# The Genetics of IBD: From Susceptibility to Drug Response and Patient Outcome

**Aleksejs Sazonovs**

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Pembroke College September 2019

# Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Aleksejs Sazonovs
September 2019

# The Genetics of IBD: From Susceptibility to Drug Response and Patient Outcome

**Aleksejs Sazonovs**

## Abstract

Inflammatory bowel disease (IBD) is a group of immune-mediated autoinflammatory disorders, primarily manifesting in the gastrointestinal tract. Affecting millions of people around the world, IBD has a severe impact on patients' quality of life. Several pharmacologic treatments have been available since the 1950s. However, the majority of patients either do not respond to a given therapy or lose response to a previously effective treatment and thus require therapeutic escalation.

In the first research chapter of my thesis, I describe the results of the Personalised Anti-TNF Therapy in Crohn's disease study. Immunogenicity to anti-TNF therapy is a major cause of loss of response, hypersensitivity reactions, and discontinuation of treatment in patients. Currently, immunogenicity cannot be predicted prior to treatment. My analysis has identified a strong dominant association in the HLA region on chromosome 6 (HLA-DQA1*05, P=5.9x$10^{-13}$; HR=1.90; 95% CI, 1.60 to 2.25). Around 40% of individuals of European ancestry carry HLA-DQA1*05, and the data suggest that around 95% of these would develop immunogenicity within the first year of infliximab monotherapy treatment (a common anti-TNF treatment regime).

In the second research chapter of my thesis, I describe a genome-wide association study of thiopurine-induced liver damage (TILI). Ultimately, the study was underpowered to detect

any associations of moderate effect size and did not detect any associations of high effect size amongst the common genetic variants. Interestingly, I was not able to replicate the association in *PTPN22*, which was reported to be a risk factor for drug-induced liver damage by Cirulli et al. [39] – suggesting that its effect might be heterogeneous depending on the therapy.

Finally, the third research chapter describes the initial analysis of the IBD 15x dataset – a whole-genome sequencing association study of around 7,000 IBD patients paired with 12,000 matching controls. I provide an overview of the sample quality control procedures and describe some of the novel challenges that sequencing studies bring in comparison to standard GWAS (e.g., sample cross-contamination due to index mismatching). Finally, I also provide the results of the initial meta-analysis of the exome-sequencing dataset produced by the Broad Institute. The results demonstrate that rare coding genetic variants play a role in IBD pathogenesis.

I would like to dedicate this thesis to my parents – Gaļina Sazonova and Vadims Sazonovs

# Acknowledgements

First of all, I would like to thank my supervisor Carl Anderson who has guided this work. I could not have wished for a more compassionate and brilliant mentor. Jeff Barrett, with whom I started my PhD, has shown me how passion for a topic can coincide with relentless rigour. The union of teams 143 and 152 – you have made me feel welcome here and I have learnt so much from each of you. The computational work (i.e. all) was enabled by the Human Genome Informatics team – Pavlos Antoniou, Chris Harrison, Colin Nolan, Alan Daly, Vivek Iyer, and Josh Randall. I've broken so many things and yet you seem to tolerate me. Nicole and other members of the Soranzo team, especially Klaudia Walter and Kousik Kundu for stimulating discussions and lively debates about 15x. Paris Litterick, Eloise Stapleton, Sally Bygraves, without whom these projects would have ground to a halt. The PANTS and TILI projects would not have been possible without the Exeter team – Nick Kennedy, Tariq Ahmad, Gareth Walker, Claire Bewshea and others. The last three years would not have been so joyful without my colleagues, who later became friends, then stopped being colleagues, but remained friends: Dan Rice, Loukas Moutsianas, Liu He, Mari Niemi, Fernando Riveros Mckay Aguilera, Scott Shooter, Tejas Shah, Arthur Gilly, and Sophie Hackinger. A heartfelt thank you to Emma Molloy for putting up with me during the final stretch of the PhD. Finally, I would like to express my gratitude to my family who have selflessly supported me on this journey. I owe so much to you.

# Table of contents

# List of figures

# List of tables

# Nomenclature

**Acronyms / Abbreviations**

| | |
|---|---|
| ADA | anti-drug antibody |
| ADR | adverse drug reaction |
| EWAS | exome-wide association study |
| GWAS | genome-wide association study |
| HLA | human leukocyte antigen |
| CD | Crohn's disease |
| IBD | inflammatory bowel disease |
| IBDU | unclassified IBD |
| UC | ulcerative colitis |
| LD | linkage disequilibrium |
| MAF | minor allele frequency |
| MHC | major histocompatibility complex |
| TNF | tumour necrosis factor |
| VEO | very early onset |
| WES | whole-exome sequencing |
| WGS | whole-genome sequencing |

# Chapter 1

# Introduction

*Parts of this chapter were previously published as a review article in Annual Review of Genomics and Human Genetics [164]*

## 1.1 Common and rare variant studies of complex traits

### 1.1.1 Genome-wide association studies

Complex diseases are disorders that are caused by a combination of genetic, environmental, and lifestyle factors. For decades, the central motivation of human complex disease genetics has been to robustly identify genetic variants associated with disease risk. After a number of false starts, a series of technical and methodological advances have enabled rapid progress toward this goal. First, the International HapMap Project [84] revealed a globally shared common genetic variation of single-nucleotide variants (SNVs) and described the detailed local correlation patterns in such variation (known as linkage disequilibrium [LD]). Next, genotyping microarrays made it possible to cheaply genotype hundreds of thousands (or indeed millions) of common variant positions in a high-throughput manner. Finally, large sample sizes were collaboratively assembled across a wide range of diseases and traits. These factors came together a decade ago [190] to mark the beginning of an era of genome-wide association studies (GWAS).

Through steady application of the basic GWAS approach and rapidly increasing sample sizes, there are now around 13,000 common genetic variants robustly associated with a wide range of traits and diseases [33]. Despite this constantly growing list of hits, there has been a great deal of discussion about how much clinical or biological insight GWAS have provided [70]. Indeed, it has become clear that most human diseases have a dramatically more complex genetic architecture than previously suspected, with at least hundreds if not thousands of distinct, and subtle, genetic risk factors [27]. While it may be disappointing that this problem is not simpler, the biological reality must be confronted if progress is to be made on better treatment for disease.

One fundamental principle of GWAS is that, by design, they concentrate on common variation. Genotyping arrays benefit from the widespread LD among common variants and thus capture nearly all of the information contained in the approximately 10 million common variants by directly measuring only 5% of that total [14]. Statistical techniques, such as genotype imputation, allow the prediction of a large number of variants not directly measured by genotyping arrays, although their accuracy for rare variation remains imperfect [120]. Therefore, the genetics community has eagerly anticipated technological developments that would allow the rapid and affordable measurement of rare variation, in order to assess how it can complement the information on common variants gleaned from GWAS.

### 1.1.2 Next-generation sequencing

Following the completion of the multibillion-dollar Human Genome Project [85] in 2003, a series of new technologies collectively referred to as next-generation sequencing have brought the price of sequencing a complete human genome down to the sub-$1,000 level. In sequencing by synthesis, currently the most popular approach, millions of short reads ($\sim$150 base pairs) are synthesised from template DNA fragments roughly randomly scattered across the genome. These reads are aligned to the reference genome, and apparent differences are identified and classified as potential sites of genetic variation. Individual reads often contain errors, and the random sampling from the genome means that many regions will not be covered from a set of reads whose total length equals the length of the genome. Thus, enough reads are generated to cover the genome several times in order to achieve redundancy at each site, and the ratio of total read length to target genome length is known as the sequencing depth.

The crucial advantage of sequencing over genotyping is the ability to detect and measure any variant in an individual's genome, rather than just the pre-specified few hundred thousand common variants on a genotyping array. While both the chemistry and informatics required to analyse sequence data are more complex, the ability to study rare variation offers new potential insights that are hidden from the GWAS approach. The rapidly decreasing cost of sequencing has now begun to make it possible to deploy this technique at the scale necessary to conduct well-powered studies of rare variation in complex disease.

This section of the Introduction focuses on the intersection of these two stories, which have previously been largely separate. As we are increasingly able to study the full range of genetic variation, from rare to common, how can we best jointly analyse different types of data to understand the genetic and biological basis of complex traits and diseases in humans?

I begin by describing the current GWAS interpretation approaches and methods in the absence of rare variation in order to better understand the outstanding challenges. Next, I consider the technical and statistical issues affecting the generation and analysis of informative rare-variant data. I then describe a variety of special study designs that may be especially informative during the transitional phase, where sequencing is still 1–2 orders of magnitude more expensive than genotyping. Finally, I consider the future outlook for joint analysis of rare and common variation.

## 1.2 Current approaches to resolving genome-wide association study signals

### 1.2.1 Statistical fine-mapping

One of the biggest challenges facing GWAS result interpretation is distinguishing between causal variants and other variants that are correlated with them. Correlated variants may show a statistically significant association but provide no insight into the underlying biology of the condition or trait. Statistical fine-mapping techniques determine likely causal variants by estimating the probability that each variant in a correlated set is causal relative to the others in the set [174]. A Bayesian approach [119] was introduced to handle the simplest case of a single disease outcome and a single causal variant (that is, one variant has a true biological effect, and association signals at all other variants in the region are due solely to their correlation with that causal variant). As it has become apparent that this simple case is often not true, this framework has been expanded to allow simultaneous testing of multiple diseases and multiple independent causal variants at the same locus (physical region on a chromosome) [82].

The fine-mapping tools described above require sample-level genotype data, which can reduce their utility. For example, for many of the largest meta-analyses of complex traits and diseases, no one individual has access to the sample-level data for all the cohorts [113]. For this reason, several methods have been developed more recently that require only summary statistics as input (e.g., effect sizes and p-values for every single-nucleotide variant) [21, 37]. For signals with only a single causal variant, these approaches should produce identical results, but fine mapping multiple causal variants from summary statistics adds two complexities. First, exhaustive conditional analysis with a set maximal number of presumed causal signals is computationally expensive. Newer methods, like FINEMAP [21], perform a stochastic search, dramatically reducing the search time and thus allowing one to test whether a signal is driven by multiple causal variants (e.g., exhaustive search in a region with 8,612 variants takes 300 years, while a stochastic search is completed in less than 30 seconds). Second, this pseudoconditional analysis requires a pairwise matrix of LD, either from the original GWAS (rarely available in practice) or from a reference panel like the 1000 Genomes Project [1]. It is important that this LD reference matches the population ancestry of the GWAS cohort and that it is large enough to provide sufficient precision in the LD estimates. Benner et al. [20] showed that an LD matrix derived from 1,000 individuals is adequate for a cohort of up to 10,000 individuals but performs poorly for a GWAS of 50,000 individuals (which achieves very good accuracy when the LD is calculated from 10,000 individuals).

Obtaining matching LD matrices for biobank-sized or less frequently studied populations may be challenging. This problem is more acute for rare-variant fine mapping. In principle, owing to the low LD with neighbouring variants, rare variants should be easier to fine map, but these variants may not be captured in the reference population or the precision of the LD estimation will be low. The absence of LD information for the specific rare variant will make it impossible to fine map it. This problem motivates the need for researchers to share LD information alongside the summary statistics of the association studies that they perform. Benner et al. ([20]) proposed LDstore, a tool for ‘efficient estimation, storage, and seamless sharing of LD information’ to simplify this process. Summary-statistics repositories, like the National Human Genome Research Institute–European Bioinformatics Institute GWAS Catalog [34], should provide the ability to deposit cohort-specific LD data.

### 1.2.2 Trans-ethnic association studies

A typical GWAS controls for population structure by either concentrating on samples of similar genetic ancestry or including genetic principal components as covariates. However, as cohorts with more diverse ancestry were genotyped and sequenced, it became possible to use this population structure to perform meta-analyses that boost statistical power and help to fine map causal variation.

Trans-ancestry studies rely on the assumption that causal variants will be shared across different populations, while the differences in the patterns of LD provide greater discrimination between causal and non-causal variants [124]. The design of early genotyping arrays was largely biased toward common variation present in European populations, making genotyping of other populations incomplete. This problem is largely resolved in modern large genotyping arrays, although other issues, like the lack of high-quality non-European imputation reference panels and less precise resources for population allele frequencies, remain.

Successful application of the trans-ethnic association approach depends on two factors. First, the sample size in different ancestries must be sufficiently large. For example, a well-mixed analysis of type 2 diabetes [118] both found more signals and resolved causal variants more precisely (23,553 individuals of European ancestry, 23,536 Japanese, 16,325 Hispanic Americans, 8,224 African Americans). Second, the methodology must be appropriate to the underlying shared or distinct genetic architecture across populations. A recent analysis in type 2 diabetes [117] showed that meta-regression accounting for ancestry provides improved fine-mapping resolution.

Trans-ethnic association studies fundamentally depend on shared causal variants across populations. This means they are best suited to common variants that arose early in human history and are therefore shared throughout the world. For this reason, trans-ethnic studies have not implicated many rare variants, which are more likely to be evolutionarily recent and hence population specific. Wang and Teo [188] pointed out that rare causal variants are in low LD with neighbouring markers and therefore do not require trans-ethnic fine mapping.

### 1.2.3 Regulatory target analysis

Even if statistical fine mapping can identify causal variants, it is still challenging to connect that variant to the gene (or genes) underlying the association [174]. This problem is, of course, straightforward if the implicated variant alters the amino acid sequence of a protein-coding gene, but it is estimated that fewer than 15% of GWAS signals are driven by such missense changes [43]. This has motivated a variety of different ways of trying to discover the genes regulated by GWAS variants, collectively referred to as regulatory target analysis.

Despite the complexities of translating associations at a genomic locus into biological knowledge, the arrival of regulatory annotation resources such as the Encyclopedia of DNA Elements (ENCODE) and the Genotype-Tissue Expression (GTEx) database has enabled in silico analyses that prioritise specific genes. Broadly, these analyses fall into two categories: those that directly measure the effect of genetic variation on gene expression and those that use functional genomics to connect regulatory elements (e.g., enhancers) to the genes they regulate.

#### Genetic variation and gene expression

Genetic variants that explain a portion of the variance in expression of a gene are known as expression quantitative trait loci (eQTLs). Local eQTLs (or *cis*-eQTLs) are typically within 1 Mb of the gene's transcription start site, while distant eQTLs (or *trans*-eQTLs) can be much farther from the transcription start site, sometimes even on a different chromosome [132] (Figure 1.1). After protein-coding changes, GWAS variants that are eQTLs have the most straightforwardly interpreted mechanism.

Public eQTL resources like the GTEx database provide tissue-specific gene expression levels and genotypes from hundreds of donors (e.g., the v7 GTEx release contains data from more than 600 donors and gene expression from 48 tissues, with at least 70 samples per tissue) [75]. These data can be combined with GWAS outputs using statistical methods such as COLOC [67] to test whether the same underlying variant likely to be causal for with both disease risk and gene expression. rather than two different variants in LD with each other, one of which is associated with disease, the other with gene expression. This is important because eQTL variants are ubiquitous, and mere physical proximity to a GWAS signal is not strong evidence that the two are related [38].

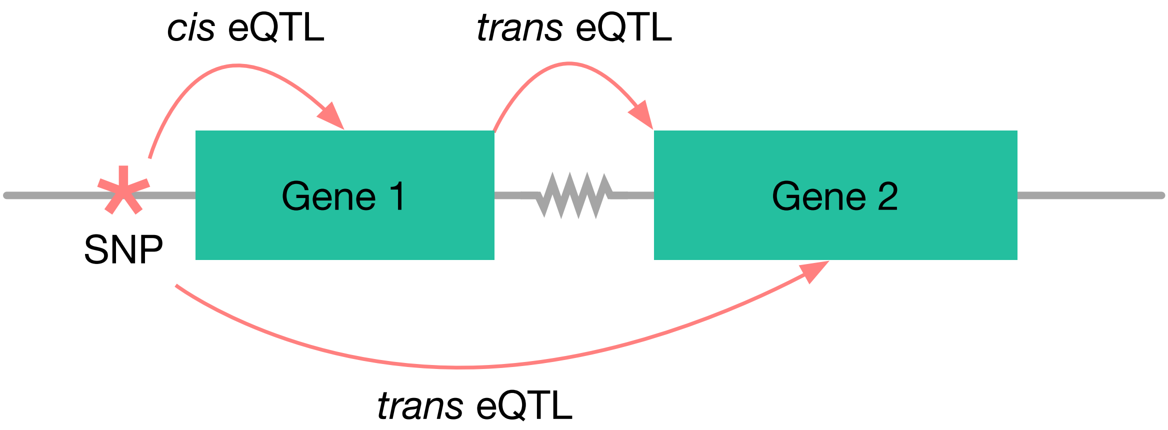


**Figure 1.1** When a genetic variant affects the quantitative level of expression of a gene, it is known as an expression quantitative trait locus (eQTL). The majority of known eQTL are caused by variants very close to the affected gene (‘*cis* eQTL’), often in the promoter. In rarer cases a variant can affect expression of a distant gene (‘*trans* eQTL’), even on a different chromosome, either indirectly through modification of a local gene (top arrow) or through an unknown mechanism (bottom arrow).

Computational methods like FUSION [76] and PrediXcan [64] leverage the data from GTEx or study-specific expression data to perform transcriptome-wide association studies. FUSION uses GWAS summary statistics to perform a gene-based association test of expression in a specific tissue. As with fine-mapping tools, transcriptome-wide association studies require accurate LD reference data. The output of a transcriptome-wide association study indicates whether the specific genes are predicted to be over- or underexpressed in affected individuals. One of the limitations of FUSION and other summary-statistics methods is the inability to make inferences based on rare-variant data owing to their poor capture by the LD reference panels. PrediXcan works with individual-level genotype data and therefore in principle is able to include rare variants in the estimates. However, Gamazon et al. [64] noted that larger and denser training data sets will be necessary to achieve accurate prediction from rare variants.

#### Functional genomics and regulatory connections

While using eQTLs is the most direct approach to proving that a specific variant regulates a specific gene, there are drawbacks. Most importantly, using eQTLs for regulatory target analysis requires that there exists a sufficiently large eQTL data set in the population and tissue or cell type where the variant is acting. When such a data set is not available, a variety of functional genomics tools can be used to try to resolve the potential function of the associated variant.

The relevant functional genomics assays can be broadly classified into two groups. First are the chromatin conformation assays, which detect physical contacts between different positions in the genome [74]. In the context of GWAS resolution, these data sets can be used to demonstrate, for example, that a disease-associated variant outside of a gene body physically interacts with a promoter (as would be expected if the variant is in an enhancer, for example). Second are methods that use cell-specific measurements of chromatin openness (e.g., via assay for transposase-accessible chromatin using sequencing (ATAC-seq) [32]) to prioritise variants that are more likely to be available to the regulatory machinery in disease-relevant tissues and cells. A statistical method [146] has been proposed that builds upon the Bayesian frameworks described above by adjusting prior probabilities for individual variants based on such chromatin openness maps.

## 1.3 Using rare variants to resolve genome-wide association studies

How are rare variants helpful in resolving the role of GWAS signals? If a disease decreases reproductive fitness, like a neurodevelopmental disorder or an autoimmune disorder with early onset, then any variant that strongly affects disease risk is kept at a low frequency by negative selection. Rare variants also tend to have occurred more recently in human history than common variants, which means that they have fewer other variants in LD with them. Together, these two factors mean that rare variants are potentially more easily interpretable than the common variants discovered by GWAS. The challenge arises in measuring and statistically analysing them.

### 1.3.1 Measuring rare variation

High-depth whole-genome sequencing is the most comprehensive approach for measuring rare variation across the genome. However, at present, its application is limited by the costs and various computational challenges, especially for large-scale cohorts. Techniques such as genotype imputation, targeted sequencing, and whole-exome sequencing are cost-effective alternatives, although each has drawbacks in terms of accuracy or scope.

#### Genotyping and genotype imputation

Modern genotyping arrays capture 200,000–2,000,000 variants across the genome. Owing to the limited number of single-nucleotide variants that an array is able to genotype, the absolute majority of these variants are common. This makes raw genotyping data poorly suited for rare-variant studies. Perhaps the most straightforward means of measuring rarer variants is to extend this existing GWAS paradigm.

Imputation is a statistical technique that allows one to infer variants that were not directly genotyped. Imputation relies on reference panels that have been more completely sequenced, with modern services that are able to impute tens of millions of sites across the entire genome based on a much sparser GWAS backbone. For dense genotyping arrays, imputation is able to predict nearly all missing common variation with high accuracy. As the variant minor allele frequency (MAF) decreases, so does the accuracy of imputation. A common practice is to exclude imputed variants with MAF $< 1\%$ (i.e., rare variants) from the GWAS genotype dataset. The authors of the Haplotype Reference Consortium imputation panel showed that the aggregate $r^2$ for imputed sites is approximately 0.85 for MAF=1% and 0.5 for MAF=0.05% [120]. The accuracy of imputation for individuals of non-European ancestry is lower owing to the current lack of large-scale ethnically diverse reference cohorts. While future panels will be able to address this issue, researchers should be cautious about the imputation quality of rare and low-frequency variants in non-European samples.

Genotype imputation is likely to remain a valuable tool, even as the cost of sequencing decreases. The abundance of existing genotyping data (e.g., 500,000 genotyped individuals in the UK Biobank) drives the demand, and the ever-improving reference panels will gradually increase the accuracy. Nonetheless, the accuracy of imputation remains poor for truly rare variants (below, say, 0.05%), meaning it cannot completely supplant direct sequencing.

### Targeted sequencing

The earliest forays into directly sequencing rare variants for disease association studies used targeted sequencing to reduce costs. Instead of sequencing the whole exome or genome, a small fraction of the genome could be prioritised based on prior knowledge. Targeted studies of common variation in candidate genes were often criticised for poor replicability and were largely surpassed by GWAS. However, the findings of GWAS themselves can be used as much more plausible hypotheses for targeted sequencing studies of rare variants. Targeted sequencing offers a cost-effective alternative to whole-genome and whole-exome sequencing, as it provides a full-resolution view of the regions of interest, including the ability to study rare variation, at a fraction of the price of WES.

There have been several successful applications of this approach, especially in the context of immune disease. Nejentsev et al. [128] used a targeted approach to sequence 10 diabetes-implicated genes, discovering four protective rare variants in *IFIH1* – a gene previously associated with type 1 diabetes via GWAS. Similarly, Rivas et al. [154] used targeted sequencing of 759 protein-coding genes that carry common variants previously associated with inflammatory bowel disease. Analysis of targeted sequences, alongside a whole-exome sequencing cohort, identified three protein-truncating variants, one of which (rs36095412, p.R179X), in *RNF186*, was significantly implicated in a follow-up analysis as protective against ulcerative colitis. The variant was replicated in genotyping and whole-genome sequencing data sets. The authors performed functional analysis and demonstrated that the variant is associated with reduced expression and altered subcellular localisation. Protective loss-of-function variants like p.R179X could be used as therapeutic targets, highlighting the value of rare-variant studies.

Targeted sequencing can also be used to investigate noncoding rare variants. Zhao et al. [197] used targeted sequencing to study 2-kb promoter regions in 410 healthy adults. The authors measured transcript abundance from peripheral blood samples using gene expression arrays. Using a burden test that evaluates the distribution of cumulative counts of rare variants in bins of expression, they observed a significant increase in the number of low-frequency and rare variants (MAF < 5%) at both high and low extremes of gene expression. They noted that the average effect size of individual variants is modest and is comparable to that of common disease-associated eQTLs. They also replicated the main findings using a smaller cohort of 75 individuals, for whom whole-genome sequencing and RNA sequencing was performed. The results were partially validated by CRISPR/Cas9 knockdowns in K562 cells. DeBoever et al. [51] used whole-genome sequencing and RNA sequencing to study 215 human induced pluripotent stem cell lines and similarly observed an enrichment of single-nucleotide variants in promoter regions, with most single-nucleotide variants having a small negative effect on gene expression.

### Whole-exome and whole-genome sequencing

As the price of sequencing has fallen, the potential cost saving of targeted sequencing has become outweighed by the benefits of hypothesis-free analyses of rare variation using whole-exome or whole-genome sequencing. These very large data sets have necessitated the development of efficient computational analyses to convert raw sequencing data into high-quality variant calls. Variant calling is used to detect variation at each locus compared with the reference genome and to determine the genotype of each individual (homozygous reference, heterozygous, or homozygous alternate). Algorithms for calling single-nucleotide variants, short insertions and deletions, and structural variants are implemented in a variety of software suites, such as the Genome Analysis Toolkit (GATK) [121], SAMtools [107], and Platypus [153]). Sets of called genotypes go through a variety of quality control filters, such as GATK's variant quality score recalibration (VQSR), to estimate the likelihood of a specific variant being 'true' by comparing its statistical properties to cohort-specific properties of known true sites (e.g., using variants found during the International HapMap Project as a truth set). Variants with low VQSR scores are typically excluded from further analysis. Alternative approaches based on random forests [116] and convolutional neural networks [148] exist. The final clean variant set is ready for association analyses, although in practice quality control is iterative: Elevated $p$-values on quantile–quantile plots and artifacts on Manhattan plots should motivate researchers to perform additional quality control.

The major experimental factor in the overall quality of the variant call set is sequencing depth. In practice, the cost of sequencing is almost linearly proportional to the sequencing depth (the small cost of DNA library preparation is constant), which makes sequencing at lower depth an appealing alternative. Lower depth leads to reduced sensitivity and higher false positive rates. Through computational simulations, Rashkin et al. [152] showed that on a per-sample basis the sensitivity to call single-nucleotide variants plateaus at 25x depth. However, reduced depth may enable researchers to sequence a larger cohort, increasing the overall statistical power to detect genetic associations. In their simulations, given a fixed budget for sequencing, the maximal power to detect rare single-nucleotide variants is achieved at 15–20x depth, while maximising the size of the sequenced cohort.

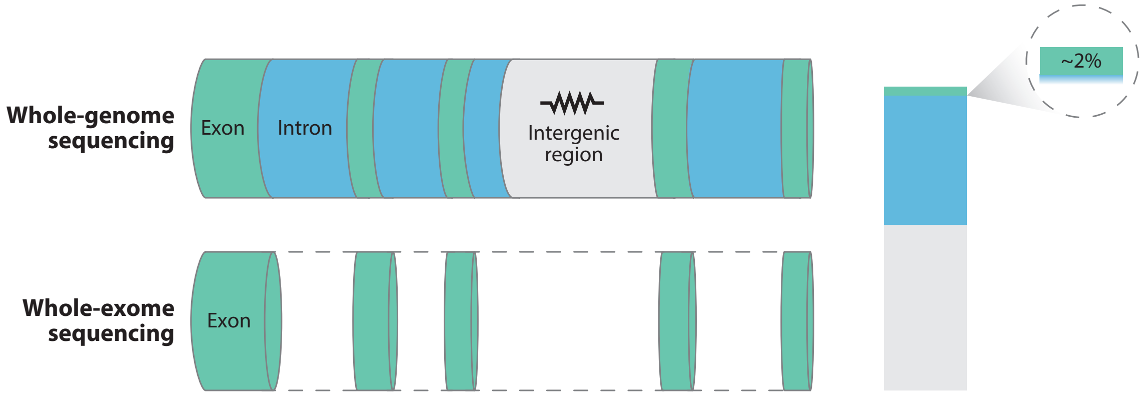


**Figure 1.2** Whole genome sequencing captures all genetic variation, whereas exome sequencing targets the 2% that encodes proteins.

Gilly et al. [68] evaluated very-low-depth sequencing at 1x as an alternative to dense array genotyping. They estimated the sequencing cost to be half of that for genotyping with a modern dense genotyping microarray. After genotype refinement with a custom imputation panel that included approximately 250 population-specific samples, low-depth sequencing achieved 97% concordance with the array data. More importantly, after imputation, low-depth whole-genome sequencing data had denser coverage of low-frequency and rare variation. However, the authors pointed out that performing variant calling and imputation is more computationally expensive in whole-genome sequencing than in genotyping, even at 1x depth. Understanding the trade-off between sample size and depth is essential for designing a rare-variant association study, and this trade-off should be considered when performing power calculations. It is also important to remember the higher false positive rate of lower-depth sequencing, which should motivate a thorough manual validation of discovered associations.

The two comprehensive sequencing approaches in widespread use are to sequence the whole genome in a completely agnostic way and to target the approximately 2% of the genome in exons that encode proteins, known as whole-exome sequencing. Whole-genome sequencing is the gold standard for performing rare-variation association studies, but the cost per sample remains much higher compared with genotyping and exome sequencing. Whole-genome sequencing allows the investigation of rare and common variation in both coding and noncoding regions of the genome (Figure 1.2). The main advantages of exome sequencing are the reduced cost (approximately one-third the cost of whole-genome sequencing in 2019) and the reduced computational burden required to process the data. Additionally, the highest effect-size associations are likely to be found within the exonic regions, justifying the use of WES for studies with small sample size, underpowered to find modest effect-size associations. The obvious drawback is the absence of variants in the noncoding regions, which may be especially relevant in the context of GWAS.

One of the issues facing exome sequencing is that the targeting of exonic sequence is imperfect. Mitigating this problem and ensuring good sensitivity and specificity for variant detection requires higher sequencing depth compared with whole-genome sequencing. For example, to sequence 85% of targeted bases at a depth of 20x or greater, the mean sequencing depth must be approximately 60x [28]. In addition, various protocols target slightly different parts of the genome, which needs to be controlled for when samples in a given analysis have used different protocols; studies that use public data sets as controls in a case–control setting should consider whether their protocol is sufficiently similar to the one used in controls.

Given these drawbacks, how useful is exome sequencing for understanding GWAS regions? It is an appropriate study design choice for conditions that have a known contribution of variants in coding regions (e.g., psychiatric and neurodevelopmental disorders) but may be less suitable for traits where the absolute majority of known variation is in noncoding regions (e.g., height). For example, Singh et al. [169] discovered that rare loss-of-function variants in *SETD1A* are associated with schizophrenia and severe developmental disorders. Subsequent analyses of those data showed that a more general burden of rare, damaging variants in patients who suffer from schizophrenia is concentrated in genes that have also been implicated in schizophrenia GWAS [170, 142]. By contrast, a large exome sequencing study in type 2 diabetes found only a modest contribution of rare coding variants [63].

### 1.3.2 Testing for statistical association with rare variants

Rare-variant association studies pose a number of statistical challenges. While the GWAS methodology can still be used, the smaller sample sizes and the intrinsic infrequency of rare variants have motivated the development of methods for variant grouping.

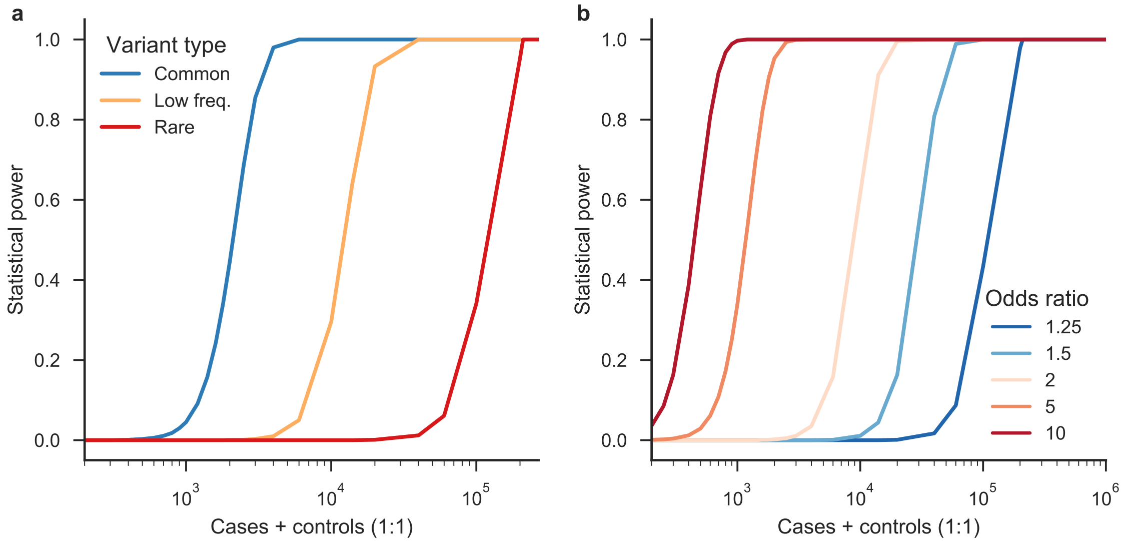


**Figure 1.3** One-stage association study power calculations for single-variant tests using the method described by Johnson et al. [88]. (a) Power to detect common (disease allele frequency = 0.25), low frequency (0.025), and rare (0.0025) with genotype relative risk of 1.5 at $5 \times 10^{-8}$ significance level. (b) Smaller-scale studies are should be well-powered to uncover large-effect semi-Mendelian variants with 1% MAF, but hundreds of thousands of samples will be required to implicate rare low-effect size variants (OR = 1.25), similar to those often found through GWAS.

### Single-variant tests

Standard GWAS statistical tests, such as linear or logistic regression, can be applied to individual rare variants. The challenge, however, is that statistical power to detect association is directly proportional to MAF. Implicating a particular rare variant therefore requires it to either have a very large effect size or be tested in a very large number of samples. While previous association studies have successfully associated single variants that are rarer than typical GWAS variants with complex disease [116], detecting single-variant association with truly rare variants will require enormous sample sizes (Figure 1.3).

