Chapter 4

Quality control and the initial analysis of the IBD 15x cohort

4.1 Introduction

Genetic association studies of inflammatory bowel disease have uncovered the vast architectural complexity of the disorder [92, 49, 116]. While some, mostly coding, genetic associations for IBD increase the risk of the disorder several-fold (e.g., variants in *NOD2* for CD, HLA-DRB1*01:03 for both CD and UC), the majority of known associations are noncoding and have a modest effect size (OR $\sim$ 1.2).

A common criticism of GWAS is whether the discovery of such low-risk associations is relevant to our understanding of disease pathogenesis and, ultimately, whether such discoveries could be translated into therapeutic targets for the next generation of IBD therapies. To counter this, one could point out the long history of IBD GWAS uncovering associations in genes and pathways that are targeted by existing and newly-developed IBD therapies: *IL23*, implicated in the pathogenesis of CD back in 2006 [57], is targeted by ustekinumab – a monoclonal antibody, approved for the treatment of CD in 2016; *TAB2* and *NFKB1* are within the *TNF* signalling pathway targeted by anti-TNF therapies; at least three known associations within the integrin genes (2017 [49]), support the efficacy of vedolizumab and etrolizumab which target the $\alpha 4 \beta 7$ dimer [49]. Neither of the of the

integrin-related variants have a high effect size (OR = 1.10–1.12 [49]), emphasising the inconclusive relationship between the disease risk effect size and the therapeutic relevance.

The success of IBD GWAS at identifying known drug targets is consistent with the observation that drugs with genetically supported targets have double the success rate in clinical development [129, 98]. It is not well understood whether drug targets that are supported by evidence from high effect size variants are more therapeutically relevant. King et al. [98] demonstrate that the drugs that target the manually curated gene-disease associations, described in the Online Mendelian Inheritance in Man dataset (OMIM), have a higher success rate compared to those that target 'GWAS to gene to trait'. One explanation for this is the difficulty in linking the noncoding GWAS variants to the causal gene, resulting in the misidentification of drug targets (see the Introduction chapter). Alternatively, the Mendelian focus of the OMIM dataset means that the majority of genetic variants that are present in it have a very high effect size, suggesting a positive relationship between the effect size and the drug target success.

The translation of GWAS association to genetic targets is nontrivial. Association studies across numerous complex diseases indicate that the majority of the identified common variants are located within the noncoding regions of the genome and cannot be always mapped to a causal variant, yet alone gene [82]. The early-day approach of mapping the noncoding variants to the nearest gene is now known to be error-prone [29] and has been largely superseded. eQTL colocolisation techniques have become a powerful instrument for linking known noncoding associations to their respective genes, but require careful consideration of the tissue type, cell type or even cell stimulation type.

Coding associations provide an easier path from GWAS to target. However, the typically stronger effect size of such variants is at odds with purifying selection: selective pressure either keeps the large effect size risk variants rare, or rapidly pushes them down the allele frequency spectrum. Uncovering further large effect genetic variants associated with IBD is nontrivial. The most recent large-scale GWAS for IBD included more than 25,000 cases and 35,000 controls and was extremely well-powered to detect common, large effect associations. Therefore, the field has likely reached a 'saturation point' when it comes to uncovering such variants. Further discoveries will require a foray into the rare allele frequency spectrum, which is poorly captured by genotyping techniques (see the Introduction chapter).

The IBD 15x study was set up to understand the role of rare coding and noncoding variation in IBD. It is a case-control cohort that includes around 19,000 subjects – 7,000 IBD patients and 12,000 controls, all whole-genome sequenced at 15x target depth. In contrast to array genotyping, short read whole-genome sequencing provides an unbiased way to study rare single nucleotide (SNP) and short insertion/deletion (INDEL) variation across approximately 95–98% of the human genome. In addition, whole genome sequencing (WGS) allows the study of structural variation [102] and accurate typing of alleles in complex regions of the genome such as HLA [55] and KIR [157].

In the past few years, various complex disease consortia have published studies performing rare-variant association studies on WGS datasets, uncovering several novel rare-variant associations for their respective traits (e.g., [63, 125]). However, these efforts are yet to result in a 'gold rush' of new associations similar to the early days of GWAS.

IBD 15x follows a previous IBD WGS study by Luo et al. [116] (4,280 cases sequenced at low-coverage and 3,652 controls) which uncovered a 0.6% frequency missense variant in *ADCY7* that doubles risk of ulcerative colitis. However, the IBD 15x builds upon the insights gathered during the low-coverage sequencing project. Firstly, the cohort includes almost exclusively Crohn's disease patients, allowing us to get more power to detect variants that have a differential effect between CD and UC (see Section 4.3.4). Secondly, it maintains an important balance between the cohort sample size and sequencing depth in order to maximise the statistical power, while not sacrificing too much sensitivity, to detect low frequency and rare variant associations (see Section 4.3.1). Lastly, it contains noticeably more cases and controls to perform the association tests. As discussed in the Introduction and in Section 4.3.1, for anything other than variants with semi-Mendelian effect sizes (OR $>$ 10; have not been previously identified for IBD) it is important to study rare variation in a dataset with at least 15–20,000 samples.

In addition, 15x is planned to be analysed in conjunction with two exome-sequencing datasets: the Broad WES cohort (early meta-analysis described in Section 4.3.4) and the upcoming Sanger IBD WES cohort. Combined, these should exceed 35,000 cases and 75,000 controls, providing a great opportunity to study the contribution of rare coding variation in IBD. In the following discussion, I describe the approaches that can be used to study the noncoding variants – a challenging task at this sample size, but that could help us to understand the genetic architecture of IBD even better.

The analysis stage of the 15x cohort began just a few months ago, and the majority of this time was spent on the sample quality control procedures that are described in the chapter below.

In this chapter, I describe the IBD 15x association study. I describe the initial efforts at variant and sample quality control, in order to enable a whole-genome association study of IBD. In addition, I describe the first results from the IBD 15x study: namely, the replication of several rare variant associations uncovered in the whole-exome cohort produced at the Broad Institute.

4.2 Methods

4.2.1 Power modelling

Sequencing depth (coverage) is the mean number of sequence reads that align to reference bases. For most of the use-cases, it is insufficient to perform sequencing at 1x depth: individual sequence reads have a high error rate and there are likely to be substantial gaps in the sequenced genome. Therefore, a higher sequencing depth is usually chosen – 10x–30x for most association studies, >50x for applications like structural variation discovery and tumour analysis. Variant calling tools, like GATK and DeepVariant, are able to use redundant reads to correct for errors, thus increasing the genotyping quality.

The default coverage for the past two generations of short read sequencers (Illumina HiSeq X, NovaSeq) is 30x, as are the majority of commercial sequencing offers (e.g., Broad Institute Genomic Services, Dante Labs). The relationship between the sequencing cost and depth is close to linear. Given a fixed budget, researchers setting up studies have to consider whether it is more beneficial to sequence a larger cohort (increasing the statistical power) or sequence a smaller cohort at a higher depth (increasing the sensitivity).

In order to increase the size of the cohort, the cases and controls for the IBD whole-genome study were sequenced at 15x target depth. In practice, the median sequencing depth in the final 15x dataset is around 18.5x due to some over-provisioning when libraries are sequenced in multiplexed mode (Figure 4.1). Early in my project, I was involved in efforts to

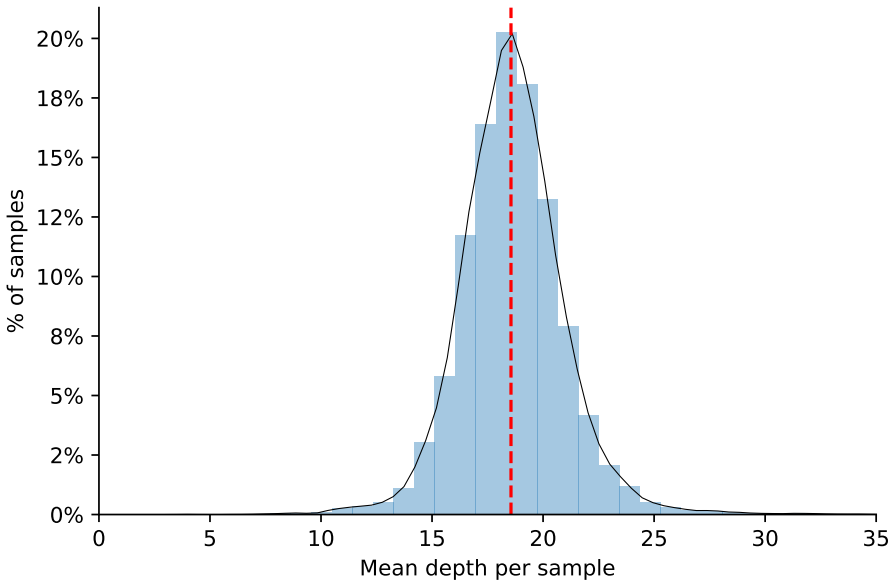


Figure 4.1 Histogram of mean per-sample coverage for samples in the combined 15x cohort. Median depth in the 15x cohort: 18.56x. N=19,374.

evaluate how the lower sequencing depth of the 15x study would influence the the ability to perform rare-variant association studies.

Variant calling sensitivity at different depths

Reduction in sequencing depth is expected to reduce the sensitivity to call variants. NA12878 is a whole-genome sample produced by the Genome in a Bottle project [44]: sequenced with extremely-high ∼300x coverage, it is considered to be the current 'gold standard' of short-read WGS data. Validated variants called at ∼300x are considered to be the truth set in various variant calling benchmarks.

The sample was downscaled by randomly discarding paired-end reads, simulating sequencing at a lower depth. At simulated 30x coverage, 98% of SNPs and 79% on INDELs from the truth set were called by the GATK 3.3 variant caller [147] (work done by Martin Pollard).

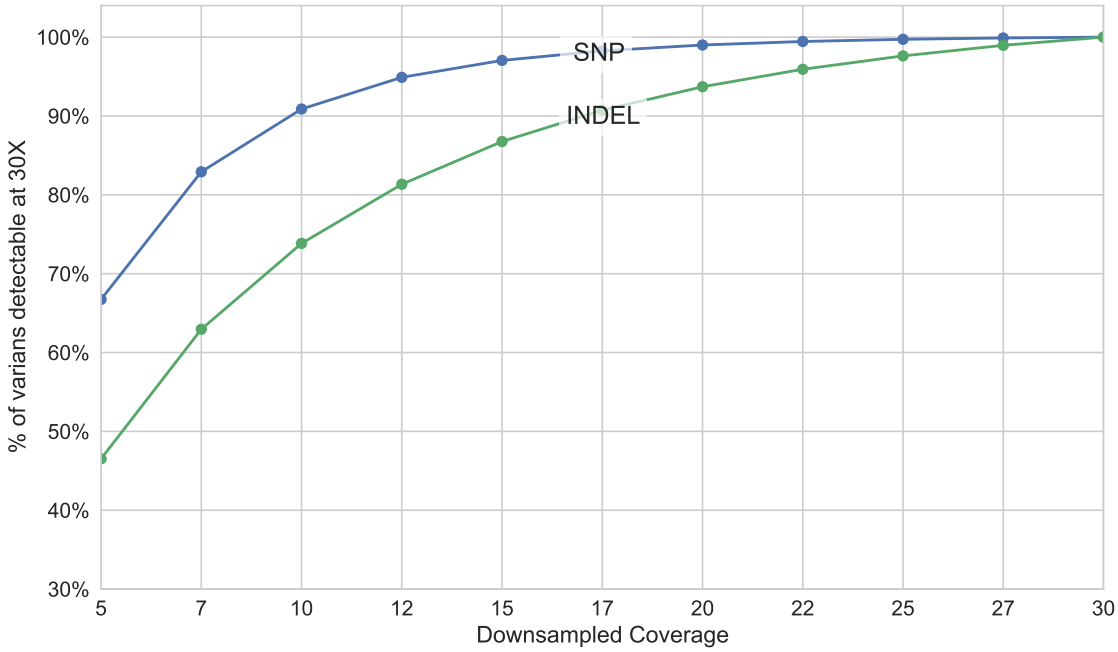


Figure 4.2 Influence of sequencing depth on the ability to detect SNPs and INDELs. Estimates provided by Martin Pollard.

The sample was further downscaled to estimate the loss of sensitivity, compared to sequencing at 30x depth (see Figure 4.2)[1].

The fraction of the called truth set variants appears to plateau around 15–17x. At 15x depth >97.5% sensitivity to discover SNPs at and >87% for INDELs is retained.

Computational model for estimating the statistical power of sequencing studies

I implemented a numeric simulation to calculate the power to detect single variant associations in case-control and quantitative trait settings. The method takes into account sensitivity to detect SNP variants at different depths. In the case-control setting, the variant is present with a probability P_{case} in cases and P_{ctrl} in controls. The model is supplied with a pre-calculated table of sensitivities to detect the variant at a depth $S(d)$. For each of the cases and controls, a random draw between 0 and 1 from a uniform distribution is made. If the draw is $\leq p_{case} * S(d)$ (or $p_{ctrl} * S(d)$ for controls), the variant is considered observed. The Fisher

[1] Variant calling and comparison to the truth set was performed by Martin Pollard [147]

exact test on 2 x 2 table of observations in cases and controls is used to calculate the p-value and the odds ratio. If the p-value is less than or equal to the significance α, the simulation run was successful. N_{sim} simulations are performed to calculate the fraction of successful associations, which is used as a measurement of the statistical power.

4.2.2 Sample selection

Cases: IBD 15x

Samples were initially selected for sequencing from previous DNA collections available at the Sanger Institute. In addition, new samples supplied by the IBD BioResource [143] and other collaborator groups from across the UK were sequenced.

