Understanding Inflammatory Bowel Disease using High-Throughput Sequencing

Katrina Melanie de Lange
Trinity College
University of Cambridge

May 2017

Dissertation submitted for the degree of Doctor of Philosophy

Understanding Inflammatory Bowel Disease using High-Throughput Sequencing

Katrina Melanie de Lange, Trinity College, University of Cambridge

For over two decades, the study of genetics has been making significant progress towards understanding the causes of common disease. Across a wide range of complex disorders there have been hundreds of associated loci identified, largely driven by common genetic variation. Now, with the advent of next-generation sequencing technology, 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, potentially offering novel insights into the biological mechanisms underlying disease pathogenesis. In this thesis I will assess the potential of this technology to further our understanding of the genetics of complex disease, using inflammatory bowel disease (IBD) as an example.

After first reviewing the history of genetic studies into IBD, I will describe the analytical challenges that can occur when using sequencing to perform case-control association testing at scale, and the methods that can be used to overcome these. I then test for novel IBD associations in a low coverage whole genome sequencing dataset, and uncover 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 genotyped cohorts, I also describe the discovery of 26 novel IBD-associated loci, including a low frequency missense variant in *ADCY7* that approximately doubles the risk of ulcerative colitis. I resolve biological associations underlying several of these novel associations, including a number of signals associated with monocyte-specific changes in integrin gene expression following immune stimulation.

These results reveal important insights into the genetic architecture of inflammatory bowel disease, and suggest that a combination of continued array-based genome-wide association studies, 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.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the contributions section of each chapter and/or specified in the text. It is not being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words.

Katrina M. de Lange
May 2017

Acknowledgements

First and foremost, I would like to thank my supervisor, Jeffrey Barrett, for the endless help, advice and support over the past few years. Your commitment to statistical rigour, reproducible research, data sharing and clear scientific writing has been inspiring, and you have taught me more things than I can count. Perhaps most importantly, you have shown me that it is possible to be an amazing scientist and still have fun. I would also like to thank my secondary supervisor, Miles Parkes, and my thesis committee members, Richard Durbin, Gosia Trynka and Trevor Lawley, for their help and guidance along the way.

I would like to extend a particular thank you to Yang Luo, for taking me under her wing and sharing her extensive knowledge on all things sequencing. Your constant smile and boundless enthusiasm made working together both enjoyable and incredibly rewarding. To everyone else who contributed so much to the projects discussed in this thesis - especially Carl Anderson, Loukas Moutsianas, and Luke Jostins - thank you for all your hard work and dedication, and for the jokes and laughter that got us through the difficult bits. Finally, this work would not have been possible without the funding of the Wellcome Trust, and the countless individuals who donated samples for us to study; thank you for your generosity and your commitment to science.

I am forever grateful to those who got me here in the first place. To Tony Smith, whose infectious enthusiasm for bioinformatics was the trigger I needed to start down this path, and to all my supervisors at the University of Waikato who gave me the opportunity to find my fit in the world of research. A big thank you also goes to the Woolf Fisher Trust, without whom I never would have had the fantastic opportunity to pursue a PhD at the University of Cambridge. Thank you not only for giving me this chance, but for continuing to believe in me and the value of my work.

I would also like to thank those people who have made my time here in Cambridge so enjoyable. To the rest of the Barrett team: thank you for the laughs, lunches, and fascinating discussions. It has been a pleasure to work with you all! To the

other PhD students I have gotten to know so well at Sanger: the experience would not have been the same without you. To Ellese, Mariel, Nicola, Sumana, John and Alice: thank you for all the fun times, and all the support in the not-so-fun times. I wouldn't have made it through without you. Finally, to Julian. For everything.

Lastly, I want to thank my family, especially my siblings, Andrew and Robyn, and my parents, Vicki and Willem. Thank you for supporting me in everything I do, and for putting things in perspective. Because, sometimes, it really is more important that you go climb a mountain rather than write your thesis.

Publications

Arising from this dissertation

de Lange, K. M. & Barrett, J. C. (2015). Understanding inflammatory bowel disease via immunogenetics. *Journal of Autoimmunity.* 64, pp. 91−100

Luo, Y.*, **de Lange, K. M.***, Jostins, L., Moutsianas, L., Randall J. et al (2017). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at *ADCY7*. *Nature Genetics* 49, pp. 186−192

de Lange, K. M.*, Moutsianas, L.*, Lee, J. C.*, Lamb, C. A., Luo, Y. et al (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* 49, pp. 256−261

Contents

Chapter 1

Background and historical perspective

The study of genetics offers a unique opportunity to uncover the causes underlying a wide range of human disease. By pinning down the genetic variations that lead to an elevated risk of developing a given disorder, we can start to elucidate some of the biological mechanisms that are contributing to disease pathogenesis. Ultimately, it is hoped that an increased understanding of disease genetics will be able to directly impact patient quality of life, by contributing to improved diagnosis, the development of novel therapeutics, and the creation of highly personalised treatment regimes.

It is an area full of promise, and we are already starting to reap some of the benefits of early genetic studies. The causal genes underlying dozens of rare disorders have been discovered, and are already being used in clinical settings for the rapid diagnosis of patients, or to aid in the development of new therapeutics. Particularly famous cases, like the identification of variants in the *BRCA* genes that can strongly predispose an individual to breast cancer (Ford et al., 1998), or the discovery of PCSK9 as a effective drug target for the treatment of cardiovascular disease (Hall, 2013), have further fuelled the excitement around using genetics to aid in disease management.

However, extending these successes to common disorders has proven to be challenging. In this chapter, I shall explain the history of genetic studies into common disease, describing both the novel findings and the unique problems that have arisen during this process. Throughout this discussion, and the remainder of this thesis, I shall be using inflammatory bowel disease (IBD) as an exemplar common disorder. Thus far, IBD has proven to be one of the most successful stories in complex disease genetics, and it therefore provides a strong setting in which to examine the successes and limitations of existing genetic studies. Looking forward, our relatively good understanding of the genetics underlying inflammatory bowel disease, compared to most other complex traits, also makes it an ideal disease with which to explore the utility of novel technologies and methods.

1.1 Inflammatory bowel disease

1.1.1 Clinical presentation

Crohn's disease (CD) and ulcerative colitis (UC), the two major subtypes of inflammatory bowel disease, are both chronic, debilitating disorders of the gastrointestinal tract (Figure 1.1). Affected individuals experience a range of symptoms associated with inflammation of the gut, including severe abdominal pain, fever, vomiting, diarrhoea, rectal bleeding, anaemia and weight loss. There is currently no cure, although symptoms can often be managed using steroids or immunosuppressants to reduce inflammation. However, many patients experience side-effects from these potent immunomodulators, and some will eventually lose response to treatment or develop complications. A subset of individuals never respond to these treatments at all. Ultimately, many patients will require major surgery to remove severely damaged portions of the bowel.

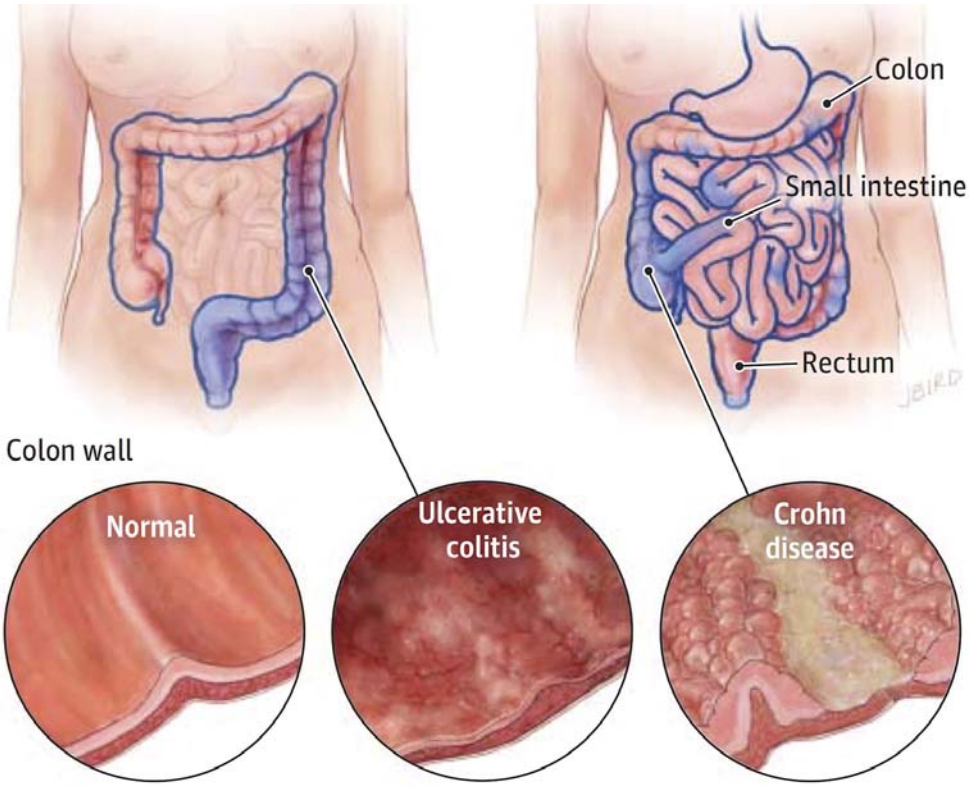

Figure 1.1: Disease localisation and appearance of Crohn's disease and ulcerative colitis, the two major forms of inflammatory bowel disease. Image sourced from Jin (2014)

Although Crohn's disease and ulcerative colitis share a number of clinical features, there are important distinctions in incidence patterns, disease localization, histopathology and endoscopic features (Table 1.1) that suggest there are differences in the underlying pathways driving each disease (Baumgart and Sandborn, 2007; Bernstein et al., 2010).

Table 1.1: Distinguishing features of the two major inflammatory bowel disease subtypes, Crohn's disease and ulcerative colitis. Adapted from Baumgart and Sandborn (2007), and Bernstein et al. (2010).

	Crohn's disease	Ulcerative colitis
Incidence patterns		
Age of onset	Incidence rates peak in the third decade of life	Stable incidence rates are seen between the third and seventh decades of life
Prevalence rates	CD is more prevalent that UC in developed countries	UC emerged before CD in developed countries, and is more prevalent in still-developing countries
Disease localisation		
Affected areas	Entire gastrointestinal tract (from mouth to anus)	Colon, plus some potential backwash ileitis
Inflammation pattern	May occur as patchy, discontinuous inflammation	Continuous inflammation in the affected area
Histopathology		
Penetrance	Transmural inflammation of the entire bowel wall	Inflammation restricted to the mucosal and submucosal layers
Appearance	Thickened colon wall with granulomas, deep fissures and a cobblestone appearance	Distorted crypt architecture, with shallow erosions and ulcers
Serological markers		
	Anti-Saccaromyces cerevisiae antibodies	Anti-neutrophil cytoplasmic antibodies
Complications		
	Fistulas, abdominal mass (lower right quadrant), colonic and small-bowel obstructions, stomatitis	Haematochezia (rectal bleeding associated with the passing of stool), passage of mucus or pus

1.1.2 Epidemiology

The prevalence of inflammatory bowel disease is currently highest in Europe (UC, 505 per 100,000 persons; CD, 322 per 100,000 persons) and North America (UC, 249 per 100,000 persons; CD, 319 per 100,000 persons), according to a systematic review by Molodecky et al. (2012). The disorder is more common in Ashkenazi Jews, who are five to eight times more likely to develop IBD compared to non-Jewish populations (Sands and Grabert, 2009). More broadly, global prevalence is rising, with rapid increases in incidence rates occurring as more countries adopt a Westernised lifestyle (Loftus, 2004). Incidence rates are also rising in younger people, which is placing an increased strain on healthcare resources, particularly as early-onset IBD has been associated with a higher risk of developing colorectal cancer (M'Koma, 2013). Overall, IBD represents a significant global health burden that is of growing concern (Figure 1.2).