Another important caveat to consider when performing single-variant association tests for rare-variant studies is the genome-wide significance threshold. For common variant association studies, a threshold of $5{\times}10^{-8}$ was adopted by the genetics community (Bonferroni correction for the number of LD-independent common variants in the genome at $\alpha$=0.05).

When additionally testing low-frequency and rare variants, the threshold should be adjusted to correct for the increased number of LD-independent variants. Through simulations, Pulit et al. [150] estimated that sequencing studies with fewer than 2,000 samples should use a genome-wide significance threshold of $5 \times 10^{-9}$ for samples of European and East Asian ancestry and $1 \times 10^{-9}$ for samples of South Asian and African ancestry.

### Burden tests

The primary approach used to date to overcome the lower power of single-rare-variant testing has been to collapse multiple variants into a single test of the burden of that class of variants. This both increases the effective frequency of the event being tested (and thus increases power) and reduces the number of tests performed (and thus relaxes the multiple-testing corrected significance threshold). A drawback of variant aggregation methods is the inability to pinpoint the specific variants associated with the disease. There are several approaches for aggregate testing of variants, broadly divided into methods for studying coding variation and methods for studying noncoding variation.

The most biologically informed group of variants for variant aggregation is based on the genes they belong to. In their simplest form, burden tests count the number of minor alleles in rare variants present within the defined region. Within genes, variants may be weighted by MAF (inversely proportional) or by the predicted function of the variants (e.g., missense, nonsense, and silent mutations). Often this takes the form of performing multiple tests per gene, such as including all nonsynonymous variants, just nonsense variants, or each group at different maximum MAF thresholds. For each set of gene-based tests, an accepted genome-wide significance threshold is $2.5 \times 10^{-6}$ (Bonferroni correction for approximately 20,000 genes).

The main drawback of classic burden tests is the assumption of a unidirectional effect of individual variants collapsed into the same group. In cases where individual variants are expected to have different directions of effect (i.e., some risk increasing and some protective), adaptive burden tests like aSum can be used [78]. Variance-component tests perform association by measuring the distribution of effects in a group of variants. Unlike standard burden tests, variance-component tests like C-alpha [126] and the sequence kernel association test (SKAT) [191] are not prone to loss of power owing to bidirectional effects of variants in the group. They are also less sensitive to the inclusion of noncausal variants into the group, allowing relaxation of the MAF cutoff.

To maximise statistical power, burden tests should be utilised when the effects in the group of variants are unidirectional and most of the variants in the group are thought to be causal. In cases when these assumptions do not hold up, variance-component tests should be used. When the genetic architecture of the trait is unknown, combined omnibus tests like the adjusted optimal sequence kernel association test (SKAT-O) should be used. Adjusted SKAT-O simultaneously performs both the burden test and the SKAT test and searches for an optimal way to combine those two statistics.

While collectively testing rare nonsynonymous variation in a particular gene makes biological sense and is especially attractive when leveraging the efficiency of the exome design, the majority of common variants implicated in GWAS are noncoding. It is reasonable to assume that functional noncoding rare variants will play a role in the same diseases, albeit possibly with smaller effect sizes than rare coding variants. Currently, the largest existing whole-exome and whole-genome association studies have tens of thousands of samples – only sufficient for implicating individual rare variants of large effect size. This motivates the development of collapsing tests for noncoding rare variation. Unfortunately, the lack of clear functional groupings (i.e., genes) makes the process of grouping noncoding variants together less straightforward.

Window-based techniques simply group adjacent variants for association testing. The distinct window approach separates them into sequential nonoverlapping regions, while the sliding window technique produces overlapping groups where each window starts with a small offset from the beginning of the previous one (Figure 1.4). The sizes of the window and the offset are typically a few kilobases. Deciding on the precise window size is nontrivial: Very small windows are effectively equivalent to single-variant testing owing to the required number of corrections, while large windows may be too noisy because they include too many variants. In addition, considering the variability of the recombination rate and LD across different regions of the genome, it is hard to tell whether fixed-sized windows are ‘biologically meaningful’ regions [18]. Several techniques for selecting the optimal window size or varying the windows across the genome have been proposed. Browning’s [31] variable-length Markov chain method accounts for the LD pattern between markers. Li et al. [108] used haplotype diversity to estimate the maximum size of the window. Tang et al. [178] proposed a technique based on principal component analysis, which does not require the

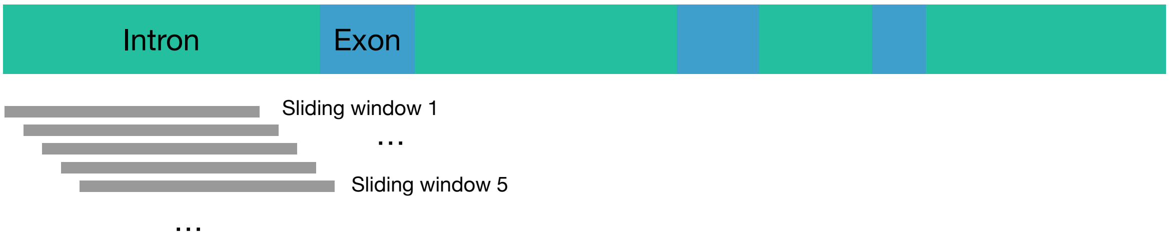


**Figure 1.4** Sliding window technique.

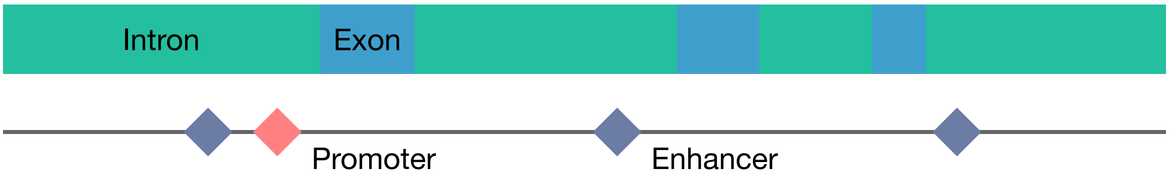


**Figure 1.5** Promoters and enhancers can be used to group non-coding variants, but with less precision than genes.

input data to be phased. Beissinger et al. [18] created a spline-based method that determines the boundaries and variable sizes for windows without requiring prior knowledge of the LD structure.

Functional annotation databases, such as ENCODE [59], can be used to group noncoding variants in an analogous way to genes (Figure 1.5). Natarajan et al. [125] used functional annotations to determine noncoding regions marked as enhancers and promoters. The same study used gene expression data from the Roadmap Epigenomics Project [22] and a chromatin-state model to connect regions annotated as enhancers to the relevant genes. The same approach could be extended to use all of the techniques discussed above for regulatory target annotation in GWAS, demonstrating that methodological development in common-variant association will be useful for understanding rare variation as well. These approaches have not yet been successful at identifying noncoding rare-variant associations, but as the functional annotation databases improve, more noncoding regions will be able to be consistently grouped.

Currently, growing exome sequence data sets are likely to provide the most information about decoding GWAS signals, even though many of the causal variants in GWAS themselves are noncoding. This is because current exome sample numbers are larger than those for whole genomes, and there are demonstrably successful approaches to grouping variants that improve the power to test for association. Of course, this approach can work only where the same locus is influenced by both rare coding variants and common noncoding variants; it is not known what fraction of all GWAS signals have this property. It will be important to develop better databases of noncoding function and to increase the number of sequenced whole genomes available for analysis.

### 1.3.3 Special study designs in the sequencing era

Most of this chapter has focused on genetic studies of unrelated individuals, either in a case–control design or a quantitative trait design. While these approaches have been the workhorse for the GWAS approach, the field of human genetics has a long history of more creative study designs that are being revisited in the era of low-cost sequencing.

#### Families and population isolates

Studying extended families with multiple individuals with a shared phenotype led to the discovery of genetic causes of hundreds of single-gene disorders, such as *CFTR* in cystic fibrosis, and a small number of high-penetrance variants for more complex diseases, like *BRCA1* in breast cancer. Whereas these studies first identified where in the genome the relevant gene is (via linkage analysis) and then sequenced only that portion, modern sequencing approaches mean that the whole genome or exome can be studied in such families. One such study in a very large inflammatory bowel disease pedigree [105] identified an associated frameshift mutation in *CSF2RB*, as well as a more general burden of GWAS risk variants in the family.

The founder effect in isolated populations provides an opportunity to uncover rare trait-associated genetic variants, that have risen to higher allele frequencies due to drift, in modestly sized cohorts. For example, Southam et al. [173] used whole-genome sequences of 250 individuals from two remote villages in Crete to build a population-specific reference panel. The panel was used to refine imputed genotypes of a larger cohort of 3,200 individuals. They were able to uncover two low-frequency cardiometabolic variants (effect allele frequencies = 0.6% and 1.3%) that were much more frequent in the founder populations compared to the rest of Italy. Tachmazidou et al. [177] uncovered an association between a variant in *APOC3* and several cardioprotective endophenotypes. The variant, R19X, is common in the studied remote Greek population (N=1,267, MAF≈2.3%) but rare in the overall European population (MAF=0.035%).

### Biobanks and risk prioritisation

Large, richly phenotyped cohorts, such as those in the UK Biobank, have enabled sweeping association studies of thousands of traits for hundreds of thousands of genotyped individuals. The power to detect rare variants in case–control and quantitative trait studies relies on the availability of large cohorts. While the cost of genotyping and sequencing is rapidly decreasing, it remains prohibitive in these types of large studies. At the same time, assembling large cohorts is challenging and time consuming.

One way to use biobanks is via polygenic risk score (PRS), which is a sum of condition-associated signals weighted by their effect size. PRS can be used to estimate the genetic component of an individual's disease risk. Jostins et al. [91] argued that individuals with known disease status and low PRS should be prioritised for recruitment into new rare-variant studies. A low PRS may be explained by previously undiscovered risk factors that drive the disease. PRS calculation requires the availability of genotype data for considered samples. The technique increases the power to uncover novel genetic variants compared with random selection.

While previously separate, the worlds of GWAS and rare-variant sequencing studies are now clearly intertwined. Some of the biggest challenges with GWAS interpretation, such as the resolution of causal variants, can be partially resolved by the careful incorporation of rare-variant data. Similarly, both methodological and biological discoveries in GWAS can inform best practices in the growing field of rare-variant studies.

## 1.4 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a group of immune-mediated auto-inflammatory disorders, primarily affecting the gastrointestinal (GI) tract. Crohn's disease (CD) and ulcerative colitis (UC) are the two most common types of IBD. Affecting millions of people around the world, IBD has a severe impact on patients' quality of life in the prime years of their lives. Neither CD or UC are currently curable and both require life-long treatment to alleviate symptoms. While IBD-tailored pharmaceutical therapies have existed since the early 1950s, the majority of patients eventually stop responding to a given treatment or never respond to it in the first place, which requires further therapy escalation. Due to therapy failure, around 70% of CD patients and 20% of UC patients eventually have to undergo life-changing surgical intervention.

The pathogenesis of IBD is not fully understood. Epidemiological factors like smoking, urban lifestyle, and westernised diet have long been associated with the risk of IBD. Physiological phenomena like the 'leaky gut' (increased intestinal permeability) are thought to play a role in the development of IBD. Despite the abundance of known risk factors and potential mechanistic hypotheses, none have been proven to be the be-all and end-all explanations of the condition.

### 1.4.1 Manifestation

Both CD and UC are characterised by chronic inflammation, which eventually results in damage to the GI tract. The two conditions differ in the way the inflammation manifests itself: CD can affect most of the GI tract, though usually is restricted to the small intestine; UC affects the colon or the rectum (Figure 1.6). Crohn's inflammation is 'patchy' with inflamed regions neighbouring areas of unaffected tissue, while inflammation in UC is more continuous. CD inflammation often spreads throughout several layers of the gastrointestinal wall, while in UC the inflammation is restricted to the the topmost layer of the wall – the colonic mucosa.

5–23% of IBD patients cannot be unambiguously diagnosed with either UC or CD, as their macroscopic and histologic features can be attributed to either of the two conditions. This condition is referred to as IBD-unclassified (IBD-U). While some of the patients are eventually reclassified and diagnosed with one of the two classic IBD sub-types as the disease progresses, 20–60% retain the IBD-U status years and decades after the initial diagnosis [99].

Despite the phenotypic differences, UC and CD are known to have similar genetic architecture (genetic correlation, $r_G$=0.68 [87]), with the majority of known significant genetic associations having a similar effect in both conditions. Recently, genetic correlations were used to demonstrate that IBD may be appropriate to reclassify as three distinct conditions – ileal Crohn's, colonic Crohn's, and ulcerative colitis [40].

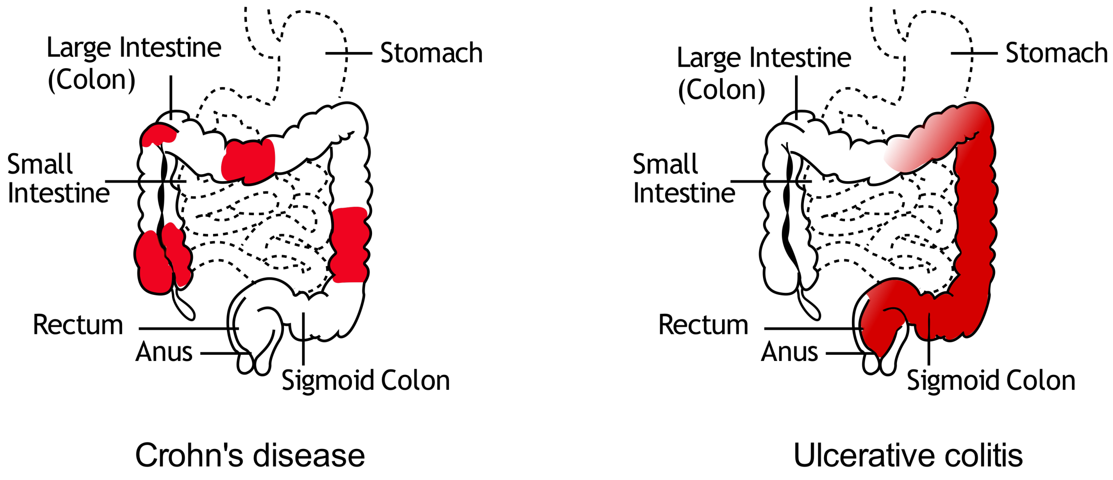


**Figure 1.6** Crohn's disease and ulcerative colitis are two common types of inflammatory bowel disease which differ in patterns and location of inflammation. License information: *Modified work, derivate of File:Patterns of Crohn's Disease.svg by Samir, vectorized by Fvasconcellos [CC BY-SA 3.0]*

Very early onset IBD (VEO IBD) is another related condition that manifests very early in the patient's life. Due to its rarity (2.2–13.3 in 100,000), the pathogenesis and the genetic architecture of VEO IBD is poorly understood. The disease is thought to be frequently monogenic, compared to the complex, polygenic paediatric and adult-onset IBD. When considering the difference in disease architecture, progression and the common non-responsiveness of VEO IBD patients to the typical IBD treatments, it is sometimes argued that VEO IBD is a distinct disorder that should not be grouped together with other IBD sub-types [168].

While the IBD type classification is an active area of research, clinically either the binary (UC/CD) or the ternary systems (UC/CD/IBD-U) are used to diagnose adult-onset patients. Throughout this thesis, I will use the ternary system when discussing the IBD sub-types.

Excluding the VEO-IBD patients, the peak age of onset of IBD is around 15–29 years old [89]. Some studies suggest that the age of onset is bi-modal with the second peak occurring at the age of 50–70, though this is not observed in some comparable cohorts with stricter inclusion criteria [52].

In addition to chronic inflammation, UC and CD share a number of other symptoms: severe abdominal pain, diarrhoea, weight loss, and chronic fatigue. These, along with some of the treatment side effects (e.g., increased infection rates due to immunosuppressant therapy) have a severe impact on most patients' quality of life.

### 1.4.2 Epidemiology

#### Prevalence and incidence

IBD affects millions of people globally. The disease incidence is population-specific and is thought to depend on environmental and genetic factors.

In Western countries, the rates of IBD have been steadily rising since around the 1850s. Globally, the increase in prevalence has been associated with westernisation of diet and lifestyle, starting in the 1950s [94].

A recent large-scale meta-analysis by Ng et al. (2017) [130] estimates that the prevalence of inflammatory bowel disease exceeds 0.3% in North America, Oceania, and many countries in Europe, though the global rates are thought to be lower. Looking at the historical data, authors estimate that the incidence is stabilising or potentially getting lower in the West, yet is still rising in the newly industrialised countries in Africa, Asia, and South America.

However, the recent work by Jones et al. (2019) [90] utilising data from the Lothian IBD Registry (Edinburgh, Scotland, UK) estimates the current prevalence to be 0.78% and predicts that it will increase up to 1% by 2028. Compared to the previous comprehensive UK-wide estimate of 0.37% from Stone et al. (2003) [175], it appears that the prevalence of IBD is still rising in some Western countries.

### 1.4.3 Disease aetiology

#### Environmental factors

The rise in IBD rates has been strongly associated with the industrialisation and urbanisation in both the Western and the newly industrialised countries [94]. A population-based study of Canadian immigrants demonstrated that the age of immigration is negatively associated with the risk of IBD (14% increased risk per earlier decade of life at immigration) [19]. A meta-analysis of IBD incidence rates between rural and urban environments indicated that urban residents have a higher risk of developing IBD (incidence rate ratio UC=1.17 and CD=1.42) [171]. However, inclusion of rural cohorts from low-incidence regions might limit the generalisability of the study [2].

### Diet and lifestyle

Despite the evident role of Westernisation as a risk factor for IBD, the exact factors influencing the pathogenesis are not well-understood.

Smoking is one of the most-studied lifestyle risk factors for IBD. Most epidemiological studies have consistently associated smoking with an increased risk of Crohn's disease [36]. Paradoxically, non-smokers and ex-smokers appear to have a higher risk of ulcerative colitis [36], though the evidence for this is arguably weaker.

At the same time, while the rates of smoking have been reducing in the Western countries since around the 1980s, the rates of Crohn's disease have grown. Moreover, a comparison of risk factors between Australia and eight newly-industrialised counties in Asia concluded that smoking quadruples the risk of Crohn's disease in Australia, yet was not a risk factor amongst the newly industrialised countries [131, 93].

The 'hygiene hypothesis' links greater urbanisation with decreased exposure to microbes in early life, resulting in abnormal bacterial recognition that leads to IBD. Kondrashova et al. compared the prevalence of transglutaminase antibodies and coeliac disease amongst children in Russian Karelia and Finland. The neighbouring locations largely share the same population history, yet differ in socioeconomic environment [101]. The Finnish cohort had a higher rate of antibodies and disease prevalence, consistent with the hygiene hypothesis. However, while coeliac disease is also an immune-mediated disorder, it is not clear whether the findings can be directly applied to IBD. In fact, a study by Ng et al. indicates that the presence of a hot water tap and flush toilet in childhood were protective against UC development [131].

Dietary preferences have been linked to a risk of inflammatory bowel disease. Ananthakrishnan et al. reported an association between higher intake of dietary fibre and lower risk of Crohn's disease [6]. In [5], the authors report a tentative association between trans-unsaturated fats and risk of ulcerative colitis.

As with the majority of epidemiological studies, establishing the causality between diet and lifestyle, and the outcome (i.e., IBD) remains challenging. Many of the epidemiologic risk factors tend to have contradicting effect (or absence of thereof) in different studies. Techniques like Mendelian Randomisation (MR), which use genetic variants as natural experiments, might be useful in verifying the causality of the factors. For example, Lund-Nielsen et al. recently reported the absence of evidence for a causal effect between vitamin D deficiency and development of IBD [115]. Projects like the UK IBD BioResource and PREdiCCt, which aim to create large cohorts of IBD patients along with their genetic information and lifestyle questionnaires, will be instrumental in enabling such studies.

### 1.4.4 Therapies and treatment options

The heterogeneity of IBD manifestation and progression makes treatment choices for IBD nontrivial. A variety of IBD therapies exist, yet the choice of therapy is often made difficult by the lack of sufficient head-to-head efficiency comparison studies. For the majority of the treatments, there is a trade-off between drug efficacy and toxicity – more effective therapies (e.g., anti-TNF) have more side effects and have a larger burden on the patients' quality of life. In addition, disagreement exists on the way the treatment should be escalated.

The 'treatment pyramid' (Figure 1.7) is often used to describe the most common therapeutic algorithm in the UK. At the bottom of the pyramid are the relatively safe treatment choices (e.g., antibiotics and 5-aminosalicylic acid [5-ASA]) while at the top are the more severe or toxic, yet effective, options (e.g., mono- or combination therapy with biologics). Conventionally, the treatment would get escalated from the safer options to the more severe ones after the patient stops responding to the current 'level' (step-up approach).

More recently, some practitioners have argued for the adoption of the top-down approach – where the more aggressive, yet more effective treatments are prescribed soon after the initial diagnosis. Early treatment with more radical treatments is thought to prevent the GI damage that occurs when a patient's current (milder) therapy stops preventing flareups.

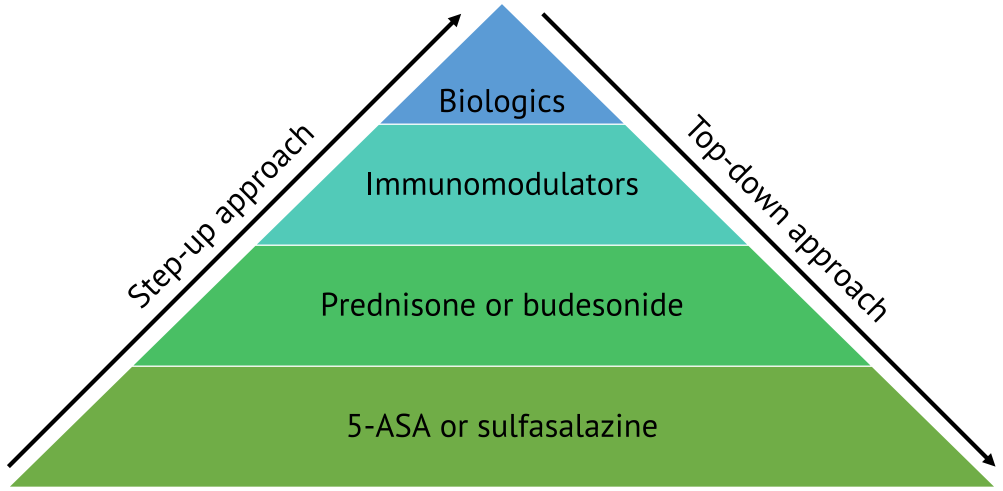


**Figure 1.7** IBD treatment pyramid. Adapted from Aloi et al. [4]. Step-up approach starts with milder, less toxic therapies. Top-down approach prioritises early aggressive treatment (e.g., with biologics).

While there are some studies demonstrating that the top-down approach might be more beneficial for patient outcomes (e.g., [53]), there is no definitive study demonstrating its overall advantage.

In some cases, combination therapy of several types of drugs is prescribed to the patients. For example, anti-TNF therapy (biologic) is frequently combined with immunomodulators like the azathioprine to reduce the risk of anti-drug antibody development [96].

### 1.4.5 Genetic component of inflammatory bowel disease

Inflammatory bowel disease is known to have a strong genetic risk component: early twin and family studies have demonstrated high heritability of the disease (>60%). Following these, linkage studies uncovered the role of the *NOD2* gene in the disease risk. Despite the high effect size, the uncovered variants explained only a small fraction of the expected total heritability, suggesting that the disease is not monogenic. The arrival of affordable genome-wide genotyping arrays around 10 years ago has enabled large-scale genome-wide association studies (GWAS).

Arguably, IBD has become one of the most well-studied conditions, with more than 240 loci currently implicated as disease risk factors. One of the most interesting findings of the GWAS era in IBD genetics is the unequivocal demonstration of the disease complexity – the known risk loci are associated with a broad variety of biological functions ranging from immune activation and defective barrier function to bacterial recognition and cell signalling.

Moreover, IBD has been shown to have non-negligible genetic correlation with a variety of other immune-mediated conditions which do not manifest in the GI tract and are not similar symptomatically (rheumatoid arthritis, primary sclerosing cholangitis). These observations underscore the complexity and heterogeneity of IBD pathogenesis. The realisation that IBD is not a single-cause and single-solution condition might be disheartening, but the unmet clinical need and the complexity of the task should motivate further research into the causes, progression and treatment of IBD.

While the largest association studies of IBD include tens of thousands of cases and controls, one could make an argument that the field is far from reaching the saturation point.

Firstly, even larger cohorts are required to implicate genetic variants with modest effect, which will be required for leveraging individual genetic profiles as a prognostic tool. Watanabe et al. [189] estimate that 0.06% of SNPs are causally associated with IBD, which would require a cohort of $\sim$1 million subjects to detect 90% of them at a genome-wide significant level.

Secondly, technological restrictions of genotyping arrays and available imputation panels do not allow us to look for associations of low-frequency, potentially high-effect genetic variants, which are thought to be more trivially translatable into drug targets. A common criticism of GWAS is that the number of known associations far exceeds the number of well-understood associations, as the post-GWAS followup (e.g., functional and model work) tends to take substantial amounts of time. However, it is not clear whether the known associations between the common variants and IBD provides us with a straightforward route to drug-targets and, ultimately, novel therapies. Two decades after uncovering the association between several variants in *NOD2* and Crohn's disease, there are no commercial therapies targeting *NOD2* [47]. The central role of *NOD2* in several important pathways relating to intestinal barrier integrity and immune homeostasis makes targeting it prone to adverse dysregulation. Recent work by O'Connor et al. [135] suggests that the extreme polygenicity of the majority of common diseases might be an artefact of purifying selection. Their analysis suggests that the genes and loci most critical to the disease pathogenesis may differ from those with the strongest common-variant associations.

Lastly, the first ten years of IBD GWAS has primarily concentrated on finding genetic associations of disease risk. While these findings are now actively used to develop the next generation of therapeutics, there are ways to utilise genetics for improving the patient care of today. Recently, the availability of phenotypic data from electronic health records and clinical trials has enabled the use of the same association techniques for finding genetic variants directly associated with disease progression and therapeutic response.

### Clues from observational epidemiology

The partially heritable nature of IBD was recognised as early as 1909 due to the prevalence of ulcerative colitis in several family members [3]. 5–23% of IBD patients are estimated to have at least one first-degree relative affected by IBD [58]. These early insights have been strong indicators of an existing genetic component of the disease pathogenicity.

Comparison of the disease amongst monozygotic and dizygotic twins observed a higher disease status concordance in monozygotes, suggesting that familial IBD is not solely caused by the shared environment. In addition, twin studies have estimated the overall heritability to be 0.75 (CD) and 0.67 (UC) [72].

Epidemiological studies were also used to study the prevalence of the disease amongst different ethnic groups. Rozen et al. (1979) reported higher incidence rates of CD amongst the Ashkenazi versus the 'non-Ashkenazi' Jews living in Tel-Aviv [159], hypothesising that, considering the largely shared environment, the higher predisposition might be linked to a 'hereditary predisposition'.

### Linkage studies

Genetic linkage analysis is a technique that allows the identification of large chromosomal segments that cosegregate through a family with a certain phenotype. Analysis of families with multiple members affected by Crohn's disease allowed the identification of a locus on chromosome 16 associated with the condition (IBD1 locus). Later, other CD-associated loci on chromosome 16 were discovered and replicated. The loci were mapped to the *NOD2* gene which remains one of the canonical IBD genes. The *NOD2* associations, uncovered via linkage analysis, have a high odds ratio ($\sim$1.5–2.5). Despite the strong effect size and high frequency, *NOD2* loci explain only a small fraction of the IBD heritability, suggesting the polygenic nature of the condition. The lack of further associations uncovered via linkage analysis was, perhaps, disappointing at the time but did indicate that the majority of other IBD associations are either less frequent or have a smaller effect size.

### Genome-wide association studies

The early success in uncovering the role of *NOD2*, was ultimately followed by a disappointing lack of new, replicated results from further linkage and candidate gene studies.

The advancements in microarray genotyping and the understanding of the linkage disequilibrium have led to the arrival of genome-wide association studies (see 1.1.1). The technique was quickly applied to studying the genetic variants associated with IBD. Yamazaki et al. [192] conducted the first genome-wide association for Crohn's disease (2005), reporting a locus in *TNFSF15* to be associated with IBD in the Japanese population. Shortly thereafter, at least eight other small-scale GWAS' (500–2,000 cases) followed (2006 to 2008) [110]. The early insights have contributed to our understanding of Crohn's: associations in *ATG16L1* and *IRGM* demonstrated the role of autophagy in CD pathogenesis. In addition, the studies have highlighted that the majority of the uncovered common association had a modest effect size (OR < 1.3).

The creation of the International IBD Genetics Consortium (IIBDGC) facilitated efforts to meta-analyse the individual cohorts. In 2008, the first meta-analysis of Crohn's disease reported 30 novel and replicated genetic associations [15]. Shortly after that, GWAS of ulcerative colitis [62] and joint analysis of IBD followed [92]. The most recent IIBDGC analysis by de Lange (2017) [49] increased the total number of IBD-associated loci to 240. The study identified three three loci which contain integrin genes, which have recently become important therapeutic targets in IBD. Two monoclonal antibodies, vedolizumab and etrolizumab, that target the $\alpha4\beta7$ dimer (encoded by *ITGA4* and *ITGB7*) have demonstrated efficacy in IBD treatment, demonstrating relevance of GWAS to therapeutic target discovery.

The IIBDGC has led other association studies that looked into specific aspects of IBD pathogenesis and genetic architecture. Liu et al. [111] meta-analysed patients of European (n=86,640) and East Asian, Indian or Iranian ancestries (n=9,846) highlighting the consistency of direction and magnitude of effect of the genetic associations across individuals of different ancestries. Goyette et al. [73] used the IIBDGC data to study the contribution of the HLA alleles in IBD pathogenesis and reported a large-effect association between a common HLA allele HLA-DRB1*01:03 and ulcerative colitis (OR=3.59).

### Whole-genome and whole-exome sequencing association studies

Luo et al. [116] used low coverage whole-genome sequencing to study the contribution of low-frequency and rare variants in IBD. The low sequencing depth (2x for the UC cases, 4x for the CD cases, and 7x for controls) limited the sensitivity to detect the full range of genetic variation present in the cohort (e.g., INDELs), but was sufficient for building a custom imputation panel. After imputing the array genotyping data of around 27,000 individuals with the custom panel, association to a low-frequency (0.7%) missense variant in *ADCY7* was detected. The variant (p.Asp439Glu) doubles the risk of ulcerative colitis, consistent with theoretical and empirical observations that rare, evolutionary recent variants may have a higher effect size in complex disease traits. In addition, the authors used gene-based tests to demonstrate a burden of rare, damaging mutations in known IBD genes.

### Prediction of IBD via polygenic risk scores

One of the main objectives in of human disease genetics is the ability to predict whether an individual will develop a disorder based on their genetic makeup. While this is possible in rare cases of monogenic late-onset disease (e.g., Huntington's disease), the uncovered polygenicity of the majority of complex disease has made this a difficult task.

Polygenic risk scores are a family of methods that leverage the results from genome-wide association studies in order to assign a score on a liability scale based on the individual's genetic makeup. In their simplest form, the odds ratios of independent genetic variants passing a certain p-value threshold (e.g., $< 5x10^{-8}$) are added up and then transformed to assess the individual's risk compared to the known cases. Recently, more advanced approaches like LDPred have emerged. LDPred builds a score based on millions of variants across the genome, while adjusting the regression betas for linkage disequilibrium. Intuitively, this is done in order to take into account the betas of genetic variants that are not passing a strict p-value threshold, while expecting the false positive and false negative variants to balance each other out contributing no false risk change.

Another recent methodological trend is to look at the ends of the liability score distribution when evaluating the utility of the model. While a marginal increase (e.g., +10%) in disease risk might not be meaningfully clinically actionable, the individuals at the end of the PRS distribution appear to be at threefold or greater risk of the tested condition [97].

Khera et al. evaluated the utility of LDPred-derived PRS for coronary artery disease (CAD), atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer [97]. For coronary artery disease the authors identified 8% of individuals at threefold or greater risk. They argue that such risk is comparable to that of familial forms of CAD for which single-variant genetic tests are carried out in many healthcare systems, thus PRS should be considered for inclusion in clinical practice. For IBD the percentage of ‘high risk‘ individuals was more modest – 3.2%, potentially due to the smaller sample size of the GWAS that the PRS was built upon.

Arguably, a nuanced approach is required when evaluating the inclusion of PRS into healthcare practice. Firstly, CAD has a much higher prevalence in most Western countries ($\sim$5%) versus $<$1% for IBD. Therefore, even a threefold increase risk of the disease is quite marginal for IBD. Secondly, compared to CAD where prophylactic therapy with statins and *PCSK9* inhibitors are known to reduce the LDL cholesterol, and ultimately reduce the incidence rate and mortality), there are no known prophylactic measures for IBD. Generic advise to stop smoking and ‘eat healthier’ could be made, but it is highly unlikely this will have a meaningful effect on the risk. Perhaps, pre-onset screening of high IBD risk individuals could be warranted, in order to rapidly diagnose the patient once the disease develops, but it is important to remember that the screening (colonoscopy) itself is quite invasive.