Throughout this chapter I will use the term 'phase' to denote large batches of samples in both IBD 15x (cases) and INTERVAL 15x (population controls). IBD and INTERVAL 15x consisted of three phases each. Within each phase the sequencing protocol remained consistent, while some protocol variability was allowed between the phases to improve the sequencing results (see 4.2.3).

I enforced several criteria for the selected samples:

- Disease type: Crohn's disease[2]
- Self-reported ethnicity: White – British, White – Irish or White – other
- DNA sample passing the QC criteria for PCR-free sequencing on Illumina HiSeq X
- Sample was not previously sequenced as a part of an earlier phase

At the later stages of the project, this was done by my colleagues, who followed a similar protocol.

Considering the high cost of whole-genome sequencing, I attempted to minimise the number of duplicate samples. Firstly, I checked the sample IDs for duplication (e.g., if centre

[2]Some of the patients in the IBD 15x Phase 1 and Phase 3 were diagnosed UC instead of Crohn's. I am finalising the list of the UC patients but it appears to be $\sim$300 cases.

x sends the DNA sample *y* twice). Secondly, I tried to account for situations where the DNA of the same individual is sent for sequencing twice with different IDs (e.g., first from a collaborating centre and then from the IBD BioResource).

I implemented a pipeline that compares the genetic fingerprints (15–25 SNPs via Fluidigm or Sequenom targeted genotyping platforms) that are produced as a byproduct of sample QC by the Sanger DNA Pipelines. The fingerprints are primarily used to detect sex discordance and to evaluate the DNA quality (higher fingerprint messiness typically indicates lower DNA quality and therefore lower quality of sequencing). The pipeline converted the fingerprints for each considered sample (previously sequenced and new candidates for sequencing) into a joint VCF file. I then ran identity by descent calculation via AKT [9] (PLINK-like kinship estimation method). Empirically, I determined that it was only possible to reliably detect duplicates (or monozygotic twins) and not first degree relatives or lower.

Controls: INTERVAL 15x

INTERVAL is a large study of 45,000 healthy blood donors that was initially set up to study the effect of blood donation frequency on subjects' health. The cohort was then used to study the effects of the genetic variation on a variety of blood cell traits [10]. All samples included in the INTERVAL 15x were previously genotyped. Prioritisation for sequencing was based on the availability of certain metabolic phenotype data (not covered in this thesis). Unrelated subjects of European ancestry were selected for further whole-genome sequencing. Sample selection was performed by the Soranzo Team members at the Sanger Institute.

4.2.3 Sequencing

The samples were sequenced at the Sanger Institute between 2016 and 2018. DNA, extracted from whole blood, underwent short-read paired-end sequencing by synthesis using the Illumina HiSeq X Ten machines. The target coverage depth was 15x. Considering the time-scale of the project, there was some variability between individual batches:

PCR versus PCR-free sequencing: INTERVAL 15x Phase 1 (controls, n=5,093) was the first batch of samples that underwent sequencing. During the DNA library preparation, an additional PCR amplification step was added due to the specifications of the library prep

kit. Libraries for the subsequent INTERVAL (Phases 2 and 3) and IBD (Phases 1–3) were prepared using a PCR-free kit.

Single versus dual indexing: IBD 15x Phase 3 (cases, n=2,530) was the latest batch to be sequenced as a part of the 15x project. Dual-indexing of the DNA fragments was applied during the library prep to minimise index misassignment. Other batches were sequenced using the standard single-indexing approach. I provide an overview of how single- and dual-indexing influences the sequencing quality in Section 4.3.2.

4.2.4 Alignment and variant calling

BWA MEM [106] software was used to align the reads to the reference genome. Genome Reference Consortium GRCh38 (with decoys) was used as a reference genome for all phases of IBD and INTERVAL 15x.

Germline variant calling was performed using the GATK4 toolkit [121] following the 'Best Practices' pipeline. Briefly, intermediate sets of SNPs and INDELs were called for each sample via local *de-novo* assembly of haplotypes using the HaplotypeCaller tool. All intermediate calls from both IBD and INTERVAL 15x were then refined during the Joint Genotyping stage.

Alignment was performed by the NPG group and variant calling was performed by the HGI group at the Sanger Institute.

4.2.5 Computational analysis pipeline

One of the biggest difficulties with conducting this project was the scale of the dataset and the computational challenges associated with analysing it. The combined size of the variant call files in compressed VCF format for was approximately 15 terabytes in size (TB = 1024 gigabytes). This is approximately seven times larger than the imputed genotypes of the 500,000 individuals in the UK Biobank cohort [35] and 1,000 times larger than the PANTS anti-TNF dataset described in Chapter 2.

The scale of the present-day association study cohorts has long passed the point at which they can be analysed on a single powerful computer within a reasonable time-frame.

This issue is addressed by distributed computing: analytical tasks are spread across several computers or servers, each with multiple central processing unit (CPU, 'processor') cores. A technique that is often used for performing distributed computation is called MapReduce [50], whereby the tasks are separated into two stages: the *map* procedure independently applies a particular function across individual parts of the dataset (e.g., counting the number of INDELs present in each partition) and the *reduce* procedure collects the outputs from the map stage and summarises them (e.g., adding up the outputs of map functions and getting the total number of INDELs in the dataset). While it is possible to implement MapReduce-like pipelines using traditional GWAS analytical tools like PLINK [151] (breaking up the dataset into per-region .BED files, running the analytic pipeline, creating a custom *reduce* function), the scale of the 15x makes the application of such an approach challenging: the majority of these tools assume that the dataset can be trivially modified and a new version saved onto disk (e.g., when a single individual is excluded during the QC). Writing an additional 15 TB of data onto a disk can take tens of hours even when using a large computational cluster and costs thousands of pounds a year to maintain. The second issue with such ad hoc distribution is that as the complexity of the analytical pipeline increases (e.g., multiple filtering stages, followed by logistic regression), so does the complexity of writing the reduce functions. In theory, tasks can be scheduled efficiently so that the processing of individual parts of the dataset can proceed independently until they need to access the outputs from other parts of the dataset, but the creation of such tree- and graph-based schedulers is nontrivial and is an active research area in computer science.

Early on in my project, I evaluated several tools that could enable the efficient and timely analysis of the 15x dataset (and also wrote a simple scheduler of my own). Ultimately, I decided to use the Hail [77] toolkit for the analysis. Hail is a 'data analysis tool with additional data types and methods for working with genomic data'. It was previously used for the all-phenotype GWAS of the UK Biobank (4,200 traits across 360,00 genotyped individuals) [127] and is currently used to produce releases of the gnomAD database (125,748 exomes and 15,708 genomes) [95]. Hail uses Apache Spark, a distributed cluster manager that builds upon the principles of MapReduce.

Unfortunately, the deployment of Hail on the internal cluster was a nontrivial task. While the actual deployment was led by the Human Genome Informatics team, as one of the early adopters (starting with the QC of the PANTS cohort) I was involved in identifying and trying to fix numerous stability and performance issues. We are still experiencing hardware-related

problems, which have limited some of the analyses (e.g., the full genome-wide logistic regressed), but the current deployment has facilitated the analyses I describe below.

The majority of the analytic pipeline for IBD 15x was written in Python, using the Hail 0.2 framework and a variety of analytical packages (`scikit-learn`, `statsmodels`, `pandas`, etc.). At different stages, it was executed across 200 to 1,350 CPU cores on the Sanger Institute's OpenStack cluster.

4.2.6 Dataset overview and pre-processing

Autosomal chromosomes (1–22) were converted into the MatrixTable format used by Hail (19,371 samples, 205,889,702 variants). Multiallelic variants were split into separate records (226,027,757 variants). Standard variant and sample quality control metrics were calculated to facilitate further filtering.

The samples were sequenced across several batches. Batch-specific features are highlighted in bold.

- IBD Phase 1 (1,427 samples, cases, PCR-free sequencing)
- IBD Phase 2 (3,060 samples, cases, PCR-free sequencing)
- IBD Phase 3 (2,530 samples, cases, PCR-free sequencing **with dual indexing**)
- INTERVAL Phase 1 (5,093 samples, controls, **PCR sequencing**)
- INTERVAL Phase 2 (5,570 samples, controls, PCR-free sequencing)
- INTERVAL Phase 3 (1,691 samples, controls, PCR-free sequencing)

4.2.7 Sample quality control

Inclusion of low-quality and outlier samples may negatively influence the results of the association study. The main goal of sample QC is to identify a set of samples that have similar high quality metrics, belong to the same ancestry group, and are not strongly related to each other.

Each of these steps helps to prevent biases that can cause spurious associations or reduce the power to detect the true ones: poorly genotyped samples are likely to contain systematic bias across many sites; due to genetic drift, ancestry outliers differ in frequency of certain common and rate variants; related samples will influence the significance of variants present in the related individuals. Below, I describe a series the QC steps that were carried out for the IBD 15x study.

Hard filters

A number of hard filters were applied to exclude low quality samples:

- Median FREEMIX across the read groups > 2% (143 samples)
- Mean depth < 12x (128 outliers)
- Call rate < 95% (9 outliers)
- Chimerism rate > 5% (38 outliers, estimated via Genome STRiP 2.0)

A total of 315 samples were excluded at this stage.

All filters were applied simultaneously, meaning that, for example, a sample with high FREEMIX and low call rate will appear on the list twice. This also applies to the distribution-based filters described in next section.

Overall, the QC metric filtering parameters and thresholds were inspired by the the filtering done to create the gnomAD database [95]. Some of the thresholds were adapted to reflect the data in our cohort (e.g., minimal depth lowered from 15x to 12x, as 15x was our target sequencing depth). It should be noted that some of the filters may be refined in the future: while gnomAD is currently the largest genetic variation database that includes hundreds of thousands of WES and tens of thousands of WGS samples, it is not used as a basis of case-control studies.

Distribution-based filters

A two-stage approach was applied. Firstly, I removed samples that were outliers within individual batches (e.g. IBD Phase 1, Interval Phase 3). In addition, it was observed that distribution of several metrics (e.g., number of INDELs) were different for samples that were sequenced using PCR (INTERVAL Phase 1) and PCR-free protocols (all other cohorts). I have repeated the outlier removal protocol grouping the samples by sequencing protocol (PCR versus PCR-free).

A sample was excluded if the value of the QC metric was four median absolute derivations (MAD) higher or lower than the median in the batch or sequencing protocol. The metrics used at this filtering stage were:

- Number of SNPs called
- Number of insertions called
- Number of deletions called
- Insertion-deletion ratio
- Transition-transversion ratio (Ti/Tv)
- Heterozygous-homozygous ratio (heterozygosity rate)
- Call rate

306 samples did not pass the distribution-based filters. The majority of outlier samples were outside the acceptable range for several metrics (Figure 4.3), indicating that the selected metrics and the applied thresholds were not needlessly excluding high-quality samples. A total of 621 samples were excluded during both stages of sample filtering (hard filters and distribution filters). The samples passing the QC criteria were brought forward for further analysis.

Identifying batch effects via metric-based PCA

Sample QC metric-based principal component analysis was used to verify the absence of hidden batch effects that may have been introduced during sequencing. The assumption was

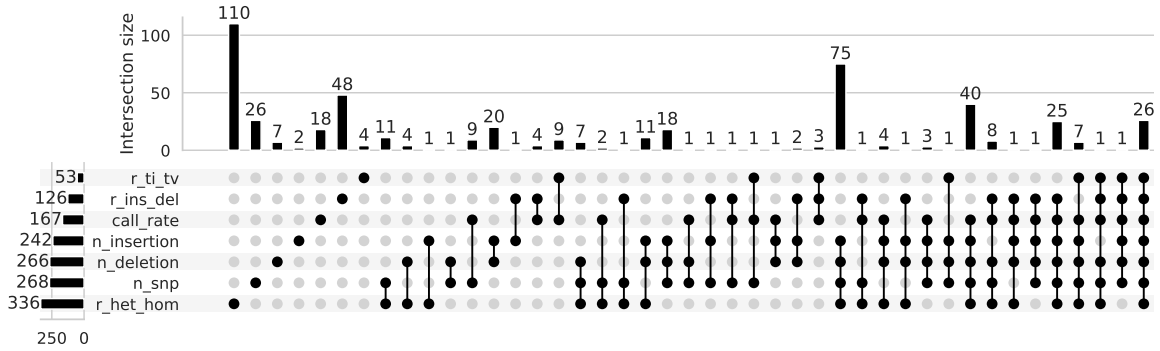


Figure 4.3 Set intersection of the samples failing different QC metric thresholds in the 15x dataset. Left bars – number of samples failing individual categories. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First seven columns (single dots) – number of samples failing only one QC metric.

that any substantial difference in the sequencing protocol for a set of samples would lead to separation of these samples on the PCs.

The sample QC metrics (mentioned in the previous two sections) were normalised and used to calculate ten principal components. Principal component eigenvalues showed that only the first 2–3 explained any substantial variance (PC1 – 5.80, PC2 – 3.06, PC3 – 1.00, PC4 – 0.43).

PC1 clearly separated samples between PCR (INTERVAL Phase 1) and PCR-free batches (all other). Inspecting the PC loadings (weights), the separation was almost equally driven by all considered QC metrics. All other PCs did not reveal any substantial clustering, suggesting that there were no major hidden batch effects that I was not aware of. PC2 was driven by depth and the number of called variants (SNPs and INDELs). PC3 was driven almost entirely by FREEMIX. PC4 was driven by depth and the number of called SNPs (Figure 4.4).

One of my concerns was that the PCA is performed on the same set of QC metrics that are used for filtering (i.e., the analysis is circular). However, when the PCs were built on the full set of the available QC metrics (e.g., adding the number of singletons and star alleles) the results were virtually the same.

