handful of sites that have been through extensive experimental follow-up. One of the few functional aspects of non-coding variation that can be predicted genome-wide is the presence of certain transcription factor binding motifs, and whether a given variant is likely to disrupt or create a known motif. The performance of other measures that have been calculated genome-wide, including the CADD score, have generally not been thoroughly evaluated in non-coding regions due to a lack of testing data.

For each robustly-defined enhancer, I therefore chose to perform two burden tests: one containing all variation overlapping with the enhancer region, and one containing just those variants predicted to disrupt or create a known transcription binding motif (TFBM). I annotated variants as TFBM-disrupting or TFBM-creating using the approach described by Huang et al. (2015), who test for variants that are likely to affect a highly conserved position in a TFBM. How conserved a position is can be determined using the information content (IC): this can be calculated using Equation 3.1, where $f_{b,i}$ is the frequency of base b at position i (D'haeseleer, 2006).

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i} \tag{3.1}$$

I considered all ENCODE transcription factor ChIP-seq motifs (Kheradpour and Kellis, 2014) that had an overall information content ≥ 14 bits (which is 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).

Each test was repeated separately for UC, CD and IBD, resulting in $121,848$ tests, with an average of 2.27 variants contributing to each test (Table 3.8).

Table 3.8: 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	Affecting a TFBM	Total
UC	28,292 (2.64)	11,532 (1.29)	39,824 (2.25)
CD	29,628 (2.75)	12,403 (1.31)	42,031 (2.32)
IBD	28,453 (2.62)	11,540 (1.29)	39,993 (2.24)

No individual enhancer contains a significant burden of rare variation (Figures 3.9 and 3.10) and passes manual quality control. It is also worth noting that, even for those variants that appear amongst the ‘froth’ of suggestively significant *p*-values, at this stage it is very difficult to draw meaningful conclusions from these individual enhancer burden tests. For the majority of enhancers in the human genome, it is not known how they are likely to affect the expression of a given gene, or even which gene they are likely to act upon.

A common approach to try and derive this information is to map expression quantitative trait loci (eQTLs), which are genomic regions statistically associated with the expression level (mRNA abundance) of a given gene (Albert and Kruglyak, 2015). Alternatively, enhancer-gene interactions can be detected directly, using conformation capture methods such as Hi-C. These methods take advantage of the fact that, during transcription, the enhancer and promoter need to be brought into close physical proximity to chemically fix chromosomal contacts. This causes fragments of DNA that are not necessarily close in the linear genome to be linked prior to sequencing, allowing long-range spatial contacts to be resolved (Belton et al., 2012).

However, regardless of the method used, identifying the role of a given enhancer requires testing in the correct cell type and under the correct conditions. For example, Fairfax et al. (2014) discover a number of important immune eQTLs that only occur in monocytes after application of specific stimuli. To try and capture some of this cell-specific expression, studies such as the GTEx consortium are mapping eQTLs across a range of tissues in multiple individuals (GTEx Consortium, 2015), while others are undertaking similar endeavours using Hi-C (Mifsud et al., 2015). As these resources continue to grow, refining of enhancer variant sets to test and interpretation of individual enhancer results may be improved in the future.

Cell- and tissue-specific enhancer set tests

Although extensive catalogues of enhancer activity across cell types and conditions are still under development, FANTOM5 does provide an estimate of cell- and/or tissue-type specific expression across 69 cell types and 41 tissues (Table 3.9). I therefore combined the individual enhancer tests into sets based on these expression patterns, looking to both improve power in an analogous fashion to the gene set tests above, and increase the interpretability of any rare variant burden that may be uncovered.

Enhancers were assigned to groups using the definition of 'positive differential expression' provided by Andersson et al. (2014). This considers the union of all significantly expressed enhancers from all samples within a given cell or tissue type (a 'facet'), and performs pair-wise comparisons between each of the facets (assessing cells and tissues separately). An enhancer is considered differentially expressed in a given facet if it has at least one pair-wise significant differential expression, plus overall positive standard linear statistics. This means that positive differential expression is therefore not the same as exclusive expression in a given cell or tissue. I obtained lists of these differentially expressed enhancer sets from http://enhancer.binf.ku.dk/presets/.

None of these cell- or tissue-specific enhancer sets had a significant burden of rare variation after correction for multiple testing (Table 3.10).