Undoubtedly, the accuracy of polygenic risk scores will continue to improve. Larger GWAS’ and the inclusion of rare pathogenic variants will allow the identification of patients with an extremely high risk of IBD. Clinical trials could be carried out to identify whether any of the existing low-burden drugs can be used as a prophylactic therapy for high-risk individuals. PRS for drug response could be developed in order to identify the therapy to which the patient is most likely to respond. Some patience and hard work will be required to achieve this goal.

### Pharmacogenetic studies

Pharmacogenetic studies aim to bring us closer towards the ultimate goal of personalised medicine: prescribing the safest and most effective drug for each patient. It is difficult to argue that this goal has been achieved for any complex disorder, yet the pharmacogenetic association studies have already uncovered a handful of common and rare variants that influence the safety or efficiency of a given therapy for a particular patient. Several pharmacogenomic association studies for drugs used in IBD have been carried out.

Around 20% of patients have to discontinue treatment with thiopurines due to adverse drug reactions (ADRs). Thiopurine-induced myelosuppression is one of the more severe ADRs. Variants in the thiopurine S-methyltransferase (*TPMT*) gene lead to a insufficient TPMT activity and result in cytotoxic 6-thioguanine nucleotide (6-TGN) metabolite formation [30].

While a pre-treatment enzyme activity test is currently more prevalent in most healthcare systems, including the NHS, a genotyping-based test is thought to be more reliable. For example, the genotyping test is not affected by recent blood transfusions and can be administered after the start of the treatment [187]. Several public and commercial providers (e.g., FDA-approved test from 23andMe) routinely offer the genotyping-based test to the patients.

More recently, several genetic variants outside *TPMT* affecting the risk of myelosuppression were reported. Walker et al. [186] performed a genome-wide and exome sequencing association study of 398 myelosuppression cases and 679 matched controls of European ancestry. The study replicated the known association in *TPMT* and showed that three variants in *NUDT15* are strongly associated with thiopurine-induced myelosuppression (OR=27.3; 95% CI, 9.3 to 116.7). The association was previously reported amongst patients of East Asian ancestry [194].

Several associations for anti-TNF response and immunogenicity were reported. They are described in detail in Chapter 1.

In this thesis, I will describe three projects that were carried out during my PhD. In Chapter 2, I will describe the largest genome-wide association study for immunogenicity to anti-TNF therapy to date. In chapter 3, I describe a genome-wide association analysis for thiopurine-induced liver injury – an adverse side effect than occurs in approximately 10% of patients who undergo thiopurine therapy. In Chapter 4, I describe the production, the quality control, and the initial findings from IBD 15x – a whole-genome sequencing study, meant to uncover the role of rare coding and non-coding variation in IBD. Finally, in Chapter 5 I discuss the future work required to further improve our knowledge of inflammatory bowel disease genetics.

# Chapter 2

# HLA-DQA1*05 is associated with immunogenicity to anti-TNF therapy

*Parts of this chapter were published in Gastroenterology [165]. This project is a collaboration of more than 400 principal investigators, research nurses, technicians, and other professions who have contributed to it. My contribution was to co-design and to run the majority of the genetic analyses, with the exception of genotype imputation, HLA to amino-acid translation, and HLA typed versus imputed concordance. I am grateful for the opportunity to be involved in this study and would like to acknowledge the contributions of other co-authors.*

## 2.1 Introduction

Biological therapies, commonly known as biologics, are typically large and complex proteins manufactured in, or derived from, living sources. Biologics have transformed the management of immune-mediated diseases and are are starting to find their application in the treatment of HIV and cancer. In 2017, biologics accounted for a global expenditure in excess of \$100 billion [25]. Compared to the traditional small molecule drugs, biologics offer greater specificity, resulting in better effectiveness and less off-target effects during the treatment [138].

Globally, one of the biggest challenges to the wider adoption of biologics is cost. Just in 2017/2018, the National Health Service (UK) spent £400 mil on a single biologic therapy – adalimumab, which is was prescribed to 46,000 patients. Generally, for the immune-mediated disorders, a course of treatment with biologics far exceeds the cost of an alternative treatment with small-molecule therapies. The high costs are explained by several factors. Like with most modern therapies, the cost of development for novel a drug ‘from lab bench to the market’ can often be in the range of billions of dollars. Expectedly, the pharmaceutical companies use the time during which the drug is protected by the patents to maximise the profit in order to cover the development costs, subsidise the development of future therapies, and to make a profit. Such a costing strategy is not unique to biologics, but it is exacerbated by the cost and complexity of the development and production of the generic version of the biologics – biosimilars.

Blackstone et al. [25] estimate that it costs \$1 million to \$4 million to bring a generic for a small-molecule drug to the market. In contrast, it takes 7 to 8 years to develop a biosimilar, at a cost of between \$100 million and \$250 million. Sadly, even the arrival of the first competitive biosimilar does not always bring down the pricing of the treatment. Dave et al. [46] estimate that, on average, the price of the first generic drug is 87% of that of the brand-name drug and 60% once four generic manufacturers have entered the market. As of March 2019, there are ten adalimumab biosimilars, produced by six manufacturers, approved for use in the European Union. According to the NHS estimates, the arrival of the adalimumab biosimilars will allow savings of ‘at least £150 million per year by 2021’. The cost of the treatment must not limit the prescription of the biologic therapies to the patients who will benefit from them, but there is a real urgency in understanding who exactly those patients are.

The high cost of treatment is not the the only issue facing the biologic treatments. The current generation of biologics rely on invasive injections because of their poor bioavailability via the oral route [8]. Some biologics, like infliximab, have to be delivered intravenously (IV) every few weeks, putting an additional burden on the patients' quality of life. Others, like adalimumab, can be used via an autoinjector pen. Future biologics, currently in development, might be delivered orally (e.g., OPRX-106 that has reached the phase 2 trial for UC).

Most importantly, just like the majority of other therapies, biologics are not always effective. Considering that biologic therapies are often used for patients refractory to other treatments, it is important to study the risk factors that influence this. In the chapter below, I describe results from the PANTS study. Our work has uncovered the first robust genetic association for development of antibodies against anti-TNF which are know to have a substantial negative impact on treatment efficacy.

Anti-tumour necrosis factor (anti-TNF) therapies are the most widely used biologics for treating immune-mediated diseases. Anti-tumour necrosis factor (anti-TNF) therapy has been effective in the treatment of UC and CD, as well as a variety of other immune-mediated complex traits, including rheumatoid arthritis and refractory asthma. Anti-TNF therapies interferes with the action of TNF, a proinflammatory cytokine that is involved in the innate immune response. In 2018, anti-TNF therapies accounted for a global expenditure in excess of $23.5 billion [86].

For Crohn's disease patients, anti-TNF therapy has been associated with mucosal healing, improved quality of life, and reduced hospitalisation and surgery rates [17]. Despite the general efficacy of the treatment, 10–30% of patients do not respond to anti-TNF and a further 23–46% of patients lose response over time [156]. As with other biologic therapies, repeated administration of anti-TNF often induces the formation of anti-drug antibodies (immunogenicity) leading to treatment failure [96].

Immunogenicity is more common in patients treated with infliximab (a murine-human chimeric monoclonal antibody) than adalimumab (a fully human monoclonal antibody) and is a major cause of low anti-TNF drug level, infusion reactions, and non-remission in patients with Crohn's disease [96, 183]. Combination immunomodulator therapy reduces the risk of immunogenicity to both adalimumab and infliximab, and for infliximab, improves treatment outcomes [96, 41, 140]. Despite these benefits, many patients are still treated with anti-TNF monotherapy because of concerns about the increased risk of adverse drug reactions, opportunistic infections, and malignancies associated with combination therapy with immunomodulators [54, 109, 139].

The ability to identify patients at increased risk of immunogenicity would direct the choice of anti-TNF treatment and the use of preventative strategies, including combination therapy with immunomodulators. However, our understanding of the cellular and molecular mechanisms underpinning immunogenicity to biologics is limited. Retrospective small-scale studies have suggested variants in *FCGR3A* [158], CD96 [11] and *HLA-DRB1* [23, 112] increase susceptibility to immunogenicity to anti-TNF therapy. These associations either did not achieve genome-wide significance [23, 112, 158] or are yet to be independently replicated [158, 11]. Both studies reporting the *HLA-DRB1* only looked at a small subset of HLA alleles, which is problematic given the long-range linkage disequilibrium within the human leukocyte antigen (HLA) region. The CD96 was uncovered in a small discovery cohort (N=62, OR = 20.2, 95% CI, 5.57–73.27, P = $1.88\times10^{-9}$), yet has a strikingly different effect size in a larger self-replication cohort (N=88, OR = 1.16, 95% CI, 1.09–1.23, P = 0.044), violating the homogeneity of odds assumption for replication. Here, I report the first genetic locus robustly associated with immunogenicity to anti-TNF therapies.

## 2.2 Methods

### 2.2.1 PANTS study: patient recruitment and phenotyping

The Personalising Anti-TNF Therapy in Crohn's disease (PANTS) study is a UK-wide, multi-centre, prospective observational cohort reporting the treatment failure rates of the anti-TNF drugs infliximab (originator, Remicade [Merck Sharp & Dohme, UK] and biosimilar, CT-P13 [Celltrion, South Korea]), and adalimumab (Humira [Abbvie, USA]) in 1,610 anti-TNF-naive patients with Crohn's disease [96].

The South West Research Ethics committee approved the study (REC reference: 12 / SW / 0323) in January, 2013. Patients were included after providing informed, written consent. The protocol is available online (www.ibdresearch.co.uk).

At inclusion, subjects were aged 6 years or over and had active luminal Crohn's disease involving the colon and/or small intestine. Choice of anti-TNF drug and use of concomitant immunomodulator therapy was at the discretion of the treating physician as part of usual care. Patients were initially studied for 12 months or until drug withdrawal. In the first year, study visits were scheduled at first dose, post-induction (weeks 12–14), weeks 30, 54, and at treatment failure. For infliximab-treated patients, additional visits occurred at each infusion. After 12 months, patients were invited to continue follow-up for a further two years. Drug persistence was defined as the duration of time from initiation of anti-TNF therapy to exit from the study due to treatment failure. Patients who exited the study for other reasons, declined to participate in the two-year extension, or were lost to follow-up were censored at the time of last drug dose or study visit.

At each visit, serum infliximab or adalimumab drug and anti-drug antibody levels were analysed using total antibody enzyme linked immunosorbent assays [145]. The total antibody, unlike the more commonly reported free antibody assay, includes a drug-antibody disassociation step that allows the assessment of anti-drug antibodies in the presence of drug. In this study, immunogenicity was defined as an anti-drug antibody concentration of $\geq$10 AU/ml, irrespective of drug level, at one or more time points.

We assembled an independent cohort to replicate significant findings from the discovery cohort. This comprised 107 Crohn's disease, 64 ulcerative colitis, and 7 IBD type-unclassified patients all with cross-sectional drug and antibody levels measured as part of routine clinical practice. The samples were genotyped using either the Illumina CoreExome array (N=164) [49] or the Affymetrix 500k array (N=14) [190]. Quality control and imputation methods were the same as in the discovery cohort (see 2.2.3).

### 2.2.2 Measurement of drug and anti-drug antibody levels

Antibody and drug level measurements were carried out by our collaborators at the University of Exeter.

Serum infliximab and adalimumab drug levels were analyzed on the Dynex (Chantilly Virginia, USA) DS2 automated Enzyme-Linked ImmunoSorbent Assay (ELISA) platform, using the Immundiagnostik (Immundiagnostik AG, Bensheim, Germany) IDKmonitor® drug (K9655 infliximab drug level and K9657 adalimumab drug level) and total antibody ELISA assays (K9654 infliximab total anti-drug antibody and K9651 adalimumab total anti-drug antibody). These assays allow quantitative determination of free infliximab and adalimumab using a sandwich ELISA technique. The IDK monitor infliximab drug level assay has a measuring range of 0.8–45 mg/L, with an intra-assay CV of <9.7% and an inter-assay coefficient of variation (CV) of <11.0%. The IDK monitor adalimumab drug level assay has a measuring range of 0.8–45 mg/L, with an intra-assay CV of <2.6% and an inter-assay CV of <13.0%. Positive anti-drug antibody status was defined in line with the manufacturer's recommendations as a concentration $\geq$10 AU/mL, irrespective of drug level. All assays, including drug and antibody levels, were tested for stability.

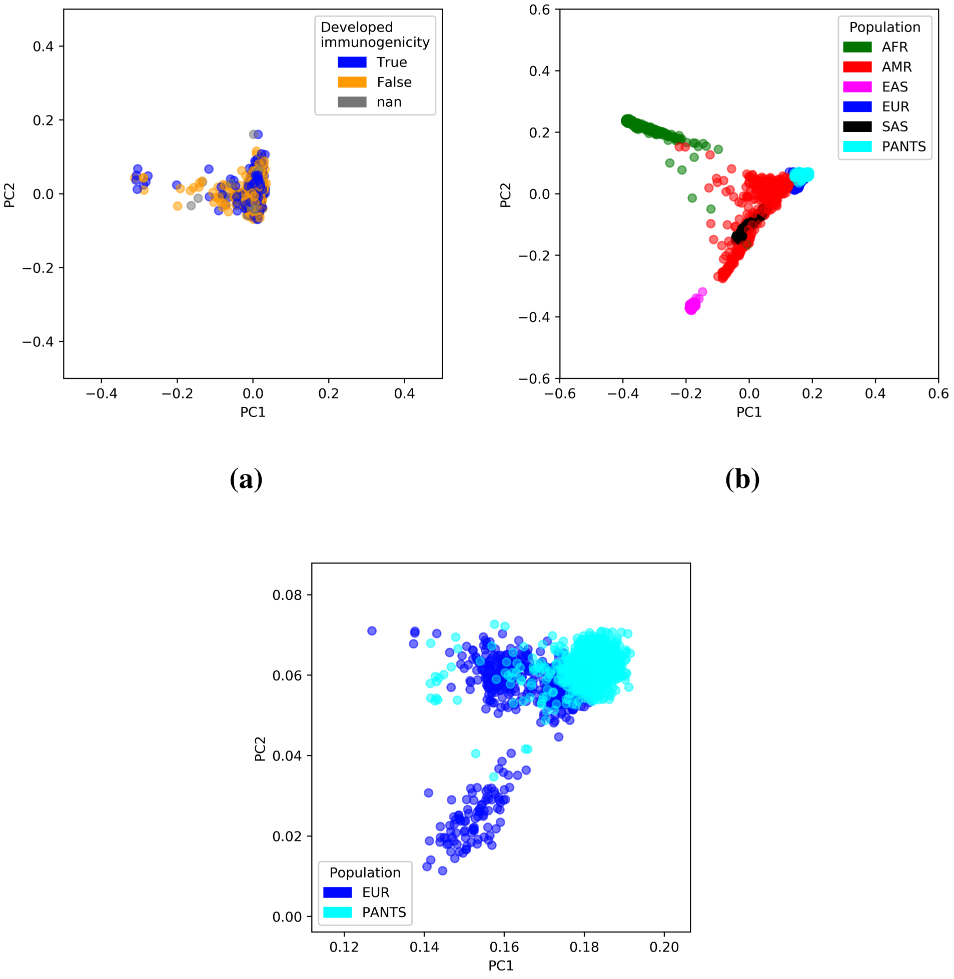
In independent experiments, our collaborators confirmed that this cut-off corresponds to the 98th percentile of the anti-drug antibody titre distribution in more than 500 drug-naïve controls.

### 2.2.3 Genotyping, quality control, and imputation

DNA was extracted from pre-treatment blood samples from 1,524 individuals in the PANTS cohort and genotyping undertaken using the Illumina CoreExome microarray (522,049 markers), with genotype calls made using optiCall [167]. Pre-imputation quality control (QC) of samples and single nucleotide polymorphisms (SNPs) was performed as previously described [49]. SNP genotypes were imputed via the Sanger Imputation Service using the Haplotype Reference Consortium (HRC) panel as a part of a larger GWAS study [120].

Following imputation, I ran additional variant and sample QC procedures. I calculated the first 15 principal components (PC) for the 1,323 samples that passed the QC criteria (Figure 2.1a). The Tracy-Widom test showed that only the first principal component explained a significant proportion of the variance. In addition, I confirmed the European ancestry of our cohort by performing principal component analyses with 2,504 samples from the 1000 Genomes Project (1KGP) [1] (Figure 2.1). Following standard practices, I discarded poorly imputed SNPs, defined as having an information score <0.4. I removed SNPs significantly deviated from Hardy-Weinberg equilibrium ($P<1\times10^{-10}$), with a call-rate $\leq$0.95 or minor allele frequency $\leq$1%, leaving 7,578,947 variants. PCA and genotype filtering were performed using the Hail framework (version 0.1).

I excluded individuals of non-European ancestry (identified using principal component analysis), one individual from each related pair (defined as a pi-hat >0.1875, halfway between the expected pi-hat for third- and second-degree relatives [7]), and those with an outlying



**(a)**

**(b)**

**(c)**

**Figure 2.1** Principal component analysis (PCA) on imputed data from 1,323 individuals in the PANTS study. (A) PCA on the PANTS cohort alone. Blue dots show individuals who develop immunogenicity and orange dots show individuals who did not. PC1 was not associated with immunogenicity status (P=0.99). (B, C) PCA analysis of the 1,323 individuals from the PANTS cohort together with the 2,504 individuals from the 1000 Genomes Project (1KGP) cohort. (B) PCA confirmed that PANTS samples clustered together with the individuals of European ancestry from the 1KGP. (C) A detailed view of PANTS samples plotted together with European samples from the 1KGP.

number of missing or heterozygous genotypes. 1,323 individuals remained in the study following quality control, of which 1,240 had drug and antibody level data available (Figure 2.2).

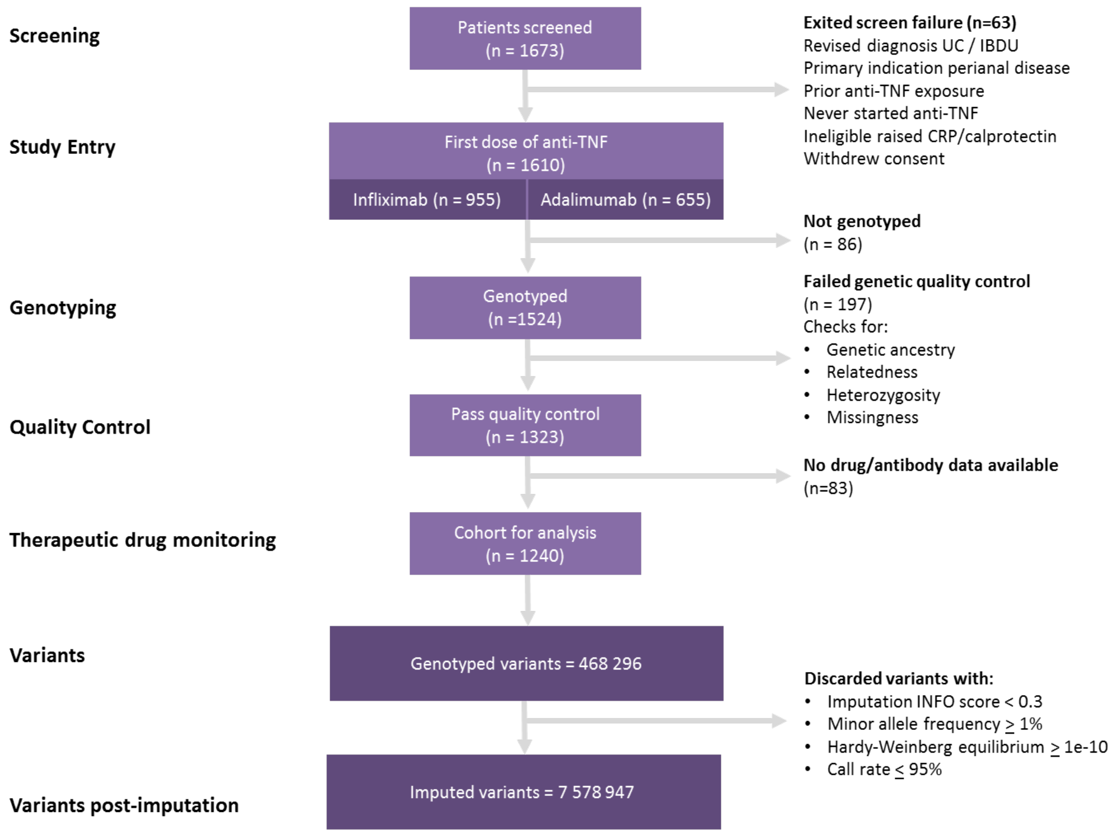


**Figure 2.2** Flowchart describing the cohort used for the time to immunogenicity genetic analysis Abbreviations: UC = ulcerative colitis, IBDU: inflammatory bowel disease unclassified, anti-TNF: anti-tumor necrosis factor, CRP: C-reactive protein, INFO: information content metric

HLA imputation was carried out using the HIBAG package [198] in R, using pre-fit classifiers trained specifically for the CoreExome genotyping microarray on individuals of European ancestry. HLA types were imputed at 2-, and 4-digit resolution for the following loci: *HLA-A, HLA-C, HLA-B, HLA-DRB1, HLA-DQA1, HLA-DQB1,* and *HLA-DPB1*. In addition, my colleague obtained amino acid sequences for all the imputed HLA alleles, using the IPD-IMGT/HLA database [155]. Following the recommended best practices for HIBAG, HLA-allele and amino acid calls with posterior probability <0.5 were set to missing for the given individual. To confirm our imputation and genetic association our collaborators carried out long-read sequencing of the HLA (Histogenetics, New York, USA).

| **Variable** | **Level** | **infliximab**<br>**(n = 742)** | **adalimumab**<br>**(n = 498)** | **p** |
|---|---|---|---|---|
| **Sex (male)** | | 353 (47.6%) | 234 (47.0%) | 0.862 |
| **Age (years)** | | 31.3 (21.2 - 46.0) | 37.6 (28.7 - 50.3) | $1.1 \times 10^{-11}$ |
| **Disease duration (years)** | | 2.6 (0.7 - 9.8) | 3.1 (0.8 - 11.5) | 0.033 |
| **Age at diagnosis (years)** | | 24.4 (16.4 - 35.9) | 29.5 (21.8 - 41.8) | $1.7 \times 10^{-11}$ |
| **Montreal location** | **L1** | 207 (28.2%) | 157 (32.0%) | 0.334 |
| | **L2** | 190 (25.9%) | 108 (22.0%) | |
| | **L3** | 332 (45.2%) | 223 (45.4%) | |
| | **L4** | 6 (0.8%) | 3 (0.6%) | |
| **Montreal L4** | | 78 (10.6%) | 23 (4.7%) | $1.8 \times 10^{-4}$ |
| **Montreal behaviour** | **B1** | 455 (61.7%) | 277 (56.3%) | $1.9 \times 10^{-5}$ |
| | **B2** | 196 (26.6%) | 184 (37.4%) | |
| | **B3** | 86 (11.7%) | 31 (6.3%) | |
| **Perianal** | | 104 (100.0%) | 40 (100.0%) | |
| **Immunomodulator** | **azathioprine** | 339 (45.7%) | 196 (39.4%) | 0.009 |
| | **mercaptopurine** | 57 (7.7%) | 32 (6.4%) | |
| | **methotrexate** | 50 (6.7%) | 24 (4.8%) | |
| | **tacrolimus** | 2 (0.3%) | 0 (0.0%) | |
| | **none** | 294 (39.6%) | 246 (49.4%) | |
| **Steroids** | | 220 (29.6%) | 133 (26.7%) | 0.275 |
| **Past resectional surgery** | | 170 (22.9%) | 122 (24.5%) | 0.539 |
| **HBI** | | 6.0 (3.0 - 9.0) | 5.0 (3.0 - 8.0) | 0.224 |
| **sPCDAI** | | 25.0 (10.0 - 50.0) | | |
| **BMI** | | 23.1 (20.0 - 27.2) | 24.6 (21.6 - 28.4) | $6.5 \times 10^{-7}$ |
| **Hemoglobin (g/L)** | | 126.0 (115.8 - 136.0) | 132.0 (122.0 - 141.0) | $2.2 \times 10^{-10}$ |
| **White cell count (×$10^9$/L)** | | 8.0 (6.2 - 10.3) | 7.9 (6.3 - 9.8) | 0.324 |
| **Platelet count (×$10^9$/L)** | | 343.0 (281.0 - 412.8) | 312.0 (256.0 - 386.0) | $3.4 \times 10^{-6}$ |
| **Albumin (g/L)** | | 39.0 (34.0 - 42.0) | 39.0 (36.0 - 43.0) | 0.051 |
| **CRP (mg/L)** | | 8.0 (3.0 - 21.0) | 6.0 (2.0 - 13.0) | $5.1 \times 10^{-5}$ |
| **Fecal calprotectin ($\mu$g/g)** | | 402 (149 - 855) | 292 (130 - 603) | $1.4 \times 10^{-4}$ |

**Table 2.1** Baseline demographic and clinical characteristics of the 1240 individuals from the PANTS cohort. P values were calculated using Fisher's exact or Mann Whitney U tests.

### 2.2.4 Statistical and genome-wide association analyses

Rates of immunogenicity were estimated using the Kaplan-Meier method. Clinical outcomes and genetic association tests with time to anti-drug antibody development were performed using multivariable Cox proportional hazards regression: sex, drug type (infliximab or adalimumab), immunomodulator use, and the first within-sample principal component, were included as covariates (Table 2.2). Patients who did not develop immunogenicity during the study were censored at the point of last observation. Post-hoc sensitivity analyses were undertaken to test our genetic findings with immunogenicity, firstly, at progressively higher antibody thresholds; secondly, to simulate a free-antibody assay and thirdly, excluding patients with a single anti-drug antibody level >10 AU/ml and subsequent negative anti-drug antibodies <10 AU/ml.

| **Covariate** | **HR** | **95% CIs** | | **P** |
|---|---|---|---|---|
| **Drug (0 – infliximab; 1 – adalimumab)** | 3.27 | 2.67 | 4.02 | $2.82\times10^{-34}$ |
| **Immunomodulators (0 – yes; 1 – no)** | 2.41 | 2.02 | 2.87 | $1.84\times10^{-22}$ |
| **HLA-DQA1*05 (dominant)** | 1.90 | 1.60 | 2.25 | $5.88\times10^{-13}$ |
| **Sex (0 – female; 1 – male)** | 0.93 | 0.78 | 1.11 | 0.41 |
| **PC1 (continuous)** | 0.33 | 0.03 | 3.60 | 0.38 |

**Table 2.2** Covariates used in the final model

The Akaike information criterion (AIC) was used to compare non-nested models to assess if the mode of inheritance was dominant or additive, and to determine whether HLA allele group, specific HLA alleles, or amino acid sequence best-explained the association. The fixed effects Q statistic was used to perform tests of heterogeneity of effect; this test is an extension of Cochran's Q-test and examines whether the observed effect size variability is larger than expected by chance. Interaction tests of the differential effects of drug type (infliximab versus adalimumab and Remicade versus CT-P13) and combination therapy (immunomodulator vs no immunomodulator) conditional on the genotype were performed. Mann-Whitney U tests were used to compare serum levels of anti-drug antibodies at week 54 stratified by anti-TNF drug and immunomodulator use.

## 2.3 Results

Within the first 12 months, 44% of patients developed anti-drug antibodies (95% CI, 0.41 to 0.48), and 62% of patients did so within 36 months (95% CI, 0.57 to 0.67). After correcting for immunomodulator use, the rate of immunogenicity was greater in patients treated with infliximab (N=742) than adalimumab (N=498) (hazard ratio (HR), 3.21; 95% CI, 2.61 to 3.95; P=$1.18\times10^{-28}$). In a model including drug-type as a covariate, rates of immunogenicity were greater in patients treated with anti-TNF monotherapy (N=544) compared to combination therapy with immunomodulators (N=696), (HR, 2.30; 95% CI, 1.94 to 2.75; P<$6.10\times10^{-21}$).

### 2.3.1 A locus within the HLA region is associated with time to immunogenicity

The time-to-event analysis identified a genome-wide significant association on chromosome 6 with time to development of immunogenicity, with the most associated SNP, rs2097432 (b38_pos: 6:32622994; HR, 1.70; 95% CI, 1.48 to 1.94; P=$4.24\times10^{-13}$), falling within the major histocompatibility complex (MHC) region (Figures 2.5, 2.3, 2.4). I replicated this association in our independent cohort of 178 patients with IBD (HR, 1.69; 95% CI, 1.26 to 2.28; P=$8.80\times10^{-4}$). A variant on chromosome 11, rs12721026 (b38_pos: 11:116835452; HR, 0.46; 95% CI, 0.33 to 0.63; P=$4.76\times10^{-8}$) also reached genome-wide significance in our discovery analysis, though the association was not replicated in our independent cohort (HR, 0.85; 95% CI, 0.49 to 1.44; P=0.51).

### 2.3.2 Fine-mapping of the signal in the HLA region

At the HLA allele group level (2-digit resolution), only HLA-DQA1*05 achieved genome-wide significance (HR, 1.90; 95% CI, 1.60 to 2.25; P=$5.88\times10^{-13}$) (Figure 2.6). At the specific allele level (4-digit resolution), no single allele reached genome-wide significance. The two most common HLA-DQA1*05 subtype alleles, HLA-DQA1*05:01 (HR, 1.57; 95% CI, 1.33 to 1.85; P=$4.24\times10^{-7}$) and HLA-DQA1*05:05 (HR, 1.48; 95% CI, 1.24 to 1.78; P=$5.54\times10^{-5}$), had similar effects on time to immunogenicity and a model containing these two 4-digit alleles was virtually indistinguishable from a model including only HLA-DQA1*05 ($AIC_{05}$=6659.07 versus $AIC_{05:01\&05:05}$=6659.50). I did not identify any

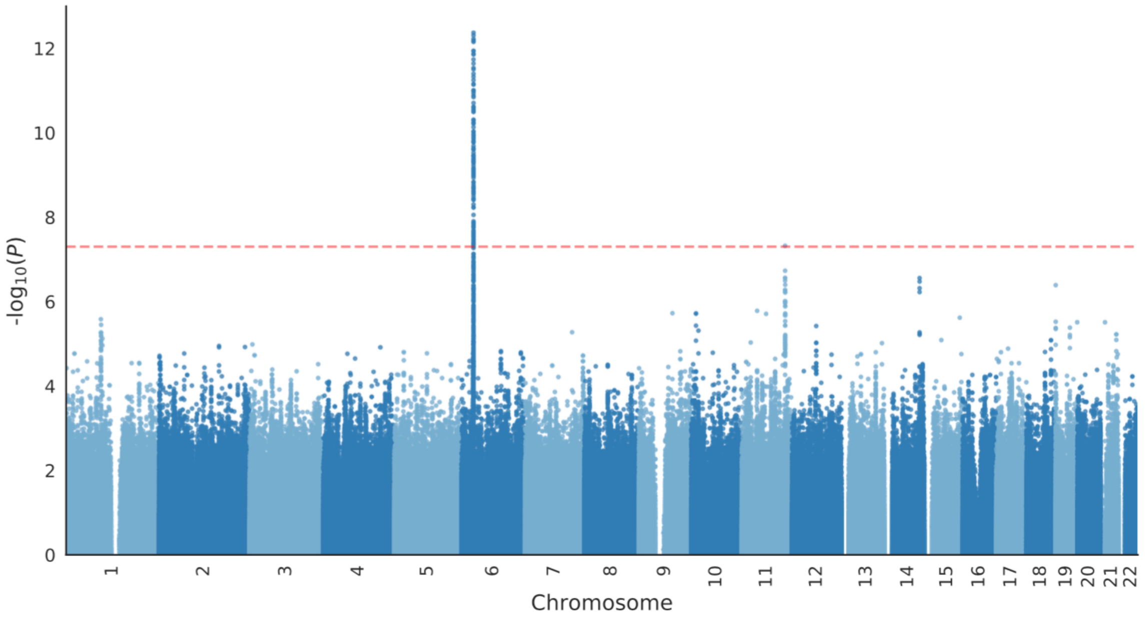


**Figure 2.3** Manhattan plot for Cox proportional hazards model analysis of time to immunogenicity

amino acids that better fit the data than HLA-DQA1*05. Our collaborators observed >99% concordance between imputed and sequenced HLA genotypes at HLA-DQA1: amongst the 1,272 overlapping samples, only one sample was discordant between HIBAG (homozygous HLA-DQA1*05) and sequenced HLA (no copies of HLA-DQA1*05).

To formally assess the inheritance pattern of HLA-DQA1*05 mediated immunogenicity, I compared the fit of additive and dominant models and found that the dominant model gave a better fit (AIC$_{DOM}$=6652.12 vs AIC$_{ADD}$=6659.07), and stronger association signal for HLA-DQA1*05 (HR, 1.90; 95% CI, 1.60 to 2.25; P=5.88×10$^{-13}$) (Figure 2.7 and Table 2.3). I also looked for non-additive effects across all other HLA alleles, but the model assuming a dominant effect for HLA-DQA1*05 remained the best fit to the data, based on both the AIC and BIC. The HLA-DQA1*05 association was confirmed in our replication cohort (HR, 2.00; 95% CI, 1.35 to 2.98; P=6.60×10$^{-4}$), again with a better fit for the dominant model (AIC$_{DOM}$=942.51 vs AIC$_{ADD}$=944.81). After conditioning on HLA-DQA1*05 I did not identify any secondary signals of association with time to immunogenicity within the MHC region (Figure 2.8).