This analysis does not substitute the genotype-based PCA that will be discussed below.

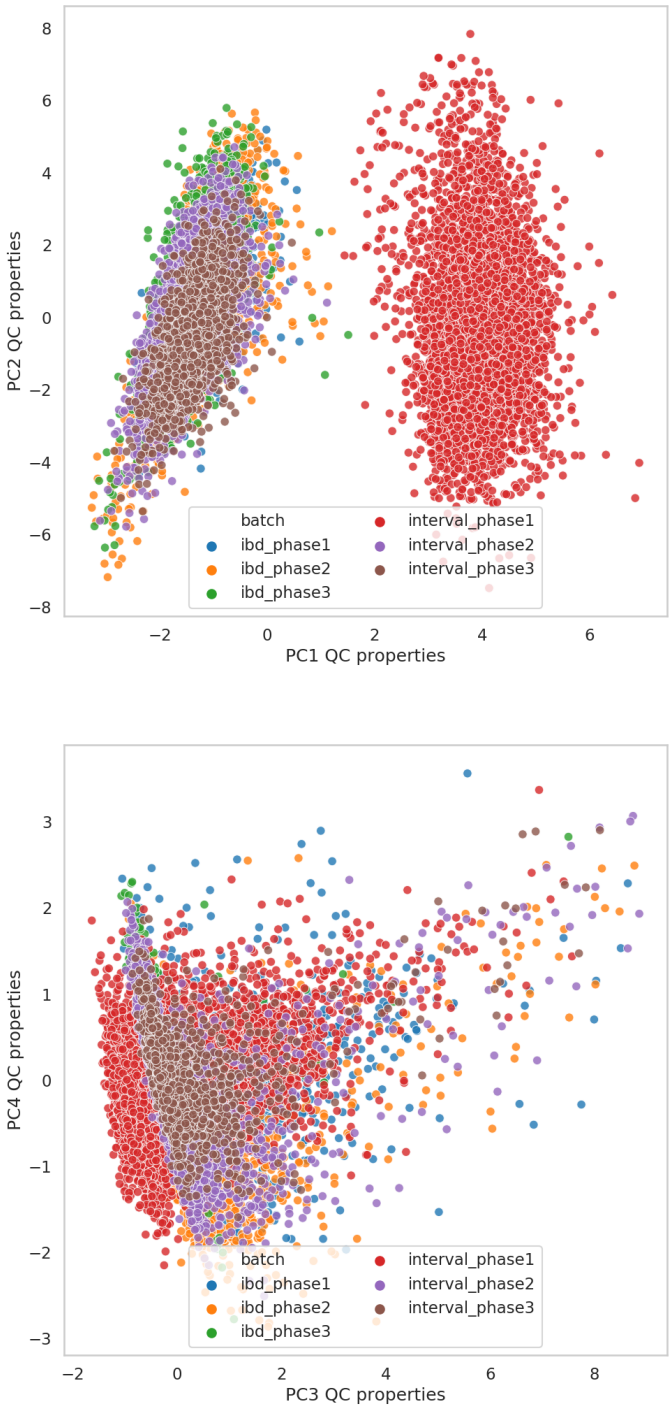


Figure 4.4 The first four principal components built based on the QC metrics of the samples in the 15x cohort. Samples coloured by sequencing phase. PC1 clearly separates samples sequenced with PCR and PCR-free library preparation protocols.

Identifying genetic ancestry outliers via 1000G PCA loading projection

Cryptic population structure may inflate the results of association analyses, especially for rare variants where the frequency of many genetic variants may vary drastically or even be entirely exclusive to a certain population. Population structure can be partially accounted for via statistical methods (e.g., PCA or generalised linear model-based methods). Alternatively, it is possible to analyse each population separately, combining the results in a trans-ancestry meta-analysis. However, these techniques require a semi-balanced distribution of each population group between cases and controls.

The majority of subjects in the 15x cohort have self-identified to be of European descent. However, self-reported ancestry is often discordant from the genomic ancestry and is insufficient for identifying cryptic population structure [123]. The PCA weight projection technique was used to estimate the genetic ancestry of the individuals in the 15x cohort.

The 1000G Project cohort includes samples from 2,504 individuals from 26 populations around the world. Principal component analysis of the 1000G cohorts reveals the complexity of the global population structure. Members of the same population or population group (e.g., South East Asian, European) cluster closely together and diverge from other clusters on the first few principal components.

The 15x dataset was filtered down to a subset of high-quality common variants that are in low linkage disequilibrium and have pre-computed population-scale loadings from 1000G. The variant set (N=17,535) was derived by the authors of the AKT package [9] and consists of common genetic variants (>5% MAF in 1000G), and is limited to balletic SNPs that are known to be present on several genotyping arrays and have been shown to be consistently called by different variant calling pipelines.

In order to estimate the genetic ancestry of each individual in the IBD and INTERVAL 15x datasets, I projected the samples onto the 1000G data. I obtained the 1000G principal component loadings for the AKT 'high confidence' variant set described above. The samples were projected, accounting for the heterozygosity and the allele frequency in 1000G.

The first ten PC projections were then inspected manually (Figure 4.5). As expected, considering the sample selection criteria, the absolute majority of the 15x samples clustered together with the European population group in 1000G. A small number of the IBD samples clustered closely with non-European population groups or were positioned between the major

clusters, suggesting admixture. Given the lack of INTERVAL samples of a similar ancestry, they had to be excluded. In addition, some samples were marginally outside the 'edge' of the European ancestry cluster, suggesting presence of admixture.

In order to identify the non-European samples within the cohort, the following procedure was followed. The 1000G cohort was subsetted to contain only individuals of European ancestry (EUR population on Figure 4.5.) For each of the ten PCs, the median and the median absolute deviation (MAD) of the 1000G European population's (N=503) PC scores were calculated. Next, the 15x samples with PC scores outside the three MAD from the median were identified. A total of 434 15x outliers were identified.

All but seven outliers were removed due to being outside three MAD in the first four PCs of 1000G (Figure 4.6). This is expected, as the PCs are ranked in terms of the explained variance (i.e., earlier PCs explain more variance and separate more genetically divergent populations).

Samples that passed the filter clustered together with the European population of 1000G (Figure 4.7). In addition, comparison of PC distributions of IBD and INTERVAL PC scores did not indicate any major shifts in the distributions, suggesting that cases and controls were well-mixed (Figure 4.8). In total, 434 samples were excluded from subsequent analyses, leaving 18,940 samples in the dataset.

Removal of duplicated and related samples

The kinship estimation technique was used to identify closely-related individuals and duplicated samples in the cohort. The kinship estimation was performed on the same 17,000 variant subset as the 1000G PCA projections. Inclusion of relatives and duplicates may bias the results of the association tests. While techniques for correcting for familial structure exist (typically based on linear mixed models, see [60]), the cohort did not have enough related individuals to justify this increase in the analysis complexity. Using the Hail implementation of the kinship estimation technique first devised for the PLINK software, 254 sample pairs with PI-HAT > 0.1 were identified. This is lower than the 0.185 exclusion threshold often used for GWAS (middle point between second- and third-degree relatives, 186 pairs), but the enrichment of distantly related individuals might cause spurious rare variant associations. In

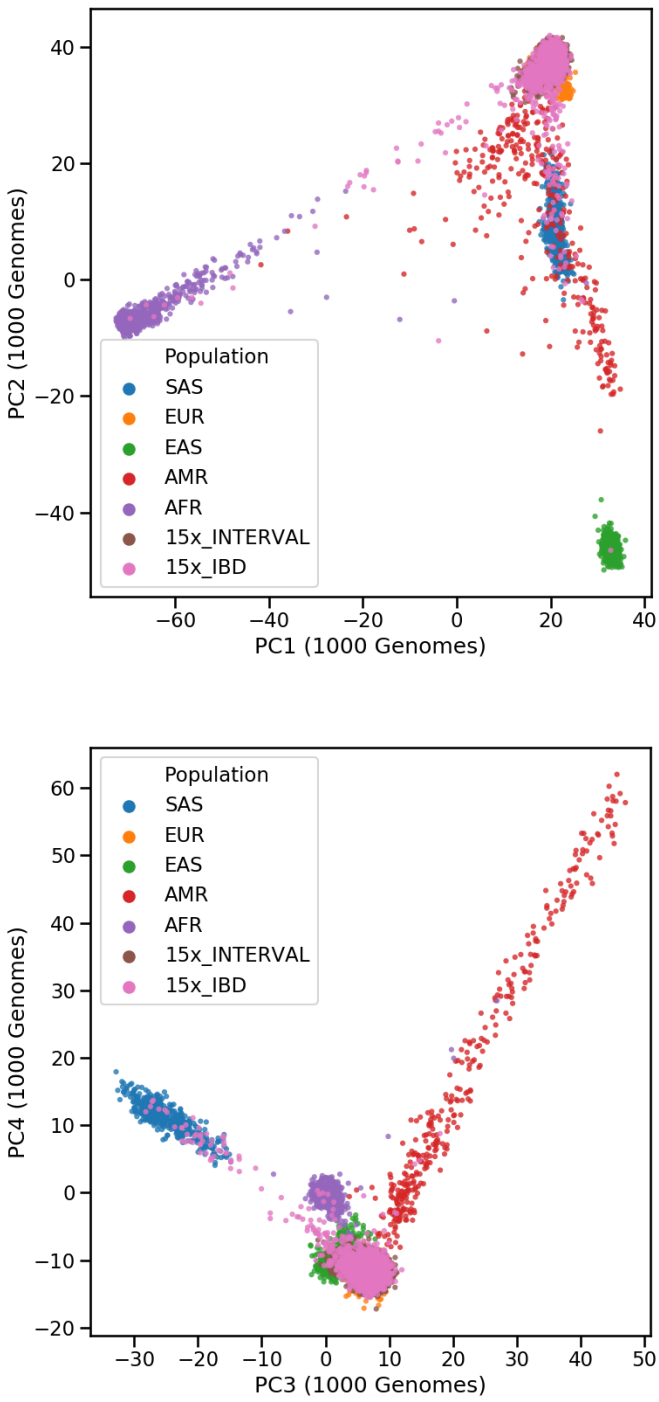


Figure 4.5 The first four principal components of the 1000G cohort, alongside the 15x cohort samples projected onto them. The absolute majority of the 15x cluster together with the European population of 1000G. The majority of the outliers were IBD samples.

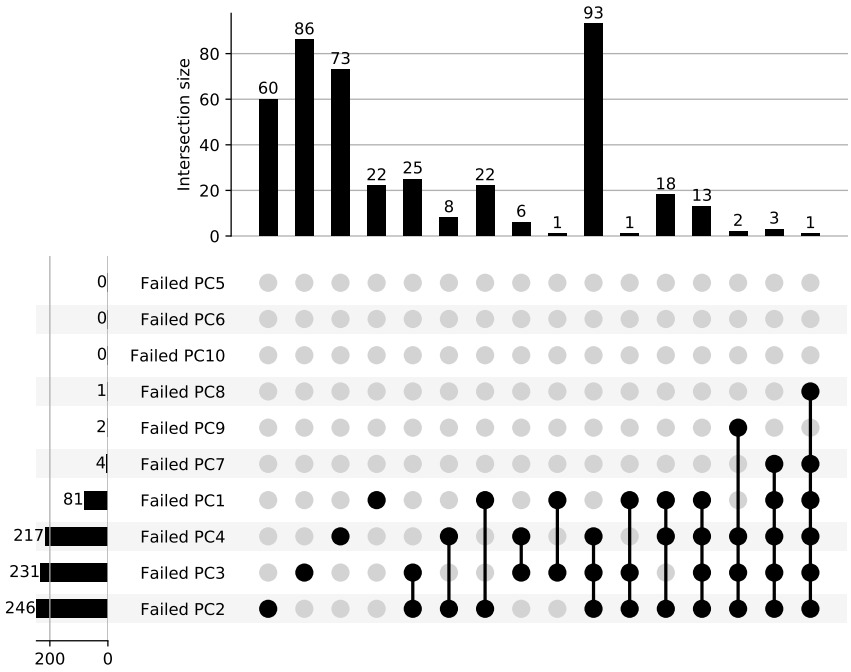


Figure 4.6 Set overlap plot for the 434 genetic ancestry outlier samples from 15x. Left bars – number of samples failing individual ancestry PC filters. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First six bars (single dots) – number of samples failing only one ancestry PC filter.