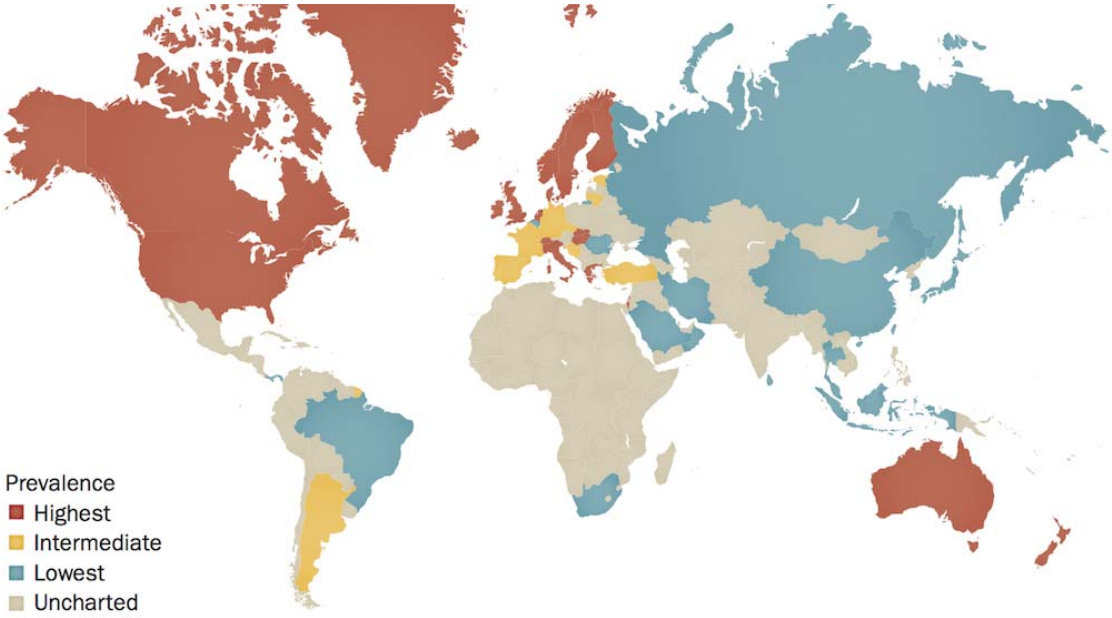

Figure 1.2: Global prevalence of inflammatory bowel disease in 2015. Image sourced from Kaplan (2015).

1.2 The early days of IBD genetics

1.2.1 Twin studies

Inflammatory bowel disease is characterised by a dysregulated immune response to unknown environmental triggers in a genetically susceptible individual, and a heritable component to the disease is well recognised. Early epidemiological observations showed clear familial clustering, which was reflected in high sibling risk ratios. Specifically, it was estimated that the siblings of an individual with ulcerative colitis are 7-17 times more likely to develop the disease themselves, compared to the general population; for Crohn's disease the elevated risk is as high as 15-42 times (Halme et al., 2006). Twin studies have since conclusively shown these observations to be at least partly attributable to genetics, rather than shared environmental factors, by comparing disease concordance rates between pairs of monozygotic (MZ) and dizygotic (DZ) twins. This assumes that both individuals in a twin pair have been exposed to roughly the same environment, and thus variation in concordance is due to genetics. It is worth noting that this assumption is not always strictly true: for example, in a prenatal setting monozygotic twins often share a placenta, while dizygotic twins do not (Marceau et al., 2016). However, using height as an example, a more recent estimation of heritability using an assumption-free model (based directly on the genetic data) has shown remarkable consistency with the original twin studies (Visscher et al., 2006). In a large meta-analysis of 6 IBD twin studies the resulting rates of 30.3% vs 3.6% for Crohn's disease (112MZ vs 196DZ), and 15.4% vs 3.9% for ulcerative colitis (143MZ vs 206DZ), support the importance of genetics in IBD risk (Brant, 2011).

Motivated by these findings, there have been a number of studies aimed at identifying the specific genomic loci that explain IBD heritability. Ideally, each of these associated loci would identify a single gene, or indeed a causative genetic variant, to help understand the biological processes involved in inflammatory bowel disease.

1.2.2 Linkage studies

Technological limitations around obtaining data on an individual's genotype at any given position has, however, been a major hurdle for these genetic studies. Although it has been possible to sequence fragments of DNA with relative ease since the advent of the dideoxy 'chain-termination' technique (widely known as Sanger sequencing) by Sanger et al. in 1977, this is a prohibitively expensive process. Initial studies therefore relied instead on restriction fragment length polymorphisms (RFLPs), which use restriction enzymes that can recognise and cut DNA at certain short sequences (Botstein et al., 1980). Where a genetic variant creates or disrupts this sequence, fragments of differing lengths will be created. If a DNA probe is then used to pull out a specific fragment, the various lengths seen amongst a group of individuals represent different alleles at that particular marker. A related method was later developed that instead tests the length of naturally varying microsatellite repeat regions, using polymerase chain reaction (PCR) primers that flank the microsatellite, followed by amplification and gel electrophoresis (Weber and May, 1989).

While these methods had the advantage of being relatively cheap, they were very low throughput. As a result, early studies into the genetics of IBD were by necessity coarse-grained, as data collection was limited to just a handful of genetic variants within a small number of individuals. To maximise the information that could be gleaned from this sort of dataset, most investigators restricted their analyses to family groups. This is because closely related individuals share longer stretches of DNA than unrelated individuals (as they are separated by fewer recombination events, where the chromosomes cross over during meiosis), and therefore fewer genetic markers are required to fully capture the pattern of DNA inheritance within a family. Maps with a density as low as one microsatellite marker every 1 or 2 centimorgans (cM) are sufficient to extract nearly 100% of the inheritance information available, and even very sparse maps of just 300-400 markers distributed roughly every 10cM across the genome can capture approximately 70% of the information content (Evans and Cardon, 2004). By using these markers to trace the DNA segments that segregate with disease status (such as variant alleles only

seen in affected individuals, and not in their unaffected relatives), sections of the genome that confer risk to the disease can be identified (Figure 1.3). This linkage analysis approach is good for detecting highly penetrant variants (i.e. those that are extremely likely to cause disease whenever present) that segregate well with disease status.

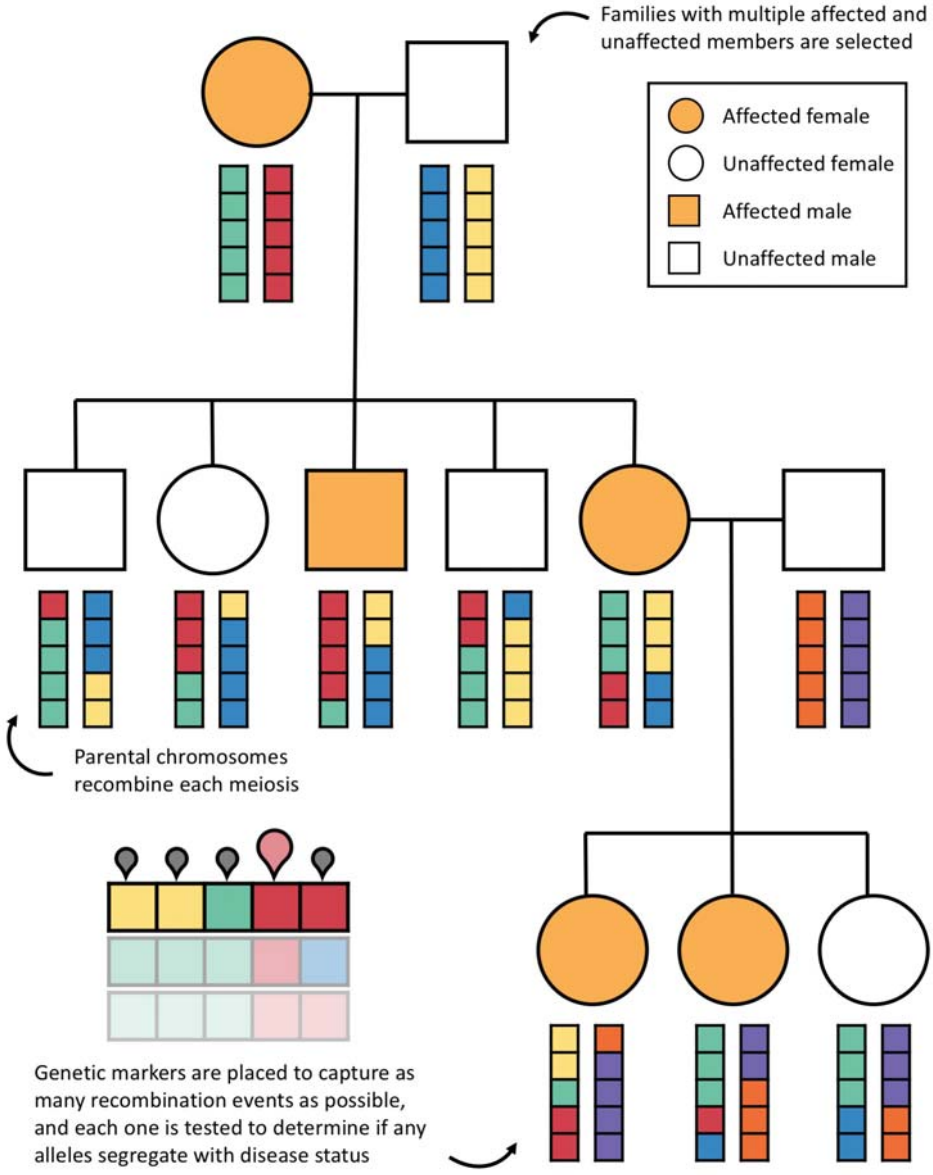

Figure 1.3: Overview of the linkage analysis study design for identifying disease-associated loci within a family containing multiple affected individuals.

Linkage studies successfully identified hundreds of highly penetrant variants for rare disorders (Gusella et al., 1983; Tsui et al., 1985; Seizinger et al., 1987; Vance et al., 1989; Siddique et al., 1991; Kandt et al., 1992; Speer et al., 1992), and were subsequently applied to a range of more common diseases. In 1996, the first such study in IBD linked a portion of chromosome 16 (dubbed IBD1) with Crohn's disease (Hugot et al., 1996), which was successfully replicated in a number of subsequent studies (Ohmen et al., 1996; Parkes et al., 1996; Curran et al., 1998; Brant et al., 1998; Cavanaugh et al., 1998; Cavanaugh and The International IBD Genetics Consortium, 2001). This finding was followed up using more closely packed markers within a small number of genes, and the IBD1 linkage on chromosome 16 was found to be caused by multiple disease risk alleles in the gene *NOD2*, whose role in the recognition of bacterial peptidoglycans and subsequent stimulation of an immune response (Figure 1.4) supports its association with the development of CD (Hugot et al., 2001; Ogura et al., 2001; Philpott et al., 2014). These variants are especially common in Ashkenazi Jews, partially explaining the increased burden of CD in that group.

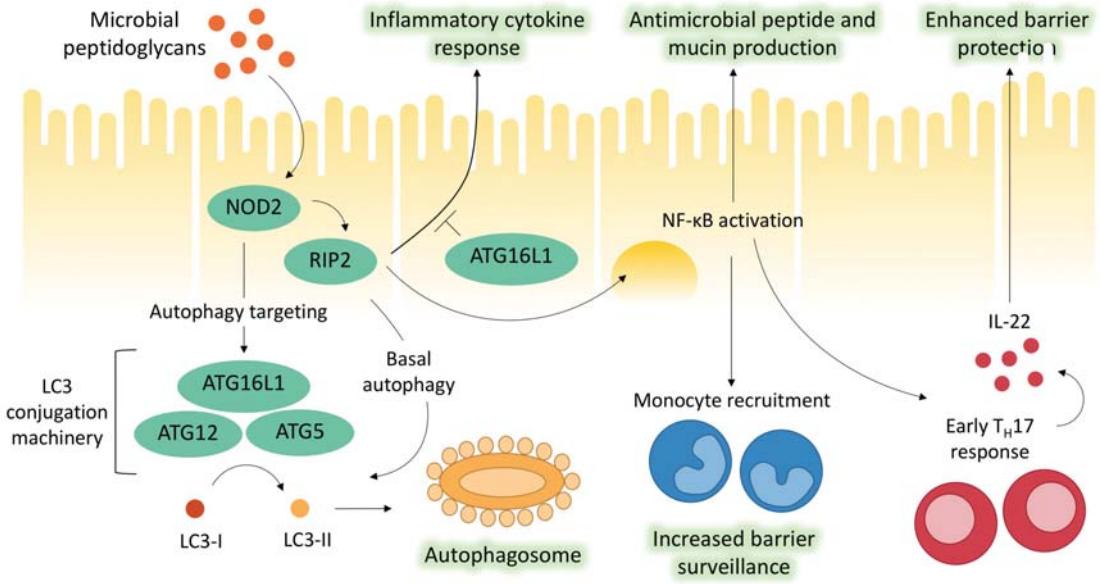

Figure 1.4: The signalling pathways through which *NOD2* responds to microbial peptidoglycan stimuli to promote innate mucosal defence and an autophagic response. Figure adapted from Philpott et al. (2014).