Sensitivity analyses showed that the effect size of HLA-DQA1*05 carriage on immunogenicity was similar across subgroups (Figures 2.9a and 2.9b): firstly, the association

| **Genetic effect included in the model** | **HR** | **95% CIs** | | **P** | **AIC** | **BIC** |
|---|---|---|---|---|---|---|
| **AA-107-ile (dominant)** | 1.9 | 1.6 | 2.26 | $4.00\times10^{-13}$ | 6649.45 | 6675.04 |
| **AA-175-lys (dominant)** | 1.9 | 1.6 | 2.26 | $4.00\times10^{-13}$ | 6649.45 | 6675.04 |
| **DQA1*05 (dominant)** | 1.9 | 1.6 | 2.25 | $5.88\times10^{-13}$ | 6652.12 | 6677.73 |
| **AA-107-ile (additive)** | 1.59 | 1.4 | 1.8 | $1.50\times10^{-11}$ | 6656.55 | 6682.14 |
| **AA-175-lys (additive)** | 1.59 | 1.4 | 1.8 | $1.50\times10^{-11}$ | 6656.55 | 6682.14 |
| **DQA1*05 (additive)** | 1.58 | 1.39 | 1.8 | $1.94\times10^{-11}$ | 6659.07 | 6684.67 |
| **DQA1*05:01 and DQA1*05:05 (additive)** | 1.61 | 1.36 | 1.9 | $1.19\times10^{-7}$ | 6659.5 | 6690.2 |
| **DRB1*03 (additive)** | 1.58 | 1.34 | 1.85 | $2.43\times10^{-7}$ | 6661.3 | 6686.89 |
| **DRB1*03 (dominant)** | 1.67 | 1.38 | 2.03 | $1.48\times10^{-7}$ | 6662.46 | 6688.06 |
| **rs2097432 (additive)** | 1.69 | 1.48 | 1.94 | $4.24\times10^{-13}$ | 6674.24 | 6699.85 |
| **rs2097432 (dominant)** | 1.92 | 1.61 | 2.28 | $4.55\times10^{-13}$ | 6674.34 | 6699.95 |
| **DQA1*05:01 (additive)** | 1.57 | 1.33 | 1.85 | $4.24\times10^{-7}$ | 6676.74 | 6702.33 |
| **DQA1*05:01 (dominant)** | 1.67 | 1.38 | 2.03 | $5.49\times10^{-7}$ | 6677.26 | 6702.85 |
| **DQA1*05:05 (dominant)** | 1.67 | 1.27 | 1.9 | $3.39\times10^{-5}$ | 6684.97 | 6710.56 |
| **DQA1*05:05 (additive)** | 1.48 | 1.24 | 1.78 | $5.54\times10^{-5}$ | 6685.84 | 6711.43 |
| **“Null” model with no genotype parameter** | – | – | – | – | 6725.43 | 6745.92 |

**Table 2.3** Comparison between different models for the observed effect in the MHC region. Models are sorted by Akaike information criterion (AIC) score (column 6), which is a measure of model fit to the data. All models included immunomodulator status, drug type, sex and PC1 as covariates (see Table 2.2 for details). HRs, CIs and p-values are calculated using SurvivalGWAS_SV and AIC is calculated using R surv package, as described in Methods. The effect assumed for the main genetic effect (additive or dominant) is shown in parentheses in the first column. Hazard ratios with 95% Confidence Intervals for the genetic effect are shown in columns 2–4, and the respective p-values in column 5. The last column contains the BIC score for each model. The ranking of the models is identical whether AIC or BIC is used.

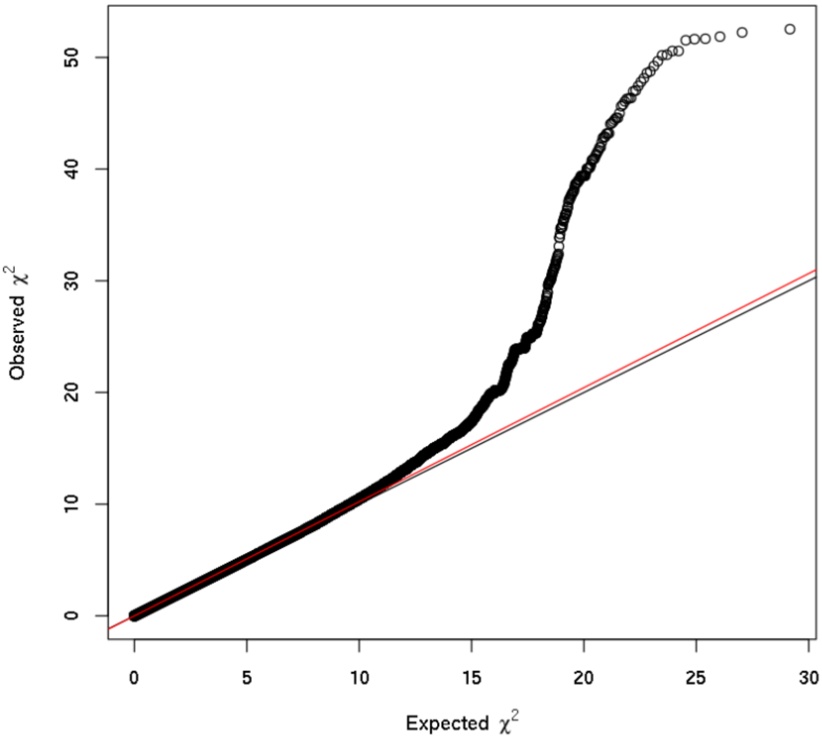


**Figure 2.4** Quantile-quantile plot for Cox proportional hazards model analysis of time to immunogenicity. Using a regression model implemented in the GenABEL package in R, I estimated the inflation factor $\lambda$ to be 1.02 (SE=$2\times10^{-5}$), suggesting a good fit to the uniform distribution. See 2.2 for the covariates used in this model.

remained significant even when the threshold for defining immunogenicity was increased from >10AU/mL to >200 AU/mL. Secondly, when I simulated a drug-sensitive instead of a drug-tolerant assay, where immunogenicity was defined as an anti-drug antibody titer $\geq$10 AU/ml without detectable drug (HR, 1.57; 95% CI, 1.23–2.01; P = $3.66 \times 10^{-4}$). Thirdly, when I removed patients with a one-off transient anti-drug antibody level $\geq$10 AU/ml (HR, 1.94; 95% CI, 1.62-2.32; P=$8.46\times10^{-13}$).

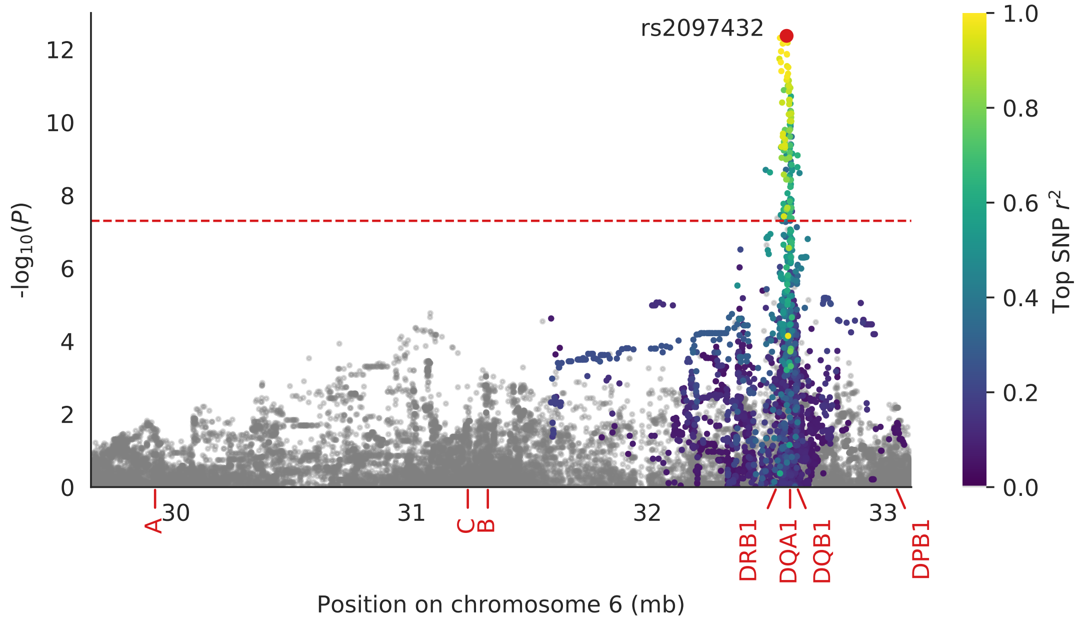


**Figure 2.5** Regional plot of the association results with the MHC region on chromosome 6. Midpoint positions of the HLA alleles across the MHC region are shown in red on the x-axis. SNPs that passed the genome-wide significance threshold (P=5×$10^{-8}$) are shown above the red horizontal dashed line with the most significant SNP in red. SNPs correlated with the lead SNP ($r^2$>0.05) are colour-coded from purple to yellow. Pairwise genotype correlation ($r^2$) between SNPs was calculated using genotype data from the non-Finnish European population of the 1000 Genomes Project.

### 2.3.3 The effect of HLA-DQA1*05 across drug and treatment regimes

While immunogenicity rates were lower with adalimumab-treated compared to infliximab-treated patients, I did not detect a significant difference in the effect of HLA-DQA1*05 on the immunogenicity rate for these two drugs (HR, 1.89; 95% CI, 1.32–2.70 in adalimumab-, HR, 1.92; 95% CI, 1.57–2.33 in infliximab-treated patients; $P_{het}$=0.91) (Fig. 2.10). I also found no significant evidence for heterogeneity of effect of HLA-DQA1*05 on immunogenicity between patients treated with the infliximab originator, Remicade, and its biosimilar CT-P13 ($P_{het}$=0.23) (Figure 2.11). Likewise, I did not detect any significant heterogeneity of effect of HLA-DQA1*05 carriage on immunogenicity for individuals on monotherapy (HR, 1.75; 95% CI, 1.37–2.22) versus combination therapy (HR, 2.01; 95% CI, 1.57–2.58) with immunomodulators ($P_{het}$=0.14). In addition, I did not identify any significant

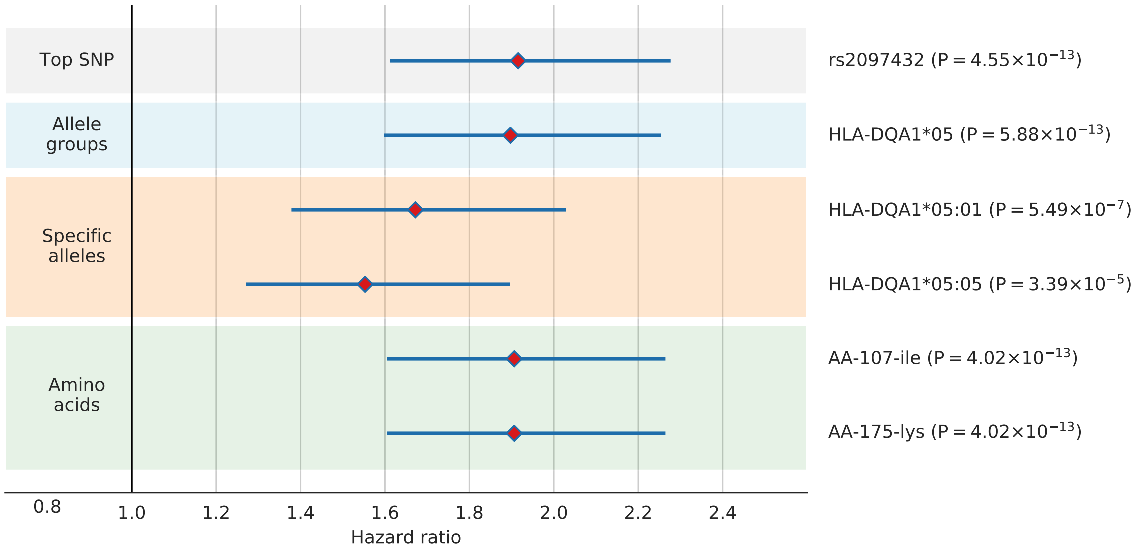
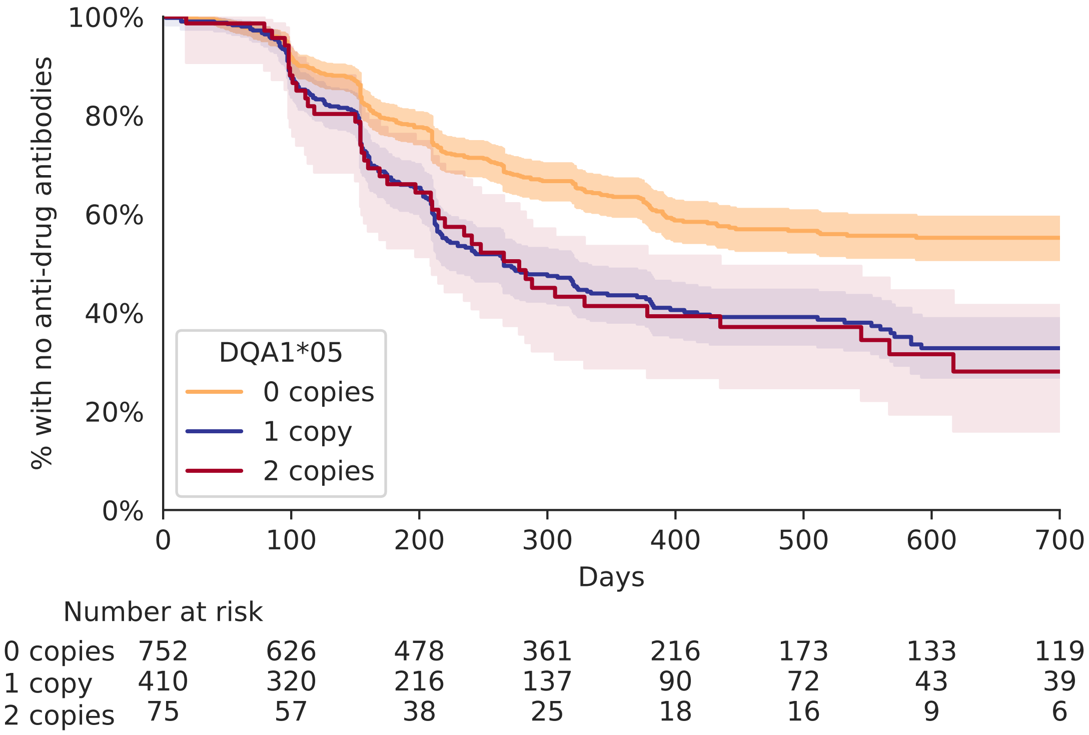


**Figure 2.6** Effect sizes of the most strongly associated SNP, HLA alleles, and amino acids of time to immunogenicity. Blue lines represent 95% CIs. Association test P-values are shown in parentheses.

interactions between HLA-DQA1*05 and the clinical covariates (drug type: P=0.83; mono- vs combination therapy: P=0.71; Remicade vs CT-P13: P=0.59).

The highest rates of immunogenicity, 92% at 1 year, were observed in patients treated with infliximab monotherapy who carried HLA-DQA1*05 (Figure 2.13a). Conversely, the lowest rates of immunogenicity, 10% at 1 year, were observed in patients treated with adalimumab combination therapy who did not carry HLA-DQA1*05 (Figure 2.13b). Our final model, which includes HLA-DQA1*05 status, sex, drug, and immunomodulator usage, explained 18% of the variance in immunogenicity to anti-TNF in our cohort.

Having demonstrated that HLA-DQA1*05 was associated with time to immunogenicity we sought associations with anti-drug antibody titers after 1 year of treatment and subsequent non-persistence on drug. Carriage of HLA-DQA1*05 was associated with higher maximal anti-drug antibody titers ($P_{\text{infliximab}}= 8\times10^{-10}$; $P_{\text{adalimumab = 0.002}}$). I observed lower drug persistence rates to year 3 in patients treated with an anti-TNF drug without an immunomodulator (Figure 2.13); the optimal model ($\text{AIC}_{interaction}$ = 5937.19 versus $\text{AIC}_{additive}$ = 5940.16) here used the interaction between immunomodulator use and HLA-DQA1*05 (DQA1*05:



Number at risk

| | 0 | 100 | 200 | 300 | 400 | 500 | 600 | 700 |
|---|---|---|---|---|---|---|---|---|
| 0 copies | 752 | 626 | 478 | 361 | 216 | 173 | 133 | 119 |
| 1 copy | 410 | 320 | 216 | 137 | 90 | 72 | 43 | 39 |
| 2 copies | 75 | 57 | 38 | 25 | 18 | 16 | 9 | 6 |

**Figure 2.7** Kaplan–Meier estimator showing the rate of anti-drug antibody development, stratified by the number of HLA-DQA1*05 alleles carried. Orange, blue and red indicate 0, 1 and 2 copies of DQA1*05 allele, respectively. Carriers of one or two copies of the allele have a similar rate of immunogenicity development, and a dominant model is a better fit for the data than an additive model (AIC$_{DOM}$=6652.12 vs AIC$_{ADD}$=6659.07). X-axis truncated at 700 days, due to the low number of observations for longer time periods.

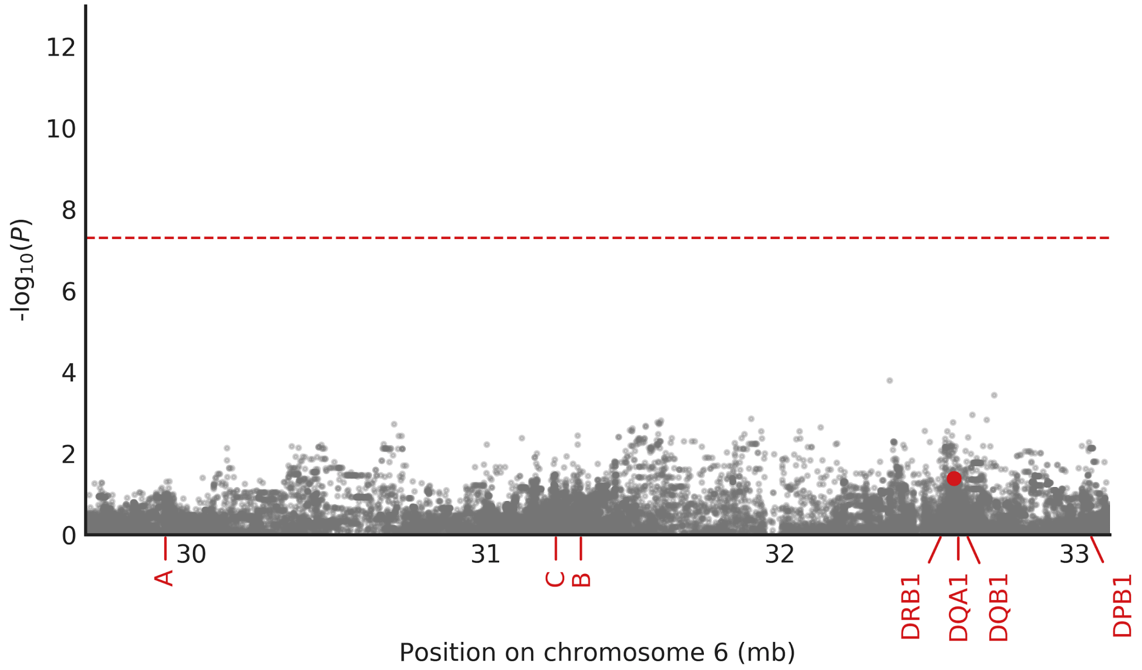
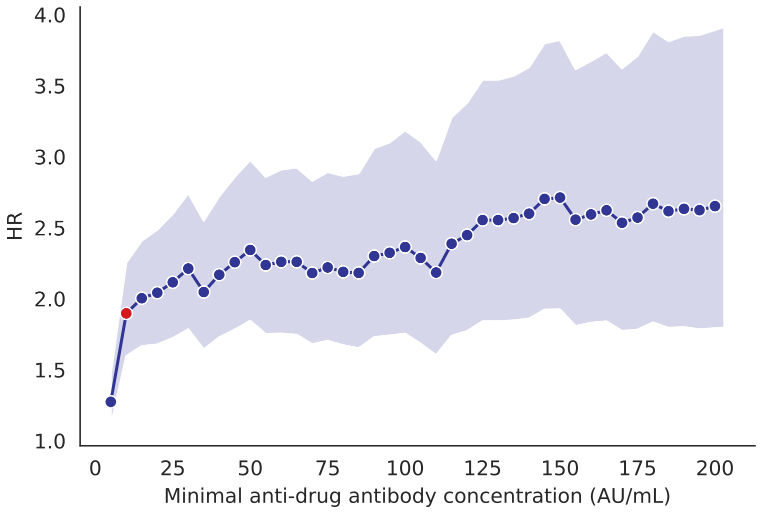


**Figure 2.8** Residual association signal in the MHC region, after conditioning on HLA-DQA1*05. Midpoint positions of the HLA alleles across the MHC region are shown in red on the x-axis. The SNP most strongly associated with time to immunogenicity (rs2097432) during the initial analysis is marked with a red dot. No other SNPs passed the genome-wide significance threshold (red dashed line), suggesting that HLA-DQA1*05 explains the chromosome 6 signal (Figure 2.5).

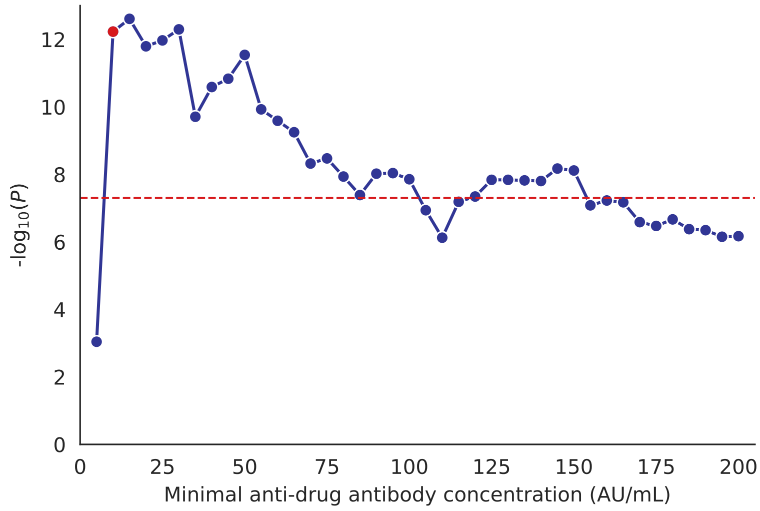
HR, 1.40; 95% CI 1.08–1.80; P=0.011, immunomodulator use: HR, 0.74; 95% CI, 0.58–0.94, P=0.014, interaction between DQA1*05 and immunomodulator use: HR, 0.65; 95% CI, 0.45-0.95; P=0.026).

## 2.4 Discussion

Immunogenicity to biologic therapies is a major concern for patients, regulatory authorities, and the pharmaceutical industry. I report the first genome-wide significant association with immunogenicity to anti-TNF therapy using the largest prospective cohort study of infliximab and adalimumab in Crohn's disease. I have demonstrated that carriage of one or more HLA-DQA1*05 alleles confers an almost two-fold risk of immunogenicity to anti-TNF therapy,



**(a)**



**(b)**

**Figure 2.9** Sensitivity analysis of the (a) effect size and (b) significance of HLA-DQA1*05 association and time to immunogenicity. In the primary analyses, immunogenicity was defined as an anti-drug antibody concentration ≥10 AU/mL, irrespective of drug concentration (red dot). I repeated the time to immunogenicity analysis varying this definition from ≥5 to ≥200 AU/mL.

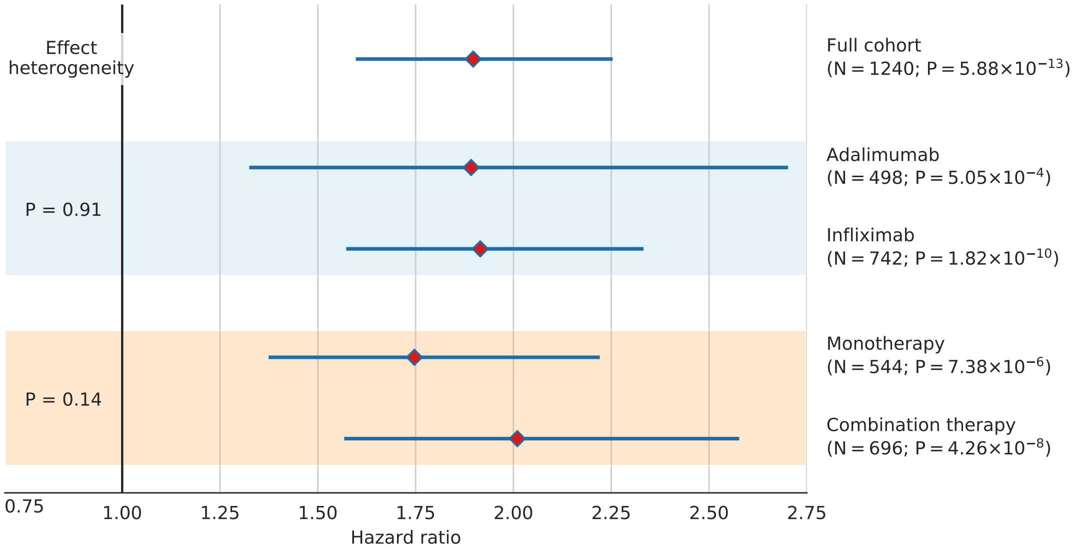


**Figure 2.10** HLA-DQA1*05 has a consistent effect on immunogenicity in different patient subgroups. Blue lines represent 95% CIs. Association test P-values are shown in parentheses. The proportional hazard association analysis was repeated, separating the full cohort into subgroups by drug and therapy type. Estimated hazard ratios and standard errors between the pairings were compared using a heterogeneity of effects test (P>0.05), suggesting that the effect of HLA-DQA1*05 on immunogenicity is not affected by these clinical covariates.

irrespective of concomitant immunomodulator use or drug type (infliximab [Remicade or CT-P13], or adalimumab). Fine-mapping and confirmatory sequencing of the HLA identified that the specific alleles HLA-DQA1*05:01 and HLA-DQA1*05:05 mediated most of this risk. Carriage of HLA-DQA1*05 was associated with higher anti-drug antibody levels and lower drug persistence rates, although further studies are needed to more accurately quantify the relationship between HLA-DQA1*05 and drug persistence. An overview of the further genetic studies that can be carried out to better to elucidate the genetics of the anti-TNF immunogenicity and response is provided in Section 5.3 of the Discussion.

Arguably, based on these data and those presented in the PANTS clinical paper [96], all patients treated with an anti-TNF should ideally be prescribed an immunomodulator to lower the risk of immunogenicity [96]. We hypothesise that for patients who carry HLA-DQA1*05 in whom immunomodulators are contraindicated or not tolerated, clinicians might advise against the use of anti-TNF drugs, particularly infliximab. In contrast, patients who do

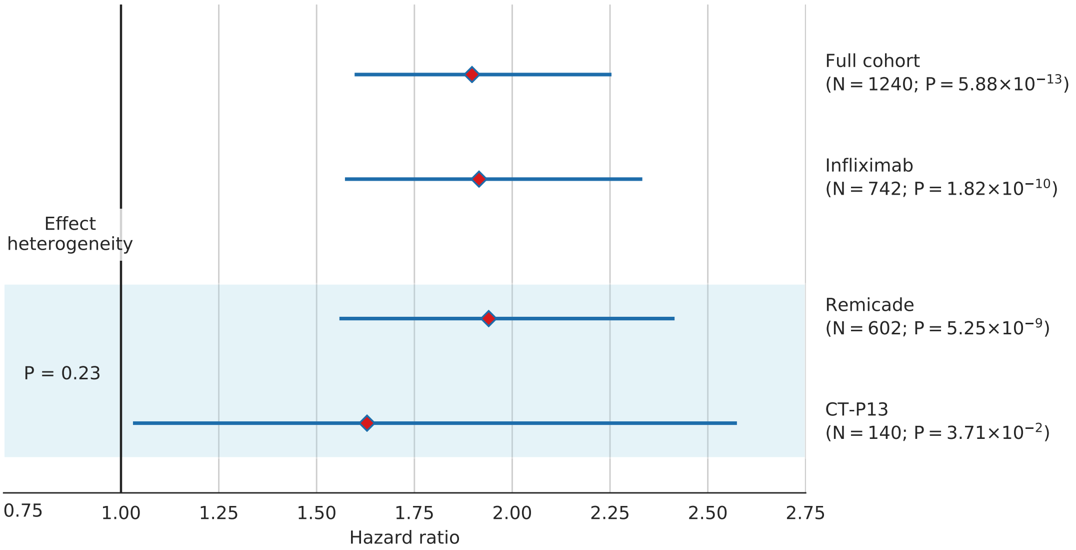
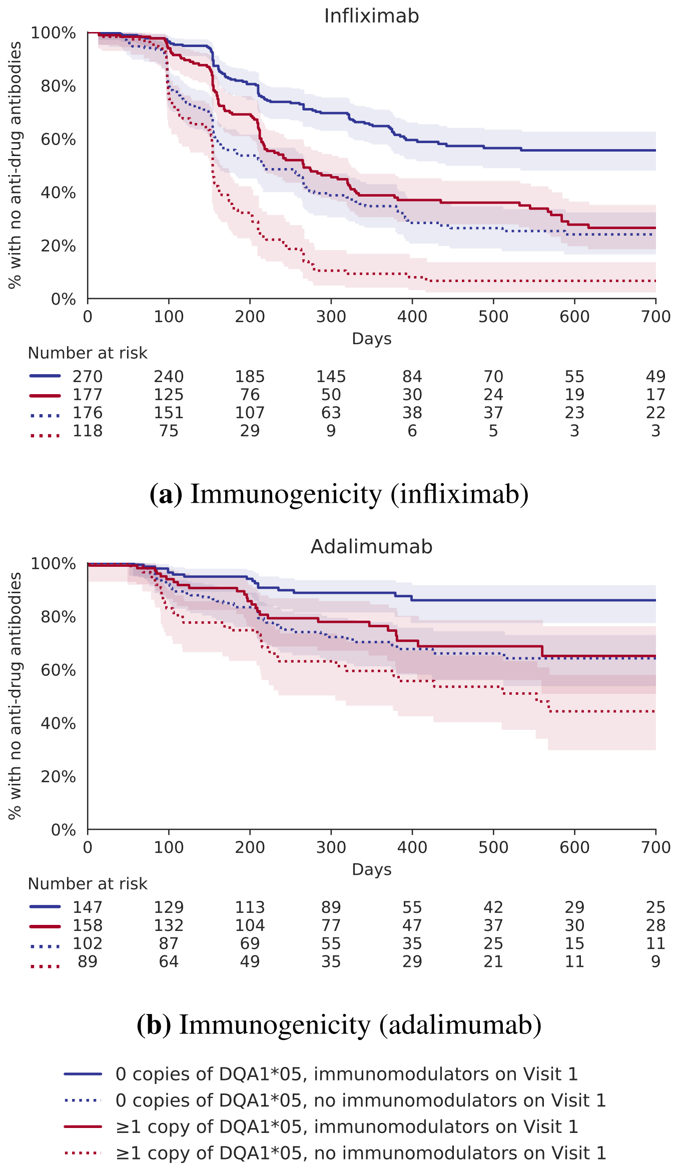


**Figure 2.11** HLA-DQA1*05 has a consistent effect on immunogenicity between patients treated with the infliximab originator, Remicade, and its biosimilar CT-P13.

not carry HLA-DQA1*05 might be given the choice between adalimumab or infliximab combination therapy. Patients without the risk allele and a history of adverse drug reactions to thiopurines and/or methotrexate or who are at high risk of opportunistic infections might be spared the additional risks of combination therapy and treated with adalimumab monotherapy. A randomised controlled biomarker trial is required to explore these hypotheses and confirm whether HLA-DQA1*05 testing may help direct treatment choices in order to improve clinical outcomes.

The shared genetic association between HLA-DQA1*05 and immunogenicity to infliximab and adalimumab may explain the widely reported diminishing returns of switching between anti-TNF therapies at the time of loss of response [141, 161]. If the immunogenic effect of HLA-DQA1*05 extends to other therapeutic antibodies, then subjects who carry the variant may be candidates for non-antibody modality therapies such as small molecule drugs.