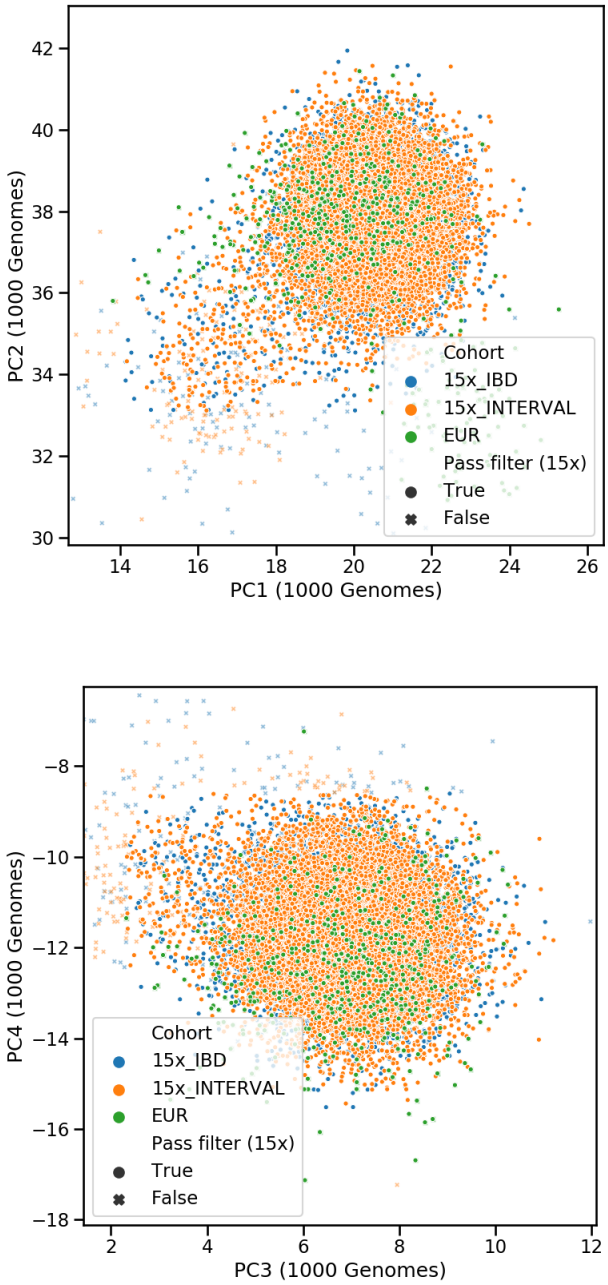


Figure 4.7 The first four principal components of the European subset of the 1000G cohort, alongside the 15x cohort samples projected onto them (zoomed in, discarding distant outliers). Blue (IBD) and orange (INTERVAL) dots pass the MAD-based filter. Opaque X's are too far removed from the 1000G EUR median and fail the filter.

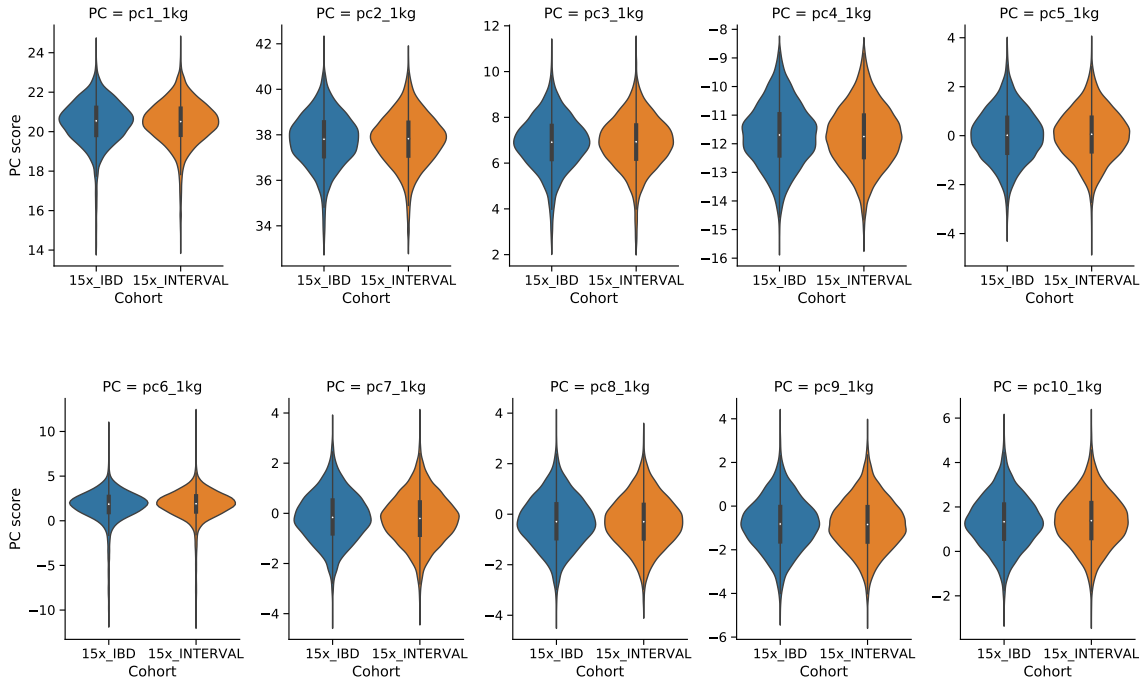


Figure 4.8 Violin plot comparing the distributions of the projected 1000G PC scores for IBD and INTERVAL samples that pass the ancestry filters. No major distribution differences across the first ten principal components are present, suggesting a good ancestry mixture between cases and controls in the 15x study.

total, 113 cases and 129 controls were removed, keeping a total of 18,165 individuals in the cohort.

Within-cohort principal component analysis

After removing the samples that failed the quality control procedures, related samples, and those outside the European population cluster, a Hardy-Weinberg-normalised principal component analysis was performed on the IBD and INTERVAL cohorts.

The 17,000 SNP subset was further filtered to include only high-quality common genetic variants (call rate $> 99\%$, $P_{HWE} > 1x10^{-10}$) and LD-pruned ($r^2 = 0.2$). In addition, regions with high LD and those harbouring known IBD associations (+/- 500 kb) were removed, retaining a total of 14,617 variants.

Ten first principal components were calculated. Principal component eigenvalues demonstrated that the first PC explained almost double the variance of PC2 (12.6 versus 6.31). Principal components 3 to 10 all had similar eigenvalues between 5.37 and 5.72.

Manual inspection of the principal component plots indicated that the samples were reasonably-well mixed between cases and controls (Figure 4.9 shows the first four PCs). A few hundred outlier samples were visible on PC1, however they did not appear to cluster with any specific sample QC metric, sequencing batch or a handful of variants that would be driving the separation.

A median absolute deviation filter, similar to the one described in the 1000G PCA section above, was applied. The MAD distance threshold was increased to four, the median and the MAD were calculated from the within-cohort PC scores rather than 1KG EUR samples. A total of 892 samples failed this filter, with the majority falling outside the MAD thresholds on PC1 and PC2. The majority of failed samples did not overlap between different PCs (Figure 4.10).

It is not entirely clear whether such a substantial number of samples should be excluded from further association studies. Firstly, the calculated principal component scores will be used as covariates for the genome-wide logistic regressions, correcting for some of the cohort heterogeneity. Secondly, I have identified that the variation on PC1 is almost entirely driven by the Southern European ancestry of some of the individuals in the cohort, captured by PC6

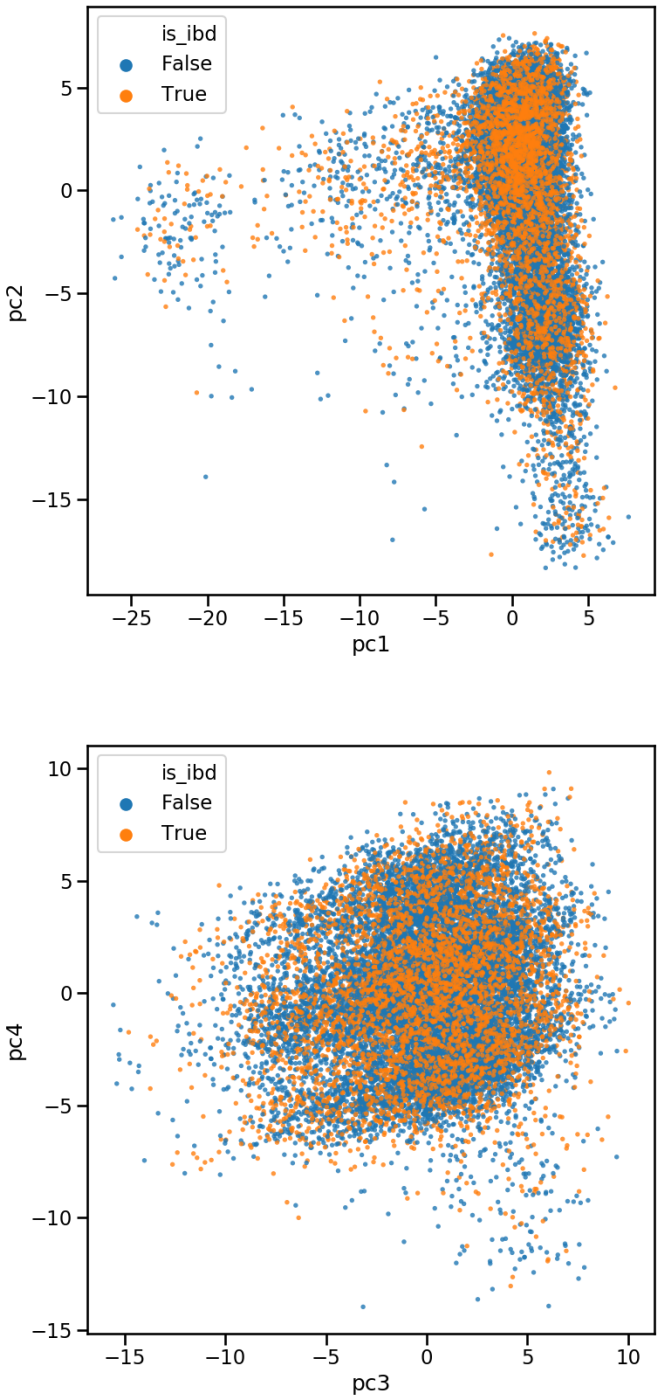


Figure 4.9 The first four principal components of 15x cohort samples. Orange dots – IBD samples, blue dots – INTERVAL samples.

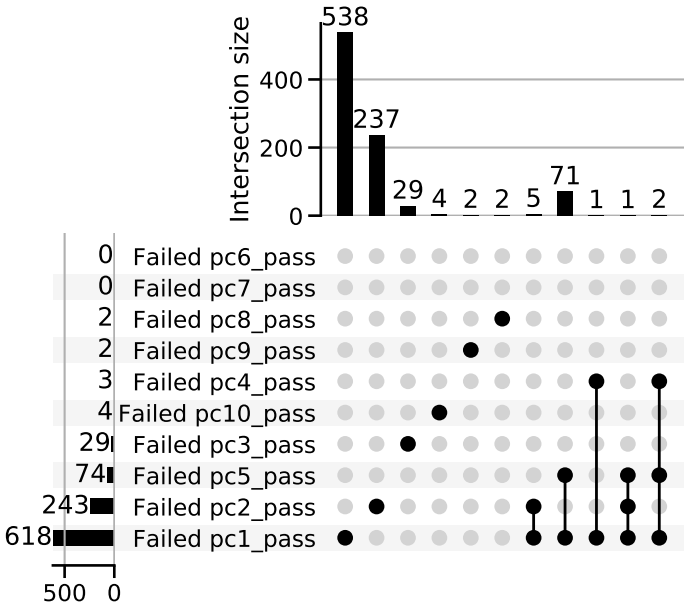


Figure 4.10 Set intersection plot of the samples failing different within-cohort PC filters. Left bars – number of samples failing individual PC filters. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First six bars (single dots) – number of samples failing only one PC.

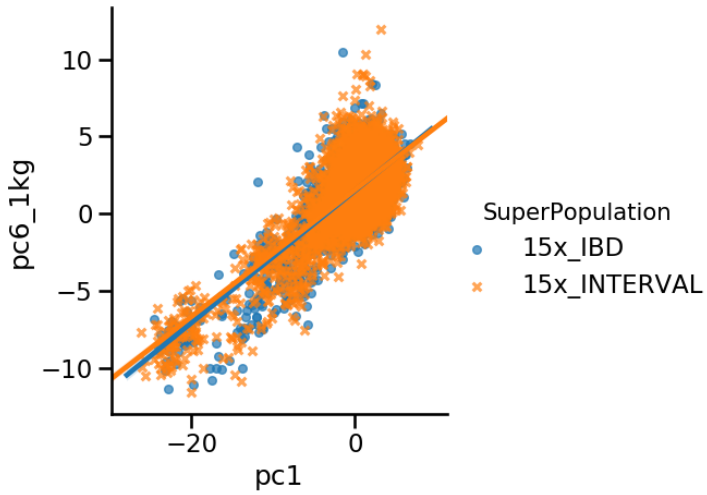


Figure 4.11 PC1 scores correlate (0.7, Pearson) with the $PC6_{1000G}$ scores from the 1000G projection PCA. This suggests that the PC1 outliers are driven by Southern European genetic ancestry.

of the 1000G analysis (Iberian and Tuscan cohorts) (Figure 4.11). Distribution of PC1 scores matched closely for IBD and INTERVAL samples, suggesting a good mixture of cases and controls (Figure 4.12). I could not identify the source of variance that drives the PC2 outliers, but, once again, the distributions of the PC scores between IBD and INTERVAL samples were very similar (Figure 4.12). PC2, overall, does not explain a lot of variance and can be corrected via covariates. I did not observe any substantial decrease in the p-value inflation in case-control association tests when the outlier samples were excluded, rather than corrected for.

Unfortunately, I was unable to calculate the PCs for a larger set of variants and including lower frequency variation (MAF $>$ 1%, rather than 5%) due to some technical issues with the cluster. It is curious that the PCR versus PCR-free separation present during the QC score-based PCA was not observed here. It is possible that calculating the the PCs based on a small and well-genotyped subset of SNPs masked the variation between the two sequencing protocols. I am planning to repeat the within-cohort PCA analyses immediately once the technical issues are resolved or a workaround is found.