1.2.3 Limitations of linkage studies and the common disease, common variant hypothesis

Unfortunately, however, successes like *NOD2* were rare: it remained one of the few robustly replicated genetic risk loci discovered via linkage, not just in IBD, but across common diseases. This widespread disappointment reflected a fundamental property of the genetic architecture of common disease: they did not have a single, highly penetrant genetic cause. Instead, it was proposed by Risch and Merikangas (1996) that complex diseases were driven by the accumulation of many risk factors of only modest effect (the common disease, common variant hypothesis). Finding associations via linkage under this scenario is difficult, as the genetic risk may be spread throughout the genome rather than concentrated in a single locus. An alternative association analysis approach (which tests if the population-level allele frequencies of cases and controls are statistically different) is much more powerful. For example, Risch and Merikangas (1996) calculated that 17,997 affected sibling pairs would be needed to detect a risk allele with 50% frequency and an odds ratio of 1.5 using linkage, as opposed to just 484 using an association analysis. However, this approach requires the right variant to be chosen for testing among the millions known to exist in the human population.

One means of choosing variants to test was to select candidate genes based on prior biological hypotheses. Unfortunately, this produced a deluge of association claims with weak statistical evidence that did not replicate in subsequent studies (Ioannidis, 2003). Genetic studies had reached an impasse: although case-control association studies could theoretically detect signals too weak to show linkage, scanning the entire genome in an unbiased way in order to identify robust genetic associations was proving difficult.

1.3 The GWAS era

1.3.1 Technological developments that made GWAS possible

Three developments upended this stasis in gene discovery, and fundamentally changed gene mapping. First, by 2005, the public database of the most common type of genetic variant, single nucleotide polymorphisms (SNPs, where a single letter of DNA is variable), contained 9.2 million sites that had been catalogued by projects such as the SNP Consortium and the International HapMap Consortium (Sachidanandam et al., 2001; The International HapMap Consortium, 2005). Second, these catalogues of population-level genetic variation had also shown that variants common in the general population (minor allele frequency [MAF] $> 5\%$), and in physical proximity, were highly correlated, or in linkage disequilibrium (LD), with each other. Human population history had left a pattern of long LD blocks of high correlation, separated by small hotspots where most historical recombination events tended to cluster (McVean et al., 2004). This uneven LD pattern meant that it was possible to test the majority of common variants by carefully selecting markers in each long LD block. Approximately 500,000 well chosen SNPs could capture nearly 5 million common SNPs in Europeans and East Asians; unsurprisingly, the more genetically diverse African populations required almost twice as many markers to capture the same amount of variation (Barrett and Cardon, 2006). Finally, in the mid-2000s, it became economically feasible to genotype hundreds of thousands of variants using new microarray technologies. These key advances opened the way for genome-wide association studies (GWAS) that could be used to detect the diverse genomic loci associated with a given complex trait. GWAS combined the hypothesis-free ability to scan the whole genome of linkage with the statistical power to detect associations of smaller effect size.

1.3.2 GWAS: a revolution in IBD genetics

Crohn's disease was among the first diseases studied using GWAS, beginning in 2006. In addition to confirming the established *NOD2* association, these early studies identified four new loci at genome-wide levels of statistical significance ($P < 5 \times 10^{-8}$), demonstrating the power of the GWAS approach (Duerr et al., 2006; Hampe et al., 2007; Libioulle et al., 2007; Rioux et al., 2007). The strongest new association was a protective low frequency allele in *IL23R* (Duerr et al., 2006), which encodes a receptor protein that is embedded in the cell membrane of many different types of immune cells and, upon binding of IL23, starts a signaling cascade that promotes inflammation and coordinates an adaptive immune response (Figure 1.5). A more surprising discovery was an association to a protein-coding variant in *ATG16L1* (Hampe et al., 2007), which encodes a protein involved in the autophagosome pathway (Figure 1.4), and provided the first strong evidence for the importance of autophagy in CD. This pathway is responsible for processing intracellular bacteria, and so the *ATG16L1* association contributed to further understanding of the dysfunction of the intestinal barrier in Crohn's disease. Finally, these early studies discovered a pair of associations on chromosomes 5p13 and 10q21 that were far from any genes (Libioulle et al., 2007; Rioux et al., 2007).

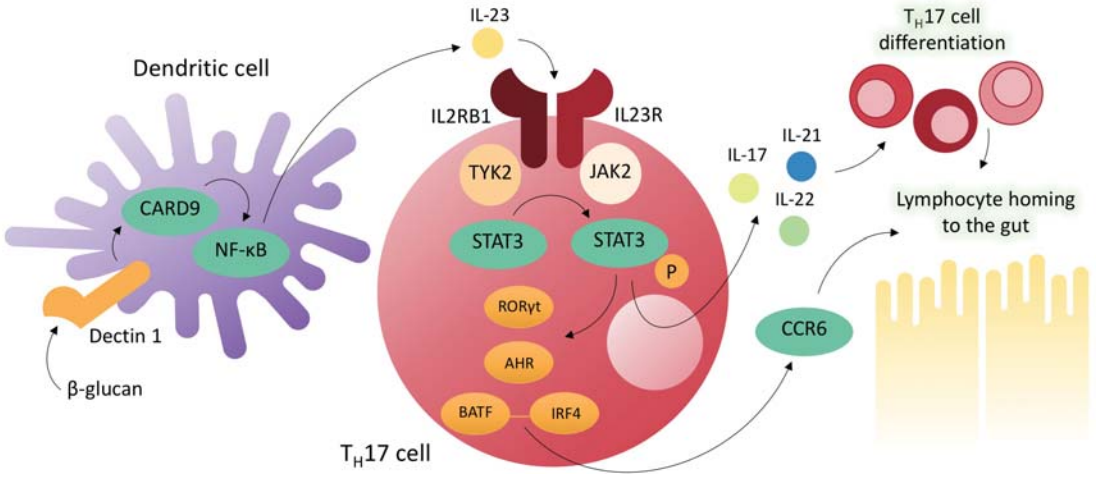

Figure 1.5: The *IL23R* signalling pathway used to activate the adaptive immune response, and the downstream T_H17 cell differentiation program (Weaver and Hatton, 2009; Zhernakova et al., 2009; Khor et al., 2011; Parkes et al., 2013).

Unlike the previous associations, these new results highlighted the important role of regulatory and non-coding elements in complex disease. Motivated by these early successes, further GWAS used increasingly larger sample sizes to implicate both the innate (*NKX2-3*, *CARD9*) and adaptive (*TNFSF15*, *PTPN2*, *IL-12B*) immune response pathways in inflammatory bowel disease, and recapitulate the role of autophagy and intracellular bacteria management (*NOD2*, *ATG16L1*, *IRGM*) in Crohn's disease (Parkes et al., 2007; Van Limbergen et al., 2009). These initial CD studies also suggested a partial overlap of genetic risk for ulcerative colitis: of the Crohn's disease associations discovered, about 30% were also found to be associated with UC via replication studies (Liu and Anderson, 2014). Additional GWAS in ulcerative colitis cohorts lead to the discovery of multiple novel UC-specific loci (Fisher et al., 2008; Franke et al., 2008; Silverberg et al., 2009; Barrett et al., 2009).

These UC-specific studies also confirmed the long-established association between UC and the classical human leukocyte antigen (HLA) locus (Satsangi et al., 1996), which contains genes encoding antigen-presenting proteins on the surface of the cell, and plays a crucial role in the regulation of the adaptive immune system. Despite the HLA being strongly associated with many other chronic inflammatory and autoimmune disorders, the association with CD is much weaker (Zhernakova et al., 2009). Overall, the pattern of association to IBD in the HLA region is the most complicated in the genome. While the most recent study of HLA in IBD conclusively showed that the HLA-DRB1*01:03 allele is the most strongly associated in both CD and UC, it also identified more than ten additional risk alleles associated with one or both diseases (Goyette et al., 2015). Most of these associations are disease-specific; HLA class I and class II variation contributes equally to CD, while class II variation is more important in ulcerative colitis. In addition, evidence of decreased heterozygosity in HLA genes was observed for ulcerative colitis only. This non-additive effect, similar to that observed by Nejentsev et al. (2007) in HLA alleles associated with Type 1 diabetes, highlights the importance of being able to detect a wide range of antigens for protective immunity.

1.3.3 Meta-analyses and the importance of sample size

While this flurry of discoveries generated new biological hypotheses for IBD, it became clear that these relatively weak associations cumulatively explained only a fraction of the heritability expected from twin studies. This missing heritability problem was universal amongst complex diseases during the early GWAS era, and was partially attributed to types of variation not captured by GWAS, such as non-European, rare and structural variants (Maher, 2008; Manolio et al., 2009). However, as Figure 1.6 shows, these early studies were in fact poorly powered, because the true genetic architecture of IBD includes many variants with odds ratios < 1.2 or even 1.1.

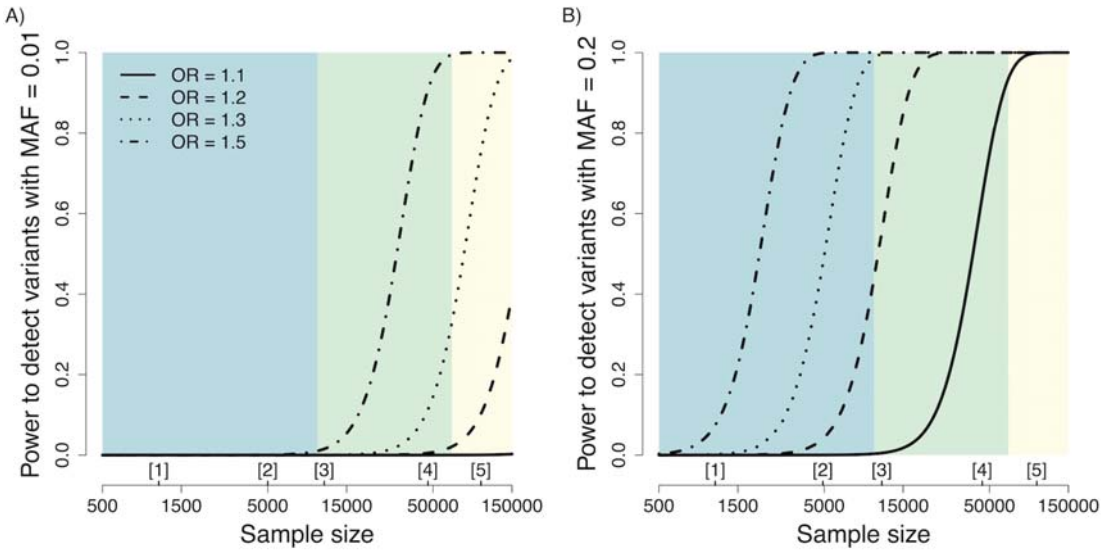

Figure 1.6: Power to detect associations of different effect size (odds ratio, OR) are compared for rare variants (MAF = 0.01, panel A) and common variants (MAF = 0.2, panel B). Effective sample sizes of several key studies are indicated along the x-axis, to reflect the power of the GWAS studies (blue), meta-analyses (green) and Immunochip-based studies (yellow). [1] Duerr et al. (2006); [2] The Wellcome Trust Case Control Consortium (2007); [3] Barrett et al. (2008); [4] Anderson et al. (2011); [5] Liu et al. (2015).

To increase power to search for these small effects, the International IBD Genetics Consortium (IIBDGC) was formed to pool thousands of already genotyped samples from previous GWAS. The merging of data from different genotyping chips was enabled by imputation, which infers missing data by comparing known genotypes

to those in a representative reference set with more complete data, such as the HapMap or 1000 Genomes resources (The International HapMap Consortium, 2005; Abecasis et al., 2010). Other between-study variation, such as population differences, could be accounted for by using a meta-analysis approach, which jointly analyses the summary statistics from each study, as opposed to the raw data.

