Chapter 5

Discussion and future directions

For over two decades, the study of genetics has been making significant progress towards understanding the causes of complex disorders such as inflammatory bowel disease. During this time, it has become evident that the substantial heritability of such traits cannot be explained by just a handful of high-impact genetic variants, arising instead through the cumulative contribution of hundreds of variants of relatively small effect. For IBD alone, well over 200 associated loci have been identified, largely driven by common variation. Now, with the advent of next generation sequencing technologies, we are able to interrogate rare and low frequency variation in a high throughput manner for the first time. This provides an exciting opportunity to investigate the role of rarer variation in complex disease risk on a genome-wide scale.

In this thesis I have described the analytical challenges that can arise when using sequencing to perform this sort of case-control association testing at scale. In particular, I focused on methods that can be used to overcome biases in the sensitivity and specificity of variant calling, as can occur when cohorts are sequenced to a different average read depth. I then applied these methods to investigate the role of rare and low frequency variation in inflammatory bowel disease, uncovering a significant burden of rare, damaging missense variation in the gene *NOD2*, as well as a more general burden of such variation amongst known inflammatory bowel disease risk genes. Through imputation into both new and existing GWAS cohorts,

I also described the discovery of a low frequency missense variant in *ADCY7* that approximately doubles the risk of ulcerative colitis. Finally, I meta-analysed these data with published GWAS summary statistics to identify a further 25 novel IBD-associated loci that are driven by common variation.

These results reveal important insights into the genetic architecture of inflammatory bowel disease. As well as the known role of common variation in disease risk, there is tantalising evidence of a potential role for rare variation affecting the same genes implicated by GWAS associations. In contrast, we observe just one high effect, low frequency variant associated with ulcerative colitis, suggesting that such variants as a class explain very little disease heritability. Overall, our results suggest that a combination of continued GWAS imputed using substantial new reference panels and large scale deep sequencing projects will be required in order to fully understand the genetic basis of complex diseases like IBD.

I then turned to the issue of how we can convert the successful identification of hundreds of disease associated loci into useful biological insights and, ultimately, directly impact the treatment and clinical diagnosis of these disorders. As an initial attempt at addressing this problem, we used fine-mapping and eQTL co-localization to resolve the biological mechanisms underlying several of the novel IBD associations identified in this study. In particular, we described likely causal missense variants in the genes *SLAMF8*, a negative regulator of inflammation, and *PLCG2*, a gene that has been implicated in primary immune deficiency. A further four signals were shown to be associated with monocyte-specific changes in integrin gene expression following immune stimulation. Interestingly, these genes encode proteins in pathways that have been identified as important therapeutic targets in IBD. Overall, we noted that new associations at common variants continue to identify genes that are relevant to therapeutic target identification and prioritization.

5.1 Studying complex genetic disease in the sequencing era

Looking forward to future experiments aimed at uncovering further risk loci for complex disease, there are two key paths that can be taken. The first is to continue to use array-based methods to cheaply genotype and impute hundreds of thousands of individuals, allowing for the detection of common variant associations of ever smaller effect size. As parallel sequencing efforts lead to the generation of improved imputation reference panels, the lower bound of the minor allele frequency spectrum that can be interrogated using this approach is likely to fall. Through the cost-effective collection of genetic information across very large samples, including expansion into non-European populations, the power to detect novel associations that may prove to be therapeutically relevant is greatly improved.

The second, complementary, approach is to perform large scale sequencing studies that focus on unearthing the role of rare variants in complex disease risk. These rare variants can be highly relevant for understanding the pathways underlying a given disease, or even identifying potential therapeutic targets. Compared to common variation, they are often more straightforward to interpret mechanistically, as they are correlated with fewer nearby variants. Although a standard protocol for performing array-based studies is well established, how exactly sequencing should be used to investigate rare variation in complex disease is not yet clear. In the following sections I will discuss some of the considerations that should be made when designing a next-generation sequencing study to investigate complex disease, and how this may change in the future.