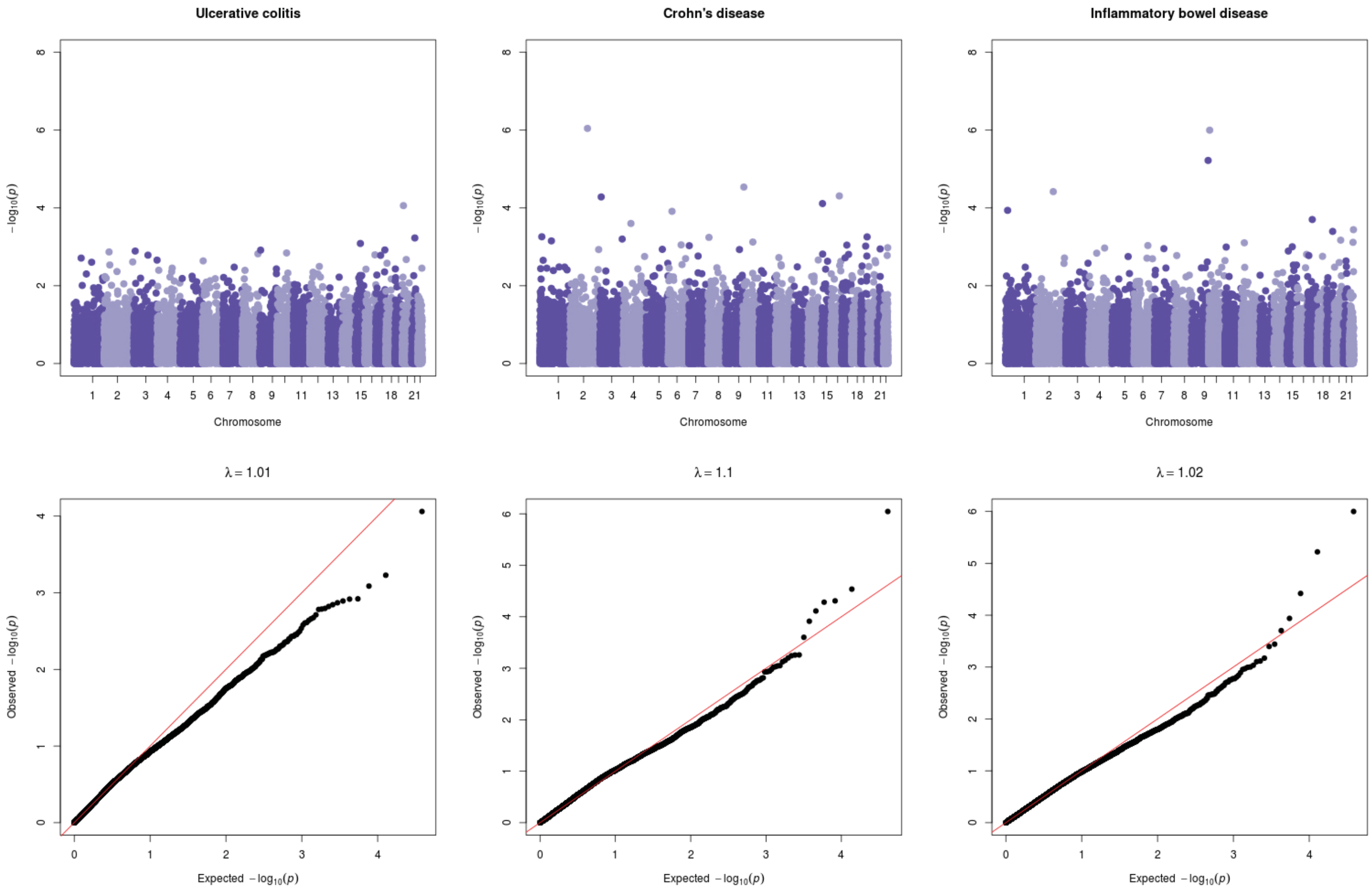

Figure 3.9: Manhattan and QQ plots showing the results of enhancer-based burden tests using all rare variation.

