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].

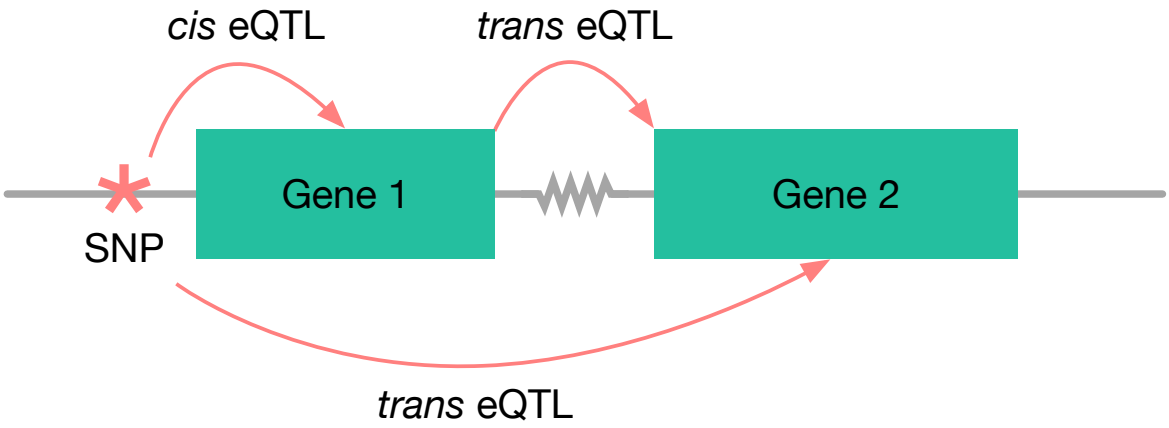

Figure 1.1 When a genetic variant affects the quantitative level of expression of a gene, it is known as an expression quantitative trait locus (eQTL). The majority of known eQTL are caused by variants very close to the affected gene ('*cis* eQTL'), often in the promoter. In rarer cases a variant can 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.

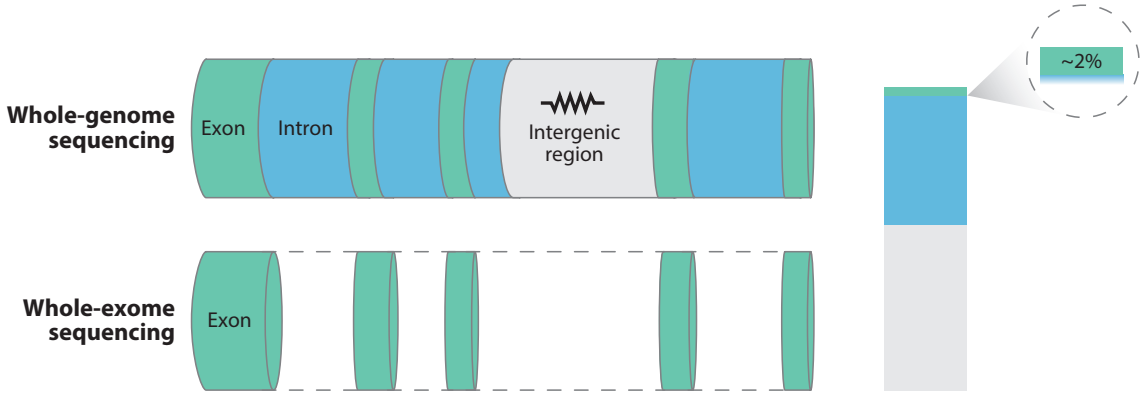

Figure 1.2 Whole genome sequencing captures all genetic variation, whereas exome sequencing targets the 2% that encodes proteins.

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.

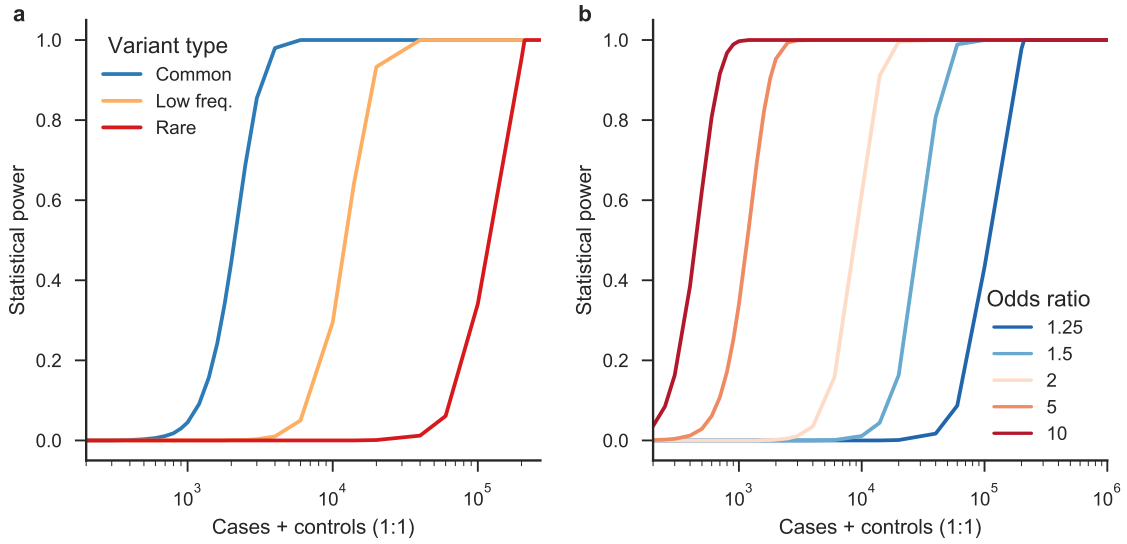

Figure 1.3 One-stage association study power calculations for single-variant tests using the method described by Johnson et al. [88]. (a) Power to detect common (disease allele frequency = 0.25), low frequency (0.025), and rare (0.0025) with genotype relative risk of 1.5 at 5×10^{-8} significance level. (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 α=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 5×10^{-9} for samples of European and East Asian ancestry and 1×10^{-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 2.5×10^{-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