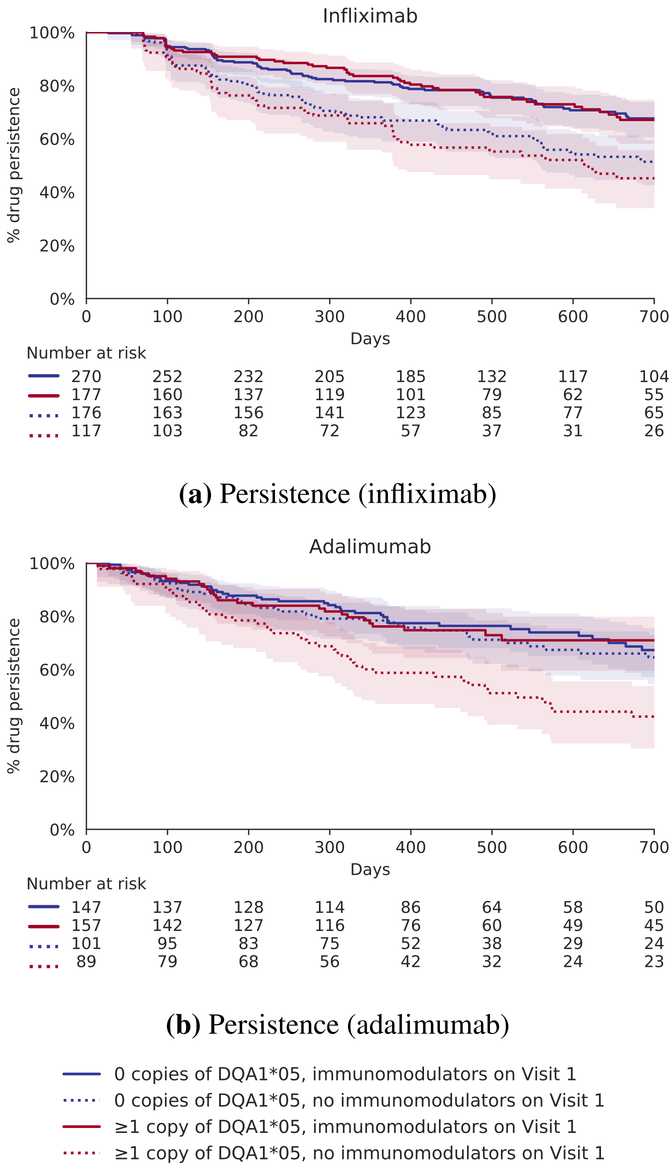
Allelic variation in the *HLA-DQA1* gene has been linked to aberrant adaptive immune responses. The HLA class II gene *HLA-DQA1* is expressed by antigen presenting cells and encodes the alpha chain of the *HLA-DQ* heterodimer that forms part of the antigen binding site where epitopes are presented to T-helper cells. Relevant to immunogenicity,



**(a)** Immunogenicity (infliximab)

**(b)** Immunogenicity (adalimumab)

**Figure 2.12** Kaplan–Meier estimator showing the rate of anti-drug antibody development (A and B), stratified by carriage of HLA-DQA1*05 alleles and treatment regime. Dotted lines indicate patients undergoing anti-TNF monotherapy; solid lines indicate combination therapy with immunomodulators. Red indicates carriers of the HLA-DQA1*05 allele (1 or 2 copies); blue indicates non-carriers. For both drugs and treatment regimes, immunogenicity is higher for HLA-DQA1*05 carriers. The X-axis was truncated at 700 days due to the low number of observations.



| Number at risk | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| — | 270 | 252 | 232 | 205 | 185 | 132 | 117 | 104 |
| — | 177 | 160 | 137 | 119 | 101 | 79 | 62 | 55 |
| ···· | 176 | 163 | 156 | 141 | 123 | 85 | 77 | 65 |
| ···· | 117 | 103 | 82 | 72 | 57 | 37 | 31 | 26 |

**(a)** Persistence (infliximab)

| Number at risk | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| — | 147 | 137 | 128 | 114 | 86 | 64 | 58 | 50 |
| — | 157 | 142 | 127 | 116 | 76 | 60 | 49 | 45 |
| ···· | 101 | 95 | 83 | 75 | 52 | 38 | 29 | 24 |
| ···· | 89 | 79 | 68 | 56 | 42 | 32 | 24 | 23 |

**(b)** Persistence (adalimumab)

**Figure 2.13** Kaplan–Meier estimator showing the rate of drug persistence (A and B), stratified by carriage of HLA-DQA1*05 alleles and treatment regime. Dotted lines indicate patients undergoing anti-TNF monotherapy; solid lines indicate combination therapy with immunomodulators. Red indicates carriers of the HLA-DQA1*05 allele (1 or 2 copies); blue indicates non-carriers. For both drugs and treatment regimes, drug persistence rates are higher for HLA-DQA1*05 carriers. The X-axis was truncated at 700 days due to the low number of observations.

carriage of HLA-DQA1*05 has been associated with coeliac disease and type 1 diabetes and protection against rheumatoid arthritis and pulmonary tuberculosis [122, 42, 182, 137]. Several hypotheses have been proposed, but exactly how specific HLA alleles contribute to disease pathogenesis or, in this case, increased immunogenicity, remains unknown.

HLA-DQ1A*05 may serve as a useful biomarker of immunogenicity risk and may impact how the next-generation of anti-TNF drugs are designed to minimise HLA-DQA1*05 mediated immunogenicity. Previous studies have shown that it is possible to map and eliminate potential immunogenic T cell epitopes with the aim of producing safer and more durable biologic drugs [48, 163]. However, caution needs to be exercised to ensure protein sequence modifications designed to reduce the risk of immunogenicity to patients carrying HLA-DQA1*05 do not put a different group of patients at risk.

Multiple assays are available to detect anti-drug antibodies and there is no universally accepted, validated threshold to diagnose immunogenicity. Our collaborators deliberately chose a total, or drug tolerant assay, that permits the measurement of anti-drug antibodies in the presence of drug, in order to minimise the number of false negative patients assigned to the control group. They then validated the manufacturer's positivity threshold in independent experiments in 500 drug-naïve controls and confirmed that the recommend cut-off of 10 AU/mL corresponds to the 99th percentile of the antidrug antibody titer distribution. In support of this threshold, our collaborators have recently demonstrated that even modestly elevated anti-drug antibodies levels (10–30 Au/ml) at weeks 14 and 54 of treatment are associated with lower drug levels at these time points, and non-remission at week 54 [96]. In addition, sensitivity analyses confirmed that the association and effect size between HLA-DQA1*05 and immunogenicity remained at progressively higher diagnostic thresholds for immunogenicity, when we simulated a free-assay, and when we removed patients with transient antibodies. Finally, HLA-DQA1*05 was associated with the quantitative trait of maximal anti-drug antibody titer.

Two important limitations of this study should be acknowledged. Firstly, we may have underestimated the contribution of HLA-DQA1*05 to immunogenicity because of the short duration of follow-up in patients who did not continue in the study beyond the first year. Secondly, because the study schedule was designed to minimise patients' inconvenience, there were fewer assessments for those treated with adalimumab than infliximab. As a result, we might have underestimated rates of immunogenicity amongst adalimumab-treated patients.

The genome-wide association study was limited to patients with Crohn's disease of European descent. Given that HLA-DQA1*05 is not associated with IBD risk [73] the percentage of carriers among our patients (39%) was similar to that reported in an independent British population cohort (38%) [71]. As such, I hypothesise that HLA-DQA1*05 will make a similar contribution to anti-TNF immunogenicity in other patient populations where the allele is not associated to disease susceptibility (e.g. ankylosing spondylitis). Due to the wide variation in the frequency of HLA-DQA1*05 across ethnic groups [71], further studies are required to assess the contribution of HLA-DQA1*05 to immunogenicity across populations. Whether HLA-DQA1*05 is also associated with immunogenicity to other biologic drugs also needs to be determined.

In this chapter, I report the first genome-wide significant association with immunogenicity to biologic drugs. Carriage of HLA-DQA1*05 almost doubles the rate of anti-TNF anti-drug antibody development, independent of immunomodulator use, for both infliximab and adalimumab. To minimise the risk of immunogenicity, pre-treatment genetic testing for HLA-DQA1*05 may help personalise the choice of anti-TNF and the need for combination therapy with an immunomodulator.

# Chapter 3

# Attempting to identify the genetic determinants of thiopurine-induced liver injury

## 3.1 Introduction

Thiopurines are a type of immunosuppressive drug that have found their application as treatments for a variety of conditions, including immune-mediated disorders and acute lymphoblastic leukaemia. They are also used as maintenance therapy for patients who have received an organ transplant to suppress the immune reaction to the graft. More recently, thiopurines have been used in conjunction with anti-TNF (‘combination therapy’) in order to improve the treatment outcomes of the biologic therapy [41, 96] (see Sections 1.4.4 and 2.1). Several thiopurines are available on the market: azathioprine (AZA), mercaptopurine (6-mercaptopurine or 6MP), and thioguanine (6-thioguanine or 6TG).

Despite the arrival of the biologic therapies, thiopurines are still used for remission maintenance in both Crohn’s disease and ulcerative colitis [103]. In addition, a mounting body of evidence suggests a significant advantage of prescribing thiopurines alongside anti-TNF in order to reduce the immunogenicity risk and, ultimately, improve treatment outcomes [41, 96]. There is also a growing interest in exploring the use of thiopurines alongside vedolizumab, although the evidence of clinical benefit is sparse [79].

Similar to other immunosuppressive treatment options for IBD, thiopurines have a burden on the patients' quality of life and may increase the risk of opportunistic infections (especially viral), lymphomas, myeloid disorders, and skin cancers [13]. In addition, patients treated with thiopurines have a high rate of adverse drug reactions of varying severity. In clinical trials, 0–15% of patients discontinued the treatment due to adverse drug effects [144, 69]. Some of the adverse treatment effects are rather trivial and can be managed – rashes, fevers, nausea. Others, like severe thiopurine-induced myelosuppression (TIM), can be potentially lethal and require utmost caution. Thiopurine-induced myelosuppression has a strong genetic component. Several risk-increasing variants in *TPMT* and *NUDT15* have been identified and are now being used for targeted clinical genotyping prior to treatment (see Section 1.4.5).

Another adverse effect causing major concern when prescribing thiopurine therapy is liver injury. The reported rates of thiopurine-induced liver injury (TILI) vary greatly across studies, with retrospective studies reporting a mean of 3% [69], while the only prospective study has reported a rate of 10% [16]. Several sub-types of TILI exist. Hepatocellular injury is the most common, most asymptomatic type of TILI that is associated with transaminase enzyme elevation. The condition occurs within 12 weeks of treatment start or after dose escalation [24, 185]. Hepatocellular TILI is usually resolved after thiopurine dosage reduction or treatment withdrawal [65]. Cholestatic liver injury is observed in 1 in 1,000 thiopurine-treated patients and is associated with jaundice, itching, and fatigue [80, 186]. Cholestatic TILI occurs between 2 and 12 months after the start of the treatment. The condition can often be mitigated by stopping the therapy, though it can continue after its cessation [80, 186]. Finally, TILI can present itself with both the symptoms and biomarkers of hepatocellular and cholestatic injury (and so will be referred to as 'mixed TILI').

Several risk factors have been associated with TILI: use of corticosteroids was associated with hepatocellular TILI; while concomitant anti-TNF appears to reduce the risk [16], nonalcoholic fatty liver disease, a condition more frequent in IBD patients, has been associated with a higher risk of hepatocellular TILI [166, 179].

As of September 2019, no genetic associations for thiopurine-induced liver damage have been reported. However, several associations, all within the HLA region, for non-thiopurine liver injury have been reported: HLA-B*57:01 for flucloxacillin (antibiotic) [45], HLA-A*02:01 and HLA-DRB1*15:01 for amoxicillin-clavulanate (AC, antibiotic) [114], HLA-B*35:02 for minocycline (antibiotic) [181], and HLA-A*33:01 for terbinafine (antifungal) [133]. The associations are thought to be drug-specific.

In a recent study, Cirulli et al. [39] performed a genome-wide association study of idiosyncratic drug-induced liver injury (DILI). They assembled a cohort of 2,048 individuals with DILI and 12,429 unmatched controls. The cohort was primarily of European ancestry (1,806 cases and 10,397 controls), but included a small number of African American and Hispanic cases with matched population controls. The authors also assembled a replication cohort, consisting of 113 individuals and 239,304 controls of Icelandic ancestry. The cases included subjects of cholestatic (26% amongst the primary European cohort), hepatocellular (41%), mixed (26%), and unknown (7%) DILI. The study included DILI cases caused by a variety of drugs, and herbal and dietary preparations.

The major finding of the study was the association in *PTPN22* – rs2476601 ($N_{cases} = 444$; OR=1.44; 95% CI, 1.28 to 1.62; P=1.2x$10^{-9}$), which was replicated in the Icelandic cohort (OR=1.48; 95% CI, 1.09 to 1.99; P=0.01). The association is primarily driven by amoxicillin-clavulanic acid combination therapy, which is used as an antibiotic (OR=1.62; 95% CI, 1.32 to 1.98; P=4x$10^{-6}$ amongst the patients of European Ancestry). The authors argue that the association is not driven by any particular category of drugs, demonstrating that association has a consistent direction of effect (OR>1) across 39 drugs. However, only seven of these reach the nominal significance level of 0.05. Some therapies, like the antibiotic flucloxacillin, do not demonstrate any evidence of association ($N_{cases} = 195$; OR=1.24, P=0.18). The replication cohort was more homogeneous, consisting of individuals with DILI due to amoxicillin-clavulanic acid and other antimicrobial drugs. As such, it is reasonable to assume that the *PTPN22* association is not a universal risk variant for drug-induced liver injury, but it only causes an adverse reaction in a subset of treatments. The effect of the association on thiopurine-induced liver injury remains uncertain, as the cohort only had 10 cases on mercaptopurine (OR=1.72; 95% CI, 0.5 to 5.97; P=0.39). This will be revisited in the chapter below.

Thiopurines continue to be a widely used therapy for treating patients who suffer from inflammatory bowel disease and a variety of other conditions. Severe drug reactions, like the thiopurine-induced myelosuppression and liver injury, pose a serious challenge and require a better understanding of the associated risk factors in order to enable better therapeutic drug monitoring and personalised prescription. In this chapter, I describe the results from a genome-wide association study of TILI subjects that were recruited as a part of the Predicting Serious Drug Side Effects in Gastroenterology (PRED4) study.

## 3.2 Methods

### 3.2.1 Genotyping, cohort assembly, and quality control

**Assembling the dataset**

Patients were recruited as a part of the Predicting Serious Drug Side Effects in Gastroenterology (PRED4) study (REC number 11/SW/0222).

The final phenotype revision of the PRED4 TILI cohort included 859 individuals – 278 TILI cases (32%) and 581 controls (68%). Cases included 126 subjects with hepatocellular TILI (45% of the cases), 41 with cholestatic (15%), 106 with mixed (38 %), and 5 with an unknown type (2%).

Unfortunately, the genotyping of the entire cohort was not performed in one batch. 1,221 genotyped samples belonging to 786 phenotyped patients were identified (up to 4 genotyped samples per patient). The samples were spread across five genotyping cohorts:

- Two cohorts (153 [`broad1`] and 485 [`broad2`] samples) genotyped at the Broad Institute as a byproduct of the G4L WES pipeline (based on the Illumina Infinium Genome-Wide Association Study array, contain ~245,000 markers).
- One produced at the Sanger institute using the Illumina HumanCoreExome-12 – 245 samples previously included in the de Lange et al. study [49] (imputed to HRC, ~11 million markers [gwas3]).
- Two cohorts produced at the Sanger Institute using the HumanCoreExome-24 array – 150 (imputed to HRC, underwent QC described in de Lange et al. [49], [newwave]) and newly genotyped 188 samples (522,049 markers [newgeno]).

The 1,221 samples were combined together, taking an overlapping set of variants across all five cohorts (1,221 samples, 230,597 variants).

In order to verify to the correctness of the phenotype-to-genotype ID matching and to remove related samples, the genetic kinship across samples was calculated. The assumption was that samples from different genotyping cohorts, tentatively assigned to the same ID, should have a high PI_HAT score.

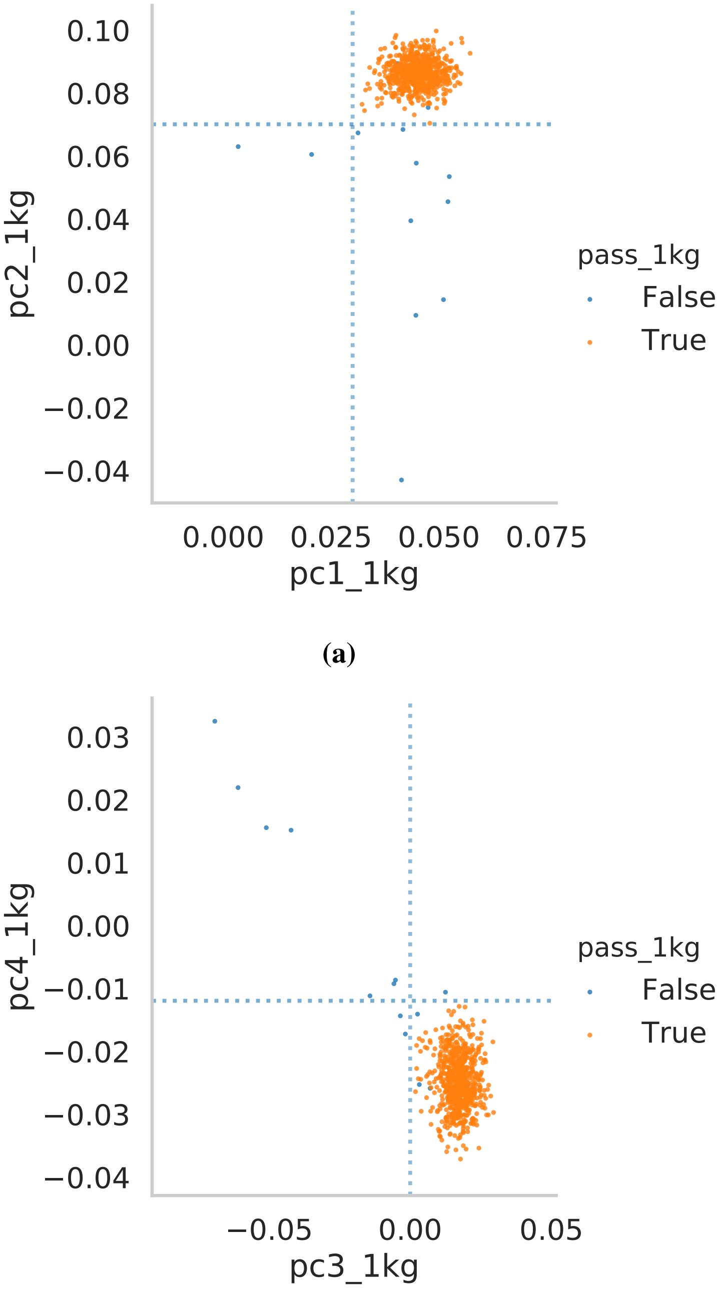
For calculating the relatedness, the full genotyping cohort was filtered to include only accurately genotyped, common SNPs (MAF > 5%, call rate > 0.95, $P_{hwe} > 10^{-6}$, 216,278 variants post-filtering). Variants in linkage disequilibrium with each other were removed via the LD-prune procedure ($r^2 = 0.2$, window = 500,000 bp, 77,053 variants post-filtering). In addition, five poorly genotyped samples were excluded (sample call rate < 0.8), as they resulted in spurious low-level relatedness with tens of otherwise unrelated samples.

The filtered cohort was used to run kinship estimation (identity by descent, similar to the method described in [151]). Pairwise relatedness was estimated for all individuals in the cohort. Pairs with PI_HAT > 0.18 were retained (473 pairs; PI_HAT = 0.1875 is halfway between the expected pi-hat for third- and second-degree relatives [7]; adjusted to be marginally lower in order to account for genotyping heterogeneity). Amongst the pairs of genotyped samples thought to belong to the same patient ID, the minimal PI_HAT was 0.96 (PI_HAT range is between 0 [unrelated] to 1 [duplicate samples or monozygotic twins]), suggesting that the ID matching was done correctly. One sample from each related pair was removed, prioritising samples with denser genotyping, to ensure that no two sample pair had a PI_HAT > 0.18. Post filtering, the cohort retained 778 samples.

### Sample quality control

Next, I have removed poor-quality samples. Minimal variant-level QC was applied (MAF > 1%, missingness < 5%, $P_{hwe}$ in controls > $10^{-6}$) to avoid the sample-level metrics being affected by low-quality variants. Samples with a minimal call rate of 99% and within three standard deviations from the median of the heterozygosity ratio and the call rate were retained. The filter was applied on both individual cohorts and on the overall dataset. 34 samples were removed, retaining 744 samples.

The absolute majority of the individuals in the cohort were of self-reported European ancestry. Weights from the 1000 Genomes project principal component analysis (PCA) were obtained from [9], and projected on the TILI samples. Thirteen outliers (Figure 3.1), based on the first four principal components were removed. The remaining 731 samples were well-mixed, forming one cluster.



**(a)**

**(b)**

**Figure 3.1** Projection of weights derived from 1000 Genome Project principal component analysis onto 778 samples from the PRED4 TILI cohort. The absolute majority of the samples are clustered together (731, orange). Thirteen outliers (blue) were removed after manually inspecting the first four principal components. Exclusion thresholds are shown with dashed blue lines.

Within-cohort Hardy-Weinberg-normalised PCA was conducted to identify outlier samples (Figure 3.2). 10 principal components were calculated using the LD-pruned variant subset (described above). Regions of the genome with long-range LD were excluded, as described in [149]. The eigenvalues were small, suggesting that there is no substantial variance across the genotyping batches ($\text{eigenval}_{PC1}$ = 1.42, $\text{eigenval}_{PC10}$ = 1.24). Thirty-three outliers, based on the first four principal components were removed. The first five principal components were used as covariates during the association tests.

The final sample set contained 698 samples – 207 TILI cases and 491 matched non-TILI controls.

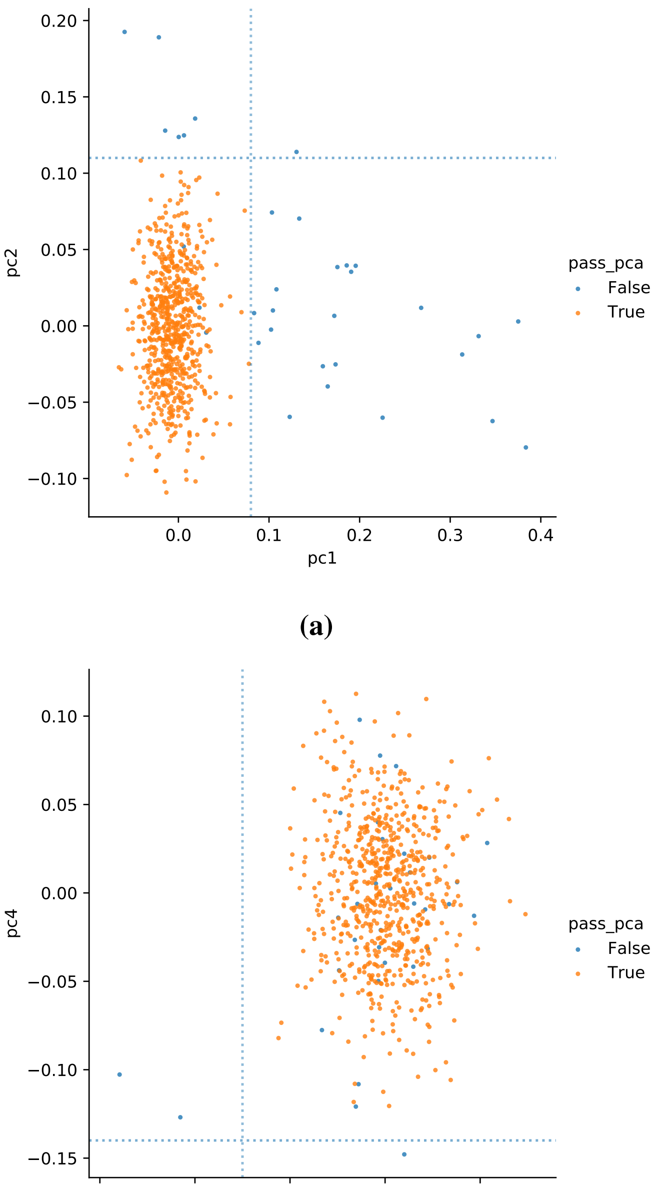
#### Variant QC

Rare and low-quality variants were removed prior to the association testing. The following filters were applied to the dataset: MAF > 1%, missingness < 5%, $P_{hwe}$ in controls > $10^{-6}$. The final dataset contained 226,337 SNPs. At this stage, I did not impute the dataset to any of the reference panels. Primarily, this was due to the low density genotyping and the strong exonic bias of the Broad G4L chip, and partially due to time constraints. A decision was made to impute the dataset only if any significant associations were uncovered (the entire GWAS signal 'peak' is unlikely to contain only imputed SNPs).

### 3.2.2 Statistical testing

For the case-control association tests, logistic regression was carried out using the Wald test. Sex, disease type (CD versus UC), and the first five principal components were used as the covariates.

After performing the draft case-control association analysis, I identified several spurious associations (single variant with no variants in LD with similar p-values). Upon further inspection, these associations were driven by samples from one of the cohorts. To correct for these batch effects, I included a series of case-case and control-control tests (e.g., cases from `GWAS3` versus cases from other cohorts). The variant was only considered significant if the p-value in neither of the batch-effect tests exceeded $\alpha$=1×$10^{-3}$.



**(a)**

**(b)**

**Figure 3.2** Principal component analysis of the PRED4 TILI cohort. Thirty-three outliers (blue) were removed after manually inspecting the first four principal components. Exclusion thresholds are shown with dashed blue lines.

Sample and variant QC, PCA, and association testing was performed using the Hail 0.2 framework [77].

### 3.2.3 Power calculation

Using the methodology described by Johnson and Abecasis [88], power to detect single-variant associations for TILI was calculated. The following parameters were used: significance threshold $\alpha$=5×10$^{-8}$, trait prevalence 10% [16], additive model, 207 cases and 491 controls. Considering that the controls were screened for TILI, the genotype relative risk ratios can be used as the odds ratio estimates. Scenarios where statistical power exceeded 0.8 were considered as ‘well powered’.

The power calculation (Figure 3.3) suggests that at the current sample size, I was poorly powered to detect associations with relative risks below 2.5 for variants of all minor allele frequencies. For rare variants (frequency of 1%), the risk would have to exceed 10. For common variants with MAF = 10% and above, I only had power to detect variants with an odds ratio of 3 and above. It should be noted that power calculations tend to overestimate the power, as they do not take into account the genotyping errors, cryptic population structure, and batch effects that have a negative impact on the ability to detect a true association.

## 3.3 Results

### 3.3.1 Case-control analysis

The genome-wide case-control analysis did not identify any robust associations for thiopurine-induced liver injury. The QQ-plot did not indicate a major inflation of the p-values. The genetic inflation factor was low ($\lambda$ = 1.01). A single variant that passed the genome-wide significance threshold – rs10935807 – appears to be significantly associated with TILI (OR = 2.32; 95% CI 2.07 to 2.57; 6.61×10$^{-11}$). In the GTEx dataset, the variant is significantly associated with the expression of the *EIF2A* gene in 18 tissues (not including liver). No formal colacolisation analysis was performed. Eukaryotic translation initiation factor 2-alpha kinase 3 (*EIF2AK3*), often referred to as protein kinase R (PKR)-like endoplasmic reticulum kinase (*PERK*), is known to be involved in phosphorylation of *EIF2A* [160]. Hopper et al.

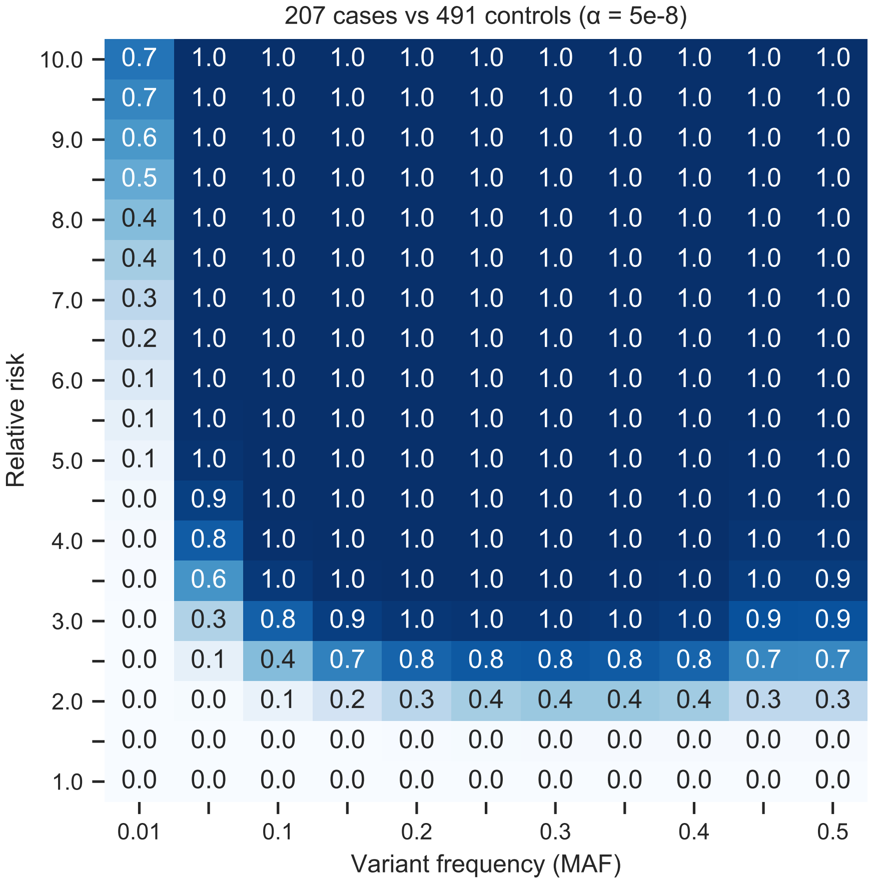


**Figure 3.3** Power to detect single-variant associations for PRED4 TILI cohort. X axis – relative risk of the variant. Y axis – variant frequency. Individual segments on the heatmap – power to detect the association at the genome-wide significance level. Methodology as described in [88]. Original code translated from JavaScript to Python in order to run power calculations for larger sets of parameters simultaneously. Parameters used: significance threshold $\alpha$=5×$10^{-8}$, trait prevalence 10% [16], additive model, 207 cases and 491 controls.

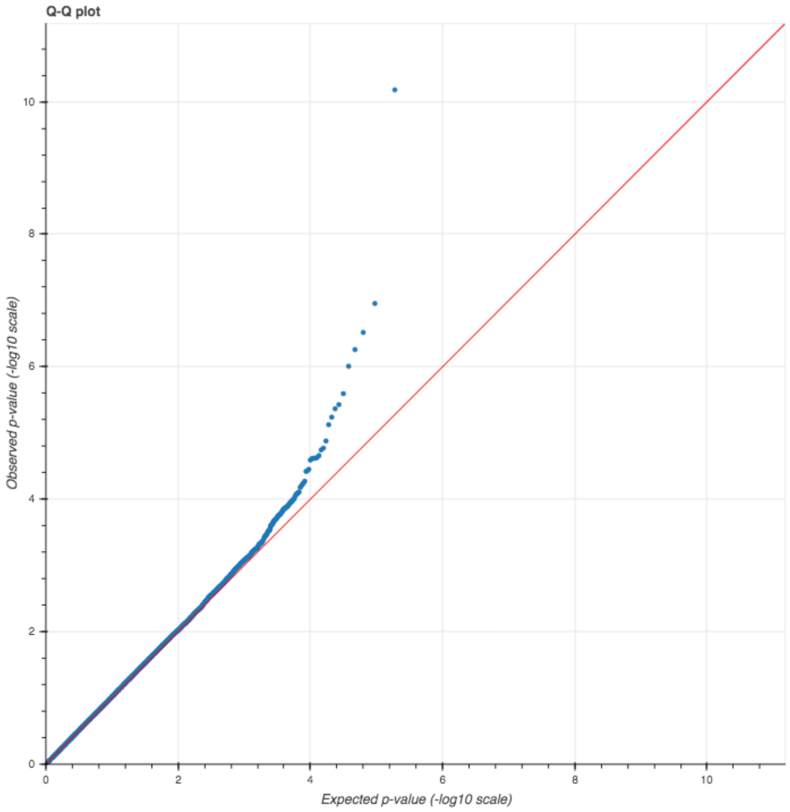


**Figure 3.4** Quantile-quantile plot for the case-control analysis of the PRED4 TILI cohort. The inflation factor $\lambda$ is 1.01, suggesting a good fit to the uniform distribution.

[81] have demonstrated that azathioprine induces autophagy, partially via stimulation of the unfolded protein response (UPR) sensor PERK. However, despite the tentative biologic explanation, I do not believe that the association is truly robust.