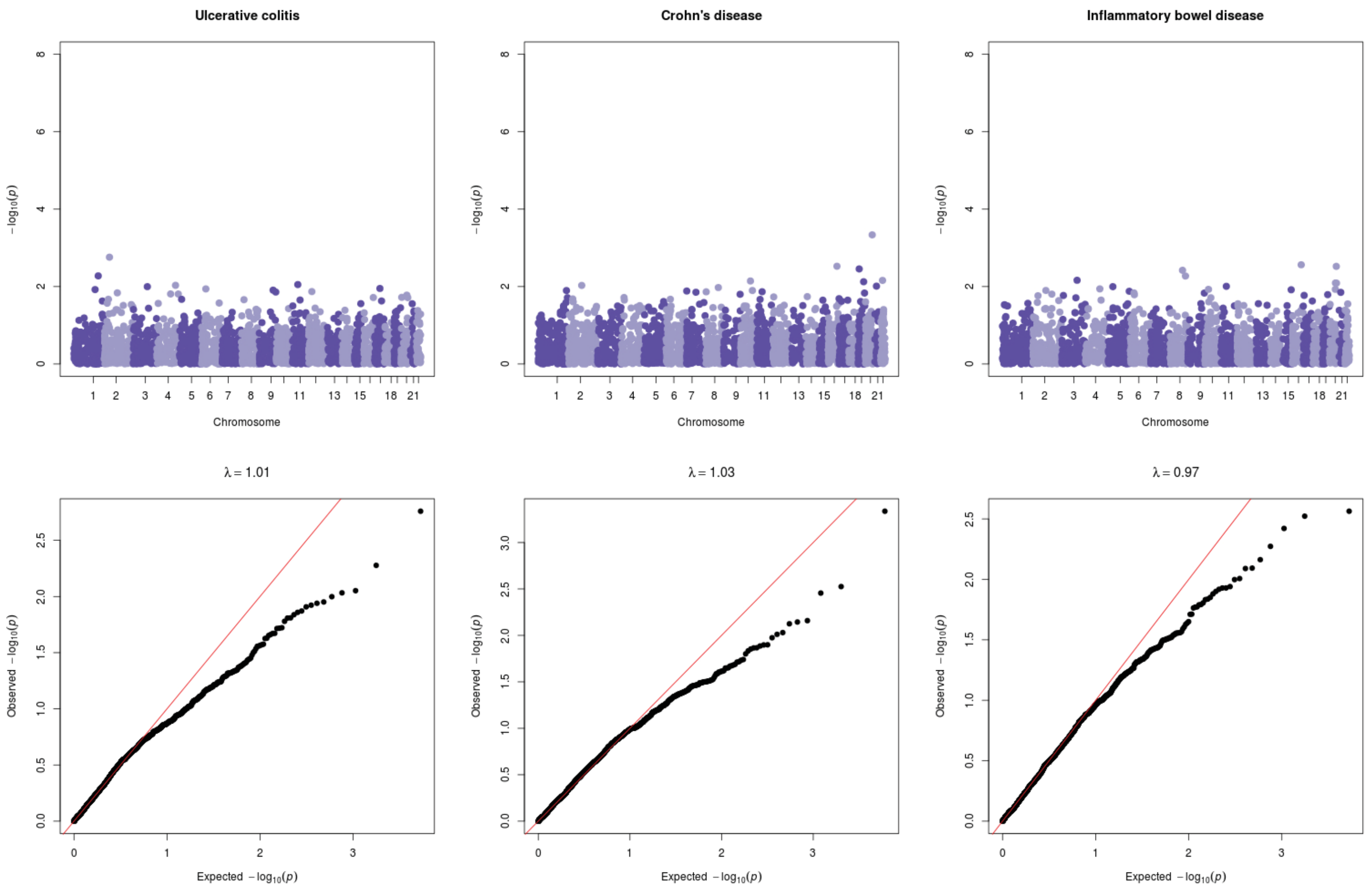

Figure 3.10: Manhattan and QQ plots showing the results of enhancer-based burden tests using rare variation predicted to disrupt or create a transcription factor binding motif. A depletion of very small p-values is observed, possibly due to the low average number of variants contributing to these burden tests (Table 3.8).

Table 3.9: The 69 cell types and 41 tissue types for which FANTOM5 defines preferentially expressed enhancer sets.

Cell types	
neuronal stem cell	endothelial cell of lymphatic vessel
myoblast	epithelial cell of Malassez
osteoblast	lens epithelial cell
ciliated epithelial cell	epithelial cell of prostate
blood vessel endothelial cell	epithelial cell of esophagus
mesothelial cell	mammary epithelial cell
T cell	preadipocyte
mast cell	keratocyte
sensory epithelial cell	trabecular meshwork cell
astrocyte	respiratory epithelial cell
mesenchymal cell	enteric smooth muscle cell
fat cell	kidney epithelial cell
chondrocyte	amniotic epithelial cell
melanocyte	cardiac fibroblast
hepatocyte	fibroblast of choroid plexus
skeletal muscle cell	fibroblast of the conjuctiva
macrophage	fibroblast of gingiva
keratinocyte	fibroblast of lymphatic vessel
vascular associated smooth muscle cell	fibroblast of periodontium
tendon cell	fibroblast of pulmonary artery
dendritic cell	hair follicle cell
stromal cell	intestinal epithelial cell
neuron	iris pigment epithelial cell
reticulocyte	placental epithelial cell
corneal epithelial cell	retinal pigment epithelial cell
monocyte	bronchial smooth muscle cell
acinar cell	smooth muscle cell of the esophagus
natural killer cell	smooth muscle cell of trachea
hepatic stellate cell	uterine smooth muscle cell
pericyte cell	skin fibroblast
urothelial cell	gingival epithelial cell
cardiac myocyte	fibroblast of tunica adventitia of artery
basophil	endothelial cell of hepatic sinusoid
neutrophil	smooth muscle cell of prostate
lymphocyte of B lineage	

Continued on next page

Table 3.9 – *Continued from previous page*

Tissue types	
lymph node	submandibular gland
large intestine	parotid gland
blood	blood vessel
throat	placenta
testis	thyroid gland
stomach	lung
heart	skin of body
brain	spleen
eye	liver
penis	small intestine
female gonad	gallbladder
uterus	kidney
vagina	spinal cord
adipose tissue	umbilical cord
esophagus	meninx
salivary gland	prostate gland
skeletal muscle tissue	thymus
smooth muscle tissue	tonsil
urinary bladder	olfactory region
pancreas	internal male genitalia
tongue	

Table 3.10: Enhancer set-based tests with $P < 0.005$. '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	Disease	Annotation	# enhancers	# 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

3.5 Low frequency variation

To investigate the role of low frequency variation in this sequencing dataset, we tested 13 million SNPs and small indels with MAF $\geq 0.1\%$ for association. It was noted that quality control had successfully controlled for systematic differences due to sequence depth ($\lambda_{1000\_UC} = 1.05$, $\lambda_{1000\_CD} = 1.04$, $\lambda_{1000\_IBD} = 1.06$, Figure 3.11), while still retaining power to detect known associations.

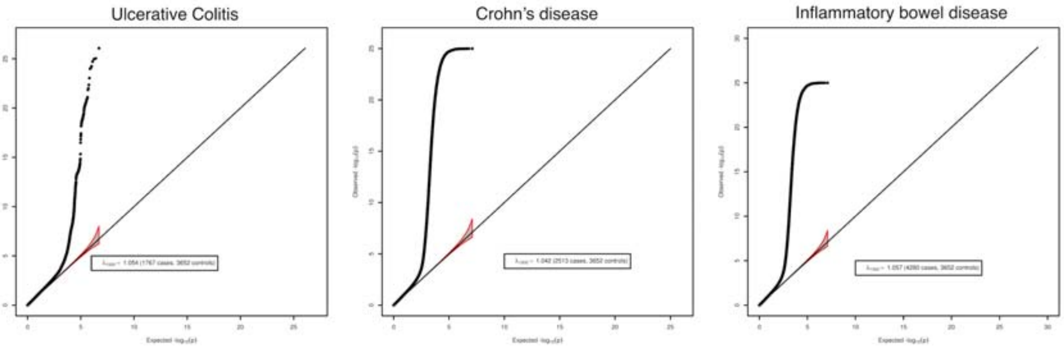

Figure 3.11: QQ plots of genome-wide association studies for variants with MAF $\geq 0.1\%$ in the sequencing dataset. λ_{1000} values are reported for the ulcerative colitis, Crohn's disease and inflammatory bowel disease analyses. Grey shapes show 95% confidence intervals. Figures produced by Yang Luo.