The first of these IIBDGC meta-analyses effectively tripled the number of known Crohn's disease susceptibility loci with the identification of 21 novel associations, including *LRRK2*, another autophagy gene (Barrett et al., 2008). This was followed by a meta-analysis of ulcerative colitis studies, which identified 29 new UC risk loci (Anderson et al., 2011), and a second Crohn's disease meta-analysis that brought the total number of CD susceptibility loci to 71 (Franke et al., 2010). This rapid accumulation of IBD risk loci culminated in 2012 with a meta-analysis containing over 75,000 cases (including both CD and UC for the first time) and controls, that brought the total number of IBD loci to 163 (Jostins et al., 2012). Numerous pathways were implicated through multiple genetic associations, including those involved in innate mucosal defence, JAK/STAT signaling, cytokine production (particularly interferon-γ, interleukin (IL)-12, tumour-necrosis-factor-α and IL10 signalling) and lymphocyte activation.

This dramatic growth in the number of IBD-associated loci, together with the first large-scale joint analyses of CD and UC, revealed that the genetic risk for Crohn's disease and ulcerative colitis substantially overlap. Although early GWAS data had suggested quite disparate underlying pathways, of the 163 loci identified in the Jostins et al. (2012) paper, 110 were associated with both phenotypes (Figure 1.7). Furthermore, of the 30 CD-specific and 23 UC-specific loci, 43 show the same direction of effect in the non-associated disease, suggesting that only a tiny minority truly have zero effect in the other disease. This considerable overlapping genetic risk implies that the two diseases are likely to share many biological mechanisms. However, the few loci that are CD- or UC-specific, as well as the relative size of effects at shared loci, might reveal clues about the distinct pathologies of the two diseases.

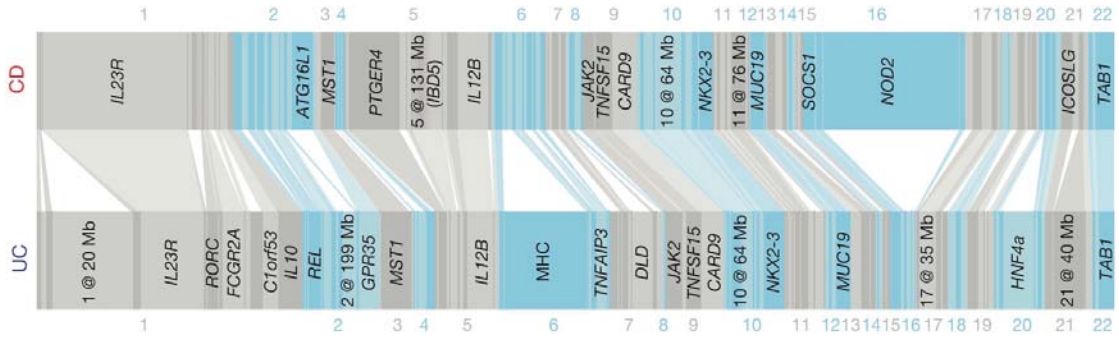

Figure 1.7: Belgravia plot of the 163 loci identified by Jostins et al. (2012), showing the shared genetic overlap between Crohn's disease and ulcerative colitis. The width of each bar is proportional to the variance explained by a given locus for the disease indicated, and bars are linked if they are associated with both CD and UC. Note the extensive genetic overlap between the two diseases, even though many of the loci with the largest effect sizes are disease-specific. Figure sourced from Jostins et al. (2012).

1.3.4 IBD genetics in the context of other diseases

Understanding both the shared and private genetics of related disorders can be useful for constructing hypotheses about the underlying biological pathways that may be driving each disease, and how distinct clinical phenotypes may arise. For example, known IBD loci are enriched for genes involved in primary immunodeficiencies, including those linked to reduced levels of circulating T cells (*ADA*, *CD40*, *TAP1*, *TAP2*, *NBN*, *BLM*, *DNMT3B*), and to T-helper cells responsible for producing T_H17, memory, and regulatory T cells (*STAT3*, *SP110*, *STAT5B*). It is interesting to note that the same genes can be affected both by the damaging protein coding variants that cause these severe disorders, and by much more subtle (presumed regulatory) variants that slightly affect risk of complex diseases like IBD.

Several studies have extended these cross-disease genetic comparisons to potentially related complex diseases, such as the common immune-mediated disorders (ankylosing spondylitis, coeliac disease, multiple sclerosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, and type I diabetes). Early analysis of GWAS results from across these diseases (together with IBD) suggested that the innate immune response, as well as the general immune pathways involved in T-cell

differentiation and signalling, are shared between many of them (Zhernakova et al., 2009), as summarised in Table 1.2.

This observed overlap of risk loci among common immune mediated diseases motivated the design of a new genotype array, called Immunochip, which contained markers densely covering loci with known associations to at least one of 11 immune-mediated diseases, or with suggestive significance in the early immune-related GWAS studies. This targeted array, which cost approximately 20% of the price of contemporary GWAS chips, made the genotyping of large samples of immune-mediated disorders possible, and also paved the way for more extensive disease subphenotype and cross-disease studies (Parkes et al., 2013). Indeed, the Immunochip formed the basis of the Jostins et al. (2012) IBD meta-analysis, which showed that 70% (113 out of 163) of the IBD loci identified are also shared with other complex diseases or traits, including 66 loci shared with other immune-mediated disorders. Sharing is particularly strong between IBD and the other seronegative diseases, ankylosing spondylitis and psoriasis. Interestingly, across the immune-mediated diseases those loci that are not shared tend to have large effect sizes, which would explain why the genetic underpinnings of CD and UC appeared so misleadingly disparate prior to the large meta-analysis efforts (Parkes et al., 2013). Extending this analysis to more distantly related diseases, Jostins et al. (2012) observed an enrichment in genes previously linked with Mendelian susceptibility to mycobacterial disease (MSMD) and leprosy (a complex mycobacterial disease): these overlaps suggest that the genetic architecture of IBD may have been shaped by selection pressures arising from mycobacterial infection.

A recent study has also exploited the overlap between IBD and other immune-mediated diseases to increase the power to detect associated loci. By jointly analysing Immunochip data from across five related disorders (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis and ulcerative colitis), Ellinghaus et al. (2016) successfully identified an additional six Crohn's disease loci.

Table 1.2: Example pathways implicated in inflammatory bowel disease pathogenesis. Genes belonging to these pathways and falling within IBD-associated loci are indicated, and cases where these overlap with other immune-mediated disorders are marked. Note however that in some cases the specific genes have not yet been identified as causal and, as many loci contain multiple candidate genes, these should not be treated as confirmed. Data sourced from Han et al. (2009), Khor et al. (2011), Jostins et al. (2012), Brown et al. (2013), Parkes et al. (2013), and Liu et al. (2015).

Pathway implicated	Pathway genes in IBD-associated loci	Ankolysing Spondylitis	Psoriasis	Coeliac	Rheumatoid Arthritis	Type 1 Diabetes	Systemic Lupus Erythematosus	Multiple Sclerosis
Innate immune response								
Epithelial barrier function and repair	*CDH1, ERRFI1, GNA12, HNF4A, ITLN1, MUC19, NKX2-3, PLA2G2E, PTGER4, REL, STAT3*	*NKX2-3, PTGER4*	*REL*	*REL*	*REL*	-	-	*PTGER4, STAT3*
Innate mucosal defense	*CARD9, FCGR2A, IL18RAP, ITLN1, NOD2, REL, SLC11A1*	*CARD9, FCGR2A*	*REL*	*IL18RAP, REL*	*FCGR2A, REL*	*FCGR2A, IL18RAP*	*FCGR2A*	*FCGR2A*
Autophagy	*ATG16L1, CUL2, DAP, IRGM, LRRK2, NOD2, PARK7*	-	-	-	-	-	-	-
Apoptosis/ necroptosis	*DAP, FASLG, MST1, PUS10, THADA*	-	-	*PUS10*	-	-	-	-
Activation of adaptive immune response								
IL23-R response pathway	*CCR6, IL12B, IL12RB2, IL21, IL23R, IL27, JAK2, STAT3, STAT4, TYK2*	*IL12B, IL12RB2, IL23R, IL27, JAK2, TYK2*	*IL12B, IL23R, STAT3, TYK2*	*IL21, STAT4*	*CCR6, IL21, STAT4, TYK2*	*IL27, TYK2*	*IL27, STAT4, TYK2*	*IL12B, STAT3, STAT4, TYK2*
NF-kB	*NFKB1, REL, TNFAIP3, TNIP1*	-	*NFKB1, REL, TNFAIP3, TNIP1*	*REL, TNFAIP3*	*REL, TNFAIP3*	*TNFAIP3*	*TNFAIP3, TNIP1*	*NFKB1*
Aminopeptidases	*ERAP1, ERAP2*	*ERAP1, ERAP2*	*ERAP1, ERAP2*	-	-	-	-	-
IL2 and IL-21 T-cell activation	*IL2, IL21, IL2RA*	-	-	*IL2, IL21*	*IL2, IL21, IL2RA*	*IL2, IL21, IL2RA*	-	*IL2RA*
Regulation of adaptive immune response								
T_H17 cell differentiation	*AHR, CCR6, IL2, IL22, IL23R, IRF4, JAK2, RORC, STAT3, TNFSF15, TYK2 IL23R, JAK2, TYK2*	*IL23R, TYK2*	-	*CCR6, TYK2*	*TYK2*	*TYK2*	*TYK2*	
T-cell regulation	*ICOSLG, IFNG, IL12B, IL2, IL21, IL23R, IL2RA, IL7R, NDFIP1, PIM3, PRDM1, TAGAP, TNFRSF9, TNFSF8*	*ICOSLG, IL12B, IL23R*	*IL12B, IL23R, TAGAP*	*ICOSLG, TAGAP*	*ICOSLG, PRDM1, TAGAP, TNFRSF9*	*TAGAP*	*PRDM1*	*IL12B, IL2RA, IL7R, TAGAP*
B-cell regulation	*BACH2, IKZF1, IL5, IL7R, IRF5*	*BACH2*	-	*BACH2*	*IRF5*	*BACH2, IKZF1*	*IKZF1, IRF5*	*BACH2, IKZF1*

The large cross-phenotype dataset described by Ellinghaus et al. (2016), containing in excess of 86,000 individuals, also offered a unique opportunity to explore the genetic basis underlying the co-morbidity of many of these diseases. The authors note that, although the overall co-morbidities of the five diseases are best explained by pleiotropy (whereby two diseases share a number of risk alleles), there is evidence that the particularly strong co-morbidity between primary sclerosing cholangitis (PSC) and ulcerative colitis may in fact be indicative of a subset of patients with a unique PSC-IBD disease. This conclusion is supported by observed clinical differences between PSC-IBD and classical inflammatory bowel disease, including an increased risk of pancolitis and colorectal cancer (de Vries et al., 2015).

1.3.5 Expanding into non-European populations

Up until this point, GWAS in IBD had largely focused on samples of European ancestry. One notable exception was a Crohn's disease study in 2005 (Yamazaki et al., 2005), performed in a Japanese population after it was noted that *NOD2* did not appear to play a significant role in the pathogenesis of CD in Japan (Yamazaki et al., 2002; Negoro et al., 2003; Yamazaki et al., 2004). This study identified a strong association between the gene *TNFSF15* and CD, despite an initial sample size of fewer than 100 patients. Additional genome wide association studies of IBD within Indian, Japanese and Korean populations showed that most IBD genetic risk is shared regardless of ancestry (Asano et al., 2009; Juyal et al., 2015; Yamazaki et al., 2013; Yang et al., 2013; Yang et al., 2014b). However many of these studies were small, preventing informative comparisons across populations.