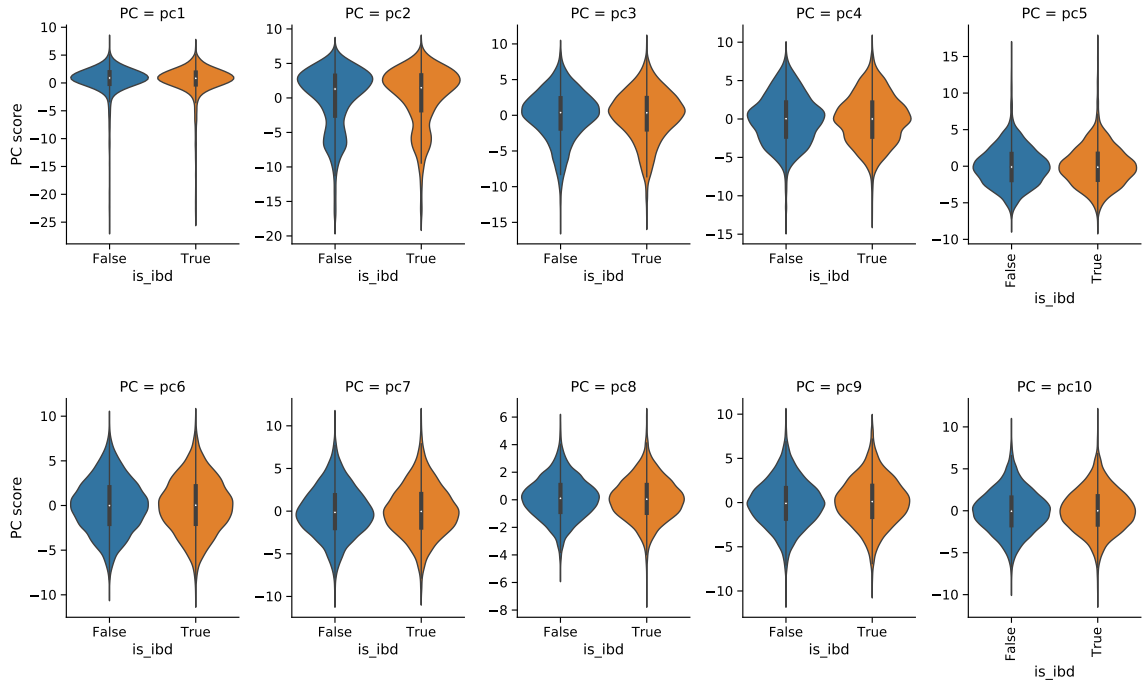


Figure 4.12 Violin plot comparing the distributions of the projected PC scores for IBD and INTERVAL (without outlier removal). No major distribution differences across the first ten principal components are present, suggesting a good case-control mixture.

4.2.8 Variant and site quality control

In addition to the manual variant filtering (described in individual sections above), I have used the VQSR method to identify poorly genotyped sites. When used in conjunction with manual filtering, VQSR does not have much effect on the common variants (the absolute majority of which pass this filter). However, the VQSR filter will be used in the future genome-wide rare variant association tests.[3]

VQSLOD (variant quality score log-odds) scores were calculated for the final set of genotype calls (separately, for SNPs and INDELs) via the VQSR method. VQSR is a machine learning based method that uses Gaussian mixture models to estimate the likelihood of a variant being true. It requires a set of variants that are considered to be 'real' (e.g., taken from a high-quality population sequencing study, such as the 1000 Genomes). VQSR then builds a model based on the distribution of the quality scores of these variants (e.g., depth at site, mapping quality rank sum, strand odds ratio). The model is then used to score the full set of variants, evaluating how close their quality scores are to the scores of the true variant set.

The final stage requires selecting the VQSLOD score cutoff. This is done using the tranche sensitivity (e.g., a VQSLOD score of 5.0 leads to the detection of 99.50% of true positive variants, compared to the VQSLOD of 3.0 that allows the detection of 99.99% of true variants[4]). For SNPs, one usually looks at how the chosen tranche influences the transition-transversion ratio (Ti/Tv), which is expected to be $\sim$2.0–2.1 for whole-genome datasets and $\sim$3.0–3.3 for whole-exome datasets (the latter will vary depending on the exome capture kit used). A good practice is to choose the highest tranche where the Ti/Tv is close to the target value, yet does not strongly differ from the Ti/Tv of the previous value (i.e., is at the rightmost side of the distribution plateau). For INDELs, the threshold choice cannot be motivated by the Ti/Tv ratio and is therefore done based on the number of novel variants each extra tranche brings (e.g., if going from the 99.8% tranche to the 99.9% tranche brings an increase of 80% percent of novel INDELs, one should be cautious about false-positives and consider picking 99.8%).

[3] Parameters and the follow up analysis performed by me. Computation set up by Allan Daly due to the availability of computing resources.

[4] VQSLOD to truth sensitivity scores mappings are provided as an example, real mappings will vary across different datasets

VQSR was performed with the parameters described in the Broad Institute's 'generic germline short variant joint genotyping' pipeline. Briefly, for SNPs I fitted 6 Gaussians and used the following fields to train the model:

- QD – QUAL score normalised by allele depth
- MQRankSum – rank sum test for mapping qualities of reads supporting REF versus reads supporting ALT
- ReadPosRankSum – rank sum test for position within reads supporting REF versus position within reads supporting ALT
- FS – Fisher's exact test for strand bias
- MQ – mapping quality
- SOR – symmetric odds ratio test
- DP – depth

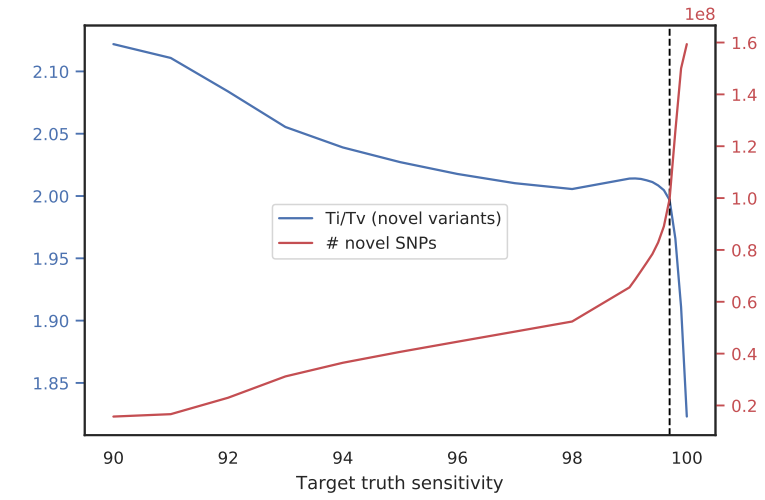
For INDELs, I fitted 4 Gaussians (reduced from 6 due to a smaller set of variants and the danger of overfitting) and used the following annotations: FS, ReadPosRankSum, MQRankSum, QD, SOR, DP.

The following resource sets were used for training:

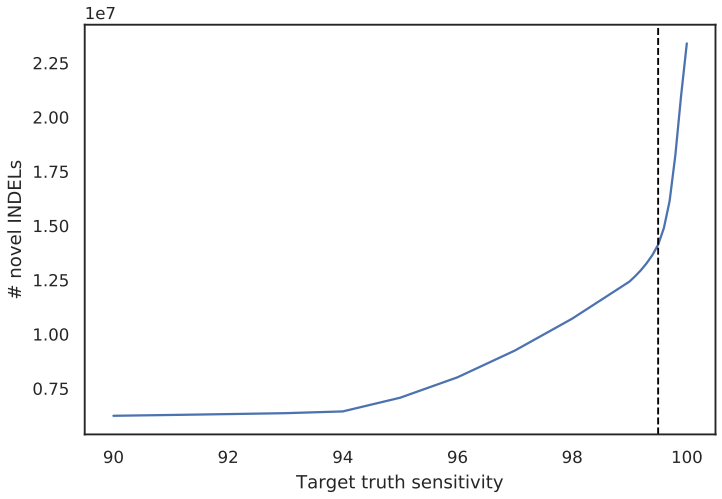
SNP True sites training resource: HapMap, Omni. Non-true sites training resource: 1000G. Known sites resource, not used in training: dbSNP.

INDEL True sites training resource: Mills. Non-true sites training resource: axiomPoly. Known sites resource, not used in training: dbSNP.

All training and testing resources were obtained from the GATK Resource Bundle. The outputs from the VQSR calibration are shown on Figure 4.13.



(a)



(b)

Figure 4.13 VQSR calibration results for IBD & INTERVAL 15x cohorts. a) Target truth sensitivity (x-axis) influences the number of novel SNP variants (blue curve and left y-axis) and their transition/transversion ratio (red curve and right y-axis). 99.7% was selected as the tranche for further filtering (i.e, in our filtered call set 99.7% of the overlapping sites present in the truth set can be detected). The tranche was selected based on the point where the Ti/Tv curve overlaps with the number of novel sites curve in order to maximise the true-positive variant. I have also verified that the Ti/Tv ratio for the tranche ($\text{Ti/Tv}_{novel} = 1.9$) closely matches the expected ∼2.0–2.1. b) Target truth sensitivity (y-axis) influences the number of novel INDEL variants. 99.5% tranche was selected for further filtering.

4.3 Results

4.3.1 Evaluating the sequencing depth and the statistical power trade-off for WGS association studies

Optimal sequencing depth for case-control experiments

I simulated power to detect rare SNP variants present in 0.25% of cases and 0.05% (OR = 5) given two experimental scenarios: unlimited budget, fixed number of samples (25,000 cases and 25,000 controls); limited budget (sufficient for sequencing 50,000 samples at 30x, 1:1 case-control ratio) with an unlimited number of cases and controls to choose from. The simulation takes into account the sensitivity of the variant calling presented in Figure 4.2 and a realistic cost estimate (cost per 'x' * depth + fixed cost per sample)[5]. Reflecting the pricing back in 2017, the cost per 'x' of depth (9Gbs) was set to 10.8; fixed cost per sample (e.g., library prep, labour) was set to 18.84.

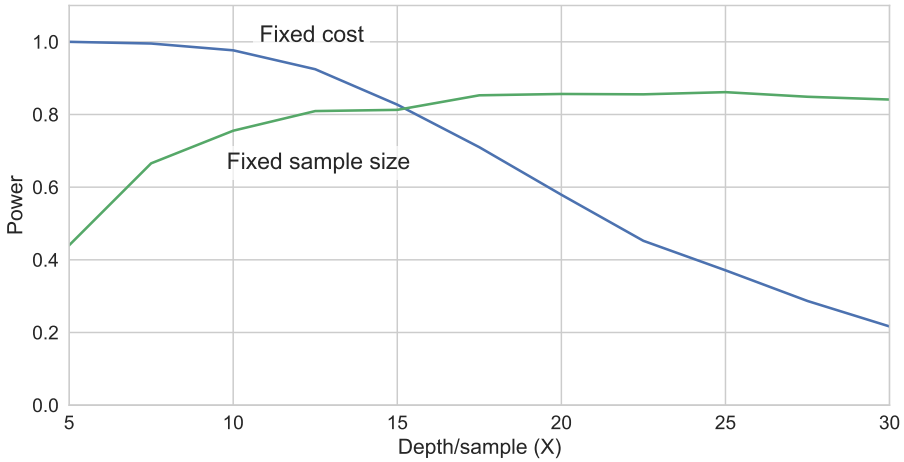


Figure 4.14 Power to discover rare SNP variants in a case-control experiment setup. Blue line simulates the scenario with a limited budget and an unlimited number of available samples, while the green line shows the unlimited budget/limited cohort scenario.

[5]Original simulation by Dr Jeff Barrett. Re-implementation and extension to support quantitative trait simulations by the author. Updated sensitivity and cost values.

In the fixed sample size scenario, the power plateaus around the 15x sequencing depth. In the fixed cost scenario, the maximal power is achieved when sequencing the largest number of samples at a minimal depth.

Realistically, the IBD 15x project was constrained by both the number of available samples and the budget. For the fixed sample size scenario, the power plateaus around the 15x-17.5x mark, which is close to the median depth of the IBD 15x and the INTERVAL datasets. Sequencing at an even lower depth (say, 10x) might have increased the power to detect SNP associations, but would have led to a severe reduction in the INDEL sensitivity and hindered future projects like structural variant calling.

Choice of sequencing depth for biobank-scale projects

I performed simulations for national biobank-scale studies (e.g. the UK Biobank, which has enrolled 500,000 participants) to see whether 15x remains the optimal sequencing depth. Statistical power to discover associations using burden tests, for a set of rare variants (Figure 4.15) at different biobank sizes (50,000 to 500,000) was estimated.

In 2018, plans to sequence the first 50,000 individuals from the UK Biobank cohort at 30x depth were announced [172] (the 'Vanguard' project). I used this model to evaluate the power to discover rare variant associations in Vanguard versus the full UK Biobank (Figure 4.16). In September 2019, plans to whole-genome sequence the whole UK Biobank were announced.

The power to detect rare variants is driven by sample size, rather than by sequencing depth. For all study sizes, except n=500,000, the power plateaus around 15x. For a (realistic) scenario, where the budget for sequencing is fixed (dotted blue line: budget sufficient for sequencing 50,000 samples at 30x), sequencing more samples is preferable to sequencing at a higher depth.

Overall, although in retrospect, I believe that sequencing at 15x depth was the right design choice for the Crohn's whole-genome sequencing association study. Considering the limited budget and the limited number of available samples, it maximised the ability to detect rare variants. For SNPs and INDELs, the sensitivity benefit of sequencing samples at a higher depth is modest, while the cost would grow semi-linearly, thus reducing the cohort size and the overall power for association studies.