However, while it was estimated that this stringent quality control produced well calibrated association test statistics for more than 99% of sites, there were also many extremely significant p-values at SNPs outside of known loci (for example, there were ~7,000 sites with $P < 1 \times 10^{-15}$). 95% of these extremely significant sites had an allele frequency below 5%. In contrast to GWAS, where basic quality control can almost completely eliminate false positive associations, the biased sequencing depths in this study makes it difficult to identify true associations from this data alone.

3.5.1 Imputation into GWAS

As was also observed by a previous study of type 2 diabetes with a similar design (Fuchsberger et al., 2016), our sequencing dataset alone is not well powered to identify new associations, even if all samples were sequenced at the same depth (Figure 3.12).

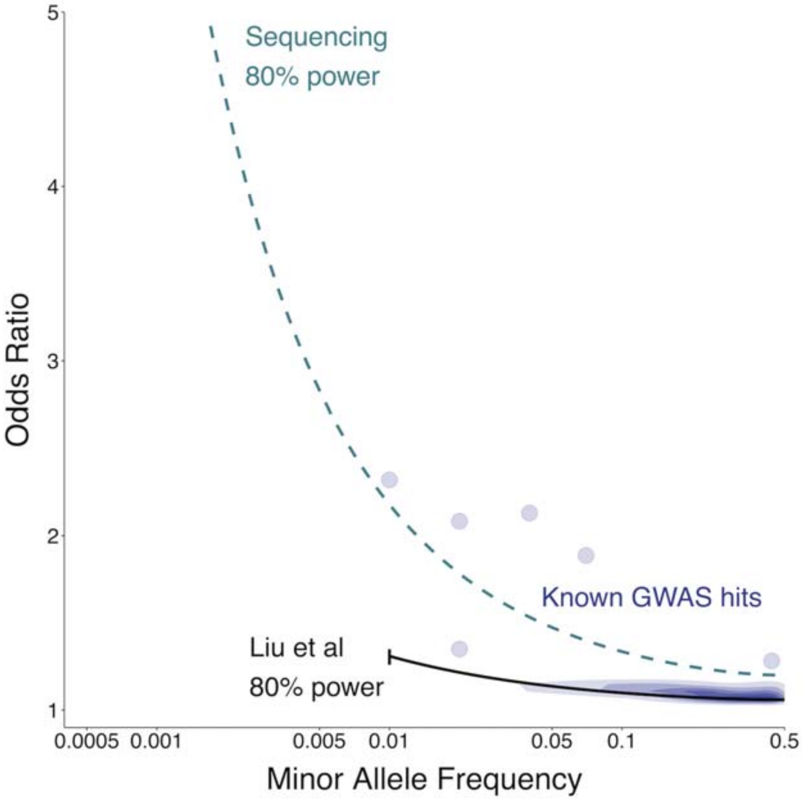

Figure 3.12: Relative power of this study compared to previous GWAS. The black line shows the path through frequency-odds ratio space where the latest International IBD Genetics Consortium (IIBDGC) meta-analysis (Liu et al., 2015) had 80% power, and the green line shows the same for this study. The earlier study had more samples but restricted their analysis to MAF > 1%.

We therefore sought to increase power by using imputation to leverage both new and existing IBD GWAS. As previous data has shown that expanded reference panels can significantly improve the imputation accuracy of low frequency variants (McCarthy et al., 2016), we built a phased reference panel of 10,971 individuals by combining our low coverage whole genomes with the 1000 Genomes Phase 3 haplotypes (1000 Genomes Project Consortium et al., 2015).

I then collected all available UK IBD GWAS data, including samples from two previous studies that did not overlap with those in our sequencing dataset (The Wellcome Trust Case Control Consortium, 2007; Barrett et al., 2009), and a novel UK IBD Genetics Consortium cohort. This new UK IBD GWAS consisted of 8,860 IBD patients without previous GWAS data and 9,495 UK controls from the Understanding Society project (www.understandingsociety.ac.uk), all genotyped using the Illumina HumanCoreExome v12 chip. I shall discuss the variant calling and quality control procedures I applied to this dataset in Chapter 4.

These genotyped samples were all imputed using the PBWT software (Durbin, 2014) and the IBD-enriched reference panel described above. We combined these imputed genomes with our sequenced genomes to create a final dataset of 16,267 IBD cases and 18,841 UK population controls (Table 3.11).

Table 3.11: Sample counts of the imputed GWAS cohorts.

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
Sequencing_CD	2, 513	3, 652	6, 165
Sequencing_UC	1, 767	3, 652	5, 419
Sequencing_IBD	4, 280	3, 652	7, 932
Total	16, 267	18, 841	35, 108

3.5.2 Quality control and association testing

I tested each GWAS cohort separately for association to UC, CD and IBD using a likelihood score test as implemented in SNPTEST v2.5 (Marchini and Howie, 2010), conditioning on the first ten principal components as computed for each cohort when excluding the MHC region (chromosome 6:28-34Mb). I then filtered all output to sites with MAF $\geq$ 0.1%, and INFO $\geq$ 0.4, before using METAL (Willer et al., 2010) 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 $\geq$ 0.4).