5.1.1 Exome vs whole genome sequencing

The high costs associated with sequencing at scale require researchers to make difficult decisions between the breadth of genomic sequence captured, and the average read depth each interrogated site is covered to. The most popular approach thus far has been to focus on just the protein-coding exome (<2% of the total), using high coverage sequencing to discover rare, coding variants. These are exactly the class of variants expected to have the largest effects on disease risk, as negative selection acts to reduce the prevalence of harmful mutations in the population (Gibson, 2011). Exome-based studies are also advantaged by the wealth of existing knowledge around the potential role of variants which disrupt protein-coding sequence, making their functional interpretation much simpler than for those in non-coding regions. However, for many complex diseases the coding genome still explains only a fraction of the common variant associations found using GWAS: the vast majority of hits (>90%) lie in non-coding regions, with presumed regulatory roles (Maurano et al., 2012). To detect this type of variation it is necessary to use whole genome sequencing, which applies an untargeted approach to capture the full breadth of genomic sequence available to current technologies.

In this thesis I presented an intermediate approach to deep whole genome sequencing, where samples were sequenced to low average depth (<10x), sacrificing individual genotype quality in order to increase overall sample size. However, falling costs now mean that, just as exome sequencing has superseded targeted gene sequencing, these low coverage whole genomes are unlikely to be widely used in the future. Although deep whole genomes are still much more expensive than exomes, the cost ratio is not as severe as might be expected from the difference in target sizes. Because of variability in exome capture technology (Figure 5.1), exomes must be sequenced to an average depth of 50-100x in order to obtain accurate calls across the target region. In contrast, whole genome sequencing is highly accurate at an average depth of ∼20x (Figure 5.2).

Furthermore, falling costs associated with sequencing, combined with the fixed costs of DNA library preparation and exome capture, mean that the overall cost differential between exome and whole genome sequencing will continue to narrow.

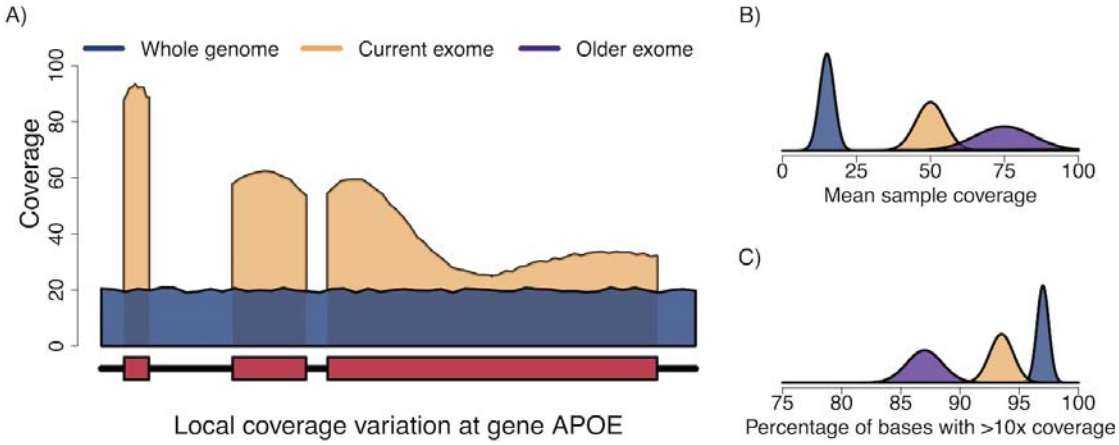

Figure 5.1: An illustration of the relative coverage that can be obtained using current exome (WES) and whole genome (WGS) sequencing techniques, and the improvements that have already been seen compared to initial WES protocols. WGS is able to produce much more even local coverage (panel A), that allows a lower global average coverage to be used (panel B) whilst still capturing the majority of sites to sufficient quality (panel C).

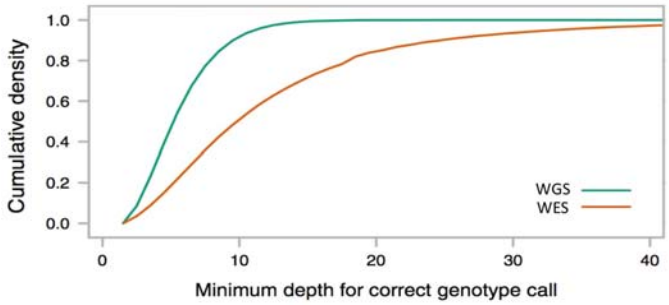

Figure 5.2: The minimum depth required to make a correct heterozygous genotype call in matched whole exome (WES) and whole genome (WGS) sequencing samples. Figure sourced from Meynert et al. (2014).

This means that deep whole genome sequencing will shortly be a viable alternative to exomes for large scale projects. Nevertheless, while each study design has their own set of advantages and disadvantages (Table 1), ultimately researchers must choose between capturing the regulatory genome and sequencing a larger number of samples.