While I could not identify any obvious issues with the genotyping quality, there are several properties that make me cautious about declaring this association to be true. Firstly, as visible on the Manhattan plot, there are no other SNPs with a p-value close to rs10935807. The SNP with the highest $R^2$ with rs10935807 (out of those currently genotyped) – rs9883613 ($R^2$ = 0.6352, D' = 0.9728) demonstrates no evidence of association (P = 0.31). In addition, amongst the two biggest case-containing batches, the variant has a substantially different minor allele frequency (`broad1` – 0.45, `newgeno` – 0.60). The frequency in the biggest control batches closely matches that reported in gnomAD for individuals of the European ancestry ($\sim$0.35). As it currently stands, I am hesitant to claim that the variant is associated with TILI. The variant could be potentially re-imputed. Unfortunately, the G4L chip does not include any of the five variants that are in high LD ($>$ 0.95) with the rs10935807, making

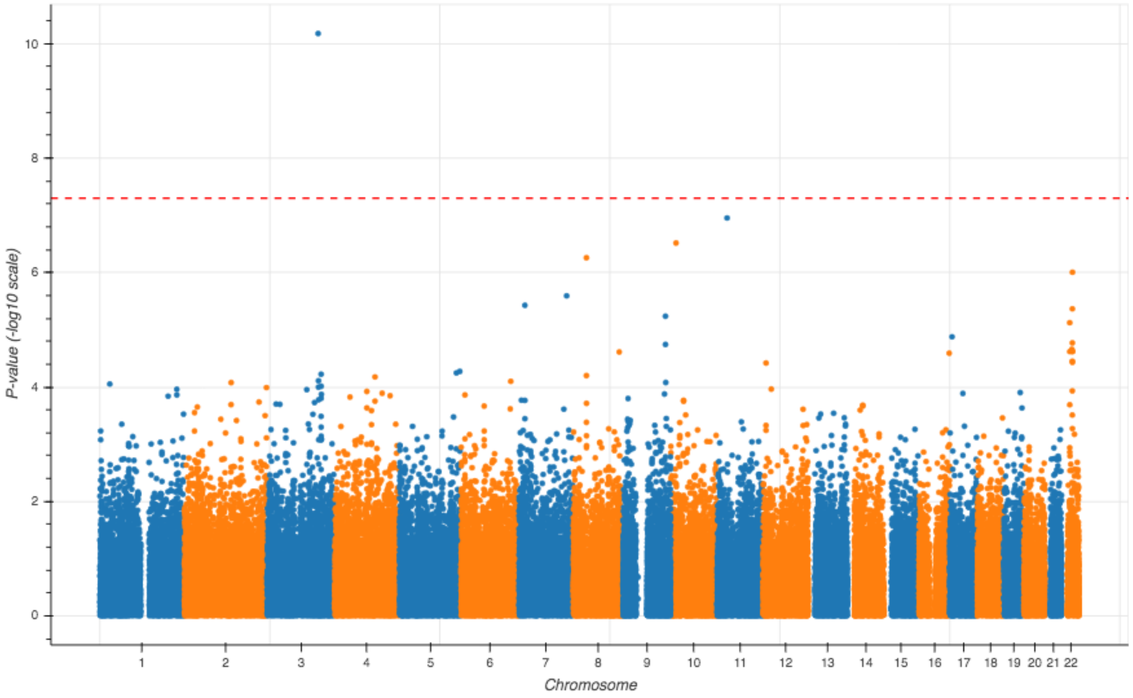


**Figure 3.5** Manhattan plot for case-control analysis of the PRED4 TILI cohort. Only one variant (rs10935807) reached the genome-wide significance level (OR = 2.32; 95% CI 2.07 to 2.57; $6.61{\times}10^{-11}$). The supporting text describes the arguments why it requires further investigation before conclusive statements can be made with regard to its association with TILI.

high-quality imputation unlikely. Alternatively, the cases and controls genotyped exclusively with the G4L chip can be re-genotyped on the CoreExome chip in order to match the rest of the cohort. Potentially, one could try validating this tentative association via replication.

### 3.3.2 rs2476601 in *PTPN22* does not appear to be associated with TILI

I have attempted to replicate the result reported by Cirulli et al. [39] (see the detailed description in the introduction). They describe a finding that a missense variant in *PTPN22* is associated with idiosyncratic drug-induced liver injury. The association is largely driven by liver injury caused by amoxicillin-clavulanic acid combination therapy, though a variety of other therapies have an effect size in the same direction. The cohort included 10 TILI cases caused by mercaptopurine which, when analysed alone, had a similar effect size to the overall association but was not significant (OR=1.72; 95% CI, 0.5 to 5.97; P=0.39). Convincingly replicating the DILI association in a TILI cohort would be of great interest and could be considered the first robust genetic association for TILI.

The rs2476601 SNP was only genotyped in the CoreExome subset of the PRED4 TILI cohort – 152 cases and 364 controls. The particular variant was well genotyped: 100% call rate, $P_{hwe}$=0.17, MAF=0.91 comparable to MAF=0.90 reported in gnomAD for Northwestern Europeans. The case-control analysis was repeated on this subset containing rs2476601.

No evidence of replication was uncovered (OR=0.88; 95% CI, 0.41 to 1.35; P=0.60). It should be noted, that I did not have the complete statistical power to replicate the association: assuming $\alpha = 0.05$ (replication-level significance) and OR=1.72 (the reported odds ratio for mercaptopurine), the power was 0.76; assuming $\alpha = 0.05$ and OR=1.44 (the pooled OR) it was 0.44.

The work of Cirulli et al. is an important step towards understanding the genetic determinants of drug-induced liver injury. It demonstrates consistent effect in DILI caused by several drugs (namely, antibiotics and antifungals). However, at this stage there is no evidence that rs2476601 is associated with liver injury caused by thiopurines.

## 3.4 Discussion

In this chapter, I have described the first genome-wide association study of thiopurine-induced liver injury. Unfortunately, at this stage, it did not result in any robust associations. I believe there are several potential reasons for this.

The success of GWAS as a technique for studying complex disease genetics comes down to the consistent application of rigorous statistical approaches to ever-growing sample sizes. The hypothesis-free nature of GWAS requires applying a stringent genome-wide significance threshold for p-values, correcting for inevitable multiple testing (typically, $\alpha$=5×$10^{-8}$, see Section 1.3.2 for a more extensive discussion). Therefore, GWAS study cohorts need to be sufficiently big to detect truly associated variants. In this chapter, I have analysed a cohort of 207 cases and 491 controls – well below the sample size that is expected for modern complex trait GWASs, which often exceed hundreds of thousands of cases and controls.

The small sample size is not unusual for pharmacogenetic studies. Cholestatic TILI occurs in 1 in 1,000 thiopurine patients who are treated for IBD, a disease that occurs in approximately 0.5–0.8% individuals in the UK. The rarity of the condition makes collecting even a few hundred cases a challenging task. Past pharmacogenetic GWASs have yielded results at modest sample sizes, due to the high odds ratios of the uncovered variants (e.g., the coding *NUDT15* variant increasing the risk of thiopurine-induced myelosuppression [OR=27.3] [186]).

An assumption often made when designing pharmacogenetic studies is that, in contrast to the majority of complex diseases, adverse drug reactions might be associated with common genetic variants (say, MAF>5%) and have a high effect size, as they have not gone through the purifying selection due to the recency of the therapy's arrival. This assumption, however, is only partially correct: therapies with severe side effects that are strongly associated with common genetic variants are not expected to pass stages II and III of clinical trials due to safety concerns. I have performed a case-control power calculation (Figure 3.3), suggesting that the study had 80% power to detect variants with relative risk of 2.5 and above for all minor allele frequencies. It is entirely possible that thiopurine-induced liver damage is a polygenic trait that is not associated with such high-effect size SNP variants.

Another limitation of the study is the heterogeneity of the genotyping arrays, resulting in only around 226,000 markers being tested for associations (discussed in Methods). The number of tested markers can be potentially increased via imputation. However, as discussed in the Methods section, the sparsity and the exonic bias of the G4L array is likely to be detrimental to the overall imputation quality. We are considering including the TILI samples into the ongoing IBD WES study in order to whole-exome sequence them at 60x depth. Exome sequencing will allow us to search for low-frequency variation that may be associated with TILI, but is poorly captured by the current genotyping arrays. Considering the small cohort size, this would be sufficient only for finding associations of a high effect size and it is likely that such variants are within the exonic regions of the genome. However, it is worth noting that, based on the power calculation, a 1% MAF variant would need to have a relative risk of 9 and above in order to be associated at least at an exome-wide significance level ($\alpha$=1×10$^{-6}$).

Finally, I believe that the TILI study can be extended further. The NIHR IBD BioResource project [143] is finalising the recruitment of its first 25,000 IBD patients, who will undergo genotyping. The participants will be given questionnaires that include timings (treatment start and discontinuation), therapy response, and information on adverse drug reactions. Assuming that 75% of those 25,000 patients underwent thiopurine therapy and a 10% incidence of TILI, one would expect 1,875 TILI cases amongst the BioResource patients. The difficulty is that TILI cases might not be encoded as such.

One way to get around the scarcity of the TILI cohorts is to perform a discovery GWAS for all-cause therapy thiopurine failure. The proxy-phenotype can be further improved by including the information on when the patients have ceased the therapy, expecting the hepatocellular TILI cases to stop the therapy within the first 12 weeks of treatment. This approach is unlikely to work for patients with cholestatic TILI since the condition occurs within the first 12 months of the treatment, which is hard to distinguish from discontinuation due to treatment ineffectiveness. Besides, considering their 1 in 1,000 frequency, the number of cholestatic patients in the entire BioResource will be very low. If the analysis results in significant associations, these can be validated in the PRED4 cohort by confirming that a variant is at least nominally significant in TILI and has a similar OR.

Ultimately, this chapter demonstrates that the search for pharmacogenetic associations remains a challenging task, largely limited by the difficulty of assembling the cohorts. The arrival of large biobanks and bioresource services could enable studies that leverage imperfect proxy phenotypes in order to maximise the statistical power at the discovery stage. However, the presence of the clinically validated datasets, like the PRED4 TILI cohort, will remain important for verifying such associations.

# 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

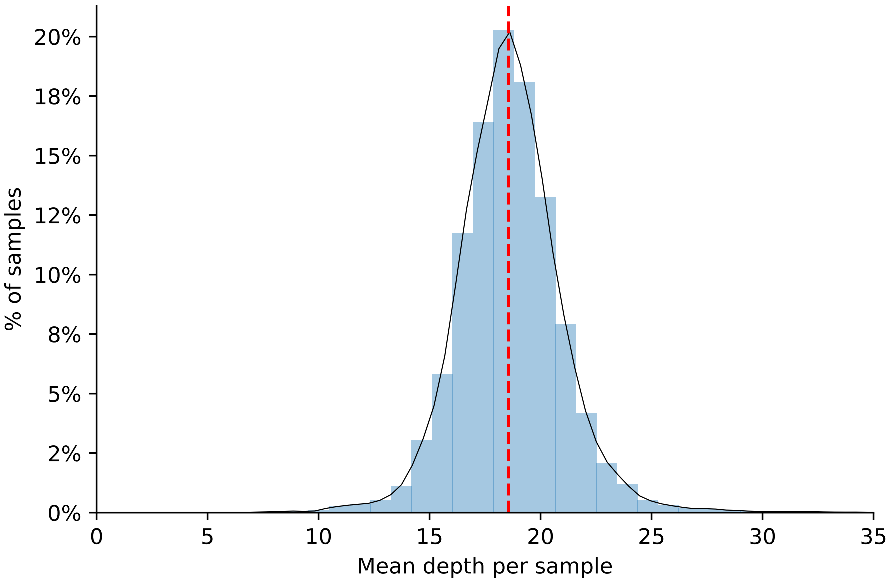


**Figure 4.1** Histogram of mean per-sample coverage for samples in the combined 15x cohort. Median depth in the 15x cohort: 18.56x. N=19,374.

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).

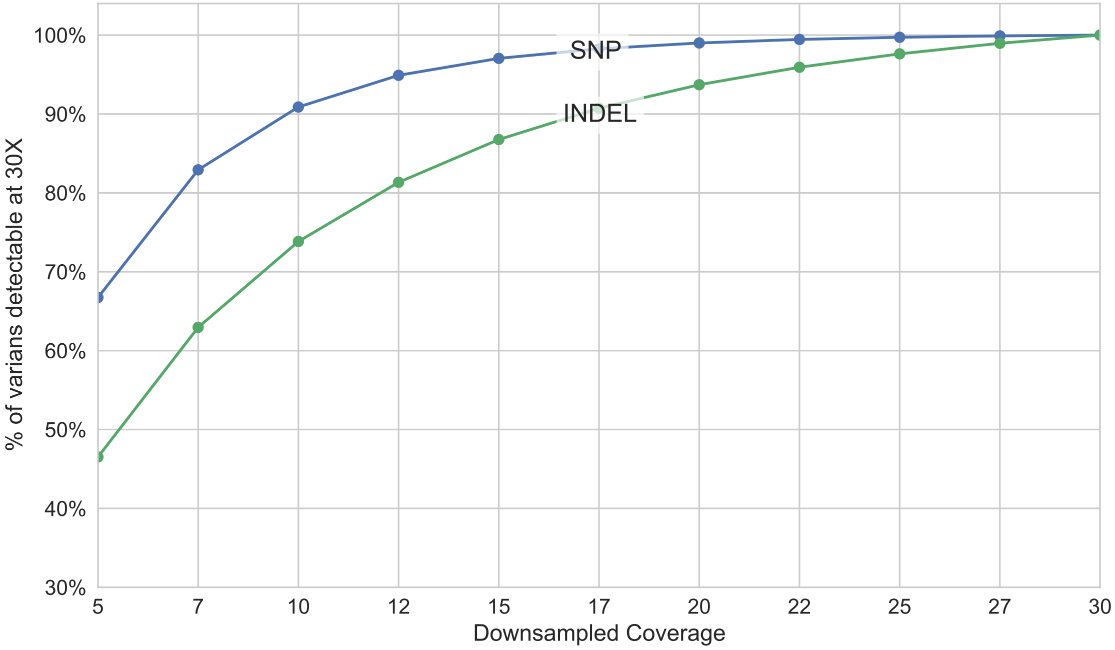


**Figure 4.2** Influence of sequencing depth on the ability to detect SNPs and INDELs. Estimates provided by Martin Pollard.

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 $\alpha$, 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 ~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

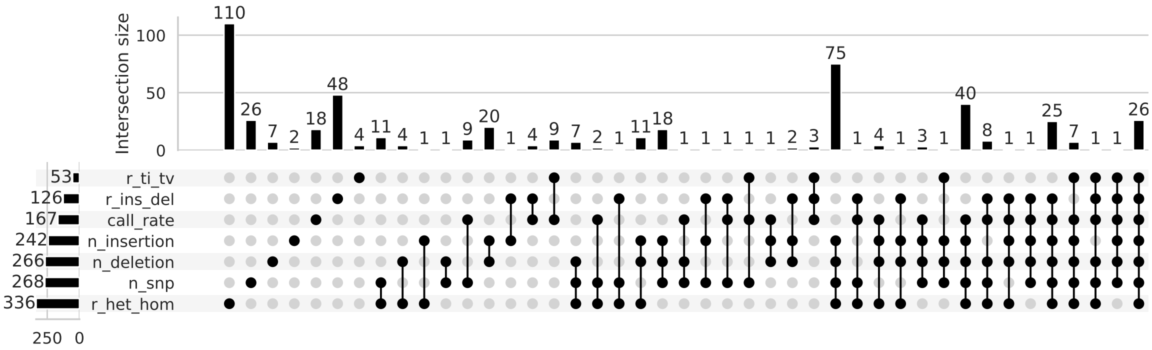


**Figure 4.3** Set intersection of the samples failing different QC metric thresholds in the 15x dataset. Left bars – number of samples failing individual categories. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First seven columns (single dots) – number of samples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

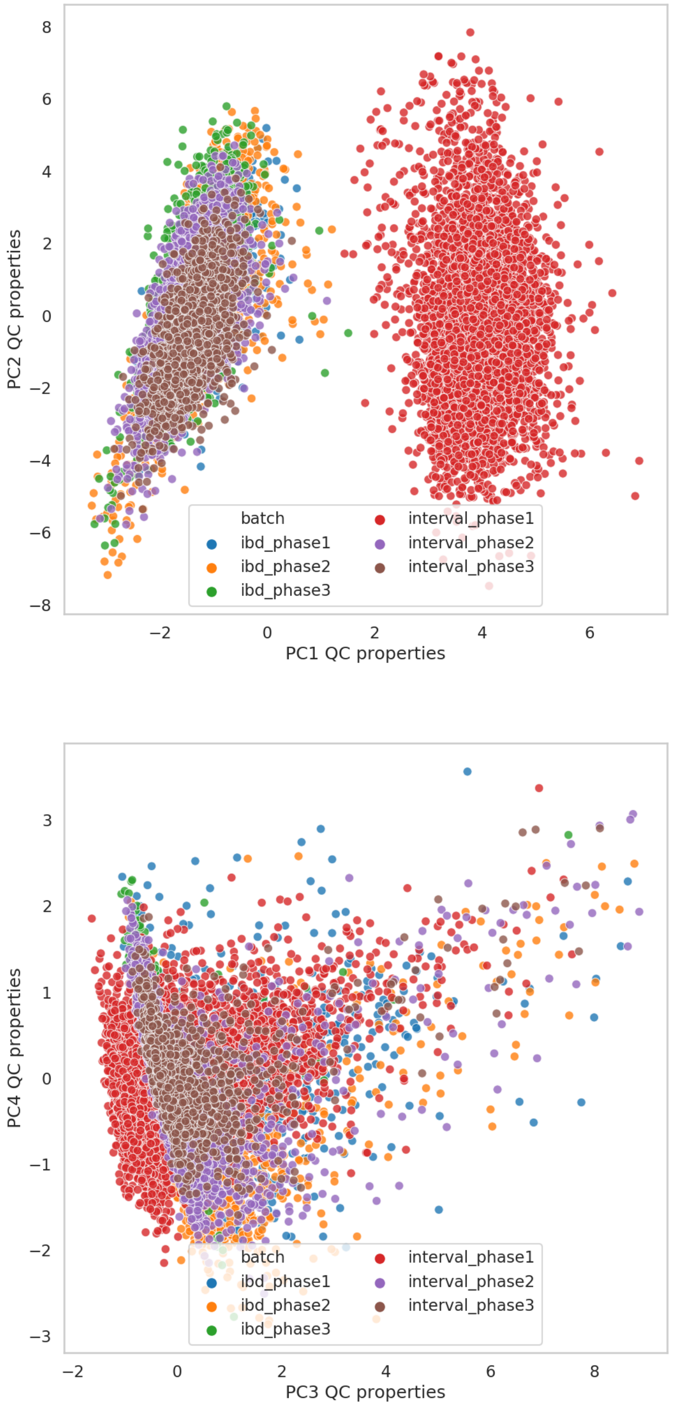


**Figure 4.4** The first four principal components built based on the QC metrics of the samples in the 15x cohort. Samples coloured by sequencing phase. PC1 clearly separates samples sequenced with PCR and PCR-free library preparation protocols.

### 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

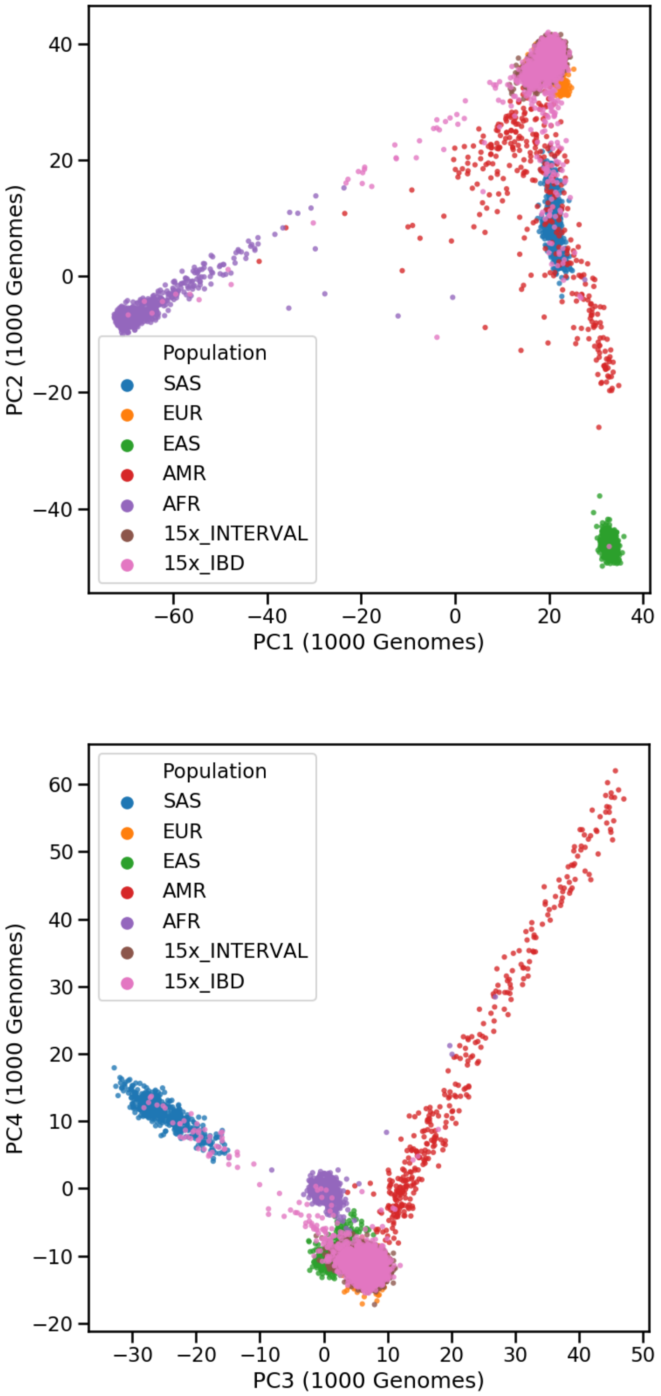


**Figure 4.5** The first four principal components of the 1000G cohort, alongside the 15x cohort samples projected onto them. The absolute majority of the 15x cluster together with the European population of 1000G. The majority of the outliers were IBD samples.

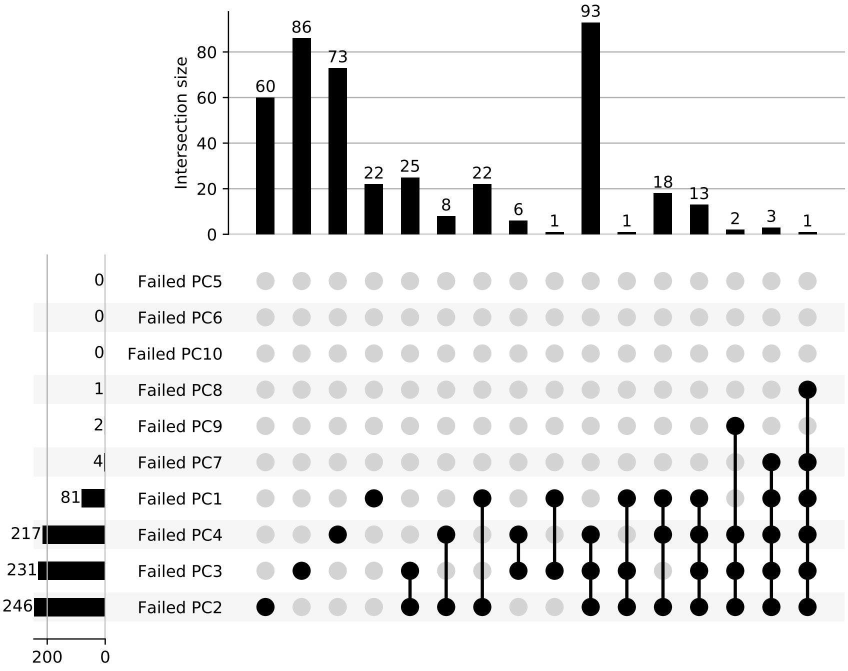


**Figure 4.6** Set overlap plot for the 434 genetic ancestry outlier samples from 15x. Left bars – number of samples failing individual ancestry PC filters. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First six bars (single dots) – number of samples failing only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

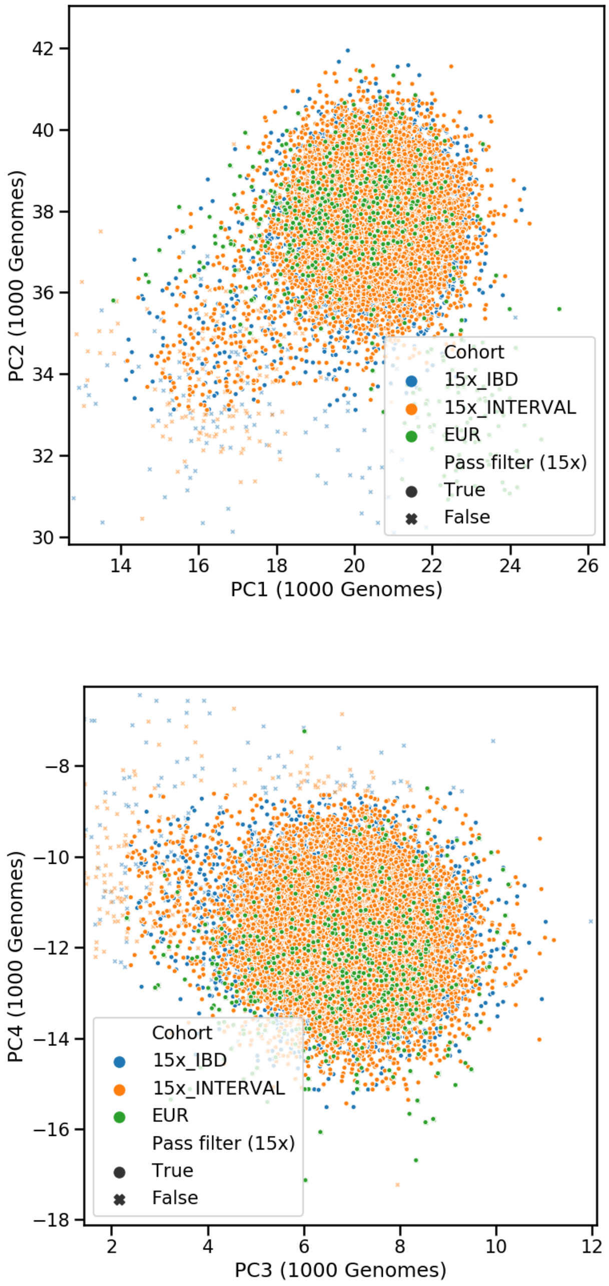


**Figure 4.7** The first four principal components of the European subset of the 1000G cohort, alongside the 15x cohort samples projected onto them (zoomed in, discarding distant outliers). Blue (IBD) and orange (INTERVAL) dots pass the MAD-based filter. Opaque X's are too far removed from the 1000G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

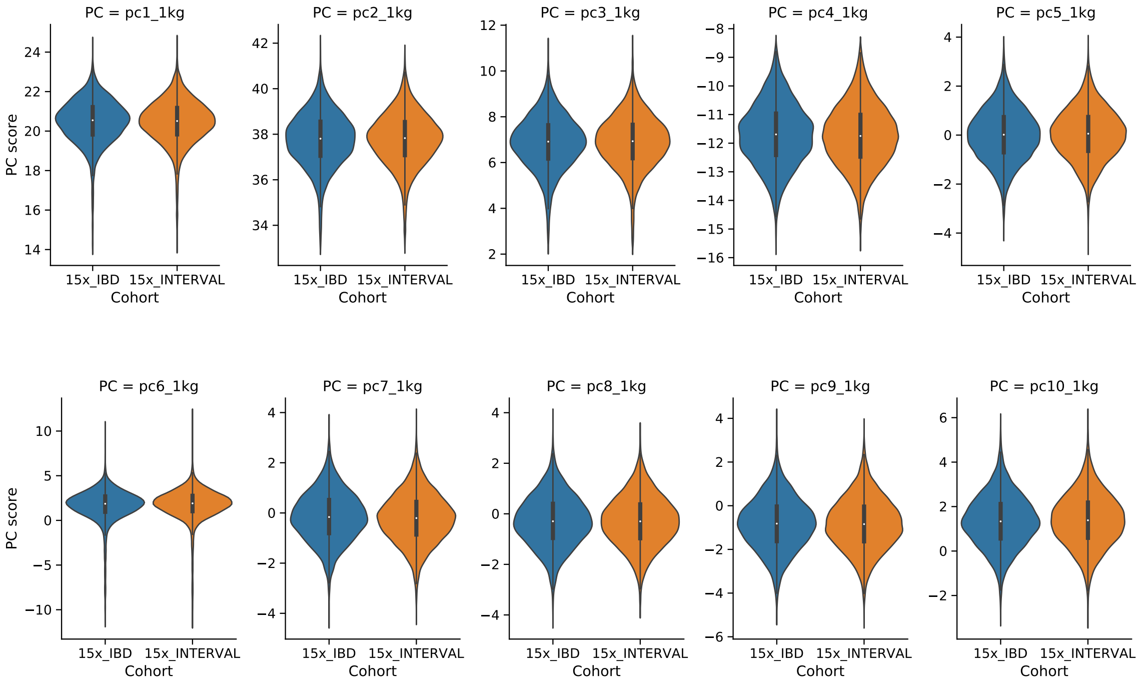


**Figure 4.8** Violin plot comparing the distributions of the projected 1000G PC scores for IBD and INTERVAL samples that pass the ancestry filters. No major distribution differences across the first ten principal components are present, suggesting a good ancestry mixture between cases and controls in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

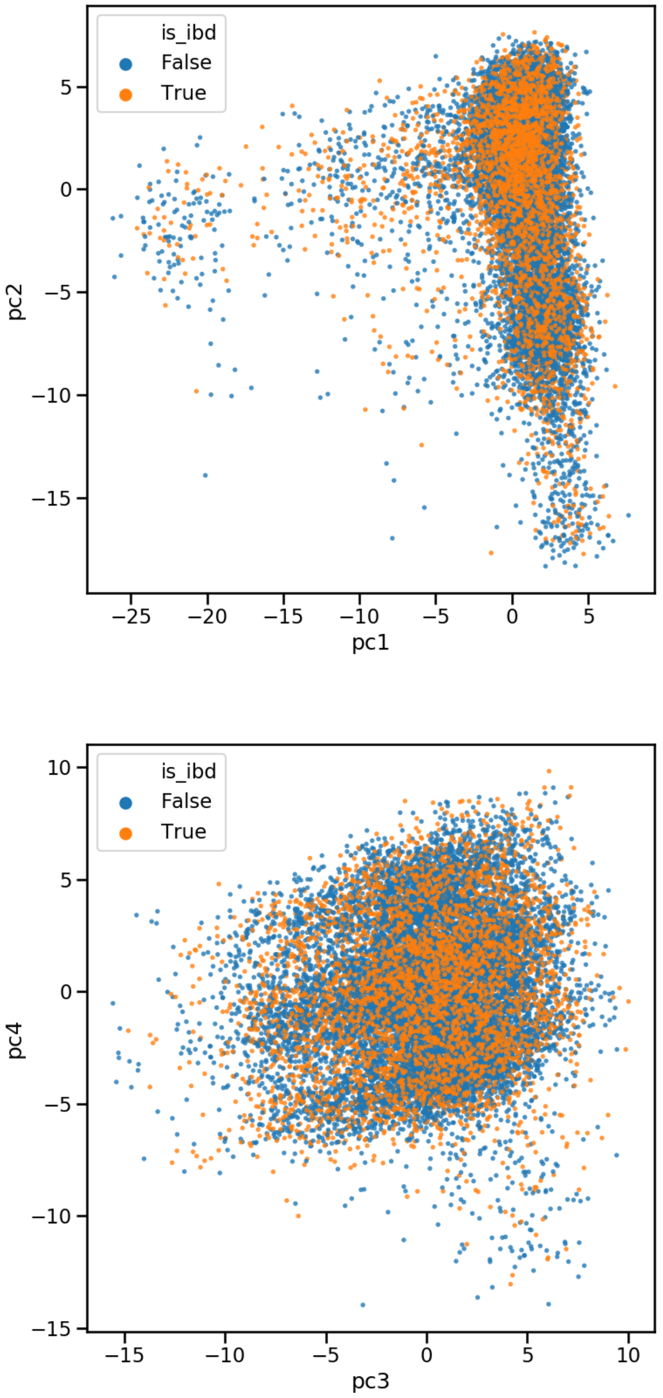


**Figure 4.9** The first four principal components of 15x cohort samples. Orange dots – IBD samples, blue dots – INTERVAL samples.

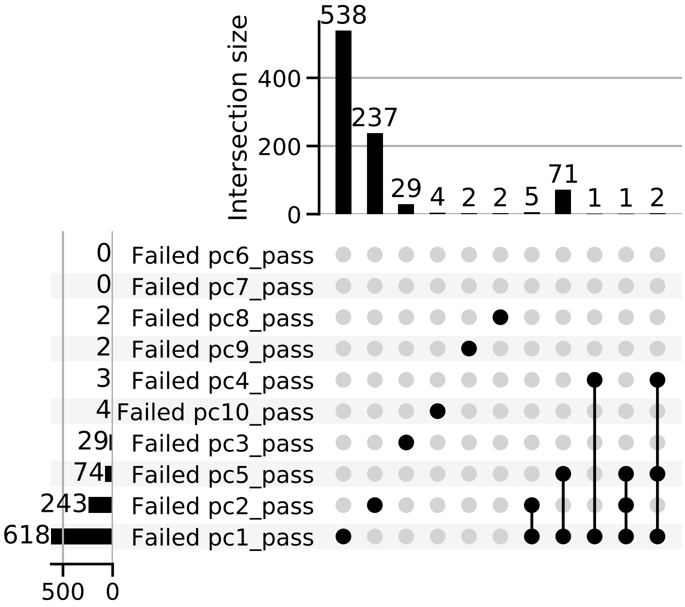


**Figure 4.10** Set intersection plot of the samples failing different within-cohort PC filters. Left bars – number of samples failing individual PC filters. Dots – set overlaps. Top bars – the number of samples overlapping between the sets. First six bars (single dots) – number of samples failing only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

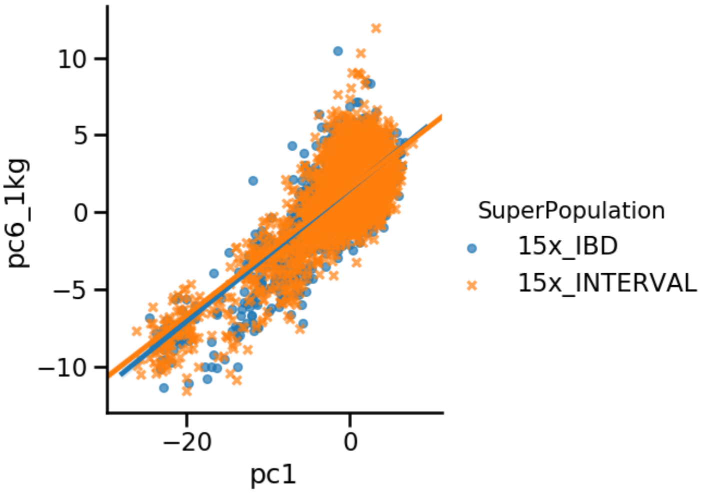


**Figure 4.11** PC1 scores correlate (0.7, Pearson) with the PC6$_{1000G}$ scores from the 1000G projection PCA. This suggests that the PC1 outliers are driven by Southern European genetic ancestry.