A large IBD study of multiple ancestries was conducted by the IIBDGC both to study IBD associations apparently unique to one population, and to boost power for detection in all populations using meta-analysis techniques that account for population stratification. GWAS and Immunochip data were analysed from 96,486 individuals of European, East Asian, Indian and Iranian descent, yielding a total of 200 IBD associated regions (Liu et al., 2015). For the vast majority of these loci, the direction and magnitude of the effect is consistent between the European and non-European cohorts, implying that the underlying casual variants

at these shared loci are likely to be common, as rare alleles are more likely to be population-specific. For the handful of associations that appear to be heterogeneous between populations, nearly all are due to differences in allele frequency between populations. For example, *NOD2* is not biologically less relevant in Japan, but rather the IBD risk variants are simply absent in that population. Only *TNFSF15*, which exhibits microbial-induced expression (Shih et al., 2009), and the autophagy gene *ATG16L1* are common in all populations but appear to have different effect sizes, possibly reflecting differences in gene-environment interactions between the populations.

1.4 Beyond GWAS

Combined, the meta-analyses and trans-ancestry study contributed to an almost 20-fold increase in the number of known IBD-associated loci (Figure 1.8). However, as with many complex diseases, this approach of analysing ever-larger genotype array-based datasets still captures only the fraction of IBD heritability explained by common variants, mostly in European populations. In fact, the latest estimates by Chen et al. (2014) suggest that common variants explain only 26% of the heritability of Crohn's disease, and 19% of the heritability of ulcerative colitis. Some of this missing heritability may be found in regions sometimes overlooked by GWAS, such as the sex chromosomes. A recent study by Chang et al. (2014) utilized X-chromosome data from existing datasets to identify a new IBD-associated gene, *ARHGEF6*, which interacts with a major surface protein on *H. pylori* (a gastric bacterium). Rare loss-of-function variants in the X-chromosome gene *XIAP*, which encodes a protein that inhibits apoptosis, have also been identified as strongly predisposing for early-onset Crohn's disease in males (Uhlig, 2013; Zeissig et al., 2015). However, uncovering rare variants associated with complex disease will require the development of new study designs, as rare variants generally have low correlation to the marker SNPs used (which usually have much higher allele frequencies, MAF>0.05, to better capture other common variation) and are therefore not well tagged (Li et al., 2013a).

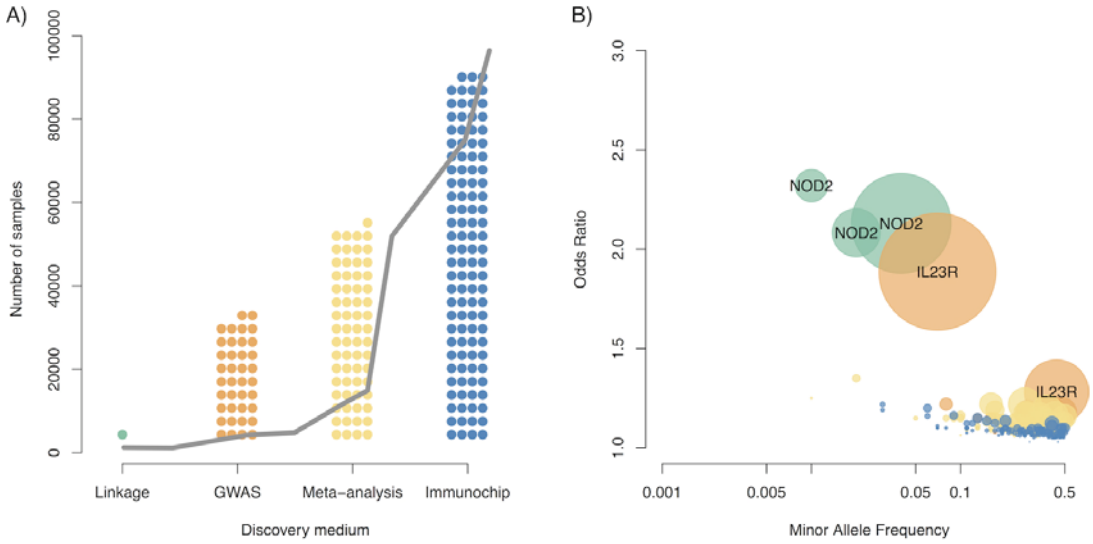

Figure 1.8: Panel A) describes the number of loci identified with different study designs (linkage=green; GWAS=orange; meta-analysis=yellow; Immunochip=blue), with each dot representing a novel locus. The number of samples required to identify these loci are shown in grey. Panel B) plots the odds ratio versus MAF for each IBD-associated variant, with the size of each circle representing the variance explained by that variant. The colours from panel A) are used to indicate the areas of the search space captured by each study design.

1.4.1 Rare and low frequency variation

To successfully identify a rare or low frequency disease-associated allele it is necessary to directly test the variant site itself, as such variants are not in high LD with many others, preventing the capture of their signal by a proxy variant (the method which drove the success of GWAS). Furthermore, because such alleles are by definition observed infrequently in the population, even the largest catalogues of human variation are unlikely to contain all variants of interest. Instead, discovery tends to require sequencing of an entire region (not just the known variable sites): something that became possible with the emergence of high-throughput (also known as 'next-generation') sequencing technology in the mid-2000s. These sequencing techniques typically produce short reads of genomic sequence, approximately 35-700 base pairs (bp) in length, which are then reassembled into a complete sequence by mapping to a reference genome (Goodwin et al., 2016). At any one position, the distribution of bases observed across a number of overlapping reads is used

to determine the presence or absence of a variant: the more contributing reads (referred to as the read depth, or coverage), the more confident the variant call will be (Sims et al., 2014).

In its infancy, next-generation sequencing was still expensive, so sequencing was limited to a handful of genes in small numbers of samples. One approach to maximize the effectiveness of IBD sequencing studies was to consider early-onset IBD, as the *XIAP* studies did. Early-onset IBD tends to be more severe, and may be more similar to single-gene, or Mendelian, disorders than adult-onset IBD. Glocker et al. (2009) identified rare recessive variants affecting IL10R protein subunits using a combination of linkage analysis and candidate gene sequencing in early-onset IBD cases from unrelated consanguineous families. Similarly, Blaydon et al. (2011) identified a rare loss-of-function mutation in the gene *ADAM17* (necessary for the cleavage of the epithelial-cell mitogen TGF-α from the cell membrane) that was homozygous in a consanguineous sibling pair affected by inflammatory bowel disease and skin lesions. As the cost of sequencing started to fall, several studies used next-generation sequencing to search for rare and low frequency variation in candidate IBD loci using case control cohorts. One of the earliest such studies sequenced 56 candidate genes identified by GWAS in 350 CD cases and 350 controls (with follow up genotyping in tens of thousands of IBD patients), identifying four additional risk variants in *NOD2*, two protective variants in *IL23R*, and a protective splice variant in *CARD9* (Rivas et al., 2011). A similar study of 55 candidate genes in 200 UC cases and 150 controls recapitulated the presence of rare variants in *CARD9* and *IL23R*, and identified a new association in *RNF186* (Beaudoin et al., 2013). This association to *RNF186* has since been followed up in a much larger cohort, where it has been shown to be highly protective for ulcerative colitis (OR = 0.30), representing the strongest association to UC seen outside of the major histocompatibility complex (Rivas et al., 2016).

Just as was seen during the GWAS era, the logical next step is to scale these candidate-gene sequencing studies up to genome-wide projects: however, deep sequencing of whole genomes across sufficiently large case/control cohorts is currently too expensive. Because the minor allele of a given rare variant is observed so infrequently, obtaining a significantly large difference in minor allele frequency

between cases and controls is not possible with achievable sample sizes. One approach is to use burden testing, which reduces the number of samples needed to detect a rare variant association by aggregating information across all variants in a given target region (such as a gene or exon). Every occurrence of a variant at any position in the region contributes to the overall count, and the difference in these counts between cases and controls is then tested as though they were from a single site of variation. In this way, rare variant associations can be detected with sample sizes that are more comparable to those used to test common variation.

Despite this, obtaining sufficiently large sequenced datasets is still difficult. Zuk et al. (2014) suggest at least 25,000 cases and an equivalent number of controls are needed for a well-powered study. While ultimately deep whole genome sequencing will become affordable, two distinct intermediate approaches exist to sequence large numbers of individuals. First, borrowing the most popular approach in Mendelian genetics, is to only sequence the so-called exome (all exons, or coding regions, in the genome), as this represents less than 2% of the complete genome (Ng et al., 2009). However, the majority of IBD-associated loci identified during the GWAS era actually implicate non-coding regions, and it is likely that rare variants affecting gene regulatory pathways will be of interest. The second design is to spread a fixed amount of sequence data across the whole genomes of many individuals. This produces lower quality data per individual, but the increased sample size improves power to detect low frequency and rare variation in a fixed-cost study (Li et al., 2011). As an added advantage, such cohorts of sequenced individuals then provide useful disease-specific reference panels for imputing rarer variants into new and existing GWAS datasets.

1.4.2 Identifying the casual mutations

With a total of 215 loci associated with Crohn's disease and ulcerative colitis over the past two decades (Parkes et al., 2007; Anderson et al., 2011; Kenny et al., 2012; Yamazaki et al., 2013; Julià et al., 2014; Yang et al., 2014b; Liu et al., 2015; Ellinghaus et al., 2016), and the promise of more to come as next-generation sequencing studies grow, attention is now turning to the identification

of casual genes and variants within these loci (a process known as fine-mapping). Historically, follow-up of genetic associations has proceeded via time-consuming experimental validation of proposed genes using cellular or mouse models. While such functional evidence is essential to fully understand the biology implicated by genetics, it is also possible to leverage the huge sample sizes put together for GWAS to improve fine-mapping before undertaking these experiments. A recent attempt was made to fine-map casual variants in a high-throughput way using the IIBDGCs large Immunochip cohorts, aiming to replicate the success seen in coeliac disease, where the densely packed markers on the Immunochip were used to narrow approximately half of the known signals to an individual gene, or in some cases even subregions of genes (Trynka et al., 2011). The IBD-focused effort was able to resolve 45 associations to a causal variant with greater than 50% certainty, and it is notable that this set is significantly enriched for variants that affect protein-coding regions, transcription factor binding sites and tissue-specific epigenetic marks. This enrichment amongst fine-mappable variants is particularly strong for non-synonymous variation, likely reflecting stronger effect sizes associated with coding variants (Huang et al., 2015).

Further prioritisation of candidate SNPs can be improved by the availability of quality functional annotations from efforts such as the ENCODE Project Consortium (2012), samples from multiple populations (as LD patterns differ between groups of differing ancestry), and combined datasets of huge sample size. Various algorithms have been developed to rank variants within a locus (Huang et al., 2015; Farh et al., 2015; Kichaev et al., 2014), but no definitive method for identifying the disease risk allele exists.

A recent study by Farh et al. (2015) highlights some of the potential challenges in fine-mapping loci given the current knowledge of the effects of different types of genetic variation, with the observation that as much as 90% of causal IBD variants may be non-coding. It was noted that, while casual variants often occurred near the binding sites of master regulators of immune differentiation and stimulus-dependent gene activation, only 10-20% alter a known transcription-factor binding motif. Gaining a more complete understanding of this regulatory code remains an important challenge in both IBD and complex disease genetics more generally.

1.5 Aims and overview

In the previous sections I have provided an overview of the history of complex disease genetics, from the twin studies that first suggested a role for the genome in disease susceptibility to the latest genome-wide association studies that have identified hundreds of associated loci, using inflammatory bowel disease as an example. Through these studies 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. While this means that the accurate genetic diagnosis of disorders like IBD is still a distant prospect, the steady collection of genetic clues has already started to offer insights into the biological mechanisms underlying disease biology, such as the role of autophagy, barrier defense and T-cell differentiation signalling in IBD. The power of sample size has repeatedly been underscored during this process, as increases in sample sizes continue to contribute to relevant disease associations of ever-smaller effect.

The course of these genetic studies over the past twenty years has been constantly shaped by attempts to maximise the scientific questions that can be answered within tight financial and technical constraints, interspersed with occasional technological advances that have produced large leaps forward in discovery. We are now in the early days of one such technological advance, as next generation sequencing offers the first opportunity to capture rarer variation in a high-throughput manner. Already, the benefit of performing large-scale sequencing studies has been demonstrated through efforts such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), which provide valuable resources of variation in the human population. However, how this ability will translate to the identification of rare variation associated with disease risk remains to be seen. In theory, the use of high-throughput sequencing in a case-control setting will offer a unique opportunity to answer unresolved questions about the genetic architecture of complex disease. In particular, the previously unexplored role of rare, low frequency and structural variation can be assessed, to determine how much missing heritability

can be attributed to these types of variation not captured using GWAS, as opposed to still more common variant associations of small effect.