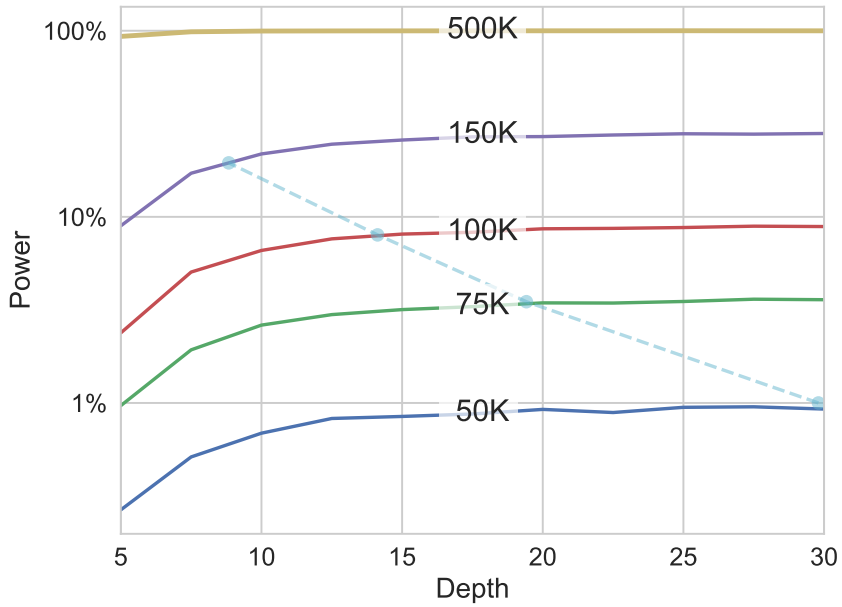


Figure 4.15 Power to discover associations using burden tests for a set of rare variants with a cumulative frequency of $5 * 10^{-4}$ in a gene, assuming $\beta = 0.5$ s.d. and $\alpha = 1x10^{-6}$. The blue dotted line shows the trade-off between depth, cost and power: sequencing 50,000 samples at 30x would result in power around 1%, while sequencing 100,000 at around 14x would provide 8% power (while keeping the cost the same, taking into account fixed costs and cost per depth).

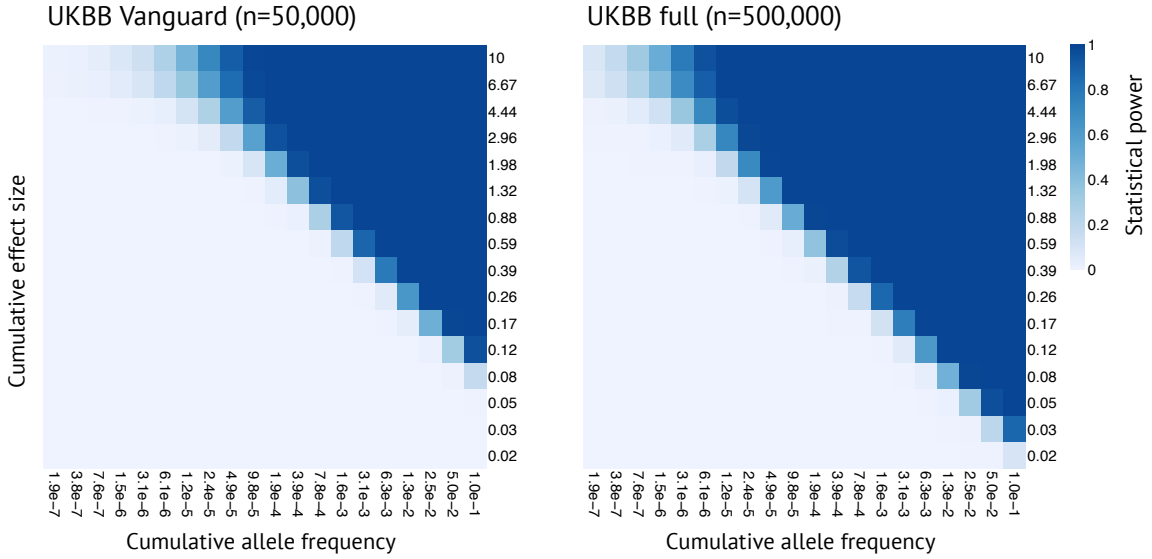


Figure 4.16 Power to discover associations by aggregated rare variants in the UK Biobank (α = $1x10^{-9}$). Pilot release ('Vanguard' project, left, n=50,000): near perfect power to discover associations for variants with a cumulative β greater than 0.6 SD and frequency of 6 : 1,000. Full UKBB WGS (right, n=500,000): near perfect power to discover associations for variants with a cumulative β greater than 0.6 SD and frequency of 4 : 10,000. Simulation by me, plot refined by Dr Klaudia Walter.

However, this simulation has several drawbacks. Lower sequencing depth leads to a higher error rate and requires much more stringent QC in order to avoid false associations (e.g., [116]). The ability to detect INDELs and CNVs, which require higher sequencing depth for accurate genotyping, should also be considered. For high-quality CNV genotyping, other techniques like long-read PacBio or Nanopore sequencing may be more appropriate, and the simulation is not currently suited for estimating the statistical power for those sequencing types.

The conclusions of my simulations match those from the work of Rashkin et al. [152], who conclude 15–20x to be the optimal depth for studies of rare variants in complex disease.

4.3.2 Index misassignment impacts multiplexed sequencing

Multiplexing allows simultaneous sequencing of several libraries during the same sequencing run. This is achieved by adding unique index sequences ('tags', 'barcodes') to DNA fragments during the library preparation stage. Multiplexing is routinely used in multi-sample studies to increase the throughput, reduce the expenses on the reagents, and, in theory, to increase the quality of the data via read group averaging. Multiplexing adds an additional level of complexity to the sequencing process, as the individual reads have to be computationally assigned to the correct target sample (demultiplexed). Reads for the same sample obtained from a single sequencing run are called a read group.

In certain cases, indexes get misassigned to the wrong read or multiple conflicting indices get attached to the same read ('chimeric' indices). The index misassignment is sometimes referred to as 'index hopping'. Reads with misassigned indexes ultimately result in low-level cross-contamination of the samples and reduce the quality of the variant calling. The exact rate of index hopping is hard to measure, as it depends on the experiment type, library preparation protocol, multiplexing factor, and other variables. The manufacturer reports the expected rates to be 1–2% [83], while some independent studies have claimed to observe index hopping rates $\sim$10%. While some protocols are thought to reduce the index hopping rate [83], it remains present in all current multiplexed studies.

I have attempted to quantify the index hopping rate in the IBD and INTERVAL 15x cohorts. To quantify the index hopping rate and, ultimately, sample cross-contamination the FREEMIX metric produced by the VerifyBamID tool [195] was used. Authors suggest

interpreting samples with FREEMIX $>$ 2–3% as potentially contaminated. The metric is often used as a sample quality control metric to exclude samples with a high level of DNA contamination, as it leads to poor genotyping quality. However, the exact threshold varies across studies. The UK10K consortium excluded samples with FREEMIX $>$ 3% [180]. The gnomAD genome aggregation database excludes whole-genome and whole-exome samples with FREEMIX $>$ 5% from the variant callset during the sample quality control stage.

I was particularly interested in batch-specific variations in contamination, given that INTERVAL Phase 1 was processed using a PCR library prep protocol (thought to lead to lower index missassignment rates [83]) and IBD Phase 3 used dual indexing. Dual indexing assigns a unique combination of indices at both ends of the read, therefore reducing the chances of read missasignment during demultiplexing (reads get discarded if the indexes mismatch). Dual indexing is planned to be used for a variety of future sequencing studies at the Sanger Institute (e.g., the IBD WES project), therefore it was important to verify whether it in fact leads to an increased misassignment rate.

I have calculated the FREEMIX (FM) per each read group (N=111,225) rather than per sample, as the rate of missasignment varies between the sequencing runs (Figure 4.17). A two-sample Kolmogorov-Smirnov test (KS) was used to evaluate whether the FM scores for two 15x batches follow the same distribution. Overall, the median FM score across all read groups was moderately low: 0.55%. 3.7% of read groups had a $FM > 2\%$ (0.36% read groups above the critically high 5%). Two of the earliest sequencing batches had the highest mean FM: INTERVAL Phase 1 = 0.96% (PCR) and IBD Phase 1 = 0.82% (PCR-free), suggesting that, on its own, the PCR-free sequencing did not negatively impact the contamination rate even at the early stages of the project (KS statistic = 0.08; p=4.59×10^{-41}). IBD Phase 3, which utilised dual indexing, had the lowest mean FM score across all batches – 0.07% and was lower than the FM in INTERVAL Phase 3 – 0.48%, which was sequenced around the same time, but utilising regular single indexing (KS statistic = -102.66; p$<1.80\times10^{-308}$). Considering the mean FM for each sample, I have identified 61 samples with $FM > 5\%$ (gnomAD threshold, used as a step in sample QC) (Figure 4.18).

I estimated the effects of index missasignment during multiplexed sequencing across two library preparation protocols (PCR versus PCR-free) and two indexing techniques (single versus dual indexing) using the FREEMIX metric. Overall, the findings suggest that only a small fraction of read groups (3.7%) had a contamination level $>$ 2%. I did not find any evidence to the manufacturer's claim that PCR-free library preparation leads

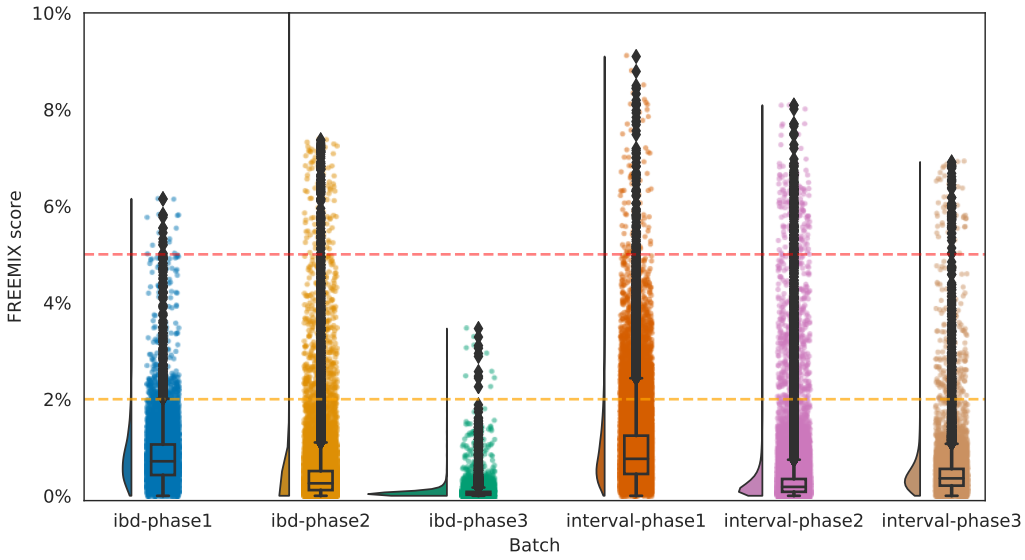


Figure 4.17 FREEMIX scores of 111,225 read groups of samples from the IBD and INTERVAL 15x (median of 6 read groups per sample). Read groups are categorised by project batch. The box plots show the median levels of contamination per batch. Scatter plots and and density plots indicate the distribution of the scores. Orange dotted line the shows level of FREEMIX (2%) that indicates potential contamination. Red dotted line shows level of FREEMIX (5%) that indicates strong contamination.

to higher index hopping [83], though there was only one batch of samples processed with PCR-including protocol. Dual-indexing appears to reduce the level of contamination by an order of magnitude and should be considered to be used in future studies. I acknowledge that FREEMIX might not be an ideal marker of index misassignment, as it will also capture sample contamination (e.g., during handling). I also noticed that the distribution of FREEMIX was not identical (although the shift was very small) between batches that followed the same sequencing protocol (INTERVAL Phase 2 versus INTERVAL Phase 3 – KS statistic = 0.30; $p<1.80\times10^{-308}$). My initial consultations with the pipelines team did not identify any cause for this.

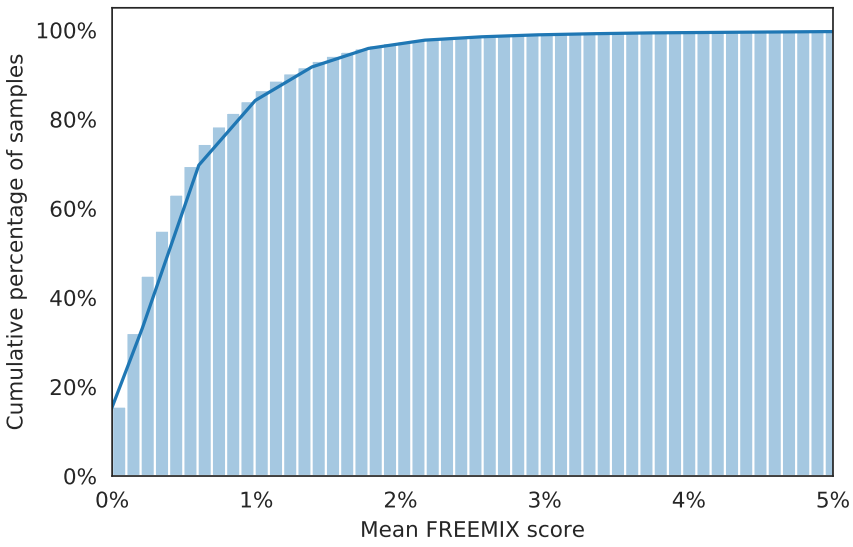


Figure 4.18 Mean FREEMIX for each sample in the INTERVAL and IBD samples. The absolute majority of the samples (96%) have a FREEMIX below the 'potential contamination' level of 2%.

4.3.3 Estimating the impact of the covariates on the power to detect associations in a case-control setting

Next, I estimated the impact of including batch and principal component covariates on the power to detect known Crohn's disease associations. I derived a list of 105 independent variants associated with CD in de Lange et al. [49]. While the variants pass the genome-wide significance threshold in those two studies, in 15x, the majority will have a much higher p-value due to the smaller number of both cases and controls. 253 known UC samples were excluded, 17,912 samples were retained. Logistic regression using the Firth test was performed to estimate the p-values. A variety of conditions were tested: not including any covariates, including 10 principal components, including additional QC metric-based principal components, explicitly correcting for sequencing batches.