Table 5.1: Study designs considerations when choosing between performing whole genome and whole exome sequencing as part of a fixed-cost study. *PCR: polymerase chain reaction; SNP: single nucleotide polymorphism; SV: structural variant.*

[1] A smaller, initial release of this dataset was first published under the name ExAC by Lek et al. (2016); however, a large dataset is now available under the name gnomAD at http://gnomad.broadinstitute.org/

[2] Although the exact resources used will vary between different technologies, these numbers reflect approximate storage requirements as detailed here: http://www.strand-ngs.com/support/ngs-data-storage-requirements.

	Whole exome sequencing	Whole genome sequencing
Coverage		
Genomic coverage	Protein coding sequence only ($< 2\%$)	All accessible sequence ($95 - 98\%$)
Average read depth	Usually $>$ 40x coverage	Most designs range from $6 - 30$x coverage
Consistency	Highly variable coverage, due to uneven probe efficiency and PCR amplification biases	Even coverage across sequence
Data quality		
Variant calling	High quality individual genotype calls for SNPs and indels, but calling SVs is difficult	High coverage studies produce quality individual genotype calls for SNPs, indels, and SVs Low coverage studies have poor individual calls, but population level data can improve accuracy
Variant interpretability	Extensive resources for predicting functional impact and severity of exonic variation	Resources indicating potential enhancers and promoters in numerous cell types and conditions are being developed, but overall the function of non-coding variation is still relatively unknown
External dataset comparisons		
Example control datasets	Publicly-available data from >120,000 exomes [1]	Publicly-available data from >15,000 genomes [1]

Continued on next page

Table 5.1 – *Continued from previous page*

	Whole exome sequencing	Whole genome sequencing
Joint analysis	A number of disease consortia have already collected large, exome-based datasets; however, variation in exome capture techniques can make joint analysis with these datasets a complex task	Relatively few large disease-specific whole genome datasets are currently available, although the exonic regions of genomes may be jointly analysed with existing exome datasets
Cost considerations		
Sample preparation	Samples must go through both library preparation and exome capture steps	Only library preparation is required
Sequencing	Exome sequencing currently costs more per read than genome sequencing, due to differences in the sequencing machines that may be used	Currently cheaper per read than exome sequencing, but it is expected that the prices will ultimately converge
Storage requirements	∼8GB per 40x exome [2]	∼150GB per 40x genome [2]
Computational resources	Smaller per-sample data tends to be easier and quicker to handle, although processing of large numbers of samples can prove very costly	High coverage whole genomes may be large and unwieldy to process Low coverage studies require genotype refinement steps that are often computationally expensive

5.1.2 Combining and analysing data across multiple studies

As briefly noted in Table 5.1, one consideration that should be made when designing next-generation sequencing studies is the availability of other datasets for joint analysis. We have already seen that the biggest complex disease discoveries of the genotyping era arose out of large, consortia-driven efforts that combined numerous studies in order to obtain very large sample sizes (Figure 1.8). Sequencing projects like ExAC have also reiterated the importance of creating these large merged datasets to better understand population diversity and interpret rare variation in a clinical setting (Lek et al., 2016).

There are several important logistical challenges that must be considered when embarking on this sort of large scale joint study, with respect to how data should best be shared and analysed. Unlike the meta-analysis approach adopted to combine the summary statistics from genome wide association studies, most large sequencing projects thus far have utilised a mega-analysis study design, where the raw data from multiple datasets is jointly called and analysed (Figure 5.3). While this requires the sharing of much more bulky raw data, and can exacerbate quality control and analysis difficulties by combining multi-source datasets, this method can also greatly improve the sensitivity and specificity of rare variant detection by joint calling across a much larger population.

Within the current limits set by the availability of sequencing data, the analytical benefits of this joint analysis have so far outweighed the computational strain of collating and re-analysing raw datasets. However, as sequencing sample sizes are currently growing much faster than computational resources, the costs involved in this process may ultimately make the sharing of raw data infeasible. Rather than reverting to the use of summary statistics, intermediate files such as the individual genotype probabilities (currently represented as gVCFs) may provide a good balance between data size and the ability to produce a consistent and well-powered study.