We are therefore faced with several key questions going forward. Firstly, how can we best use the available technologies to better understand the genetic architecture of complex disease, and eventually capture the full breadth of genetic variation contributing to an individual's risk. Furthermore, how can we 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. In this thesis I will begin to address some of these questions, continuing to use inflammatory bowel disease as an exemplar complex trait.

In chapter 2, I will describe some of the challenges of performing large-scale sequencing studies in a case-control setting. In particular, I focus on the bias in sensitivity and specificity of variant calling that can arise when cohorts are sequenced to a different average read depth, and the methods that can be used to overcome this. Through the implementation of a new association test statistic, and the development of several sequencing-specific filtering metrics, I show that it is possible (albeit difficult) to perform large-scale association testing in sequencing data that suffers from widespread systematic biases between cases and controls. This opens up the opportunity for researchers to perform case-control analyses on datasets that have been obtained from multiple sources, such as can often occur when merging datasets in large-scale efforts by disease consortia, or when looking to maximise sample sizes in a fixed-cost study through the use of publicly available control datasets.

In chapter 3, I analyse such a dataset, which consists of low coverage whole genome sequences from 4,280 IBD cases and 3,652 controls sourced from the UK10K project. In order to maximise the number of IBD patients included in this study, the cases were sequenced to a lower average depth (2-4x) than the controls (7x). Using the methods described in chapter 2, I investigate the role of rare, low frequency and structural variation in inflammatory bowel disease risk. Notably, I observe a significant burden of rare, damaging missense variation in the gene *NOD2*, as well as a more general burden of such variation amongst known inflammatory bowel

disease risk genes. Through imputation into both new and existing GWAS cohorts, I also describe the discovery of a low frequency missense variant in *ADCY7* that approximately doubles the risk of ulcerative colitis.

In chapter 4, I meta-analyse these low coverage whole genomes and imputed GWAS datasets with publicly available summary statistics to perform the largest genome-wide association study of common variation in IBD to date. This leads to the identification of 25 novel IBD susceptibility loci, which I then evaluate using fine-mapping and eQTL co-localization in order to resolve the biological mechanisms underlying several of these associations. In particular, I describe 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 are 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, I note that new associations at common variants continue to identify genes that are relevant to therapeutic target identification and prioritization.

Finally, in chapter 5, I turn to the future of studies into the genetics underlying complex diseases such as IBD. I outline some thoughts on the role of next-generation sequencing in understanding disease risk, and consider the implications of these types of study for translation into clinical practice. To conclude, I then present potential opportunities for improving our understanding of environmental risk factors, such as the human microbiota, in the context of complex disease genetics.

Chapter 2

Case-control association testing using sequencing data

2.1 Introduction

The emergence of 'next-generation' technology has caused the cost of DNA sequencing to plummet over the last ten years. This has already led to a number of very successful large-scale sequencing studies using healthy human populations, such as the 1000 Genomes, UK10K, and Exome Aggregation Consortium projects. However, researchers are now looking to extend this success to the identification of disease risk variants using case-control cohorts. Through the direct capture of millions of rare and low frequency variants, such studies offer an unprecedented opportunity to better understand the genetic architecture of complex disease, uncover novel associations underlying disease risk, and further resolve signals down to causal variants of potential therapeutic relevance.

Despite the promise offered by such studies, in practice they are hampered by the high costs associated with sequencing at scale, and the complexity of analysing such data. One cost-saving approach that has been used very successfully in array-based genome wide association studies is to borrow control samples from publicly-available datasets, allowing a maximal number of disease cases to be assayed. However,

attempts to use the same study design in a sequencing setting are faced with a number of difficulties associated with combining multi-source sequencing data at scale. In particular, systematic biases in exome capture technology and sequencing depth lead to crucial sensitivity and specificity differences when performing variant calling; for case-control studies, the effects of these systematic biases can be observed as a slew of false associations.

2.1.1 Chapter overview

In this chapter, I shall describe methods that can be used for the case-control analysis of sequencing data in the presence of a known bias in sensitivity and specificity between the cohorts, as may arise through systematic differences in, amongst other things, sequencing depth. Existing methods to approach this problem include the incorporation of population-level information, through the use of joint calling, genotype refinement, and imputation into GWAS datasets, in order to improve the ability to test for association at sites of low frequency variation.

For rare variation, where the minor allele is observed too infrequently for population-based methods to be effective, I implement a new statistic proposed by Derkach et al. (2014) that is able to account for systematic biases between cases and controls directly in the association test. In order to obtain a well-behaved test statistic on real data, I develop a number of additional filtering recommendations that can be used to identify both errors and variants that are likely to be true sites of variation but have been poorly captured in one of the groups due to systematically lower sequencing depth.

Together, these methods demonstrate that it is possible, albeit difficult, to perform large-scale association testing in sequencing data that suffers from widespread systematic biases between cases and controls. This opens up the opportunity for researchers to perform case-control analyses on datasets that have been obtained from multiple sources, such as can often occur when merging datasets in large-scale efforts by disease consortia, or when looking to maximise sample sizes in a fixed-cost study through the use of publicly available control datasets.

2.1.2 Contributions

In order to test the methods described here, I used a low coverage sequencing study of inflammatory bowel disease performed by the UK IBD Genetics Consortium. Variant calling, genotype refinement, and many of the quality control analyses on this dataset were performed by Yang Luo. Further details on this dataset, and those who contributed to preparing it, will be provided in Chapter 3. Of particular relevance to the work in this chapter, the analysis of low quality sites using support vector machines was performed by Yang Luo. Unless stated, I carried out all other analyses.

2.2 Next-generation sequencing studies

2.2.1 Study design considerations

Next-generation sequencing offers an exciting opportunity to improve our understanding of the genetics underlying complex traits. However, in reality this excitement is tempered by the high costs still associated with sequencing. Because expenditure increases approximately linearly with the number of short sequencing reads produced, a crucial design decision in a fixed cost study revolves around how best to distribute these reads to maximise information: towards increased sample size, increased individual coverage, or an increased number of interrogated sites.

To date, the majority of sequencing studies have focused on the exome. This cost-effective approach to sequencing captures just the protein-coding portion of the genome to high coverage, which makes it well suited for use in clinical diagnostics and the discovery of rare, coding disease variants. Initial studies were therefore focused on individuals or small family groups with unexplained Mendelian disorders. However, exome sequencing has seen an explosion in popularity over the past decade, culminating in the recent release of over 60,000 exomes by the Exome Aggregation Consortium (Lek et al., 2016). During this time, exome studies have offered important insights into a number of aspects of human health and disease, ranging from the identification of causal mutations in rare disorders (Choi et al., 2009; Ng et al., 2010; Wright et al., 2015) and driver mutations in cancers (Barbieri et al., 2012; Stephens et al., 2012), through to more general characterisations of rare coding variation across large cohorts (Walter et al., 2015; Lek et al., 2016).

An alternative study design involves redistributing the sequencing reads to capture the whole genome, but to much lower coverage. This allows for large sample sizes, and the detection of potentially interesting non-coding variation, but comes at the cost of data quality at the individual sample level. This type of study has proven to be a valuable way of obtaining comprehensive genome-wide catalogues of variation across human populations, via studies such as the 1000 Genomes and UK10K projects (1000 Genomes Project Consortium et al., 2015; Walter et al., 2015). Furthermore, through the cost-effective collection of large whole genome

cohorts, such studies have led to the development of a haplotype reference panel containing over 32,000 individuals, providing a very important public resource that can be used for the accurate imputation of low frequency variants from existing genotyping arrays (McCarthy et al., 2016).

2.2.2 Challenges of performing case-control analyses

These large-scale exome and low coverage whole genome efforts have highlighted not only the importance of generating very large sequencing cohorts, to reveal patterns of human population biology and provide vital resources for interpreting the clinical relevance of variation, but also the practical difficulties in managing multi-source data at this scale. The lack of a standardised approach for the generation of sequencing data has resulted in a number of slight variations on the basic study design, whether it be high-coverage exomes or low-coverage whole genomes, as investigators try to fine-tune their designs to answer a variety of scientific questions. As a result, when combining data from 14 different studies, the Exome Aggregation Consortium pointed out that variations in exome capture technology and sequencing depth across their 60,706 exomes required a joint analysis of such computational intensity and analytical complexity that it would be impossible using the limited resources available to most research centres (Lek et al., 2016).

I will note here that the systematic differences between cohorts being referred to here are not the same as the batch effects that can arise through the course of an experimental study. Just as is often seen with genotyping data, sequencing studies are still plagued by such issues: the specific reagents and machines used, slight variations in experimental conditions, or even the day on which a sample was processed can all lead to differences in the quality of the data produced (Figure 2.1). Naturally, these problems are important to consider, and indeed if samples are processed at multiple sequencing facilities then these effects can become even more pronounced. However, generally, these sorts of batch effects can be accounted for using careful quality control.

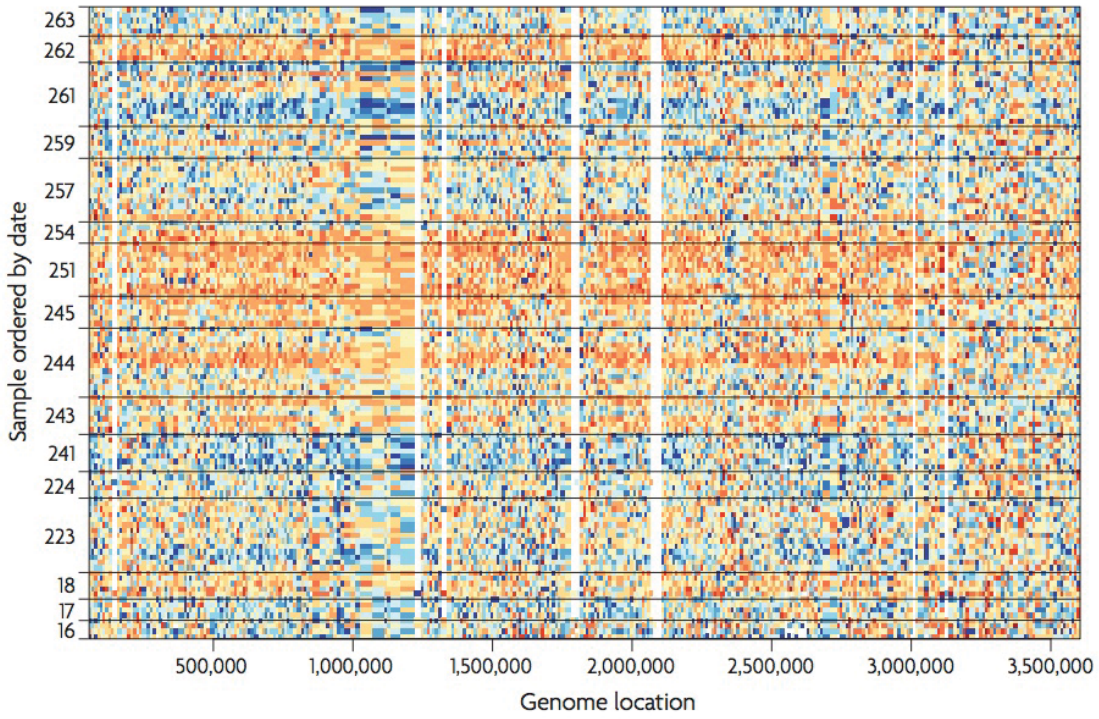

Figure 2.1: Batch effects observed in the 1000 Genomes project sequence data. Each row repesents a different HapMap sample, sorted by processing date. Every sample was processed in the same sequencing facility, using the same platform. Colours represent the standardised coverage data for each sample: blue indicates three standard deviations below average, and orange indicates three standard deviation above average. A large batch effect is observed between days 243 and 251. Figure sourced from Leek et al. (2010).

Of greater concern when combining sequence data from multiple sources are more widespread systematic differences that have arisen due to variation in the study designs. One example of this is the exome capture kit used, which defines the regions of the genome that will be sequenced and (through variable probe efficacies) the relative read depth that is likely to be obtained for certain regions. Systematic differences in read depth can also be observed on a more global scale, when data has simply been collected to different average coverages.