The conditions were compared against the base-case – not including any covariates at all. Amongst the tested conditions, the strongest p-values were obtained when controlling for 10 within-cohort PCs, closely followed by the no-covariate setting. Inclusion of the first QC metric-based PC, which effectively separates the PCR and PCR-free cohorts, had

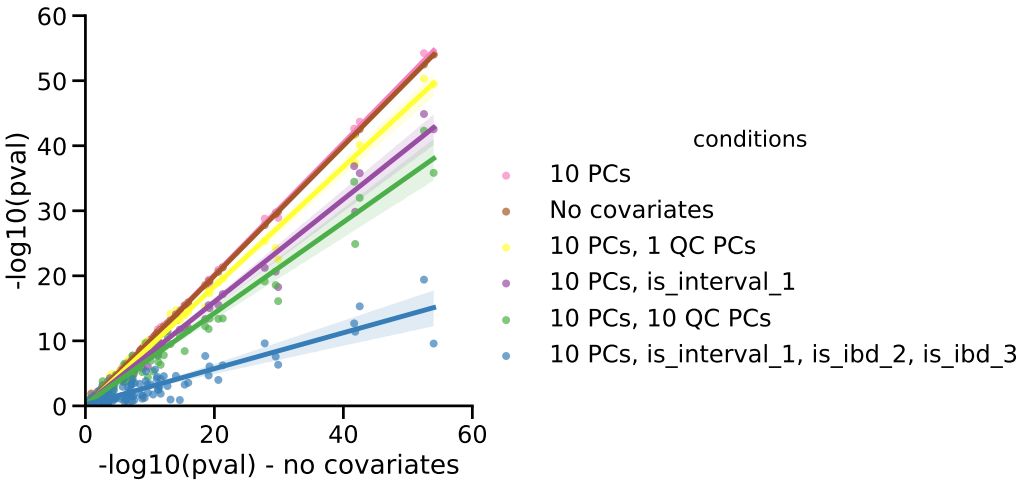


Figure 4.19 Influence of the inclusion of different covariate types on the power to identify known Crohn's associations in the IBD 15x cohort. X axis – p-values when replicating known Crohn's disease associations when performing logistic regression with no covariates. Y axis – p-values when replicating known Crohn's disease associations when using a particular set of covariates.

a smaller detrimental impact on the power than explicitly controlling for PCR via a binary covariate. Given that the PCR vs PCR-free sequencing seems to be the largest source of sequencing heterogeneity in the cohort, one should consider including this PC in future regression analyses. Overall, the batch-based covariates strongly reduced the power, as they effectively regress out case-control status of the samples in that cohort. In addition, binary covariates that are only positive in cases or controls require switching from the Wald or the LRT test to use the more computationally 'expensive' Firth test (2x–3x greater execution time), which should be considered when testing 200 million variants. The betas of known CD associations from de Lange et al. were compared to the betas estimated in 15x. In all covariate scenarios the betas were very strongly correlated (Peason $r > 0.95$). This suggests the absolute majority of the cases in the IBD 15x were, in fact, Crohn's disease patients (Figure 4.20).

In addition, I performed a case-control analysis on the LD-pruned subset of variants, with IBD-associated regions excluded. The variants were filtered quite stringently (MAF $> 5\%$, depth > 10, genotyping quality > 10, call rate $> 99\%$). The genetic inflation factor λ was calculated for each covariate-control scenario described above to estimate the presence of cryptic population structure and batch effect. Overall, I identified the inflation factor to be

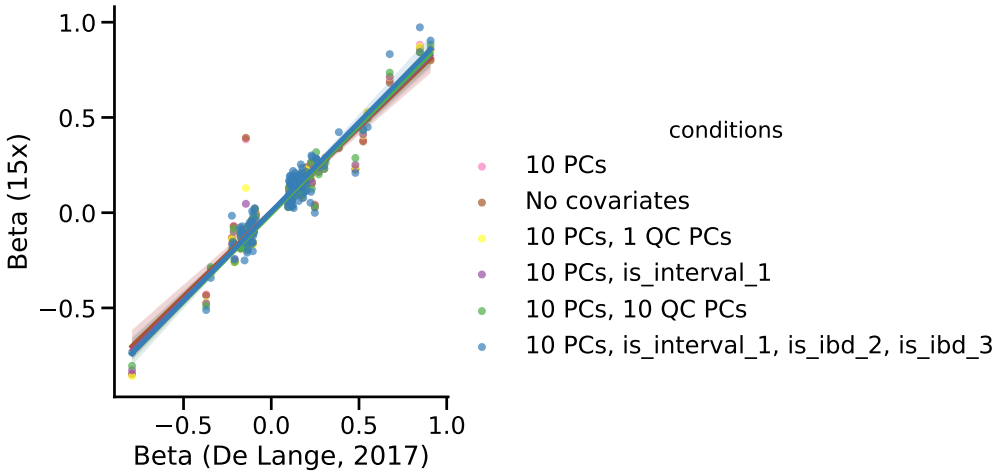


Figure 4.20 Betas of the known Crohn's disease associations estimated in the 15x cohort are strongly concordant to the ones reported in de Lange et al. [49]: minimal Peason r=0.95 (no covariates), maximal r=0.97 (10 PCs, 10 QC PCs).

between 1.13 (10 PCs, 10 QC PCs) and 1.19 (no covariates and 10 PCs, 1 QC PC). One notable exception was the scenario where I controlled for the 10 principal components, two IBD batches and the Interval Phase 1 batch – $\lambda = 1.02$. Interpretation of the absolute lambda values is not entirely straightforward, given that any polygenic trait will have $\lambda > 1.00$ and some published GWAS have $\lambda = 1.42$ (although, with many more samples) [193].

I believe there are several potential explanations for this: Perhaps the performed sample QC was insufficient and the outlier samples or the unidentified batch effects could be driving the moderate p-value inflation. This will require further investigation. Alternatively, poorly-genotyped variants could be driving the inflation. However, the rather stringent variant QC and lack of genome wide significant associations suggest that this is not the case. Alternatively, while the regions around the known IBD hits were excluded, IBD is thought to be a highly polygenic trait: Watanabe et al. [189] estimate that 0.06% of SNPs are causally associated with IBD. Therefore, despite excluding the known IBD variants, the inflation factor might be capturing some of the unknown causal variants.

4.3.4 Meta-analysis with the Broad IBD WES results

Our collaborators at the Broad Institute are currently finalising the production of a large multi-ethnic whole-exome sequencing cohort of IBD patients and matched population controls. The current data freeze contains around 10,000 non-Finish European cases and 17,000 controls. In addition, approximately 2,000 African American cases and a similar number of controls; 2,600 Ashkenazi Jewish cases paired with 4,000 controls; 1,500 American Hispanics with 1,000 controls (split into two groups due to admixture); 1,300 Finnish cases and 8000 controls were exome sequenced as a part of the same project. A number of 'promising' rare variants which reach a lenient significance threshold α=1×10^{-5} in an internal meta-analysis of all population cohorts were identified. Several variants have reached genome-wide significance level in past GWASs (bold in Table 4.1). The variants were annotated as 'GWAS' if they were within close proximity to known IBD associations or 'novel' if they fell outside such regions. Variant effect sizes and p-values calculated for the Crohn's disease subset of the Broad WES cohort were considered.

My goal was to meta-analyse the nominally-significant Broad WES results together with the summary statistics from the 15x study, to verify the feasibility of a future exome-wide meta-analysis and to evaluate the homogeneity of our results. Fixed effects meta-analysis was performed to combine the results from the individual WES cohorts with 15x. In addition, the I^2 metric was calculated to evaluate the heterogeneity of effect sizes from the Sanger 15x and the Broad non-Finnish European cohorts. I^2 metric across all populations was also calculated. As expected, it was marginally higher than the WES NFE vs 15x metric – both due to additional power to estimate heterogeneity, and, potentially, due to the heterogeneous effect across populations. Liu et al. demonstrated [111] that the effects of most IBD-associated variants are not heterogeneous across different global populations. However, this assumption will need to be revisited for rare-variant associations when the Broad WES dataset is finalised.

The 15x cohort was subsetted to 17,912 Crohn's disease cases and controls that passed the previously described sample QC. Logistic regression using the Firth test was performed, controlling for 10 principal components. Variants that passed the exome-wide significance threshold (α=4.3×10^{-7} [176] for coding variants) are listed in Tables 4.1 (GWAS-implicated regions) and 4.2 (novel).

Amongst the variants within the known IBD regions (Table 4.1), the strongest association was with the frameshift insertion in *NOD2* – rs199883290 (b37_pos: 16:50763778:G:GC,

OR meta = 3.04; 95% CI meta: 2.84 to 3.26; P meta = 7.25×10^{-220}, I^2 EUR = 0; MAF (NFE gnomAD) = 2.6%; MAF (INTERVAL) = 1.8 %). The variant had p-values lower than 0.05 across all cohorts and had a consistent effect across all ancestry groups $I^2 = 0$. The particular variant appears to be 3020insC, described by Ogura et al. [136].

Interestingly, some large effect size variants are found in regions that were previously only known to harbour low effect size variation. For example, a frameshift deletion in *TNFRSF6B* – rs54058315 (b37_pos: 20:62328248:CAG:C; OR meta = 2.95; 95% CI meta: 2.03 to 4.28; P meta = 1.54×10^{-8}) is around 1 Mb away from an intronic variant rs6062496 that has an odds ratio ∼1.13–1.15 in past GWAS (lead variant in a signal mapped to TNFRSF6B) [111, 49]. This indicates that rare large effect size variants are not limited to the regions with known common large effect associations (e.g., *NOD2*).

Finally, four significant associations outside the known IBD regions were identified (Table 4.2). One of the variants (8:144995964:G:A) was within PLEC – a gene that encodes plectin, a cytolinker protein which is involved in maintaining cell and tissue integrity [26]. Another missense (14:81972441:T:C) variant was within *SEL1L* that is thought to be required for the maintenance of intestinal homeostasis [61].

One of the significant variants is in *PKD1* (16:2142083:C:G) – a gene previously implicated in intestinal immune regulation. Administration of the PKRD1 protein is thought to induce down-regulation of TNF-α expression in macrophages [134]. However, despite the potential biological relevance, the variant appears to be entirely driven by the signal in the WES NFE cohort ($p = 4.11\times10^{-10}$). The variant was not even nominally significant in the 15x cohort (p = 0.38) or any WES cohort (apart from NFE). It shows strong heterogeneity effect between the Sanger 15x and WES NFE cohorts (I^2 EUR = 89.11). Therefore, I believe that the association is false. This underscores the importance of tests for heterogeneity of effects when performing meta-analysis.

The associations outside of the known IBD regions, despite the smaller sample size compared to the biggest IBD GWASs, are interesting in light of the recent work by O'Connor et al. [135] who hypothesise and provide some evidence that the extreme polygenicity of complex traits is a byproduct of purifying selection that purges high-effect variants from 'critical' genes and loci, leaving behind common-variant associations in critical regions of the genome. However, further work is required to formally evaluate this.

At this stage, I have only meta-analysed the variants that show some evidence of association in the Broad WES cohort. Fifteen rare variant associations across twelve genes were identified at the exome-wide significance level. Some of these were of extremely low frequency – down to 4 in 10,000 (16:50750810:A:G in *NOD2*) and not passing the significance threshold in any of the individual cohorts. This demonstrates the utility of meta-analysis to increase the statistical power for identifying rare variant associations in IBD and other complex diseases. Considering that the 15x sample size is comparable to that of the WES NFE cohort (which drives the majority of associations), the next logical step is to run a full-meta analysis of the two cohorts.

4.4 Discussion

In this chapter, I described the IBD 15x study – the largest IBD whole-genome sequencing association study to date. The study will help understand what role rare and low-frequency variation, largely missed during the GWAS era, plays in the pathogenesis of IBD. Ultimately, I hope that the uncovered genetic associations will inform potential IBD drug targets, and perhaps lead to the development of new IBD therapies. In addition, the study includes several thousand richly-phenotyped individuals from the NIHR IBD BioResource – and will be used to study the subphenotypes of IBD, and enable the extension of the pharmacogenetics studies described in the first two research chapters.

The variant calling was completed in July 2019, which meant that the time I was able to spend on analysing the final dataset was fairly limited. The scale of the WGS dataset, complexity of the quality control procedures, and the difficulty of differentiating between false and true positive associations made rapid progress quite difficult. Therefore, I have decided to concentrate on the sample QC procedures – a step which will be crucial to ensure the quality of the future association studies that use the 15x cohort.