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.

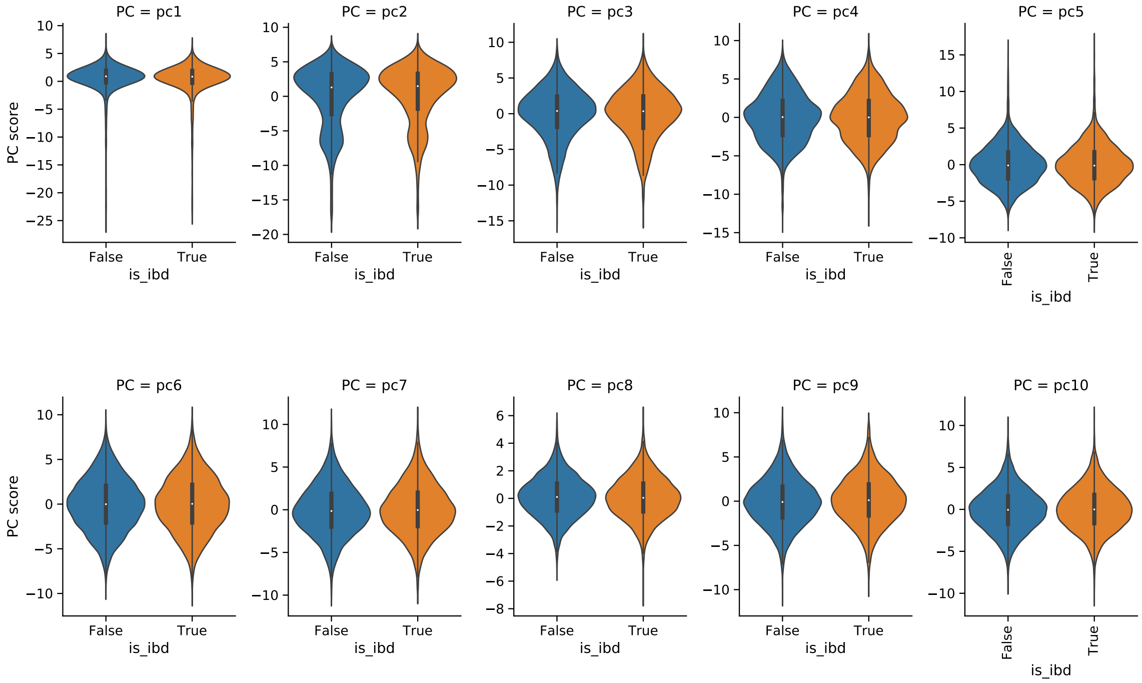


**Figure 4.12** Violin plot comparing the distributions of the projected PC scores for IBD and INTERVAL (without outlier removal). No major distribution differences across the first ten principal components are present, suggesting a good case-control mixture.

### 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 ∼2.0–2.1 for whole-genome datasets and ∼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

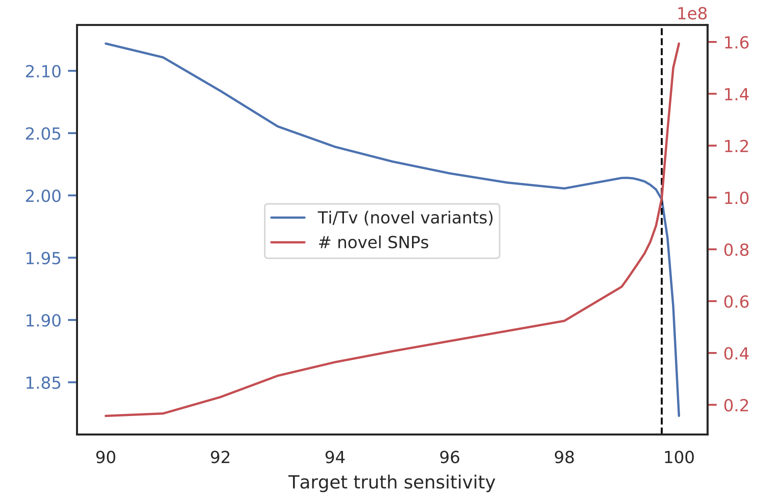
For INDELs, I fitted 4 Gaussians (reduced from 6 due to a smaller set of variants and the danger of overfitting) and used the following annotations: FS, ReadPosRankSum, MQRankSum, QD, SOR, DP.

The following resource sets were used for training:

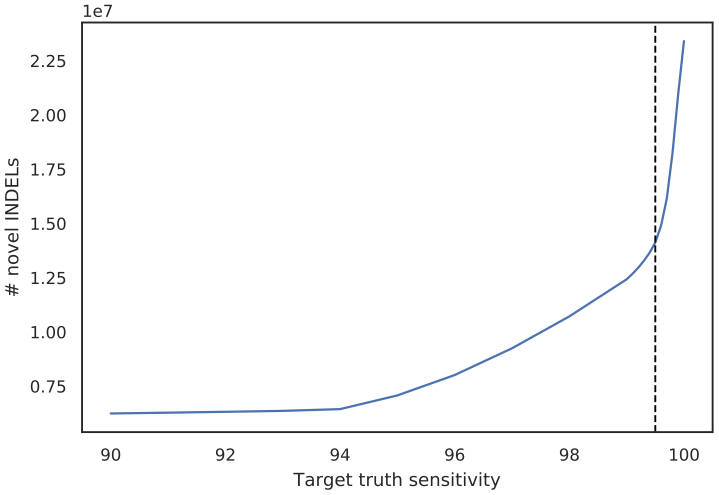
**SNP** True sites training resource: HapMap, Omni. Non-true sites training resource: 1000G. Known sites resource, not used in training: dbSNP.

**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.

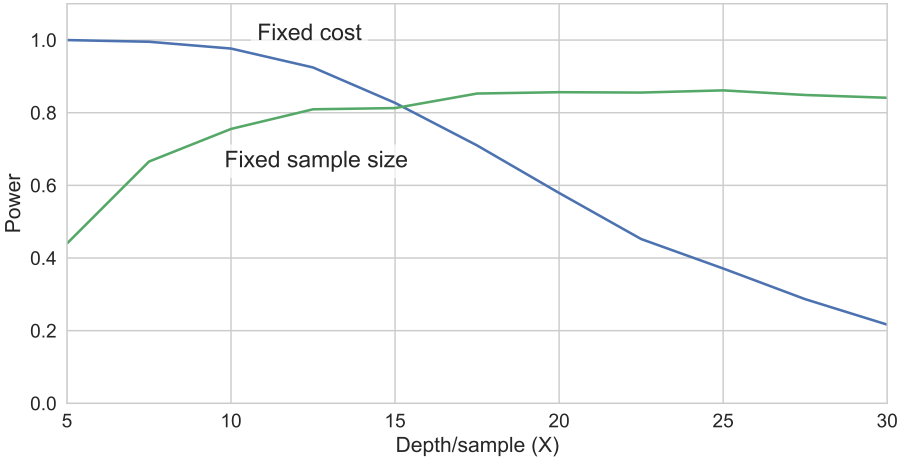


**Figure 4.14** Power to discover rare SNP variants in a case-control experiment setup. Blue line simulates the scenario with a limited budget and an unlimited number of available samples, while the green line shows the unlimited budget/limited cohort scenario.

[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.

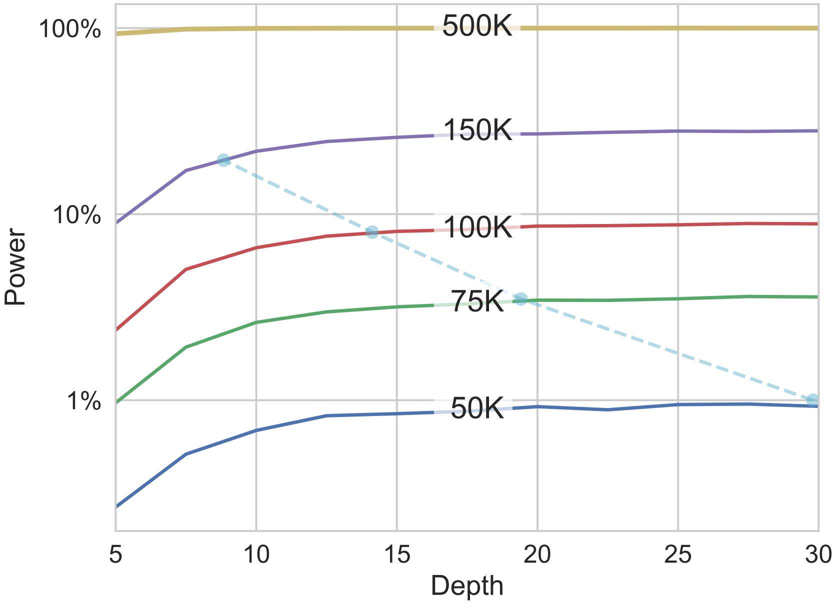


**Figure 4.15** Power to discover associations using burden tests for a set of rare variants with a cumulative frequency of $5 * 10^{-4}$ in a gene, assuming $\beta = 0.5$ s.d. and $\alpha = 1x10^{-6}$. The blue dotted line shows the trade-off between depth, cost and power: sequencing 50,000 samples 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).

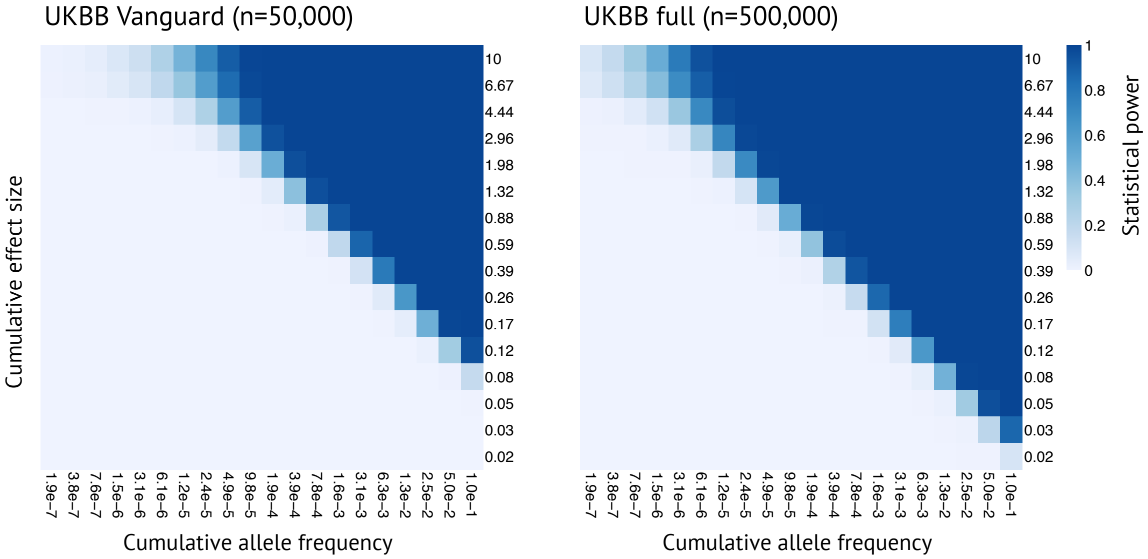


**Figure 4.16** Power to discover associations by aggregated rare variants in the UK Biobank ($\alpha$ = $1x10^{-9}$). Pilot release (‘Vanguard’ project, left, n=50,000): near perfect power to discover associations for variants with a cumulative $\beta$ greater than 0.6 SD and frequency of 6 : 1,000. Full UKBB WGS (right, n=500,000): near perfect power to discover associations for variants with a cumulative $\beta$ 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=4.59x10$^{-41}$). 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<1.80x10$^{-308}$). 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

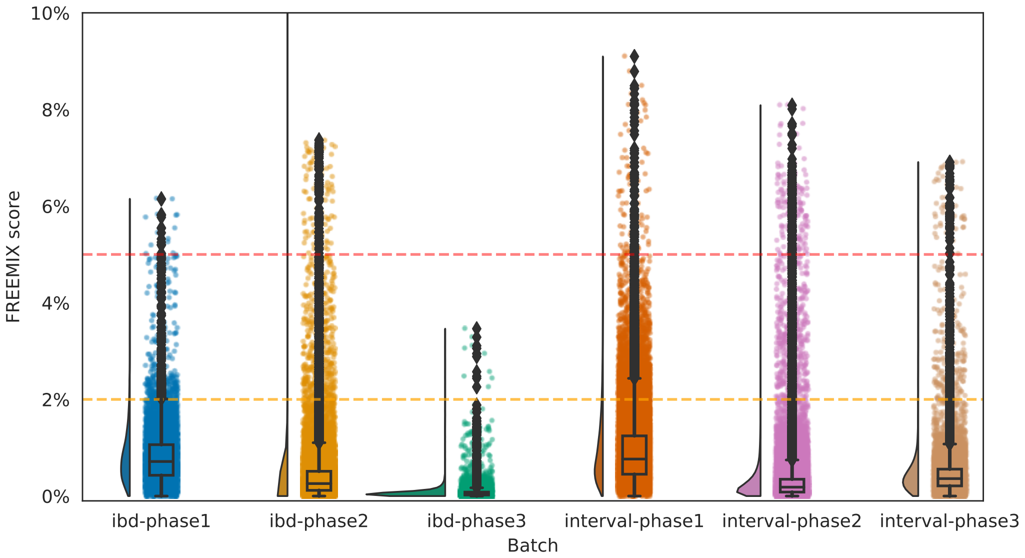


**Figure 4.17** FREEMIX scores of 111,225 read groups of samples from the IBD and INTERVAL 15x (median of 6 read groups per sample). Read groups are categorised by project batch. The box plots show the median levels of contamination per batch. Scatter plots and and density plots indicate the distribution 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.

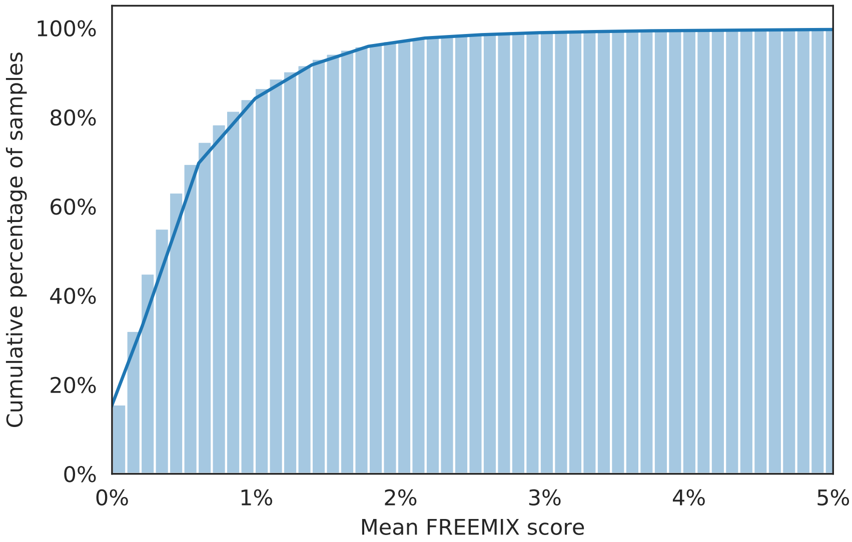


**Figure 4.18** Mean FREEMIX for each sample in the INTERVAL and IBD samples. The absolute majority of the samples (96%) have a FREEMIX below the 'potential contamination' level of 2%.

### 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

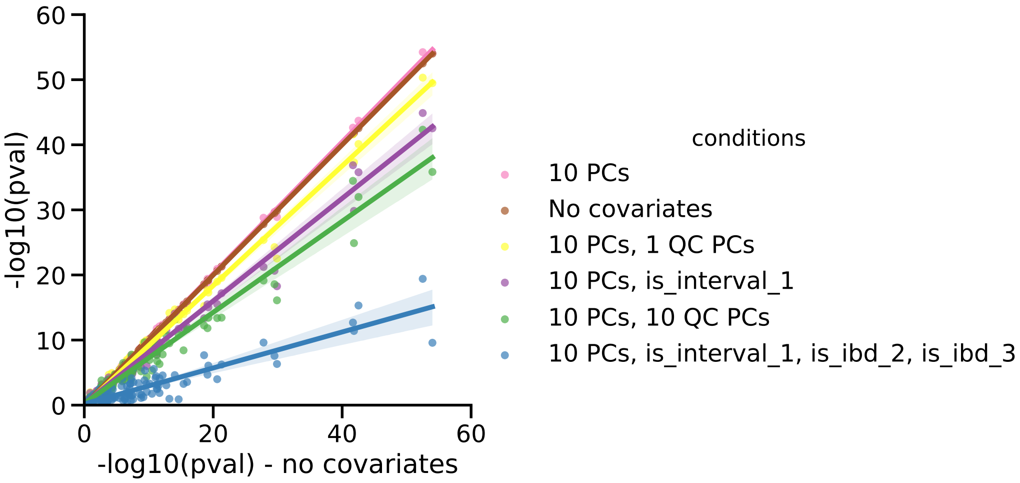


**Figure 4.19** Influence of the inclusion of different covariate types on the power to identify known Crohn's associations in the IBD 15x cohort. X axis – p-values when replicating known Crohn's disease associations when performing logistic regression with no covariates. Y axis – p-values when replicating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

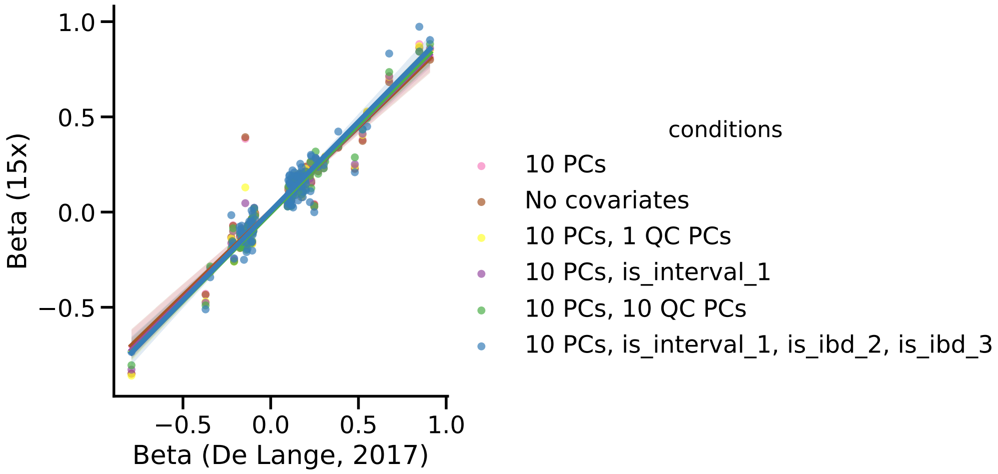


**Figure 4.20** Betas of the known Crohn's disease associations estimated in the 15x cohort are strongly concordant to the ones reported in de Lange et al. [49]: minimal Peason r=0.95 (no covariates), maximal r=0.97 (10 PCs, 10 QC PCs).

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 $\alpha$=1x$10^{-5}$ 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 ($\alpha$=4.3x$10^{-7}$ [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 = $7.25{\times}10^{-220}$, $I^2$ 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 = $1.54{\times}10^{-8}$) is around 1 Mb away from an intronic variant rs6062496 that has an odds ratio $\sim$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-$\alpha$ 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 = $4.11{\times}10^{-10}$). 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 | OR | OR | OR | 95% CI | I2 | P | P | P |
|---|---|---|---|---|---|---|---|---|---|---|
| | | INT | NFE | 15x | meta | meta | EUR | NFE | 15x | 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 | OR | OR | 95% CI | I2 | P | P | P |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | NFE | 15x | meta | meta | EUR | NFE | 15x | 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 $\lambda$ 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 × 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.

# Chapter 5

# Discussion

This dissertation describes three projects that explore different aspects of IBD genetics and pharmacogenetics of therapies used to treat IBD. In essence, all three were association studies of array genotyping or sequencing data. However, each posed unique challenges. The anti-TNF immunogenicity project, described in Chapter 2, required a non-standard genome-wide proportional hazards analysis followed by a scrupulous examination of the only significant association in order to understand which HLA allele it maps to and how it influences immunogenicity across different treatment regimes. In Chapter 3, I attempted to uncover the genetic variation associated with thiopurine-induced liver damage. While this project did not result in any robust associations, it underscores the importance of considering the quality of the dataset, data normalisation, and sample size when conducting GWAS. Lastly, in Chapter 4, I discuss the sample quality control and the initial association analysis of a large whole-genome sequencing dataset – IBD 15x. The scale of WGS datasets poses new computational challenges, which had to be addressed to enable further analyses. In addition, rare variant association studies require stringent quality control to avoid spurious genetic associations.

More than ten years since the first genome-wide association study [192], the genetics of IBD is far from being 'solved'. Large scale GWAS have demonstrated the genetic complexity of the disorder, underpinned by both the number of the associated loci and the complexity of resolving them down to a single variant and gene. Shortly thereafter, the GWAS techniques, and often the same datasets, were applied to study a variety of IBD-related traits, including disease progression, phenotype heterogeneity, and drug response. Below, I provide an overview of several projects that use the techniques developed for identifying and elucidating trait-associated variation in order to understand novel aspects of the genetics of IBD.

## 5.1 Longitudinal studies for drug response leveraging expression data

Longitudinal studies follow the participants over a prolonged period of time, recording events of interest (e.g., treatment complication, remission, flareup). In addition, longitudinal studies often include a perturbation event at the beginning of the observation period (e.g., start of treatment). Such studies are widely used in epidemiological research and clinical trials, yet remain quite novel in the field of complex disease genetics.

The association between HLA-DQA1*05 and immunogenicity was, in part, established due to the longitudinal design of the PANTS study: in a purely case-control setting the association just about passed the genome-wide significance threshold. However, when performing a time to event analysis using the Cox proportional hazards regression, the association became much more robust. Additional statistical power allowed me to investigate the effects of the allele across different treatment regimes and to identify the non-additive nature of the association. As I will discuss in Section 5.3, for pharmacogenetic GWAS this approach is rarely used due to the difficulty of collecting longitudinal cohorts at scale.

However, smaller scale longitudinal studies (e.g., clinical trials) can leverage a combination of genotyping and gene expression or microbiome data to study the biological processes behind drug response.

Gene expression analysis is used to quantify the level of gene product across all genes. Gene expression was previously measured using microarrays, which have now largely been superseded by RNA-seq and a plethora of new single cell sequencing methods. Gene expression is measured at various points in time, including before the treatment is administered for the first time (base expression).

Once the outcome of the trial is known (e.g., responders versus non-responders, normal versus adverse drug reaction), longitudinal gene expression data can be analysed to derive response signatures that associate the expression at base level to the measured outcome. The composition of the expression signatures can be investigated to understand which individual genes are up- and downregulated, providing additional insights into the biology of drug response.

Furthermore, eQTL mapping can be utilised to understand the role of genetic variation in drug response. Expression of individual genes is mapped to genetic variants in close (cis-EQTLs) or distant proximity (trans-eQTLs) from the gene. When the study utilises a longitudinal design, it is possible to map eQTLs at multiple time points; of particular interest are the eQTLs that exhibit a differential magnitude of effect across time points as they point out the likely mechanisms behind drug response.

RNA-seq at different time points was recently performed for around 400 individuals from the PANTS study – 200 responders and 200 non-responders. My colleague will shortly start analysing these data.

## 5.2 Host-microbiome interactions

IBD, being a disorder of the gastrointestinal tract, has long thought to be associated with changes in the gut microbiome. However, these changes are not well characterised. For example, dysbiosis (microbial imbalance) is frequently reported amongst IBD patients. However, no single microorganism has been consistently reported as being the one that dominates this imbalance [100]. Govers et al. describe the presence of dysbiosis amongst treatment-naive CD patients, suggesting that it is not entirely driven by microbiome alternations caused by therapies [66]. At the same time, it is unknown whether dysbiosis is caused by some of the symptoms of IBD themselves (e.g., diarrhoea), or driven by genetic variation, or is in fact itself a potential 'trigger' of the disease. Several approaches currently used in human disease genetics can be used to elucidate the causal relationship between the the host genetic variation, microbiome and IBD.

QTL mapping techniques can be used to identify genetic variants associated with the abundance levels of specific microorganisms. This analysis should be followed by a colocalization comparison with known IBD loci. Colocalization of the known IBD risk variants with the microbiome eQTLs would suggest that the bacterial makeup in the gut is partially driven by host genetics. One of the advantages of this approach is that the eQTLs can be mapped in the large-scale non-IBD microbiome datasets.

Sanna et al. [162] have used the two-sample bidirectional Mendelian randomisation technique to evaluate the causal relationship between host genetics, production of SCFA butyrate and insulin response, establishing a causal link. Similar approaches can be applied to IBD.

Recent work by Zimmermann et al. [199] indicates that microbiome-encoded enzymes can influence the metabolism of a broad variety of drugs. Future IBD pharmacogenetic studies can perhaps evaluate whether host genetics influence the microbiome composition, thereby modifying the drug metabolism, leading to drug non-response and adverse drug reactions.

## 5.3 Extending anti-TNF pharmacogenetic analysis

The association study of the PANTS cohort has resulted in the first known robust genetic association for immunogenicity to anti-TNF. While this was not the first attempt to do this, the previous studies were, arguably, subject to several methodological issues: small sample size (N < 500); or a targeted approach (e.g. specific HLA alleles, TNF); or the lack of self- or external replication.

The lessons learnt during the early stages of complex disease and trait GWAS still apply to pharmacogenetics – ‘good enough’ phenotyping combined with a decent sample size and adequate genotyping quality is more optimal than maximising either of the three components.

At the current stage of the field, pharmacogenetics appears to be limited largely by phenotype availability. While the GWAS of many complex diseases are now exceeding several hundreds of thousands of disease-affected subjects, genetic association studies of drug response for these conditions rarely exceed a thousand samples. The analysis described in the PANTS was based on a clinical trial that ran between 2013 and 2019 and was managed by 400 principal investigators and nurses across 150 research centres. While it is possible to meta-analyse these data with other anti-TNF trial cohorts in the future, it is evident that increasing the sample size by an order of magnitude would require finding alternative approaches for phenotype collection. One approach that could be worth exploring is to leverage the retrospective data available in bioresources and biobanks.

### 5.3.1 Analysing retrospective data from the NIHR IBD BioResource

The IBD BioResource project [143] has so far recruited 27,000 IBD patients. Based on the medication questionnaire, $\sim$8,000 of them have undergone treatment with anti-TNF and approximately an additional 800 have been treated with vedolizumab (a biologic drug targeting integrin $\alpha_4\beta_7$). The patients have consented to participate in research studies.

The BioResource participants are asked to fill out questionnaires on a variety of topics, including diet, lifestyle, and drug response. The drug response questionnaire captures the approximate dates (month and year) of when they have started and finished certain therapies – such as thiopurines, anti-TNF, and vedolizumab – and whether these have worked. The questionnaire responses can be used to reconstruct the approximate treatment timelines that can be analysed with a survival model. In addition, a blood sample is taken upon enrolment into BioResource. While this does not allow the measurement of antibodies longitudinally (like in PANTS), the analysis of the PANTS data indicates that patients who develop immunogenicity will maintain it for multiple observation time-points. The difference between the immunogenicity time-point and the blood draw may reduce the power of the time-to-even analyses. However, when I analysed the retrospective replication cohort used in the PANTS project (using the delta between start date and sampling date as time to event/censoring), the replication results were very consistent with the main prospective cohort. Therefore, I believe that the BioResource dataset can be used to study both the anti-TNF response and immunogenicity.

I prioritised the patients on anti-TNF for inclusion in our 15x and WES cohorts and $\sim$3,000 of them were already sequenced. In addition, efforts to genotype all 25,000 patients are ongoing. These data can be used for future anti-TNF pharmacogenetic projects. Increasing the sample size $\sim$7x (assuming all 8,000 anti-TNF patients get genotyped) will likely yield additional associations between genetic variation and response to anti-TNF, further improving our understanding of therapeutic response.

Some preliminary work was carried out to evaluate the feasibility of this project. I trialled imputing HLA alleles from sequencing data via HLA-LA. Compared to the genotype-based imputation techniques (like HIBAG), HLA-LA is a graph-based method that uses sequence-level data (e.g., aligned BAMs or CRAMs) to achieve higher imputation resolution and accuracy. This method imputes the HLA alleles at the G group-level resolution which matches the resolution of clinical HLA typing (compared to the more crude 2- and 4-digit level imputation that I used in Chapter 2). Increased accuracy may enable the identification of novel associations within the HLA, though it appears that the immunogenicity is fully driven by the HLA-DQA1*05.

When examining an early release of the BioResource anti-TNF phenotype data, it became apparent that the real-world treatment histories are much more heterogeneous than those of the patients in the PANTS cohort: patients change the the types of biologic therapies, start and stop immunomodulators, and have gaps in the treatment history. During the analysis of the PANTS dataset, I used the right-censored Cox proportional-hazards regression to analyse time to immunogenicity. The model used a number of covariates, including immunomodulator usage, type of the anti-TNF drug (infliximab versus adalimumab), assuming they remain constant throughout the enrolment. In order to use the BioResource data to perform time-to-even analyses, one would need to control for changing covariates during the observation period. Time-varying survival regression can be used to achieve this [196]. To the best of my knowledge, no current GWAS survival analysis tools support it. In order to address this issue, I have written a prototype software package that is able to perform the regular Cox proportional-hazard and the time-varying regressions on the genotype data.

### 5.3.2 Prescription records from the UK Biobank

The UK Biobank (UKBB) is a population-scale cohort that has recruited 500,000 individuals from all across the United Kingdom. The participants were genotyped, and are currently being whole-exome (first 50,000 samples available) and whole-genome sequenced. Recently, anonymised GP prescription records of 222,000 individuals were released. Although the GPs are unlikely to prescribe anti-TNF therapy themselves, their records *should* contain records of all drug prescriptions that the patient receives. The average age of the UKBB participants was 57 years upon recruitment, meaning that the cohort is enriched for individuals suffering from IBD, rheumatoid arthritis, and other conditions for treatment of which anti-TNF is used.

It would be interesting to analyse whether there are any genetic variants associated with shorter prescription length of anti-TNF. A reasonable assumption can be made that the patients that are prescribed anti-TNF for a couple of months did not respond to the treatment. I acknowledge that this phenotype is quite heterogeneous and may result in spurious associations. If any associations for ‘short anti-TNF prescription time’ are uncovered, these should be replicated in a well-phenotyped clinical trial cohort like PANTS.

Similar to anti-TNF, the same approach can be adopted in the study of thiopurine-induced myelosuppression and liver injury, where the typical time of the adverse reaction is known (e.g., ten weeks for hepatocellular TILI).

Overall, I believe that future pharmacogenetic studies will have to leverage retrospective data from both specialised bioresourses and population-scale datasets. Here, I have described two of such datasets. In addition, one could leverage insurance records (especially, in the US) and national drug prescription databases (such as the one that operates in Finland). Validation of the results from such proxy phenotypes could be challenging, which means that clinical trial datasets (like PANTS) will be required for validation.

# References

[1] 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

[2] Abegunde, A. T., Muhammad, B. H., Bhatti, O., and Ali, T. (2016). Environmental risk factors for inflammatory bowel diseases: evidence based literature review. *World Journal of Gastroenterology*, 22(27):6296.

[3] Allchin, W. H. (1909). A discussion on 'ulcerative colitis': introductory address. *Proceedings of the Royal Society of Medicine*, 2(Med Sect):59.

[4] Aloi, M., Nuti, F., Stronati, L., and Cucchiara, S. (2014). Advances in the medical management of paediatric IBD. *Nature Reviews Gastroenterology & Hepatology*, 11(2):99.

[5] Ananthakrishnan, A. N., Khalili, H., Konijeti, G. G., Higuchi, L. M., de Silva, P., Fuchs, C. S., Willett, W. C., Richter, J. M., and Chan, A. T. (2014). Long-term intake of dietary fat and risk of ulcerative colitis and Crohn's disease. *Gut*, 63(5):776–784.

[6] Ananthakrishnan, A. N., Khalili, H., Konijeti, G. G., Higuchi, L. M., de Silva, P., Korzenik, J. R., Fuchs, C. S., Willett, W. C., Richter, J. M., and Chan, A. T. (2013). A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. *Gastroenterology*, 145(5):970–977.

[7] Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564.

[8] Anselmo, A. C., Gokarn, Y., and Mitragotri, S. (2018). Non-invasive delivery strategies for biologics. *Nature Reviews Drug Discovery*.

[9] Arthur, R., Schulz-Trieglaff, O., Cox, A. J., and O'Connell, J. (2016). AKT: ancestry and kinship toolkit. *Bioinformatics*, 33(1):142–144.