Because variants are detected in sequence data using the distribution of alleles across all the reads that overlap at a given position, sequencing depth has a direct impact on the sensitivity and specificity of variant calling. In particular, increased read depth leads to both improved sensitivity (the detection of true variant sites)

and improved specificity (the ability to distinguish true variants from sequencing errors). As a result, a cohort sequenced to higher depth (whether that be globally or locally) can be expected to contain more sites of true variation, and fewer errors, than a cohort sequenced to lower average depth across the same regions.

This observation is likely to be a serious problem as we extend the success of sequencing-based studies in healthy human populations to explore disease associations in case-control cohorts. I shall describe one such effort in Chapter 3, where we use low coverage sequencing to search for rare and low frequency variation associated with IBD. In that example, the cases were sequenced to a lower average depth than the controls (which were sourced from the UK10K project), in order to maximise sample size and therefore power to detect associations. Although this study may represent a particularly extreme example of differing read depths between cases (2-4x) and controls (7x), we envision that similar issues are likely to arise in other studies that use publicly available controls to save on costs. In this sort of case-control setting, any systematic differences between sequencing data from different sources is likely to heavily bias attempts to perform association testing.

In the following sections, I shall describe a range of methods that can be used to overcome systematic differences in sequencing depth between cases and controls. These consist of two broad approaches, depending on the prevalence of the variant of interest in the population. Firstly, for more common variants, population level information can be used to improve the overall sensitivity of both datasets and reduce differences between cohorts, thereby allowing standard association testing methods to be used. For rare variants, where this information is not available, I instead describe the development of a new approach to perform association testing in the presence of coverage bias between cases and controls.

2.3 Low frequency and common variants

2.3.1 Joint calling across samples

A powerful means of overcoming systematic sequencing differences between cases and controls at sites of low frequency and common variation is to perform joint variant calling (Figure 2.2). This method uses population-level detail about a given site to improve sensitivity to detect variation in carriers that have only intermediate levels of sequence support. It also allows for better specificity in variant detection: essentially, when more information is incorporated, it becomes easier to model errors and detect false positives. This is particularly important for sequencing data where, unlike the extensively curated variant lists that are included on genotyping arrays, there has been no pre-selection for true sites of variation.

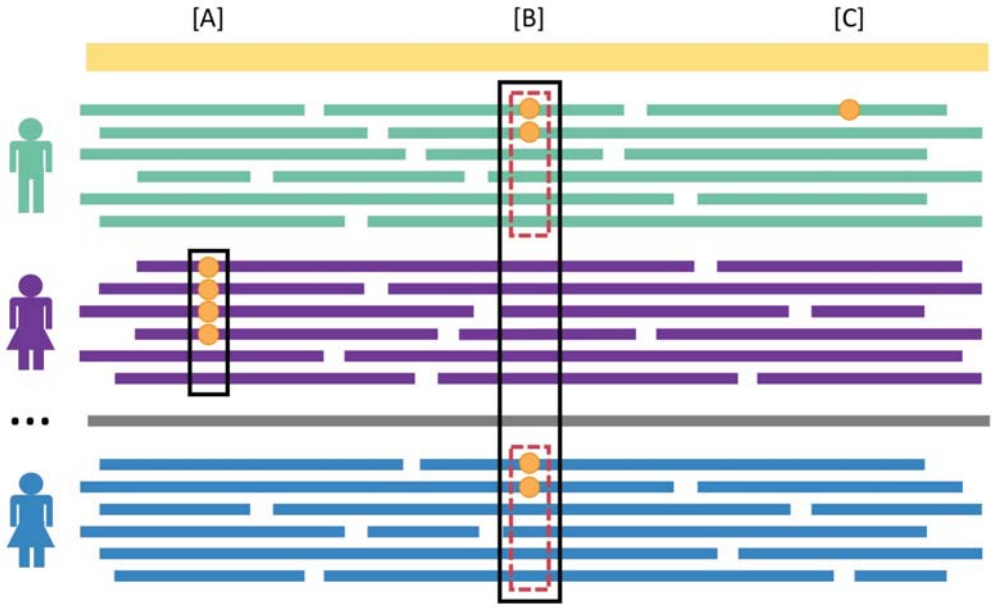

Figure 2.2: Calling variants jointly across a number of individuals can increase both the sensitivity and specificity of variant detection. While some variants may be accurately captured through variant calling on a single sample (A), for some individuals a lack of sequence support can cause the variant to be missed (B). However, if the variant is jointly called across reads pooled from a number of samples, these variants can be more accurately detected (B). Joint calling also helps to improve the detection of errors (C).

By performing variant calling jointly across the entire case-control cohort, the genotype calls for all samples will utilise information from reads accumulated over both cases and controls. This can greatly improve the sensitivity and specificity of variant calling for both groups, and reduce calling differences that may have arisen due to variations in average sequencing depths.

2.3.2 Genotype refinement

After joint calling, some variants that have been poorly captured for a given individual can be improved using genotype refinement (Figure 2.3), which infers specific genotypes by imputing from other individuals and neighbouring variation. As Li (2011) explains, this method improves the genotype call for an individual, I, who happens to have poor sequence coverage at the site of interest, S_0. If there are other samples that have high coverage at S_0 then, if there exists a second site S_1 which is in high linkage disequilibrium with S_0, and for which both I and the other samples have sufficient sequence support, the likely genotype for individual I at position S_0 can be inferred.

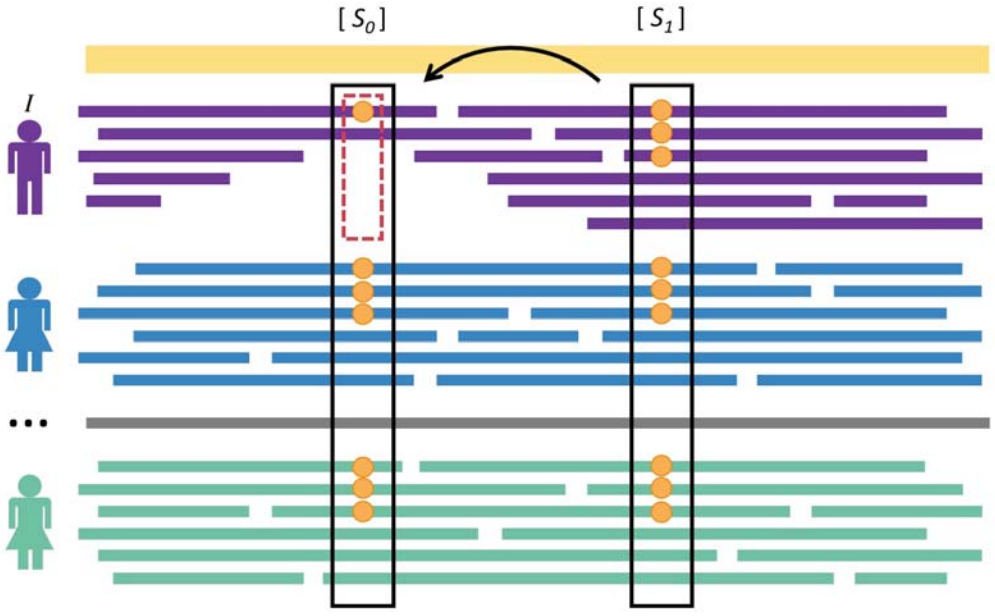

Figure 2.3: Genotype refinement through imputation, where the poor quality genotype at position S_0 for individual I is improved by imputing from position S_1.

2.3.3 Imputation of GWAS cohorts

A combination of joint variant calling and genotype refinement is an effective way of improving variant calls in sequencing data, particularly when the average read depth is low. Both methods were used successfully in the 1000 Genomes and UK10K projects to generate high-quality variant call sets, and when applied simultaneously to both case and control cohorts they are also able to help alleviate the variable sensitivity and specificity that can arise from systematic differences in sequencing coverage (Figure 2.4).

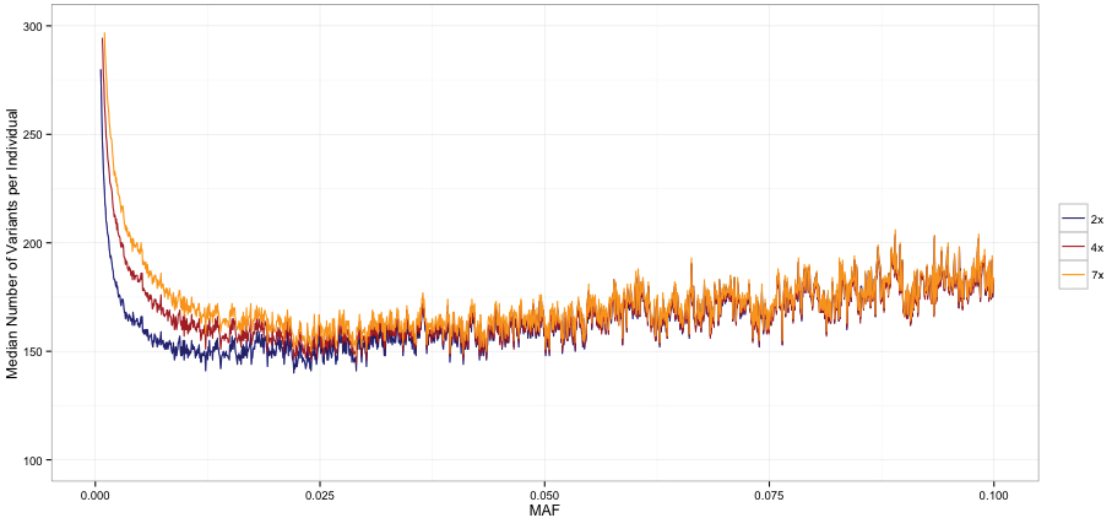

Figure 2.4: I investigate the effect of read depth on sensitivity and specificity across the allele frequency spectrum, for a range of average sequencing depths as shown with blue (2x), red (4x) and yellow (7x) lines. Variants have been jointly called across three cohorts (1,767 2x, 2,513 4x, and 3,652 7x samples), followed by genotype refinement. Sensitivity is then approximated as the median number of variants called per individual. Compared to rare variant calls, which do not have sufficient population-level information to be improved through joint calling and genotype refinement, the differences in sensitivity between each cohort have been notably improved for low frequency and common variation.

However, association testing using low frequency and common variation (MAF $\geq$ 0.1%) is still susceptible to residual bias due to sequencing depth. As will be discussed in more detail in section 3.5, despite using joint calling, genotype refinement, and very stringent quality control on our low coverage IBD sequences, there was still an excess of extremely significant sites ($P < 1 \times 10^{-15}$) falling outside

of known IBD-associated loci, the majority of which had a MAF $< 5\%$. Most (if not all) of these are likely to be false associations that simply reflect the greater number of observations in the higher coverage group due to better sensitivity, rather than any true effect on disease risk.

Although residual bias from sequencing depth differences can prevent case-control association testing of low frequency variation in differentially sequenced cohorts alone, these datasets still provide valuable imputation reference panels. With quality variant call sets produced using joint calling and genotype refinement, a set of haplotypes from across both cases and controls can be used to impute these variants into large panels of genotyped individuals. This approach not only increases sample size, and therefore power to detect associations, but will also produce case-control datasets that are not affected by the original coverage bias present in the sequenced reference panels. For example, imputation into GWAS was used successfully by a recent case-control association study of Type 2 diabetes to increase the utility of their low-coverage whole genome sequences (Fuchsberger et al., 2016).

2.4 Rare variant association testing

Because the minor allele of a given rare variant is observed so infrequently, methods that rely on the incorporation of population-level information, such as joint calling, genotype refinement, and imputation, cannot be usefully applied (Figure 2.4). This leads to two major issues when performing rare variant association studies in case-control cohorts. Firstly, testing can only be performed in directly sequenced individuals, limiting sample sizes. Given the scarcity of these variants in the population, obtaining a significantly large difference in minor allele frequency between cases and controls is simply not possible with achievable sample sizes. Secondly, any systematic bias in read depth between the cohorts cannot be overcome by processing the data prior to association testing, requiring new association test statistics that are tailored to this specific situation. I shall discuss the development

of an approach that can be used to address each of these problems in the following sections.