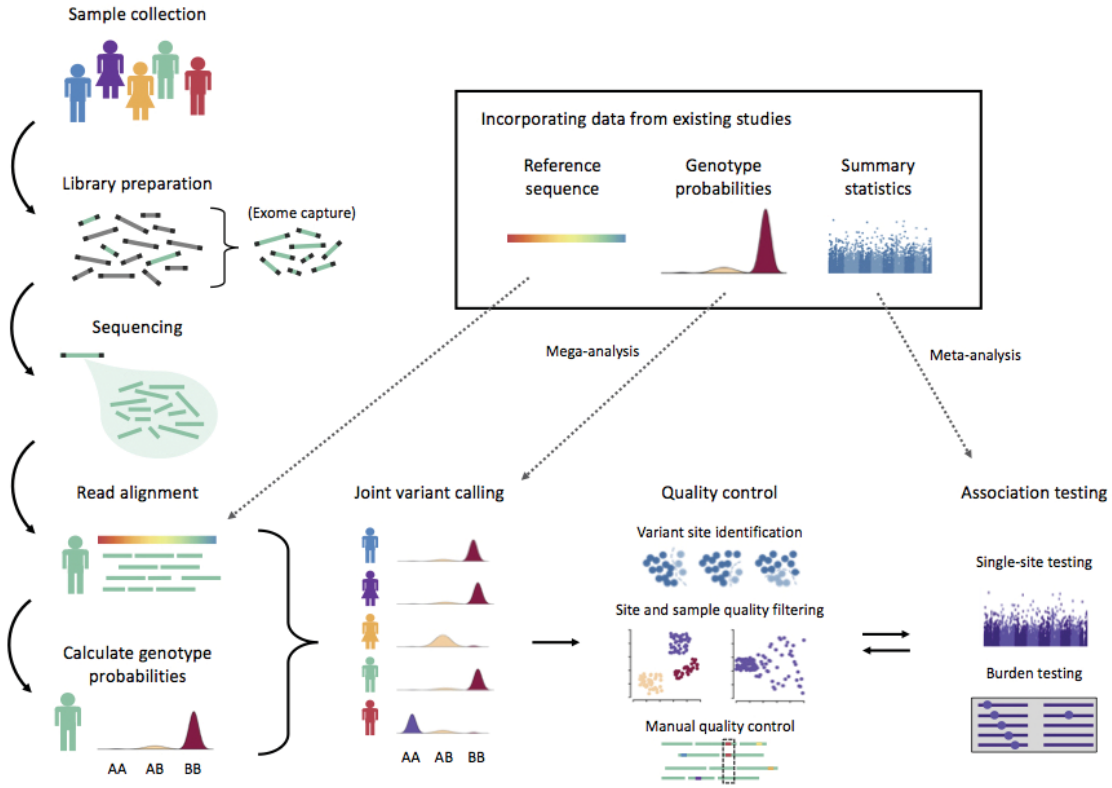

Figure 5.3: An overview of the key features of a sequence-based study, from initial sample collection and sequencing through to the final association testing and joint analysis with external datasets.

However, the sharing of intermediate files demands a degree of stability in the protocols being used; stability that is currently sorely lacking. The race to evolve our methods to keep up with the ever-growing deluge of data often leads to changes that can make the incorporation of old data difficult and can cause lengthy delays in the adoption of new techniques. A striking example of this is the reference genome used for alignment: despite GRCh38 being released over three years ago, most studies being published today still use the outdated GRCh37 reference in order to maintain compatibility with existing datasets and functional annotation resources. Until we are able to settle on a gold-standard protocol for performing sequencing studies, we will plausibly face the repeated re-analysis of thousands of DNA sequences in our attempts to generate large, quality datasets.

5.1.3 Overcoming computational limitations

The development of a gold-standard sequencing study protocol is likely to require significant changes in the way the genetics community as a whole tackles the storage and analysis of data. The sequencing analysis methods described in this thesis generally represent only incremental updates to those that are used for the analysis of genotyping data, and this slow rate of evolution is already struggling to keep up with the rapid changes occurring in the underlying sequencing technology. At this stage there is no evidence that the rate of technological advance is slowing down, particularly with the appearance of new competitors on the market, such as the long-read Pacific Biosciences system and the portable nanopore sequencing offered by Oxford Nanopore. Ultimately, revolutionary improvements in the wet-lab technology are demanding similar revolutions in our software and analysis techniques.

One of the major changes that is already starting to gain momentum in the field is the transition to cloud-based systems, which provide general access to very large computational resources. These systems are designed to perform tasks using massively parallel processing, and novel genetics software will be required to fully exploit this. Early developments in this area include Hail (Seed et al., 2017), a scalable analysis framework for genetic data, and Cromwell (https://github.com/broadinstitute/cromwell), a workflow execution engine that can run the existing Genome Analysis Toolkit in the cloud. Both tools have been designed for easy incorporation into scripted pipelines, which can greatly improve the reproducibility of analyses. The centralised nature of a cloud-based system also means that these pipelines may be easily shared between users to improve the consistency of datasets generated across a range of facilities.