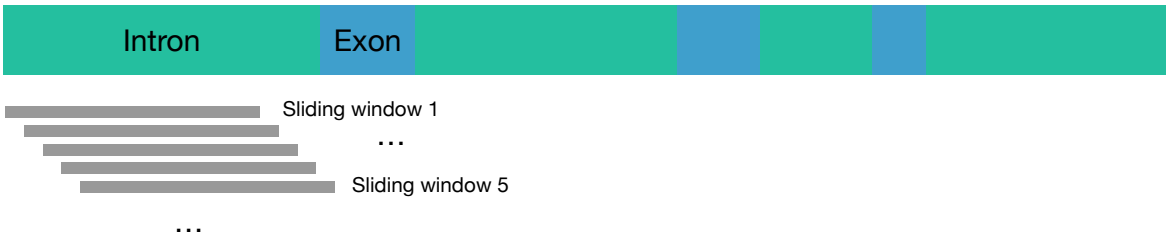

Figure 1.4 Sliding window technique.

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].

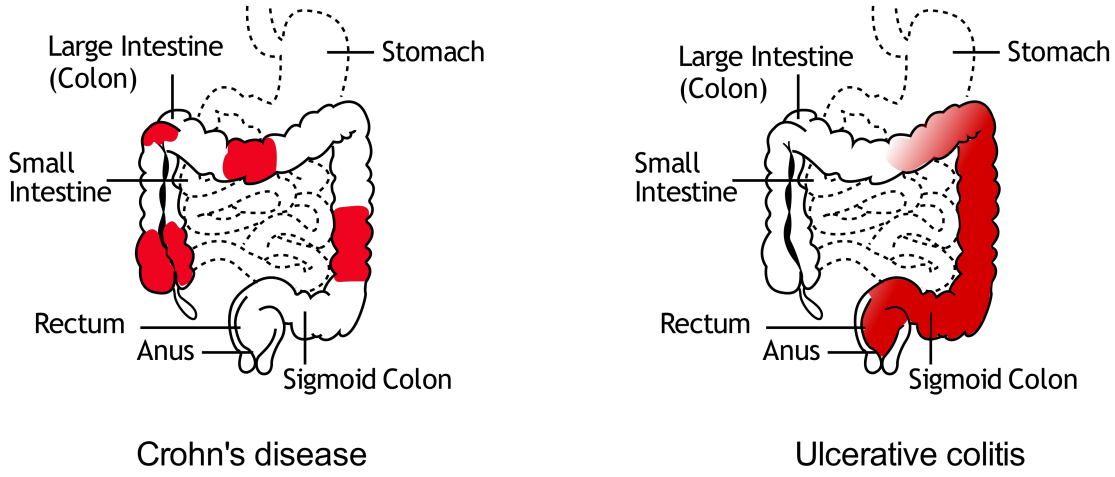

Figure 1.6 Crohn's disease and ulcerative colitis are two common types of inflammatory bowel disease which differ in patterns and location of inflammation. License information: *Modified work, derivate of File:Patterns of Crohn's Disease.svg by Samir, vectorized by Fvasconcellos [CC BY-SA 3.0]*

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.

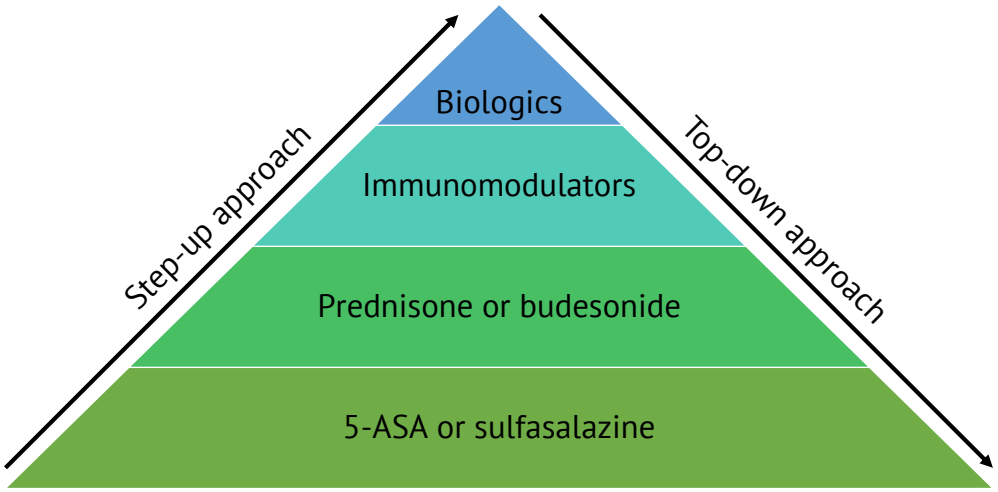

Figure 1.7 IBD treatment pyramid. Adapted from Aloi et al. [4]. Step-up approach starts with milder, less toxic therapies. Top-down approach prioritises early aggressive treatment (e.g., with biologics).

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 $\alpha 4\beta 7$ 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.