The output of the fixed-effects meta-analysis was then further filtered to remove sites with:

- INFO< 0.8 in at least 1/3 (CD,UC) or 2/4 (IBD) of the cohorts included in the meta-analysis
- High evidence for heterogeneity ($I^2 > 0.90$) or deviations from HWE in controls ($P_{HWE} < 1 \times 10^{-7}$) in any of the cohorts
- A meta-analysis p-value higher than all of the cohort-specific p-values
- No evidence of association with IBD in these datasets, but present in the Immunochip or IIBDGC datasets

This produced high quality genotypes at 12 million variants, which represented more than 90% of the sites with MAF > 0.1% that we could directly test in our sequences. Compared to the most recent meta-analysis by the IIBDGC (Liu et al., 2015), which used a reference panel almost ten times smaller than ours, we tested an additional 2.5 million variants for association to IBD. Furthermore, because the GWAS cases and controls were genotyped using the same arrays, they should be not be differentially affected by the variation in sequencing depths in the reference panel, and thus not susceptible to the artifacts observed in the sequence-only analysis. Indeed, compared to the thousands of false-positive associations present in the sequence-only analysis, the imputation based meta-analysis revealed only four previously unobserved genome-wide significant IBD associations. Three of these had MAF > 10%, so were carried forward to a meta-analysis of our data and published IBD GWAS summary statistics as will be discussed in Chapter 4.

3.5.3 p.Asp439Glu in *ADCY7* doubles risk of ulcerative colitis

The fourth new association ($P = 9{\times}10^{-12}$) was a 0.6% missense variant (p.Asp439Glu, rs78534766) in *ADCY7* that doubles risk of ulcerative colitis (OR=2.19, 95% CI =1.75-2.74), and is strongly predicted to alter protein function (SIFT=0, PolyPhen=1, MutationTaster=1). This variant was associated ($P = 1 \times 10^{-6}$) in a subset of directly genotyped individuals, suggesting the signal was unlikely to be driven by imputation errors. However, to further validate this finding we obtained two replication cohorts:

- *450 UC cases and 3,905 controls (p=0.0009)*
  We genotyped an additional 450 UK ulcerative colitis cases and obtained 3,905 population controls (Dupuytren's contracture cases) from the British Society for Surgery of the Hand Genetics of Dupuytren's Disease consortium, both genotyped using the Illumina Human Core Exome v12 array. I applied the same quality control procedure to this replication dataset as the new UK IBD GWAS dataset (see Chapter 4).

- *982 UC cases and 136,464 controls from the UK Biobank (p=0.0189)*
  We extracted an additional 982 additional UC samples and 136,464 controls from the UK Biobank, genotyped on either the UK Biobank Axiom or UK BiLEVE array. Standard Biobank quality control was used (http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf), and non-British or Irish individuals were excluded from further analysis. Cases were defined as those with self-reported ulcerative colitis or an ICD10 code of K51 in their Hospital Episode Statistics (HES) record. Controls were defined as those individuals without a self-diagnosis or hospital record of ulcerative colitis or Crohn's disease (HES = K50).

Logistic regression conditional on 10 principal components was carried out in both replication cohorts. A meta-analysis of all three directly genotyped datasets showed genome-wide significant association ($p = 1.6 \times 10^{-9}$), no evidence for heterogeneity ($p = 0.19$) and clean cluster plots (Table 3.12, Figure 3.13).

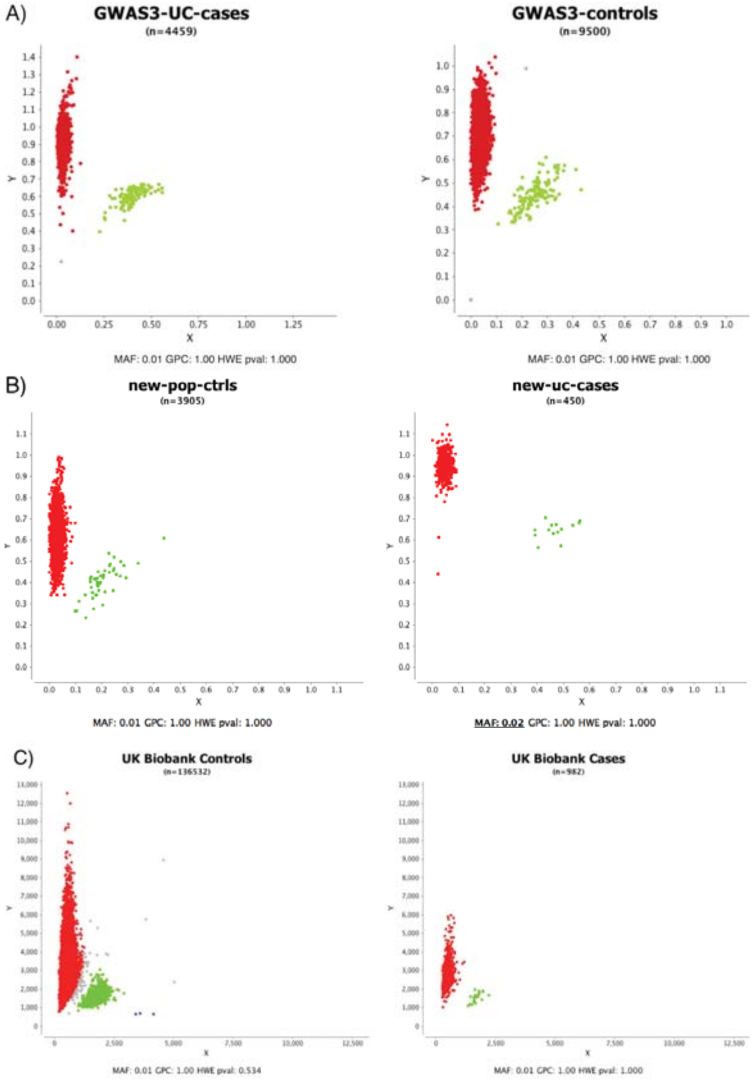

Figure 3.13: Cluster plots are shown for rs78534766 (chr16:50335074, *ADCY7* p.Asp439Glu) for the A) new UK IBD GWAS, B) replication and C) UK Biobank samples that passed quality control. The SNP genotypes have been assigned based on cluster formation in scatter plots of normalized allele intensities X and Y. Each circle represents one individual's genotype. Blue and red clouds indicate homozygote genotypes for the SNP (CC/AA), green heterozygote (CA) and grey undetermined. Figures generated by Daniel Rice.