In this way, cloud-based systems and massively parallel computing work to help solve the issues we currently face with scalability, reproducibility, and data sharing. Other efforts are focussing instead on improving the updatability of genomic datasets, in hopes of ensuring their continued relevance as we rapidly collect more information on global human variation. Amongst the more advanced lines of research in this area are graph-based genomes, which represent a collection of

sequences as a series of alternative paths through a mathematical graph. This system can better encode indels and other complex genomic features, and new sequences may be easily added to extend the underlying reference graph or customise it to population-specific variation (Dilthey et al., 2015; Dilthey et al., 2016). Initial developments around read alignment and variant calling using this approach have been favourable compared to current analysis methods (Novak et al., 2017).

These are just a couple of examples of new approaches that are being developed to support the imminent influx of sequencing data. However, to fully realise the potential of this sequencing era, a concerted effort will need to be made by the genomics community to not only develop these, and other, novel techniques, but also to ensure their timely incorporation into standard analysis protocols. As we start to adapt to the new scale at which sequencing studies will now need to operate, we can expect to see a number of other advances in how we process and analyse genetic data over the next few years.

5.1.4 The future of locus discovery

Despite the technical advances that will be required to manage sequencing data at scale, it is not difficult to imagine a world where whole genome sequencing is routine. With a $100 genome tantalisingly close (already, Illumina have promised that this will be achievable 'soon', with the introduction of their new NovaSeq technology in January 2017), sequencing costs will soon be on par with other standard medical diagnostic tests. The prospect of patients routinely having their genomes sequenced is an exciting one, as we could see the rapid generation of datasets containing millions of individuals.

A particularly exciting aspect of routine whole genome sequencing within a clinical setting is the ability to tie genetic data to electronic health records across millions of individuals, providing a very rich and multi-faceted dataset for mining. A wide range of traits and phenotypes could theoretically be tested, simply using standard clinical notes and diagnostic tests that are performed on a regular basis. From the perspective of complex disease analysis, integration of medical records provides a

fantastic opportunity to investigate sub-phenotypes, including specific features such as disease location, complications, response to treatment, or disease progression.

Eventually, it is conceivable that we may one day have access to genetic data from nearly every individual with inflammatory bowel disease in the country. Such a dataset would make it possible to plausibly capture the complete contribution of genetics to disease risk. Not only would we be able to detect low frequency and common variants of very small effect, but we could thoroughly characterise structural and high-impact rare variation. With the likely availability of parental genomes, *de novo* mutations and highly-penetrant rare variation within families could also be uncovered. Finally, a dataset of this size would be very well powered to fine-map associations down to the precise causal variants, aiding in the translation of genetic associations to biological hypotheses.

5.2 Prospects for translation into the clinic

Through a combination of large-scale sequencing studies and the cost-effective genotyping and imputation of hundreds of thousands of samples, we are likely to see the rapid accumulation of loci associated with complex traits like IBD over the next ten years. Ultimately, it is hoped that we will be able to complete the picture of heritability for these traits, fully explaining the role of genetics in disease risk. However locus discovery in itself, whilst interesting from a scientific standpoint, is of little direct benefit to those individuals suffering from these disorders. It is therefore important that we also look to interrogate these associated loci for insights that can allow us to directly inform treatment, better understand the biology underlying disease pathogenesis, and aid in the development of novel therapeutics.

5.2.1 Integration with functional datasets

Just as the size of genetic datasets is expected to grow rapidly over the next decade or so, we can also expect to see similar growth in functional datasets that aim to determine the downstream impact of genetic changes. Large eQTL studies that

investigate the changes in gene expression associated with a given genetic variant can be used to predict the likely function of non-coding variation, while enhancer-gene interactions can be directly captured using conformation capture approaches like Hi-C. Over time, these studies will describe gene expression changes across an extensive range of specific cell types and environmental conditions. Further datasets that describe methylation profiles, chromatin modifications, transcription factor binding, and other epigenetic markers are also likely to grow in size and coverage. For some of the more informative functional assays, it is conceivable that they may also be incorporated into a clinical setting, increasing both the availability of data and also allowing for this information to be evaluated in a disease-specific setting. By integrating this functional data with genetic associations, we may eventually be able to resolve the biological mechanisms underlying the majority of disease-associated variants.