[10] Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429.

[11] Aterido, A., Palau, N., Domènech, E., Mateu, P. N., Gutiérrez, A., Gomollón, F., Mendoza, J. L., Garcia-Planella, E., Barreiro-de Acosta, M., Muñoz, F., et al. (2019).

Genetic association between *CD96* locus and immunogenicity to anti-TNF therapy in Crohn's disease. *The Pharmacogenomics Journal*, 19(6):547–555.

[12] Avitzur, Y., Guo, C., Mastropaolo, L. A., Bahrami, E., Chen, H., Zhao, Z., Elkadri, A., Dhillon, S., Murchie, R., Fattouh, R., et al. (2014). Mutations in tetratricopeptide repeat domain 7A result in a severe form of very early onset inflammatory bowel disease. *Gastroenterology*, 146(4):1028–1039.

[13] Axelrad, J. E., Roy, A., Lawlor, G., Korelitz, B., and Lichtiger, S. (2016). Thiopurines and inflammatory bowel disease: current evidence and a historical perspective. *World Journal of Gastroenterology*, 22(46):10103.

[14] Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6):659.

[15] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, 40(8):955.

[16] Bastida, G., Nos, P., Aguas, M., Beltrán, B., Rubín, A., Dasí, F., and Ponce, J. (2005). Incidence, risk factors and clinical course of thiopurine-induced liver injury in patients with inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics*, 22(9):775–782.

[17] Beigel, F., Deml, M., Schnitzler, F., Breiteneicher, S., Göke, B., Ochsenkühn, T., and Brand, S. (2014). Rate and predictors of mucosal healing in patients with inflammatory bowel disease treated with anti-TNF-alpha antibodies. *PLoS One*, 9(6):e99293.

[18] Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., and de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution*, 47(1):30.

[19] Benchimol, E. I., Mack, D. R., Guttmann, A., Nguyen, G. C., To, T., Mojaverian, N., Quach, P., and Manuel, D. G. (2015). Inflammatory bowel disease in immigrants to Canada and their children: a population-based cohort study. *The American Journal of Gastroenterology*, 110(4):553.

[20] Benner, C., Havulinna, A. S., Järvelin, M.-R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics*, 101(4):539–551.

[21] Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

[22] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045.

[23] Billiet, T., Vande Casteele, N., Van Stappen, T., Princen, F., Singh, S., Gils, A., Ferrante, M., Van Assche, G., Cleynen, I., and Vermeire, S. (2015). Immunogenicity to infliximab is associated with *HLA-DRB1*. *Gut*, 64(8):1344–5.

[24] Björnsson, E. S., Bergmann, O. M., Björnsson, H. K., Kvaran, R. B., and Olafsson, S. (2013). Incidence, presentation, and outcomes in patients with drug-induced liver injury in the general population of Iceland. *Gastroenterology*, 144(7):1419–1425.

[25] Blackstone, E. A. and Joseph, P. F. (2013). The economics of biosimilars. *American Health & Drug Benefits*, 6(8):469.

[26] Bouameur, J.-E., Favre, B., and Borradori, L. (2014). Plakins, a versatile family of cytolinkers: roles in skin integrity and in human diseases. *Journal of Investigative Dermatology*, 134(4):885–894.

[27] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

[28] Broad Institute Genomic Services (2019). Human whole exome sequencing. http://genomics.broadinstitute.org/products/whole-exome-sequencing. Accessed: 26.06.2019.

[29] Brodie, A., Azaria, J. R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? *Nucleic Acids Research*, 44(13):6046–6054.

[30] Broekman, M. M., Coenen, M. J., Wanten, G. J., van Marrewijk, C. J., Klungel, O. H., Verbeek, A. L., Hooymans, P. M., Guchelaar, H.-J., Scheffer, H., Derijks, L. J., et al. (2017). Risk factors for thiopurine-induced myelosuppression and infections in inflammatory bowel disease patients with a normal *TPMT* genotype. *Alimentary Pharmacology & Therapeutics*, 46(10):953–963.

[31] Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *The American Journal of Human Genetics*, 78(6):903–913.

[32] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213.

[33] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2018). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

[34] Burdett, T., Hall, P., Hastings, E., Hindorff, L., Junkins, H., Klemm, A., MacArthur, J., Manolio, T., Morales, J., Parkinson, H., et al. (2016). The NHGRI-EBI catalog of published genome-wide association studies. *Available at: www.ebi.ac.uk/gwas*.

[35] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203.

[36] Calkins, B. M. (1989). A meta-analysis of the role of smoking in inflammatory bowel disease. *Digestive Diseases and Sciences*, 34(12):1841–1854.

[37] Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, 200(3):719–736.

[38] Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., de Jager, P. L., Sunyaev, S. R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics*, 49(4):600.

[39] Cirulli, E. T., Nicoletti, P., Abramson, K., Andrade, R. J., Bjornsson, E. S., Chalasani, N., Fontana, R. J., Hallberg, P., Li, Y. J., Lucena, M. I., et al. (2019). A missense variant in *PTPN22* is a risk factor for drug-induced liver injury. *Gastroenterology*, 156(6):1707–1716.

[40] Cleynen, I., Boucher, G., Jostins, L., Schumm, L. P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J. M., Annese, V., Brand, S., et al. (2016). Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *The Lancet*, 387(10014):156–167.

[41] Colombel, J. F., Sandborn, W. J., Reinisch, W., Mantzaris, G. J., Kornbluth, A., Rachmilewitz, D., Lichtiger, S., D'haens, G., Diamond, R. H., Broussard, D. L., et al. (2010). Infliximab, azathioprine, or combination therapy for Crohn's disease. *New England Journal of Medicine*, 362(15):1383–1395.

[42] Concannon, P., Rich, S. S., and Nepom, G. T. (2009). Genetics of type 1A diabetes. *New England Journal of Medicine*, 360(16):1646–1654.

[43] Cooper, G. M. and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628.

[44] Cornish, A. and Guda, C. (2015). A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International*, 2015.

[45] Daly, A. K., Donaldson, P. T., Bhatnagar, P., Shen, Y., Pe'er, I., Floratos, A., Daly, M. J., Goldstein, D. B., John, S., Nelson, M. R., et al. (2009). HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nature Genetics*, 41(7):816.

[46] Dave, C. V., Hartzema, A., and Kesselheim, A. S. (2017). Prices of generic drugs associated with numbers of manufacturers. *New England Journal of Medicine*, 377(26):2597–2598.

[47] de Bruyn, M. and Vermeire, S. (2017). *NOD2* and bacterial recognition as therapeutic targets for Crohn's disease. *Expert Opinion on Therapeutic Targets*, 21(12):1123–1139.

[48] de Groot, A. S., Knopp, P. M., and Martin, W. (2005). De-immunization of therapeutic proteins by T-cell epitope modification. *Developments in Biologicals*, 122:171–94.

[49] de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, 49(2):256.

[50] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

[51] DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., et al. (2017). Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. *Cell Stem Cell*, 20(4):533–546.

[52] del Val, J. H. (2011). Old-age inflammatory bowel disease onset: a different problem? *World Journal of Gastroenterology*, 17(22):2734.

[53] D'Haens, G., Baert, F., van Assche, G., Caenepeel, P., Vergauwe, P., Tuynman, H., de Vos, M., van Deventer, S., Stitt, L., Donner, A., et al. (2008). Early combined immunosuppression or conventional management in patients with newly diagnosed Crohn's disease: an open randomised trial. *The Lancet*, 371(9613):660–667.

[54] D'Haens, G., Reinisch, W., Panaccione, R., Satsangi, J., Petersson, J., Bereswill, M., Arikan, D., Perotti, E., Robinson, A. M., Kalabic, J., Alperovich, G., Thakkar, R., and Loftus, E. V. (2018). Lymphoma risk and overall safety profile of adalimumab in patients with Crohn's disease with up to 6 years of follow-up in the pyramid registry. *The American Journal of Gastroenterology*, 113(6):872–882.

[55] Dilthey, A. T., Mentzer, A. J., Carapito, R., Cutland, C., Cereb, N., Madhi, S. A., Rhie, A., Koren, S., Bahram, S., McVean, G., et al. (2019). HLA*LA – HLA typing from linearly projected graph alignments. *Bioinformatics*, 35(21):4394–4396.

[56] Drubay, D., Gautheret, D., and Michiels, S. (2018). A benchmark study of scoring methods for non-coding mutations. *Bioinformatics*, 34(10):1635–1641.

[57] Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463.

[58] Ek, W. E., D'Amato, M., and Halfvarson, J. (2014). The history of genetics in inflammatory bowel disease. *Annals of Gastroenterology*, 27(4):294.

[59] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57.

[60] Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M., Blackwell, J. M., Cordell, H. J., Wellcome Trust Case Control Consortium 2, et al. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics*, 10(7).

[61] Fevr, T., Robine, S., Louvard, D., and Huelsken, J. (2007). Wnt/$\beta$-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Molecular and Cellular Biology*, 27(21):7551–7559.

[62] Fisher, S. A., Tremelling, M., Anderson, C. A., Gwilliam, R., Bumpstead, S., Prescott, N. J., Nimmo, E. R., Massey, D., Berzuini, C., Johnson, C., et al. (2008). Genetic determinants of ulcerative colitis include the *ECM1* locus and five loci implicated in Crohn's disease. *Nature Genetics*, 40(6):710.

[63] Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., et al. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41.

[64] Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098.

[65] Gearry, R. B., Barclay, M. L., Burt, M. J., Collett, J. A., and Chapman, B. A. (2004). Thiopurine drug adverse effects in a population of New Zealand patients with inflammatory bowel disease. *Pharmacoepidemiology and Drug Safety*, 13(8):563–567.

[66] Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe*, 15(3):382–392.

[67] Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5).

[68] Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R., Melloni, G., Hatzikotoulas, K., Farmaki, A.-E., Ritchie, G., Schwartzentruber, J., et al. (2018). Very low depth whole genome sequencing in complex trait association studies. *bioRxiv*.

[69] Gisbert, J. P., González-Lama, Y., and Maté, J. (2007). Thiopurine-induced liver injury in patients with inflammatory bowel disease: a systematic review. *The American Journal of Gastroenterology*, 102(7):1518.

[70] Goldstein, D. B. et al. (2009). Common genetic variation and human traits. *New England Journal of Medicine*, 360(17):1696.

[71] González-Galarza, F. F., Takeshita, L. Y., Santos, E. J., Kempson, F., Maia, M. H. T., da Silva, A. L. S., Silva, A. L. T. e., Ghattaoraya, G. S., Alfirevic, A., Jones, A. R., and Middleton, D. (2015). Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research*, 43(D1):D784–D788.

[72] Gordon, H., Trier Moller, F., Andersen, V., and Harbord, M. (2015). Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflammatory Bowel Diseases*, 21(6):1428–1434.

[73] Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E. S., Annese, V., Hauser, S. L., Oksenberg, J. R., Thomsen, I., Leslie, S., Daly, M. J., Van Steen, K., Duerr, R. H., Barrett, J. C., McGovern, D. P. B., Schumm, L. P., Traherne, J. A., Carrington, M. N., Kosmoliaptsis, V., Karlsen, T. H., Franke, A., and Rioux, J. D. (2015). High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature Genetics*, 47(2):172–179.

[74] Grob, S. and Cavalli, G. (2018). Technical review: a hitchhiker's guide to chromosome conformation capture. In *Plant Chromatin Dynamics*, pages 233–246. Springer.

[75] GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204.

[76] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252.

[77] Hail Team (2019). Hail. https://github.com/hail-is/hail/releases/tag/0.2.20. Accessed: 26.06.2019.

[78] Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54.

[79] Hedin, C. and Halfvarson, J. (2018). Should we use vedolizumab as mono or combo therapy in ulcerative colitis? *Best Practice & Research Clinical Gastroenterology*, 32:27–34.

[80] Hoofnagle, J. H. (2013). Livertox: a website on drug-induced liver injury. In *Drug-Induced Liver Disease*, pages 725–732. Elsevier.

[81] Hooper, K. M., Casanova, V., Kemp, S., Staines, K. A., Satsangi, J., Barlow, P. G., Henderson, P., and Stevens, C. (2019). The inflammatory bowel disease drug azathioprine induces autophagy via mTORC1 and the unfolded protein response sensor PERK. *Inflammatory Bowel Diseases*, 25(9):1481–1496.

[82] Huang, H., Fang, M., Jostins, L., Mirkov, M. U., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662):173.

[83] Illumina (2017). Effects of index misassignment on multiplexing and downstream analysis. https://emea.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf. Accessed: 26.06.2019.

[84] International HapMap Consortium et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299.

[85] International Human Genome Sequencing Consortium et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860.

[86] IQVIA Institute (2019). The Global Use of Medicine in 2019 and Outlook to 2023.

[87] Ji, S.-G., Juran, B. D., Mucha, S., Folseraas, T., Jostins, L., Melum, E., Kumasaka, N., Atkinson, E. J., Schlicht, E. M., Liu, J. Z., et al. (2017). Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nature Genetics*, 49(2):269.

[88] Johnson, J. L. and Abecasis, G. R. (2017). Gas power calculator: web-based power calculator for genetic association studies. *bioRxiv*.

[89] Johnston, R. D. and Logan, R. F. (2008). What is the peak age for onset of IBD? *Inflammatory Bowel Diseases*, 14(2):S4–S5.

[90] Jones, G.-R., Lyons, M., Plevris, N., Jenkinson, P. W., Bisset, C., Burgess, C., Din, S., Fulforth, J., Henderson, P., Ho, G.-T., et al. (2019). IBD prevalence in Lothian, Scotland, derived by capture-recapture methodology. *Gut*, 68(11):1953–1960.

[91] Jostins, L., Levine, A. P., and Barrett, J. C. (2013). Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments. *PLoS One*, 8(10).

[92] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., et al. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119.

[93] Kaplan, G. G. (2015). The global burden of IBD: from 2015 to 2025. *Nature Reviews Gastroenterology & Hepatology*, 12(12):720.

[94] Kaplan, G. G. and Ng, S. C. (2017). Understanding and preventing the global increase of inflammatory bowel disease. *Gastroenterology*, 152(2):313–321.

[95] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*.

[96] Kennedy, N. A., Heap, G. A., Green, H. D., Hamilton, B., Bewshea, C., Walker, G. J., Thomas, A., Nice, R., Perry, M. H., Bouri, S., et al. (2019). Predictors of anti-TNF treatment failure in anti-TNF-naive patients with active luminal Crohn's disease: a prospective, multicentre, cohort study. *The Lancet Gastroenterology & Hepatology*, 4(5):341–353.

[97] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219.

[98] King, E. A., Davis, J. W., and Degner, J. F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *bioRxiv*.

[99] Kirschner, B. S. (2016). Indeterminate colitis/inflammatory bowel disease unclassified (IBD-U). In *Textbook of Pediatric Gastroenterology, Hepatology and Nutrition*, pages 335–340. Springer.

[100] Knox, N. C., Forbes, J. D., van Domselaar, G., and Bernstein, C. N. (2019). The gut microbiome as a target for IBD treatment: are we there yet? *Current Treatment Options in Gastroenterology*, 17(1):115–126.

[101] Kondrashova, A., Mustalahti, K., Kaukinen, K., Viskari, H., Volodicheva, V., Haapala, A.-M., Ilonen, J., Knip, M., Mäki, M., Hyöty, H., et al. (2008). Lower economic status and inferior hygienic environment may protect against celiac disease. *Annals of Medicine*, 40(3):223–231.

[102] Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20(1):117.

[103] Lamb, C. A., Kennedy, N. A., Raine, T., Hendy, P. A., Smith, P. J., Limdi, J. K., Hayee, B., Lomer, M. C., Parkes, G. C., Selinger, C., et al. (2019). British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. *Gut*, 68(Suppl 3):s1–s106.

[104] Lee, P. H., Lee, C., Li, X., Wee, B., Dwivedi, T., and Daly, M. (2018). Principles and methods of in-silico prioritization of non-coding regulatory variants. *Human Genetics*, 137(1):15–30.

[105] Levine, A. P., Pontikos, N., Schiff, E. R., Jostins, L., Speed, D., Lovat, L. B., Barrett, J. C., Grasberger, H., Plagnol, V., Segal, A. W., et al. (2016). Genetic complexity of Crohn's disease in two large Ashkenazi Jewish families. *Gastroenterology*, 151(4):698–709.

[106] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

[107] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[108] Li, Y., Sung, W.-K., and Liu, J. J. (2007). Association mapping via regularized regression analysis of single-nucleotide–polymorphism haplotypes in variable-sized sliding windows. *The American Journal of Human Genetics*, 80(4):705–715.

[109] Lichtenstein, G. R., Feagan, B. G., Cohen, R. D., Salzberg, B. A., Diamond, R. H., Langholff, W., Londhe, A., and Sandborn, W. J. (2014). Drug therapies and the risk of malignancy in Crohn's disease: results from the treat registry. *The American Journal of Gastroenterology*, 109(2):212–23.

[110] Liu, J. Z. and Anderson, C. A. (2014). Genetic studies of Crohn's disease: past, present and future. *Best Practice & Research Clinical Gastroenterology*, 28(3):373–386.

[111] Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9):979–986.

[112] Liu, M., Degner, J., Davis, J. W., Idler, K. B., Nader, A., Mostafa, N. M., and Waring, J. F. (2018). Identification of *HLA-DRB1* association to adalimumab immunogenicity. *PLoS One*, 13(4):e0195325.

[113] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197.

[114] Lucena, M. I., Molokhia, M., Shen, Y., Urban, T. J., Aithal, G. P., Andrade, R. J., Day, C. P., Ruiz-Cabello, F., Donaldson, P. T., Stephens, C., et al. (2011). Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology*, 141(1):338–347.

[115] Lund-Nielsen, J., Vedel-Krogh, S., Kobylecki, C. J., Brynskov, J., Afzal, S., and Nordestgaard, B. G. (2018). Vitamin D and inflammatory bowel disease: Mendelian randomization analyses in the Copenhagen studies and UK Biobank. *The Journal of Clinical Endocrinology & Metabolism*, 103(9):3267–3277.

[116] Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C., et al. (2017). Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nature Genetics*, 49(2):186.

[117] Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T.-G. C., and Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human Molecular Genetics*, 26(18):3639–3650.

[118] Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C., Prokopenko, I., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234.

[119] Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294.

[120] McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279.

[121] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[122] Megiorni, F. and Pizzuti, A. (2012). HLA-DQA1 and HLA-DQB1 in Celiac disease predisposition: practical implications of the HLA molecular typing. *Journal of Biomedical Science*, 19:88.

[123] Mersha, T. B. and Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Human Genomics*, 9(1):1.

[124] Morris, A. P. (2011). Trans-ethnic meta-analysis of genome-wide association studies. *Genetic Epidemiology*, 35(8):809–822.

[125] Natarajan, P., Peloso, G. M., Zekavat, S. M., Montasser, M., Ganna, A., Chaffin, M., Khera, A. V., Zhou, W., Bloom, J. M., Engreitz, J. M., et al. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications*, 9(1):3391.

[126] Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3).

[127] Neale Lab (2018). GWAS of the UK Biobank. http://www.nealelab.is/uk-biobank/. Accessed: 26.06.2019.

[128] Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. A. (2009). Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)*, 324(5925):387–9.

[129] Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8):856.

[130] Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., Panaccione, R., Ghosh, S., Wu, J. C., Chan, F. K., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: A systematic review of population-based studies. *The Lancet*, 390(10114):2769–2778.

[131] Ng, S. C., Tang, W., Leong, R. W., Chen, M., Ko, Y., Studd, C., Niewiadomski, O., Bell, S., Kamm, M. A., de Silva, H., et al. (2015). Environmental risk factors in inflammatory bowel disease: a population-based case-control study in Asia-Pacific. *Gut*, 64(7):1063–1071.

[132] Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362.

[133] Nicoletti, P., Aithal, G. P., Bjornsson, E. S., Andrade, R. J., Sawle, A., Arrese, M., Barnhart, H. X., Bondon-Guitton, E., Hayashi, P. H., Bessone, F., et al. (2017). Association of liver injury from specific drugs, or groups of drugs, with polymorphisms in HLA and other genes in a genome-wide association study. *Gastroenterology*, 152(5):1078–1089.

[134] Nielsen, D. S. G., Fredborg, M., Andersen, V., and Purup, S. (2017). Administration of protein kinase D1 induces a protective effect on lipopolysaccharide-induced intestinal inflammation in a co-culture model of intestinal epithelial Caco-2 cells and RAW264. 7 macrophage cells. *International Journal of Inflammation*.

[135] O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476.

[136] Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., et al. (2001). A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603.

[137] Oliveira-Cortez, A., Melo, A. C., Chaves, V. E., Condino-Neto, A., and Camargos, P. (2016). Do HLA class II genes protect against pulmonary tuberculosis? A systematic review and meta-analysis. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(10):1567–80.

[138] Oo, C. and Kalbag, S. S. (2016). Leveraging the attributes of biologics and small molecules, and releasing the bottlenecks: a new wave of revolution in drug development. *Expert Review of Clinical Pharmacology*, 9(6):747–749.

[139] Osterman, M. T., Sandborn, W. J., Colombel, J.-F., Robinson, A. M., Lau, W., Huang, B., Pollack, P. F., Thakkar, R. B., and Lewis, J. D. (2014). Increased risk of malignancy with adalimumab combination therapy, compared with monotherapy, for Crohn's disease. *Gastroenterology*, 146(4):941–9.

[140] Panaccione, R., Ghosh, S., Middleton, S., Márquez, J. R., Scott, B. B., Flint, L., van Hoogstraten, H. J., Chen, A. C., Zheng, H., Danese, S., et al. (2014). Combination therapy with infliximab and azathioprine is superior to monotherapy with either agent in ulcerative colitis. *Gastroenterology*, 146(2):392–400.

[141] Panaccione, R., Loftus, E. V., Binion, D., McHugh, K., Alam, S., Chen, N., Guerette, B., Mulani, P., and Chao, J. (2011). Efficacy and safety of adalimumab in Canadian patients with moderate to severe Crohn's disease: results of the adalimumab in Canadian subjects with moderate to severe Crohn's disease (ACCESS) trial. *Canadian Journal of Gastroenterology = Journal canadien de gastroenterologie*, 25(8):419–25.

[142] Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, 50(3):381.

[143] Parkes, M. (2019). IBD BioResource: an open-access platform of 25,000 patients to accelerate research in Crohn's and colitis. *Gut*, 68(9):1537–1540.

[144] Pearson, D. C., May, G. R., Fick, G. H., and Sutherland, L. R. (1995). Azathioprine and 6-mercaptopurine in Crohn disease: a meta-analysis. *Annals of Internal Medicine*, 123(2):132–142.

[145] Perry, M., Bewshea, C., Brown, R., So, K., Ahmad, T., and McDonald, T. (2015). Infliximab and adalimumab are stable in whole blood clotted samples for seven days at room temperature. *Annals of Clinical Biochemistry*, 52(6):672–674.

[146] Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573.

[147] Pollard, M. (2016). Variant discovery, accuracy and coverage on the Illumina HiSeq X. Unpublished presentation.

[148] Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983.

[149] Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., et al. (2008). Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, 83(1):132–135.

[150] Pulit, S. L., de With, S. A. J., and de Bakker, P. I. W. (2016). The multiple testing burden in sequencing-based disease studies of global populations. *bioRxiv*.

[151] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[152] Rashkin, S., Jun, G., Chen, S., Abecasis, G. R., Genetics and Epidemiology of Colorectal Cancer Consortium, et al. (2017). Optimal sequencing strategies for identifying disease-associated singletons. *PLoS Genetics*, 13(6).

[153] Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., Wilkie, A. O., McVean, G., Lunter, G., WGS500 Consortium, et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912.

[154] Rivas, M. A., Graham, D., Sulem, P., Stevens, C., Desch, A. N., Goyette, P., Gudbjartsson, D., Jonsdottir, I., Thorsteinsdottir, U., Degenhardt, F., et al. (2016). A protein-truncating R179X variant in *RNF186* confers protection against ulcerative colitis. *Nature Communications*, 7:12342.

[155] Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. (2014). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, 43(D1):D423–D431.

[156] Roda, G., Jharap, B., Neeraj, N., and Colombel, J.-F. (2016). Loss of response to anti-TNFs: definition, epidemiology, and management. *Clinical and Translational Gastroenterology*, 7(1):e135.

[157] Roe, D. and Kuang, R. (2019). Predicting KIR structural haplotypes with novel sequence signatures from short-read whole genome sequencing. *bioRxiv*.

[158] Romero-Cara, P., Torres-Moreno, D., Pedregosa, J., Vílchez, J. A., García-Simón, M. S., Ruiz-Merino, G., Morán-Sanchez, S., and Conesa-Zamora, P. (2018). A *FCGR3A* polymorphism predicts anti-drug antibodies in chronic inflammatory bowel disease patients treated with anti-TNF. *International Journal of Medical Sciences*, 15(1):10.

[159] Rozen, P., Zonis, J., Yekutiel, P., and Gilat, T. (1979). Crohn's disease in the Jewish population of Tel-Aviv-Yafo: epidemiologic and clinical aspects. *Gastroenterology*, 76(1):25–30.

[160] Rozpedek, W., Pytel, D., Mucha, B., Leszczynska, H., Diehl, J. A., and Majsterek, I. (2016). The role of the PERK/eIF2$\alpha$/ATF4/CHOP signaling pathway in tumor progression during endoplasmic reticulum stress. *Current Molecular Medicine*, 16(6):533–544.

[161] Sandborn, W. J., Rutgeerts, P., Enns, R., Hanauer, S. B., Colombel, J.-F., Panaccione, R., D'Haens, G., Li, J., Rosenfeld, M. R., Kent, J. D., and Pollack, P. F. (2007). Adalimumab induction therapy for Crohn disease previously treated with infliximab. *Annals of Internal Medicine*, 146(12):829.

[162] Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vila, A. V., Võsa, U., Mujagic, Z., Masclee, A. A., Jonkers, D. M., Oosting, M., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics*, 51(4):600.

[163] Sathish, J. G., Sethu, S., Bielsky, M.-C., de Haan, L., French, N. S., Govindappa, K., Green, J., Griffiths, C. E. M., Holgate, S., Jones, D., Kimber, I., Moggs, J., Naisbitt, D. J., Pirmohamed, M., Reichmann, G., Sims, J., Subramanyam, M., Todd, M. D., van der Laan, J. W., Weaver, R. J., and Park, B. K. (2013). Challenges and approaches for the development of safer immunomodulatory biologics. *Nature Reviews Drug Discovery*, 12(4):306–24.

[164] Sazonovs, A. and Barrett, J. (2018). Rare-variant studies to complement genome-wide association studies. *Annual Review of Genomics and Human Genetics*, 19:97–112.

[165] Sazonovs, A., Kennedy, N. A., Moutsianas, L., Heap, G. A., Rice, D. L., Reppell, M., Bewshea, C. M., Chanchlani, N., Walker, G. J., Perry, M. H., et al. (2020). HLA-DQA1*05 carriage associated with development of anti-drug antibodies to infliximab and adalimumab in patients with Crohn's disease. *Gastroenterology*, 158(1):189–199.

[166] Schröder, T., Schmidt, K. J., Olsen, V., Möller, S., Mackenroth, T., Sina, C., Lehnert, H., Fellermann, K., and Büning, J. (2015). Liver steatosis is a risk factor for hepatotoxicity in patients with inflammatory bowel disease under immunosuppressive treatment. *European journal of Gastroenterology & Hepatology*, 27(6):698–704.

[167] Shah, T., Liu, J., Floyd, J., Morris, J. A., Wirth, N., Barrett, J. C., and Anderson, C. (2012). optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics*, 28(12):1598–1603.

[168] Shim, J. O. (2019). Recent advance in very early onset inflammatory bowel disease. *Pediatric Gastroenterology, Hepatology & Nutrition*, 22(1):41–49.

[169] Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., et al. (2016). Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nature Neuroscience*, 19(4):571–577.

[170] Singh, T., Walters, J. T., Johnstone, M., Curtis, D., Suvisaari, J., Torniainen, M., Rees, E., Iyegbe, C., Blackwood, D., McIntosh, A. M., et al. (2017). The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nature Genetics*, 49(8):1167–1173.

[171] Soon, S., Molodecky, N. A., Rabi, D. M., Ghali, W. A., Barkema, H. W., and Kaplan, G. G. (2012). The relationship between urban environment and the inflammatory bowel diseases: a systematic review and meta-analysis. *BMC Gastroenterology*, 12(1):51.

[172] Soranzo, N. (2018). Whole genome sequencing in the UK Biobank. http://www.ukbiobank.ac.uk/wp-content/uploads/2018/07/1145-Soranzo-UPDATED-1.pdf. Accessed: 26.06.2019.

[173] Southam, L., Gilly, A., Süveges, D., Farmaki, A.-E., Schwartzentruber, J., Tachmazidou, I., Matchan, A., Rayner, N. W., Tsafantakis, E., Karaleftheri, M., et al. (2017). Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nature Communications*, 8(1):1–11.

[174] Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–R119.

[175] Stone, M. A., Mayberry, J. F., and Baker, R. (2003). Prevalence and management of inflammatory bowel disease: a cross-sectional study from central England. *European Journal of Gastroenterology & Hepatology*, 15(12):1275–1280.

[176] Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., Holm, H., Kong, A., Thorsteinsdottir, U., Sulem, P., et al. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics*, 48(3):314.

[177] Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G. R., Xifara, D. K., Matchan, A., Hatzikotoulas, K., Rayner, N. W., Chen, Y., et al. (2013). A rare functional cardioprotective *APOC3* variant has risen in frequency in distinct population isolates. *Nature Communications*, 4:2872.

[178] Tang, R., Feng, T., Sha, Q., and Zhang, S. (2009). A variable-sized sliding-window approach for genetic association studies via principal component analysis. *Annals of Human Genetics*, 73(6):631–637.

[179] Tran-Minh, M.-L., Sousa, P., Maillet, M., Allez, M., and Gornet, J.-M. (2017). Hepatic complications induced by immunosuppressants and biologics in inflammatory bowel disease. *World Journal of Hepatology*, 9(13):613.

[180] UK10K consortium et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82.

[181] Urban, T. J., Nicoletti, P., Chalasani, N., Serrano, J., Stolz, A., Daly, A. K., Aithal, G. P., Dillon, J., Navarro, V., Odin, J., et al. (2017). Minocycline hepatotoxicity: clinical characterization and identification of HLA-B*35:02 as a risk factor. *Journal of Hepatology*, 67(1):137–144.

[182] Ursum, J., van der Weijden, M. A. C., van Schaardenburg, D., Prins, A. P. A., Dijkmans, B. A. C., Twisk, J. W. R., Crusius, J. B. A., and van der Horst-Bruinsma, I. E. (2010). IL10 GGC haplotype is positively and HLA-DQA1*05-DQB1*02 is negatively associated with radiographic progression in undifferentiated arthritis. *The Journal of Rheumatology*, 37(7):1431–8.

[183] Vermeire, S., Gils, A., Accossato, P., Lula, S., and Marren, A. (2018). Immunogenicity of biologics in inflammatory bowel disease. *Therapeutic Advances in Gastroenterology*, 11.

[184] Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., et al. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*.

[185] Walker, G. and Ahmad, T. (2019). Drug toxicity: personalising IBD therapeutics – the use of genetic biomarkers to reduce drug toxicity. In *Biomarkers in Inflammatory Bowel Diseases*, pages 257–269. Springer.

[186] Walker, G. J., Harrison, J. W., Heap, G. A., Voskuil, M. D., Andersen, V., Anderson, C. A., Ananthakrishnan, A. N., Barrett, J. C., Beaugerie, L., Bewshea, C. M., et al. (2019). Association of genetic variants in *NUDT15* with thiopurine-induced myelosuppression in patients with inflammatory bowel disease. *JAMA*, 321(8):773–785.

[187] Wan, Y. (2009). TPMT testing before azathioprine therapy? *Drug and Therapeutics Bulletin*, 47(1):9.

[188] Wang, X. and Teo, Y.-Y. (2015). Trans-ethnic fine-mapping of rare causal variants. In *Assessing Rare Variation in Complex Traits*, pages 253–261. Springer.

[189] Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348.

[190] Wellcome Trust Case Control Consortium et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

[191] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

[192] Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T., et al. (2005). Single nucleotide polymorphisms in *TNFSF15* confer susceptibility to Crohn's disease. *Human Molecular Genetics*, 14(22):3499–3506.

[193] Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M., et al. (2011). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807.

[194] Yang, S.-K., Hong, M., Baek, J., Choi, H., Zhao, W., Jung, Y., Haritunians, T., Ye, B. D., Kim, K.-J., Park, S. H., et al. (2014). A common missense variant in *NUDT15* confers susceptibility to thiopurine-induced leukopenia. *Nature Genetics*, 46(9):1017.

[195] Zhang, F., Flickinger, M., Taliun, S. A. G., Abecasis, G. R., Scott, L. J., McCaroll, S. A., Pato, C. N., Boehnke, M., Kang, H. M., InPSYght Psychiatric Genetics Consortium, et al. (2020). Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Research*, 30(2):185–194.

[196] Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine*, 6(7).

[197] Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T., Lee, C. M., Banskota, S., Marigorta, U. M., Bao, G., and Gibson, G. (2016). A burden of rare variants associated with extremes of gene expression in human peripheral blood. *The American Journal of Human Genetics*, 98(2):299–309.

[198] Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., and Weir, B. S. (2014). HIBAG—HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, 14(2):192.

[199] Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature*, 570(7762):462.