Table 3.12: Association statistics for rs78534766 (chr16:50335074, *ADCY7* p.Asp439Glu) across UC cohorts. Missingness in cases and controls is zero for the sequenced data due to the genotype refinement step, and there is also zero missingness in the imputed data. Table compiled by Loukas Moutsianas.

Cohort	Cases	Controls	OR [95% CI]	*P*-value	MAF (controls)	Method	Info	Missingness (cases/controls)	P_{het}
WTCCC2	1,921	2,918	2.62 [1.63-4.22]	7.03×10^{-05}	0.0061	Imputed	0.82	N/A	
GWAS3	4,072	9,495	2.05 [1.53-2.75]	1.43×10^{-06}	0.0065	Genotyped	N/A	0.00025/0.0024	
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]	9.20×10^{-12}		(Meta-analysis)			0.69
UK Biobank	982	136,464	1.70 [1.18-2.44]	0.0189	0.0061	Genotyped	N/A	0.0000/0.0004	
Replication	450	3,905	4.10 [1.76-9.51]	0.0009	0.0069	Genotyped	N/A	0.0000/0.0044	
All directly genotyped	5,504	149,864	2.06 [1.63-2.60]	1.62×10^{-09}		(Meta-analysis)			0.19
All cohorts	13,264	165,929	2.16 [1.77-2.62]	1.17×10^{-14}		(Meta-analysis)			0.39

A previous study described an association between an intronic variant in *ADCY7* and Crohn's disease (Li et al., 2015), but our signal at this variant ($P = 2.9 \times 10^{-7}$) vanishes after conditioning on the nearby associations at *NOD2*, (conditional $P = 0.82$). By contrast, we observed that p.Asp439Glu shows nominal association with Crohn's disease after conditioning on *NOD2* ($P = 7.5 \times 10^{-5}$, OR=1.40), while the significant signal remains for ulcerative colitis (Figure 3.14). Thus, one of the largest effect alleles associated with UC lies, apparently coincidentally, only 300 kilobases away from a region of the genome that contains multiple large effect CD risk alleles (Figure 3.14).

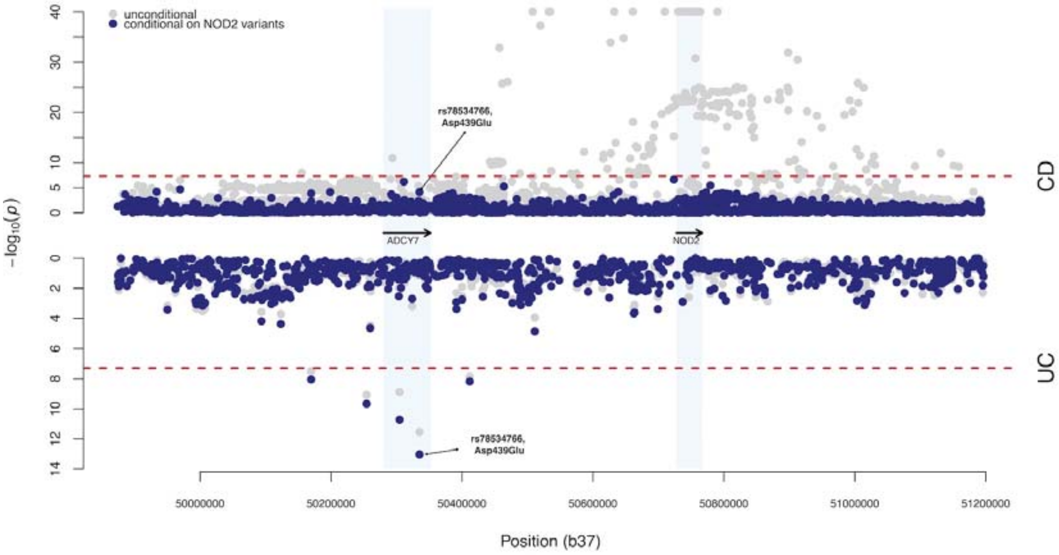

Figure 3.14: Association analysis for the *NOD2*/*ADCY7* region of chromosome 16. Results from the single variant association analysis are presented in gray, and results after conditioning on seven known *NOD2* risk variants in blue. Results for Crohn's disease (CD) are shown in the top half, and ulcerative colitis (UC) in the bottom half. The dashed red lines indicate genome-wide significance, at $\alpha = 5 \times 10^{-8}$. Figure produced by Loukas Moutsianas.

ADCY7 encodes adenylate cyclase 7, part of a family of ten enzymes responsible for the conversion of ATP to the ubiquitous second messenger cAMP. Our associated variant, p.Asp439Glu, affects a highly conserved amino acid within a long cytoplasmic domain that lies immediately downstream of the first of two active sites, and may affect the function of the enzyme by causing misalignment of these active sites (Pierre et al., 2009).

Each adenylate cyclase has distinct tissue-specific expression patterns, with *ADCY7* being expressed in haemopoietic cells (Figure 3.15). Here, cAMP has an important role in the modulation of both innate and adaptive immune functions, including the inhibition of the pro-inflammatory cytokine TNFα, which is the target of the most potent current therapy in IBD (Dahle et al., 2005). In human THP-1 (monocyte-like) cells, siRNA knockdown of *ADCY7* has been shown to increase TNFα production (Risøe et al., 2015). While constitutive Adcy7 knockout mice die in utero, myeloid-specific knockouts have been shown to be viable. These mice exhbit higher production of TNFα by macrophages upon stimulation, as well as impairment of both B cell function and T cell memory, increased susceptibility to lipopolysaccharide-induced endotoxic shock, and a prolonged inflammatory response (Duan et al., 2010; Jiang et al., 2013).