5.2.2 Informing treatment

As described in section 4.3.4, a number of IBD susceptibility genes have been shown to have important applications in the development of new treatments. A notable case is the associated locus near *SMAD7*, which has been shown to reduce the activity of TGF-β1 (an immunosuppressive cytokine) when present at high levels. In a recent phase 2 trial of an oral *SMAD7* antisense oligonucleotide, mongersen, Crohn's disease patients receiving the drug had significantly higher remission rates than those given a placebo (Monteleone et al., 2015). Similarly, the drug efalizumab targets the product of *ITGAL*, an integrin αL subunit of lymphocyte function-associated antigen 1 (LFA-1), and has been used to treat psoriasis. A brief, open-label study of efalizumab for treating Crohn's disease showed evidence of a clinical response in the majority of subjects (James et al., 2011). Notably, the effect sizes of these clinically relevant genes are relatively small (Figure 5.4), highlighting the importance of continuing to catalogue IBD-associated loci to build up a complete picture of disease pathogenesis and susceptibility.

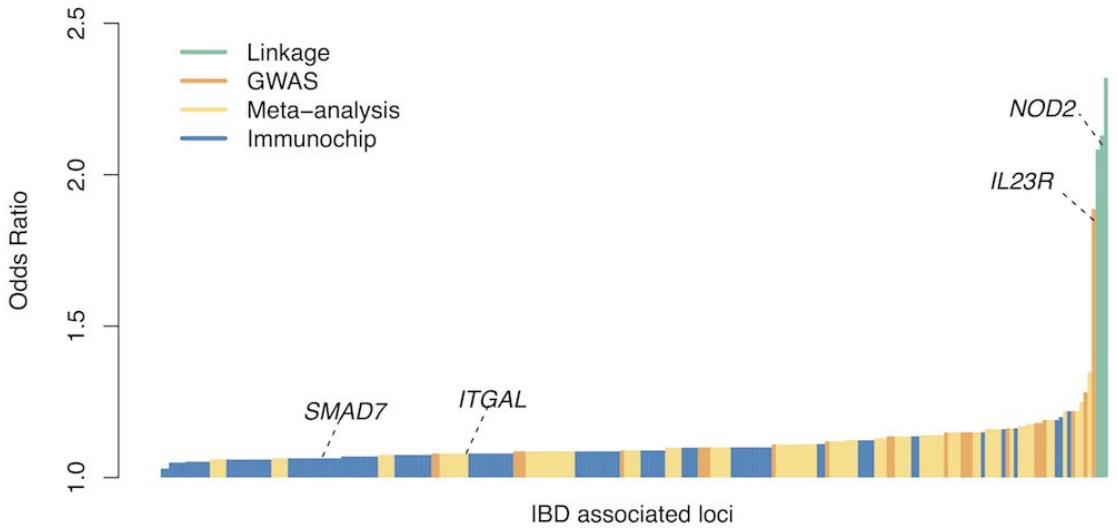

Figure 5.4: Effect sizes of IBD-associated loci identified using various study designs. The largest effect sizes are seen for the first two genes associated with IBD, *NOD2* and *IL23R*. Nevertheless, the genes *SMAD7* and *ITGAL*, which have relatively small effect sizes, are both confirmed drug targets.

However, as well as uncovering potential targets for therapeutic development, identified genetic associations can also prove useful in determining clinical subphenotypes and predicting disease course. For example, in Crohn's disease, associations have been found between the HLA and colonic CD (Silverberg et al., 2003), while *NOD2* variants have been shown to predict ileal location and the need for CD-related surgery (Cleynen et al., 2013). Several other genetic variants have been found that, despite not contributing to disease risk, are associated with a more favourable prognosis in Crohn's disease (Lee et al., 2017). Similarly, for ulcerative colitis the HLA is associated with extensive disease and colectomy (Haritunians et al., 2010).

Such information can be used to construct individual genetic risk scores, which summarize predictions about disease risk and likely progression based on a patient's specific genetic profile. Techniques like this can then help to identify misdiagnosed patients and drive more personalized treatment approaches. For example, a recent study by Cleynen et al. (2016) used genetic risk scores to show how inflammatory bowel disease can be represented as a continuum of disorders based on disease

location, which may be better represented using three groups (ileal Crohn's disease, colonic Crohn's disease, and ulcerative colitis), as opposed to the two-scale CD and UC definitions used now (Figure 5.5). They also note that disease location, which is in part genetically determined, is not only an intrinsic component of an individual's disease, but also represents a major driver of changes in disease behaviour over time. Correct identification of the subtype of IBD affecting a patient can therefore be an important factor in determining the course of treatment.