Overall, post sample-QC, the dataset can be used to finalise the site and variant QC, and, finally, start running the association studies. After removing the outlier samples, I was able to replicate 91% of the known CD associations (96 out of 105) to the $\alpha = 0.05$ significance level and with variant betas closely matching those described in the largest Crohn's disease GWAS (r=0.97). Inevitably, given the iterative nature of the association studies and depending on the results from the first genome-wide analyses, some of the sample QC thresholds may be

V	MAF	Gene INT	OR NFE	OR 15x	OR meta	95% CI meta	I2 EUR	P NFE	P 15x	P meta
16:50763778:G:GC	0.018	NOD2	3.07	2.97	3.04	(2.84, 3.26)	0.00	2.85e-106	1.48e-70	7.25e-220
1:67705958:G:A	0.068	IL23R	0.44	0.43	0.43	(0.40, 0.46)	0.00	4.73e-43	5.44e-55	1.19e-121
16:50745926:C:T	0.047	NOD2	1.91	2.01	1.96	(1.85, 2.07)	0.00	8.23e-54	5.89e-55	1.67e-121
16:50756540:G:C	0.013	NOD2	2.45	2.42	2.42	(2.22, 2.63)	0.00	4.33e-37	1.63e-29	3.41e-89
16:50750842:A:G	0.0013	NOD2	2.76	2.75	2.94	(2.38, 3.62)	0.00	1.08e-04	8.71e-06	1.06e-23
19:10463118:G:C	0.051	TYK2	0.74	0.64	0.69	(0.64, 0.75)	60.19	1.14e-06	8.91e-15	1.18e-20
4:103188709:C:T	0.076	SLC39A8	1.26	1.26	1.24	(1.18, 1.30)	0.00	2.04e-08	3.41e-09	4.12e-17
16:50746086:C:T	0.0043	NOD2	2.08	1.82	2.10	(1.76, 2.49)	0.00	1.70e-09	3.59e-05	7.86e-17
9:139259592:C:G	0.006	CARD9	0.30	0.37	0.37	(0.29, 0.47)	0.00	8.29e-11	1.42e-07	1.15e-16
16:50827518:C:T	0.07	CYLD	1.21	1.16	1.21	(1.15, 1.27)	0.00	4.70e-06	4.66e-04	3.03e-14
19:10469975:A:C	0.095	TYK2	1.15	1.20	1.19	(1.14, 1.25)	0.00	3.08e-04	5.11e-07	6.90e-14
4:3449652:G:A	0.067	HGFAC	1.25	1.13	1.22	(1.15, 1.28)	63.19	2.69e-07	4.29e-03	1.86e-13
12:40740686:A:G	0.017	LRRK2	1.47	1.36	1.39	(1.27, 1.52)	0.00	1.10e-06	1.14e-04	4.17e-13
22:21998280:G:A	0.014	SDF2L1	1.52	1.33	1.46	(1.32, 1.62)	14.49	1.15e-06	1.49e-03	4.50e-13
1:67705900:G:A	0.015	IL23R	0.61	0.70	0.67	(0.59, 0.75)	0.00	2.56e-06	4.53e-04	7.32e-11
19:10464843:G:A	0.0077	TYK2	0.43	0.60	0.53	(0.43, 0.65)	57.24	2.04e-07	5.19e-04	1.62e-09
19:10600418:G:A	0.018	KEAP1	1.35	1.29	1.30	(1.19, 1.42)	0.00	5.72e-05	7.93e-04	2.87e-09
11:65425764:C:T	0.0043	RELA	2.00	1.51	1.74	(1.45, 2.08)	46.88	2.31e-07	6.77e-03	3.38e-09
16:50750810:A:G	0.00047	NOD2	3.34	2.73	2.15	(1.66, 2.78)	0.00	4.72e-03	8.74e-03	5.02e-09
9:139358899:C:T	0.029	SEC16A	0.77	0.78	0.75	(0.69, 0.83)	0.00	3.65e-04	3.03e-04	1.02e-08
20:62328248:CAG:C	0.00068	TNFRSF6B	2.73	2.60	2.95	(2.03, 4.28)	0.00	9.36e-04	3.55e-03	1.54e-08
2:234436069:C:T	0.049	USP40	0.82	0.81	0.82	(0.76, 0.88)	0.00	7.39e-04	1.22e-04	3.37e-08
16:50745929:C:T	0.0048	NOD2	1.54	1.69	1.63	(1.37, 1.95)	0.00	8.07e-04	1.81e-04	3.46e-08
22:21800049:G:A	0.0034	HIC2	1.94	1.43	1.52	(1.30, 1.78)	44.05	1.71e-05	3.52e-02	1.32e-07
1:161496178:G:A	0.097	HSPA6	1.13	1.08	1.13	(1.08, 1.18)	0.00	1.42e-03	4.23e-02	3.29e-07

Table 4.1 Summary statistics for the meta-analysed variants within the known IBD-associated regions. Only variants that pass the exome-wide significance threshold are shown. Variants previously reported in other GWAS are highlighted in bold.

V	MAF	Gene	OR NFE	OR 15x	OR meta	95% CI meta	I2 EUR	P NFE	P 15x	P meta
1:117122269:GGTC:G	0.008	IGSF3	0.52	0.37	0.38	(0.29, 0.49)	21.60	7.34e-03	1.17e-09	2.91e-13
16:2142083:C:G	0.0015	PKD1	0.25	0.77	0.42	(0.31, 0.57)	89.11	4.11e-10	3.82e-01	3.46e-08
8:144995964:G:A	0.07	PLEC	1.14	1.14	1.15	(1.09, 1.21)	0.00	1.79e-03	1.43e-03	6.41e-08
14:81972441:T:C	0.014	SEL1L	1.42	1.32	1.36	(1.21, 1.53)	0.00	3.82e-05	1.93e-03	1.39e-07

Table 4.2 Summary statistics for the meta-analysed variants outside of the known IBD-associated regions. Only variants that pass the exome-wide significance threshold are shown. The variant in *PKD1* is likely to be a false association, driven entirely by one of the meta-analysed cohorts.

adjusted and some extra steps added. However, I believe the implemented QC pipeline works robustly with WGS data and can be extended fairly easily.

In addition, I have described my earlier work on power modelling for sequencing association studies. The modelling results suggest that sequencing more samples at around 15x to 17x depth provides more statistical power to detect rare, single variant associations in case-control and quantitative trait settings, compared to sequencing a smaller cohort at full 30x depth. The conclusions match those published by Rashkin et al. [152].

I have provided an overview of the index missassignment issue, widely reported to be affecting the last two generations of Illumina short read sequencing machines. I confirmed the presence of cross-sample index missassignment across all 15x batches. However the results indicate that only a small fraction of read-groups are strongly affected (3.7%). In addition, I confirmed that dual indexing greatly reduces the missassignment levels and should be considered for all future WGS and WES studies.

Finally, I have provided early single-variant association results from the 15x cohort by spot meta-analysing some ‘promising’ variants, found by our collaborators at the Broad Institute in a large whole-exome sequencing cohort. The majority of the significantly associated rare variants appear to be harboured in known IBD genes like *NOD2*, *TYK2*, and *IL23R*. Some of these variants appear to have a much higher effect size than their previously-known common variant counterparts (for example, rs540583157 in *TNFRSF6B*. In addition, the meta-analysis indicates that variants in *PLEC*, *SEL1L*, and *IGSF3* play a role in the pathogenesis of IBD, and, to my knowledge, no previous IBD associations have reported variants linked to them.

The spot meta-analysis will be followed by a full-scale joint association study that combines more than 35,000 cases and 95,000 controls across several global populations.

Ultimately, while the spot meta-analysis has already provided some interesting results, this is just the beginning of work on the IBD 15x association study.

Single-variant tests association tests should be performed genome-wide to estimate the effects and the significance values of individual variants. Almost certainly, during the first few iterations these results will contain plenty of artefacts – spurious false associations. QC metric properties of such variants should be observed to refine the variant and site filters to make them more stringent. In addition, the p-value inflation metric λ and QQ-plots should be used to validate the absence of population structure, which often leads to an abundance of marginally significant variants across the entire genome. It is important to get the single variant association tests done to a good standard, as the spurious associations can negatively influence the outcome of the gene and noncoding burden tests (described below), where it is even harder to identify false results driven by false associations.

Separating true and spurious rare variant associations may be nontrivial. When conducting traditional GWAS, a known heuristic approach is to create a locus zoom plot and observe neighbouring associations which should have p-values close to the top SNP (due to the LD). Unfortunately, for many rare variants such an approach is futile – there may be no neighbouring variants in high LD. However, other techniques can be used to validate rare variant associations. Firstly, one should verify that the variant is not present in only one of the sequencing batches, or, ideally that the allele count per batch matches the expected one given the batch sizes. Secondly, given the presence of the summary statistics from the Broad WES cohort, an exome-wide meta-analysis could be conducted and used as a QC tool. QC metrics of variants with high evidence of heterogeneity should be inspected, potentially informing the filtering thresholds. Thirdly, large-scale frequency databases like gnomAD can be used to verify that the frequency of the variant closely matches that reported in the database. At the time of this thesis completion gnomAD was not available for genome build 38 data, but should be updated for the next release. Lastly, all reported single-variant associations should be verified by manually inspecting the track plots produced by tools like IGV – these visualise the reads that went into the variant call, helping to understand whether a calling error has occurred. Ideally, for the reported associations, targeted Sanger sequencing of a few carriers should be performed.

Once the QC is complete, I would expect the absolute majority of the novel rare associations to be coding. This is expected, as, at the current sample size, we are well-powered to detect rare variant associations with an odds ratio of around 1.5 and above ($\sim$80% power to detect rare-variant associations of 1% frequency variants with relative risk of 1.5 and above). In order to increase the number of novel rare-variant associations even further, the summary statistics from the single-variant tests will be used to perform the joint coding region meta-analysis with other cohorts: Broad WES and Sanger WES.

In the initial spot meta-analysis a fixed-effect model was used. Fixed-effect meta-analysis assumes that the differences in the observed effects are due to sampling errors. This is justified given the observation that the majority of known common IBD associations have a non-heterogeneous effect across global populations (Cochran's Q test for heterogeneity, $p > 0.05$) [111]. In the current meta-analysis, some heterogeneity of effects was observed. Therefore, analysis with a random-effects model should be considered. Mixed-effect models allow for the true effect size to be different between the groups – accounting for potential ancestry-specific effects, gene $\times$ environment interactions, and for heterogeneity of recruitment. It is unclear whether one would expect the rare variant associations to have a similar effect across different ancestry groups: isolate population studies have consistently uncovered pathogenic variants which have similar effects in both the isolate and the global populations (see Introduction chapter). However, the heterogeneity of rare variant associations has never been been studied systematically and warrants further investigation. Additional meta-analysis techniques, like the Bayesian MCMC-based methods, can be considered.

WGS and WES datasets provide an opportunity to study extremely rare, almost private genetic variation. However, single variants tests are not sufficiently powered to robustly associate these variants with the phenotype. To overcome this, techniques that group together the effects of ultra-rare, typically deleterious, variants (LoFs) exist (see the Introduction chapter). The variants are usually grouped together on the per-gene (gene-based tests) or a per-exon level. The burden of the rare variants is compared between cases on controls. Since less elements are tested, compared to the single-variant association tests, the multiple-testing significance threshold is adjusted accordingly. Luo et al. [116] used gene-based tests to detect a burden of very rare, damaging variants in known Crohn's disease risk genes. It would be interesting to see whether the burden tests performed on 15x and the WES cohorts allow us to identify new IBD-associated genes not previously implicated via single variant tests.

Burden tests can be used in a more targeted, hypothesis-driven way. Instead of testing the burden across all genes, one could evaluate groups of genes united by some biologic function (pathways, groups of genes associated with a disorder, etc.). One of the less explored questions in IBD genetics is the architecture of the neonatal ('infantile-onset') and very early onset IBD. For neonatal IBD, 60 monogenic defects that cause IBD-like colitis have been identified [168]. Very few of these overlap with genes implicated in common-variant GWAS. It is not well-understood whether the phenotypic similarity of neonatal and adult-onset IBD is underpinned by overlapping biologic mechanisms, though some of the monogenic variants are in genes involved in epithelial barrier function (e.g., *TTC7A* [12]). The interest in the architecture is not just driven by scientific curiosity, but by the fact that monogenic IBD patients are often refractory to conventional therapies. It would be interesting to see if the genes involved in neonatal IBD are enriched for a burden of rare, pathogenic variants in the adult-onset 15x cohort. If this is the case, IBD and neonatal IBD are genetically overlapping disorders, with some adult IBD patients carrying rare, pathogenic variants within the neonatal IBD-implicated genes. A lack of burden might suggest that the neonatal IBD patients are often refractory to conventional treatments, due to these treatments targeting different biological processes.

Arguably, the most challenging task of the future analysis is uncovering rare associations within the noncoding regions of the genome. Assuming low or moderate effect size of such variants, the cohort is not big enough to find many of these during single-variant tests. Noncoding variants can be grouped together and used for association tests. The significantly associated groups can be the then examined to identify individual variants that are driving the signal. A detailed overview of the grouping techniques is provided in the Introduction chapter. The majority of these approaches either groups the variants in an unbiased way (e.g., sliding windows across the genome) or tries to link them to the target gene). Grouping the noncoding variants to the gene is nontrivial. One of the approaches is to link the noncoding variants within the enhancers and promoters of that gene. Gene expression data can be used to refine these groupings. More recently, a variety of methods for *in-silico* prioritisation of noncoding regulatory variants have emerged [104], yet their predictive value remains imperfect [56].

Finally, the rare coding and noncoding variation can be used to improve the predictive value of the polygenic risk scores (PRS) for IBD. Currently, the predictive value of PRS is quite low (AUC = 0.633) [97]. The predictive value typically correlates with the percentage of the explained variance ('SNP-based heritability') which is low even in the biggest IBD GWAS

and does not match the traditional twin-based heritability. The same discrepancy is observed for the absolute majority of complex traits (the 'missing heritability' problem). Recent work by Wainschtein et al. [184] shows that by including the rare variation from 20,000 whole-genome sequenced individuals, it is possible to to 'recover' this missing heritabiliy for BMI and height. Rare, especially coding, variants in low LD with neighbouring variants were enriched for heritability. It is unknown whether the same effect holds true for complex disease, but the IBD 15x cohort provides a great opportunity to study this. If rare variants are enriched for IBD heritability, a WGS-based polygenic risk score can be derived and evaluated.

The role of rare variation in IBD pathogenesis remains largely unknown. Uncovering rare pathogenic variants in known and novel IBD regions will improve the ability to prioritise drug targets. The 15x study and the adjoining whole-exome datasets will be instrumental in this task.