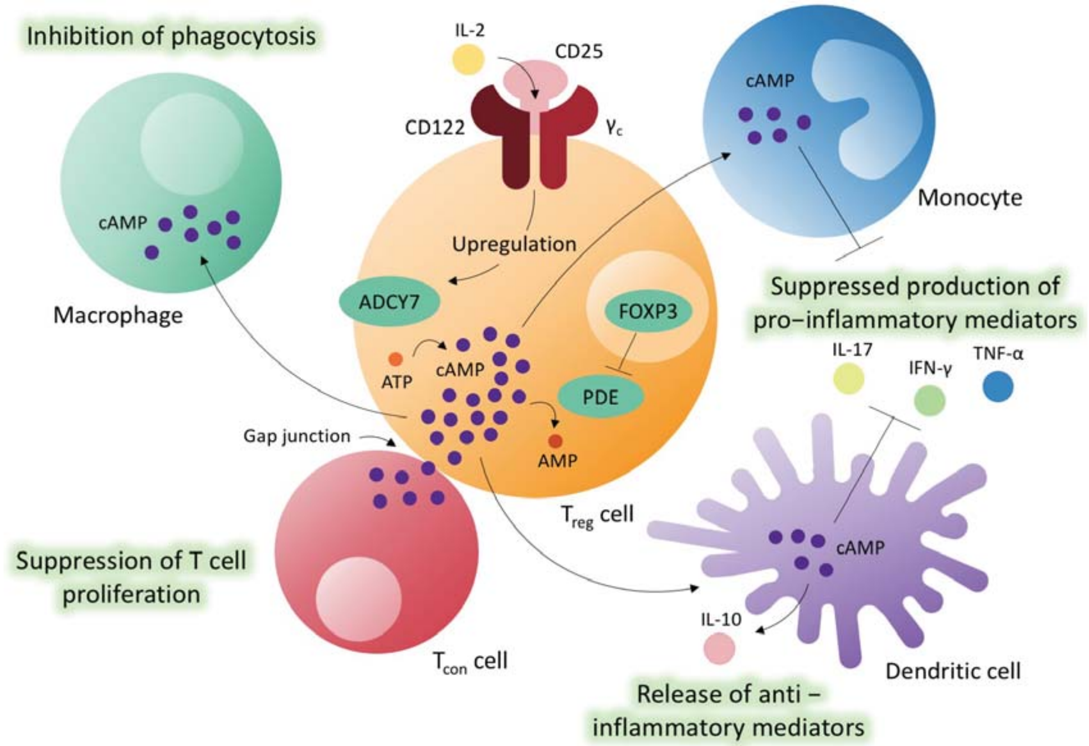
Inhibition of phagocytosis
IL-2
CD25
CD122
γc
cAMP
Monocyte
cAMP
Macrophage
Upregulation
ADCY7
FOXP3
Suppressed production of pro–inflammatory mediators
ATP
cAMP
PDE
IL-17
IFN-γ
TNF-α
AMP
Gap junction
T_reg cell
cAMP
Suppression of T cell proliferation
IL-10
T_con cell
Dendritic cell
Release of anti – inflammatory mediators

Figure 3.15: An overview of the role of ADCY7 in the inflammatory response, where it is responsible for the conversion of ATP to cAMP in haemopoietic cells. A subset of the immune-related functions performed by the secondary signalling molecule cAMP are depicted here (Rossi et al., 1998; Tiemessen et al., 2007; Duan et al., 2010; Boyman and Sprent, 2012; Raker et al., 2016; Rueda et al., 2016).

3.6 Discussion

In this chapter I have described an investigation into the role of rare and low frequency variants in IBD risk, using a combination of low coverage whole genome sequencing and imputation into GWAS data (Figure 3.16). The sole low frequency association uncovered by this study was a missense variant in *ADCY7* that, with an odds ratio of 2.19, represents one of the strongest ulcerative colitis risk alleles outside of the major histocompatibility complex. One possible mechanistic explanation for this association is that a loss of *ADCY7* function leads to reduced production of cAMP in haemopoietic cells, leading to an excessive inflammatory response. Interestingly, a previous study has investigated the use of general cAMP-elevating agents as a potential therapy for intestinal inflammation, with results suggesting that action upon multiple adenylate cyclases in this way may in fact worsen IBD (Zimmerman et al., 2012). Others have looked into targeting specific members of the adenylate cyclase family as potential therapeutics in different contexts (Pierre et al., 2009), but specific upregulation of *ADCY7* has not been attempted. Our association between *ADCY7* and ulcerative colitis raises an intriguing question as to whether altering cAMP signalling in a leukocyte-specific way may be of therapeutic benefit in inflammatory bowel disease.

Although we collected low coverage whole genome sequences specifically to investigate both coding and non-coding variation, our sole new association is a missense variant. This is not particularly surprising: the only previously discovered IBD risk variants with similar odds ratios (Figure 3.16) are all protein-altering changes (affecting the genes *NOD2*, *IL23R* and *CARD9*). The observation that the alleles with the largest effect sizes at any given frequency tend to be coding has been made more generally (Huang et al., 2015), explaining why coding variants are often the first to be discovered when novel technologies allow for new areas of the minor allele frequency spectrum to be explored.