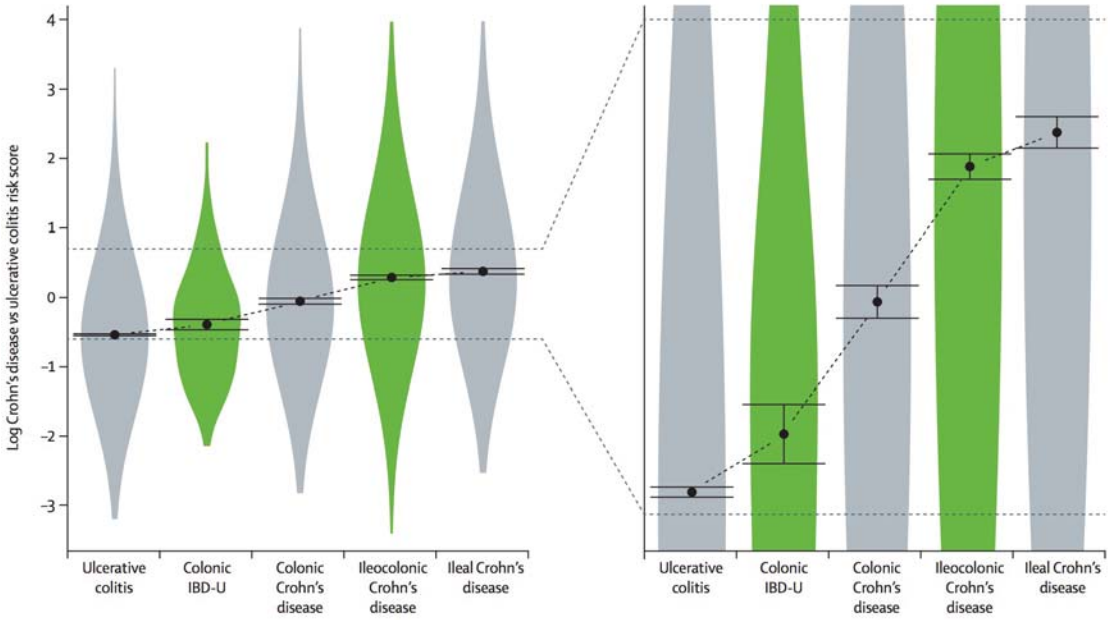

Figure 5.5: The genetic substructure of inflammatory bowel disease location, identified using genetic risk scores. A continuum of disorders based on disease location can clearly be seen, from those largely affecting the colon (UC, and colonic CD) to those largely affecting the ileum (ileal CD). Figure sourced from Cleynen et al. (2016).

More direct predictions about likely response to current IBD therapeutics may also be possible using genetic associations. For example, a common variant in the gene *NUDT15* is shown to be strongly associated with an elevated likelihood of developing life-threatening leukopenia (the loss of white blood cells) amongst Crohn's disease patients treated with thiopurine (Yang et al., 2014a). The UK IBD Genetics Consortium is currently undertaking a similar study to investigate genetic risk factors that may predict a patient's response to anti-TNF therapy. It is hoped

that one day these genetic insights will allow clinicians to immediately prescribe patients the most suitable therapeutic for their particular genetic profile, helping to minimise the development of adverse side-effects. Targeted therapy in this way can also be used to ensure those individuals predicted to have a mild disease course are not given stronger treatments than necessary, while patients with poor prognosis can be rapidly escalated to the most effective treatments.

5.2.3 Environmental factors: the microbiome

Another area that offers particular promise for the translation of genetic findings into clinical practice is investigation into the interaction between an individual's genome and their environment. In the case of IBD, loci identified to date have provided strong evidence of a role for the gut microbiota in disease pathogenesis, with the epithelial barrier and autophagy pathways repeatedly implicated (Khor et al., 2011). Microbiome studies in IBD have shown there are distinct differences in the composition of the gut flora in diseased and healthy individuals, such as a decrease in bacteroides, firmicutes, ruminococcaceae and bifidobacterium, and an increase in the presence of *Escherichia coli* and fusobacterium (Kostic et al., 2014). However, cause and effect are difficult to disentangle: did the disturbed microbiome arise as a result of the extensive inflammatory response, or did it trigger it? The effects of therapeutics on the intestinal environment further complicate such questions, as treatments such as antibiotics are known to affect the gut microbial community (Dethlefsen et al., 2008; Antonopoulos et al., 2009). Finally, even amongst healthy individuals the precise composition of the microbiome is extremely sensitive to diet and other unknown environmental factors: family observations show that sharing both genetics and a living space is no guarantee of a completely shared microbiome, and even within the same individual temporal variations are observed (Schloss et al., 2014).