We observe this same pattern when investigating the role of rare variation in IBD risk, where a significant burden of very rare coding variants is seen in previously implicated IBD genes, but no signal is observed across the enhancer regions tested. Although our results imply that rare variants are likely to play an important role in

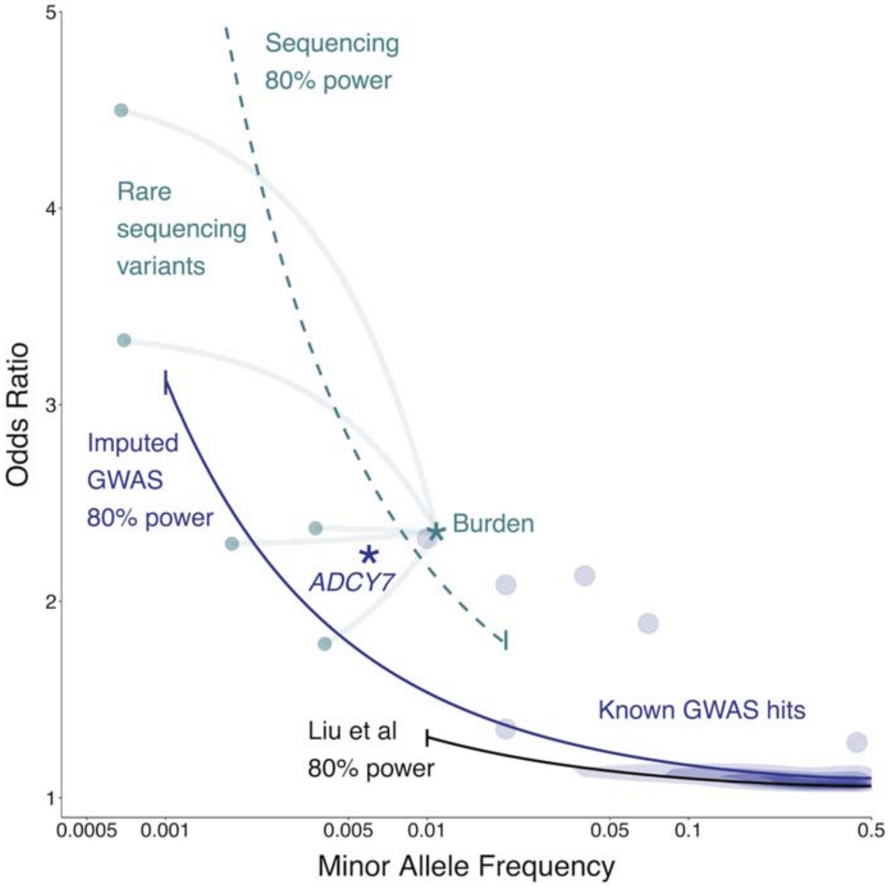

Figure 3.16: The frequency-odds ratio space investigated by this study, comparing the latest IIBDGC meta-analysis (black line) to the sequencing (green) and imputed GWAS (purple) used in this study. The earlier study had more samples but restricted their analysis to MAF > 1%. Purple density and points show known GWAS loci, with our novel *ADCY7* association (p.Asp439Glu) highlighted as a star. Green points show a subset of our sequenced *NOD2* rare variants, and the green star shows their equivalent position when tested by gene burden, rather than individually.

IBD risk, making real progress on rare variant association studies will require much larger numbers of deeply sequenced exomes or whole genomes. Extrapolating for *IL23R*, the known IBD gene with the most significant coding burden (p=0.0005) after *NOD2*, we would require roughly 20,000 cases to reach genome-wide significance (Zuk et al., 2014).

The challenge of detecting a burden of variation in regulatory regions is further compounded by our current inability to clearly distinguish likely functional variation from neutral mutations in non-coding sequence. The importance of being able to make this distinction is highlighted when considering a burden test across known IBD genes: if we include all rare coding variants (MAF $\leq 0.5\%$ in controls, N=136) in IBD genes the p-value is 0.2291, compared to $P = 0.0045$ when using just the subset of 54 coding variants predicted to have a damaging effect. Therefore, identifying the role of rare variation in the non-coding genome is likely to not only require the sequencing of tens of thousands of samples, but also much better discrimination between functional and neutral variants in regulatory regions.

During the course of this work, we noted a number of complexities associated with analysing sequencing data, and in particular with combining data from different studies. The most obvious issue was that, in order to maximise the number of IBD patients that could be sequenced, our cases were sequenced at lower depth that the UK10K control samples. Although very careful joint analysis of the datasets was able to largely overcome this bias, it became clear that the analysis of sequencing datasets at scale will require the development of many novel tools and techniques. Furthermore, these challenges are not just restricted to low coverage whole-genome sequencing designs: the Exome Aggregation Consortium recently noted that variable exome capture technology and sequencing depth across their 60,000 exomes required a joint analysis of such computational intensity that it would be impossible to carry out using the limited resources available to most research centres (Lek et al., 2016).

Therefore, if sequence-based rare variant association studies are to be as successful as common variant GWAS, computationally efficient methods and accepted standards for combining these novel datasets need to be developed. An example of one such effort is the Haplotype Reference Consortium (HRC), which has collected whole genome sequences from more than 32,000 individuals (including the IBD samples discussed here) in order to create a reference panel that can be used for imputation of low frequency and common variants (McCarthy et al., 2016). Imputation into GWAS using this large HRC panel is as accurate as low-coverage sequencing down to MAF $\sim$0.05% (McCarthy et al., 2016), suggesting that in the future the most effective way to discover low frequency variants associated with complex disease will be to impute the huge resources of existing GWAS data with large new reference panels. Thus, while projects such as this one provide valuable resources in the form of publicly available reference panels, it is unlikely that there will be much need for low coverage whole genome sequencing in the future. Together, our results suggest that a combination of continued GWAS imputed using substantial new reference panels and large scale deep sequencing projects will be required in order to fully understand the genetic basis of complex diseases like IBD.