The importance of understanding the role of the microbiome is reflected in the recent success of fecal microbiota transplants (FMTs) as a treatment for inflammatory bowel disease. FMTs aim to reduce dysbiosis in the bowel by modifying the microbiome using stool from a healthy donor. Although the idea was first intro-

duced over five decades ago by Eiseman et al. (1958) to treat pseudomembranous enterocolitis, it has only recently gained popular attention from the IBD community. An initial study by Suskind et al. (2015) showed temporary remission in seven of nine patients, and more extended remission in five of those cases. Efficacy of the FMT depended on whether it successfully engrafted or not, and on how similar the recipient's original microbiome was to the donor one. Despite this early success, further clinical studies are required to properly evaluate the safety and efficacy of this method.

Genetics provides a valuable opportunity to unravel the role of the microbiome in inflammatory bowel disease. In particular, genetic variation provides a useful starting point when trying to determine the casual relationships between environmental factors like the gut microbiota and the development of disease phenotypes like IBD. This is because germline genetic variation is unaffected by environmental factors, meaning it can act as a causal 'anchor' when considering relationships. Essentially, an individual's genotype can affect their phenotype, and their environment and phenotype can both influence each other, but environment and phenotype will not affect genotype (except when considering somatic mutations). This observation led to the development of Mendelian randomization techniques (Figure 5.6), which can test for causal effects between correlated traits (such as IBD and the gut microbiota) even in the presence of confounders (Davey Smith and Hemani, 2014).

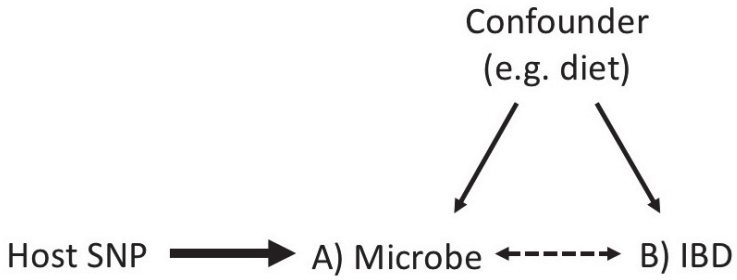

Figure 5.6: Mendelian randomization can be used to infer a causal relationship between two correlated traits, A and B (in this case the microbiome and IBD). If this correlation has arisen because A causes B, then it follows that any variable that affects trait A should also affect trait B (but not vice versa). If we can determine genetic variants that are associated in a known direction with A (e.g. genetic variants that are associated with changes in the microbiome in healthy individuals) we can then test for a causal relationship with B.

Using genetics to understand how changes in the gut microbiota can influence the host response holds promise for identifying the role of the microbiome in IBD, and may even allow us to uncover some of the reasons why some genetically-susceptible individuals develop disease, while others do not. Ultimately, it may lead to better understanding of why therapies such as fecal microbiota transplants appear to offer some relief in IBD, and contribute to the development of new, more targeted treatments.

5.3 Concluding remarks

It is an exciting time for the field of complex disease genetics. Over the past twenty years there have been dramatic advances in our understanding of the genetic causes underlying complex disorders, with common variation across hundreds of loci associated with disease risk. Now, a series of impressive technological developments have given us the ability to collect DNA sequences on an unprecedented scale, opening the door to expand this locus discovery effort into rare and low frequency variation. As sample sizes continue to grow, it is becoming a very real possibility that we will be able to resolve the complete picture of heritability in complex traits, fully capturing the contribution of genetics to disease risk. Through this steady accumulation of genetic clues, we are now starting to uncover the biological mechanisms that underlie disease pathogenesis, offering insights that can be used to directly impact treatment and inform the development of new therapeutics. Overall, these advances in the field of genetics hold promise for understanding the causes of complex disorders such as inflammatory bowel disease, which can ultimately lead to tangible improvements in the lives of people suffering from these debilitating disorders.