2.4.1 Increasing power using burden testing

Single-variant association tests can only be successfully applied to rare variants if the sample sizes are sufficiently large, or the variant effects are particularly strong. Because of this, rare variant association testing generally relies on the aggregation of signals from across multiple variants in order to increase power. The most common methods by which variants are aggregated and their cumulative effects are tested can be broadly broken into three categories: burden tests, variance-component tests, and combined tests (Lee et al., 2014b; Moutsianas and Morris, 2014). Depending on the underlying genetic architecture of the disease being tested, different methods will be better powered to detect an association (Table 2.1).

The simplest approach is to perform a burden test, which combines information across a number of variants in a target region (e.g. by counting the number of occurrences of each minor allele) and then tests the resulting summary score. However, such methods only work well if the majority of variants included are causal, and all have the same direction of association with the trait. One way to overcome these limitations is to use a variance component test, which compares the observed variance with the expected variance of the distribution of allele frequencies in a target region. If the variance is over-dispersed, meaning an increase from the expected binomial variance, this can indicate a subset of variants that are preferentially observed in either cases or controls (Figure 2.5). In this way, it is possible to efficiently test for a combination of effect directions (risk, neutral or protective), although this does come at the cost of reduced power if all variants do in fact act in the same direction (Neale et al., 2011).

Table 2.1: A comparison of current rare variant association testing methods, adapted from Lee et al. (2014b).

Method	Advantages	Disadvantages	Examples
Burden tests			
Collapse multiple variants into a single summary score	Well-powered when a large proportion of the variants included are causal, and all have the same direction of association with the trait	Performs poorly in the presence of both risk, protective and null variants	CAST (Morgenthaler and Thilly, 2007) CMC (Li and Leal, 2008) ARIEL(Asimit et al., 2012) MZ test (Morris and Zeggini, 2010) WSS (Madsen and Browning, 2009)
Variance component tests			
Test for over-dispertion in the variance of genetic effects	More robust to the presence of both risk and protective variants, or if only a small proportion of included variants are causal	Less powerful than burden tests if most variants are in fact causal and operating in the same direction of effect	C-alpha (Neale et al., 2011) SKAT(Wu et al., 2011) SSU test(Pan, 2009)
Combined tests			
Linear combinations of burden and variance component tests	Useful for datasets when the genetic architecture is unknown, reduces the power loss associated with applying the incorrect model	Is less powerful than applying either a burden test or a variance component test to data where their respective assumptions about the genetic architecture hold	SKAT-O (Lee et al., 2012b) Fisher method (Derkach et al., 2013) MiST (Sun et al., 2013)

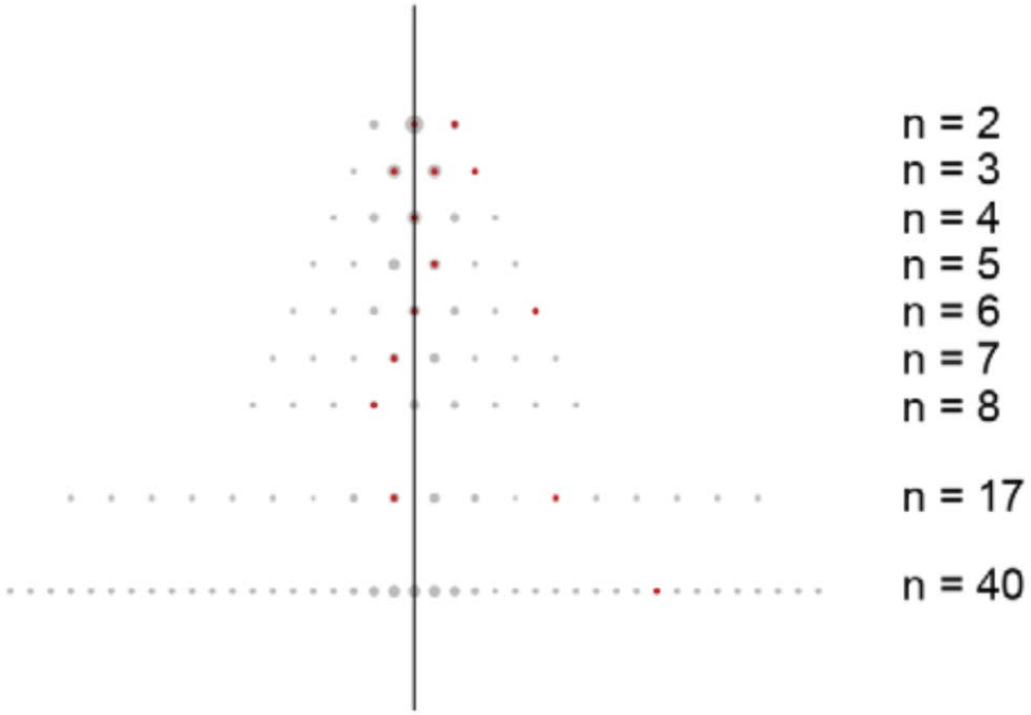

Figure 2.5: An example of the distribution of recurrent, low frequency non-synoymous variants in *NOD2*, comparing 350 CD cases to 350 controls. Each row defines variants observed n times in the dataset, with the observations split between controls (left of the vertical line) and cases (right of the vertical line). As an example, the n = 3 row describes three observed variants in red, one seen in 3 cases and 0 controls, one seen in 2 cases and 1 control, and one seen in 1 case and 2 controls. The variance component test determines if there is a difference in the variance of the observed data (red) and the binomial probability distribution (grey). Figure sourced from Neale et al. (2011).

While variance component tests are generally the preferred approach when faced with the aggregation of variable effect sizes and directions, their loss of power compared to simple burden tests when effect direction is consistent means that many people who are testing data of unknown genetic architecture will turn to tests that combine both burden and variance component approaches. Rather than simply applying each test separately and taking the minimum p-value, which can lead to an inflated type I error rate, these combined tests attempt to find the optimal linear combination of both the burden and variance-component tests (Lee et al., 2012b).

2.4.2 Accounting for differences in sensitivity and specificity between cases and control

In general, the rare variant association tests discussed above assume the case and control datasets have been well matched. In particular, the minor allele frequencies to be tested are derived directly from genotype calls, thereby assuming that these calls are equivalent for the two datasets. Unfortunately, when there are systematic biases in coverage between the cohorts this assumption does not hold. In practice, there is increased sensitivity to detect variation in the higher coverage group, and decreased specificity to avoid errors in the lower coverage group (Figure 2.6). This can lead to two types of false association signals: an excess of erroneous variants that have been called in the lower coverage group, and an excess of true variant calls in the higher coverage group that failed to be detected in the lower coverage cohort. Depending on how different subsets of these variants (which have opposing false signals) are selected for aggregation into a burden test, it is possible that significant false associations may be observed.

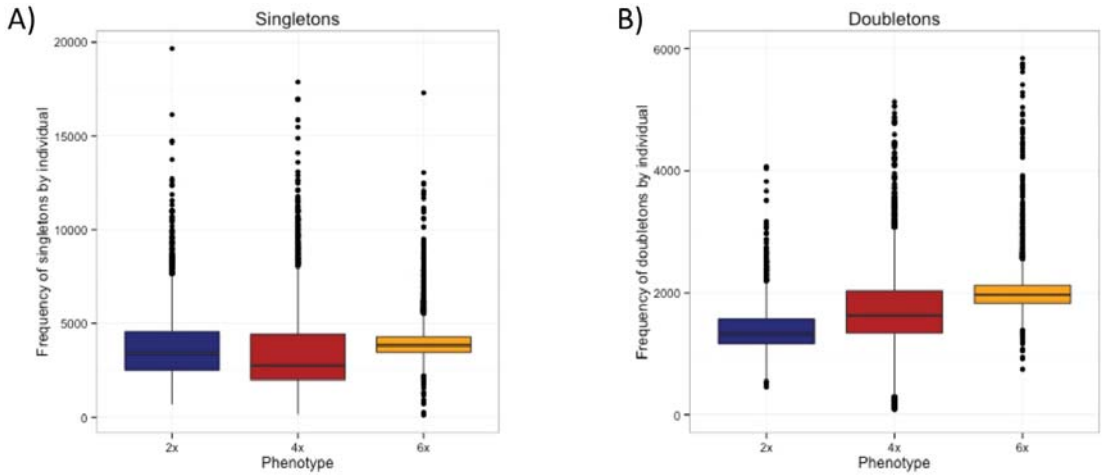

Figure 2.6: The effect of read depth on the sensitivity and specificity of calling genotypes for rare variants. Variants were jointly called across three cohorts (1,767 2x (blue), 2,513 4x (red), and 3,652 7x (yellow) samples), followed by genotype refinement. For singletons, which are observed just once in the population, there is an excess of variants called in the 2x cohort, indicating a loss of specificity at low coverage (panel A). For doubletons, which are observed twice in the population, we see a more general pattern of increasing sensitivity as read depth is increased (panel B).

One way to overcome this issue would be to down-sample the higher coverage group so that the average read depth is consistent across both cases and controls prior to variant calling, and then perform association testing using one of the standard methods from Table 2.1. However, this requires the removal of potentially useful sequence information. To avoid the loss of valuable data, another commonly proposed solution is to test for association using a logistic regression analysis that includes the read depth as a covariate (Garner, 2011), or weights variants based on quality scores (Daye et al., 2012). However, if the cases and controls can be perfectly separated by read depth then it cannot be used as a covariate, as it will cause the parameters of the logistic regression to no longer be estimable (Derkach et al., 2014).

Instead, the solution I use here is to account for known differences in the sensitivity and specificity of variant calling by replacing the hard genotype calls with genotype dosages. Rather than discrete counts of the minor allele, such that a genotype call for individual i at position j can be defined as $G_{ij} \in \{0, 1, 2\}$, the dosage is calculated as the expected genotype given the sequencing data D, such that $E\left(G_{ij}|D_{ij}\right) = \sum_{g=0}^{2} gP(G_{ij} = g|D_{ij})$. Here, $P(G_{ij} = g|D_{ij})$ is the probability of each genotype given the sequencing data. The resulting dosage estimate better reflects the confidence of a variant call, allowing for the effects of read depth to be incorporated into the test.

Association testing using genotype dosage

Skotte et al. (2012) developed a score statistic that performs association testing using this genotype dosage data. Their statistic is derived from the joint likelihood of phenotype and sequencing data across all individuals at a given locus (Equation 2.1). This assumes that, across n samples, for any one individual i their phenotype Y depends on the observed sequencing data D through the unobserved genotype G at locus j.

$$P\left(\mathbf{Y} = (Y_1, ..., Y_n), \mathbf{D} = (D_{1j}, ..., D_{nj})\right) = \prod_{i=1}^{n} \left(\sum_{g=0}^{2} P\left(Y_i|G_{ij} = g\right) P\left(G_{ij} = g, D_{ij}\right) \right) \quad (2.1)$$

The main component of interest in this likelihood is the relationship between the phenotype and the genotype, $P(Y_i|G_{ij} = g)$: if we were to consider $logit\left(P\left(Y_i|G_{ij} = g\right)\right) = B_0 + B_1 g$ then a test to determine if the slope is null ($H_0 : B_1 = 0$) can be used to indicate if there is any association between the two. S_j, the score statistic for B_1, has been derived in Equation 2.2, and has the variance as shown in Equation 2.3. The corresponding test statistic $T_j = \frac{{S_j}^2}{Var(S_j)}$ is chi-squared, with one degree of freedom. Under the null hypothesis, $S_j = 0$.

$$S_j = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) E\left(G_{ij}|D_{ij}\right) \tag{2.2}$$

$$Var\left(S_j\right) = \sum_{cases} \left(1 - \bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) + \sum_{controls} \left(\bar{Y}\right)^2 Var\left(E\left(G_{ij}|D_{ij}\right)\right) \tag{2.3}$$

Importantly, the variance of $E(G_{ij}|D_{ij})$ is read depth dependent. Intuitively, as read depth increases the data will better reflect the true genotype, so that $E(G_{ij}|D_{ij})$ will approach the true G_{ij} while $Var(E(G_{ij}|D_{ij}))$ approaches the true $Var(G_{ij})$. This is because we obtain less information about the true genotype at lower coverages, and thus the expected variance of the genotype given the data, $E(Var(G_{ij}|D_{ij}))$, is greater. At sufficiently high coverage, when we can consider the data to perfectly reflect the true genotype, this value should converge to 0. Therefore, by the law of total variances (Equation 2.4), estimating the variance of the true genotypes using $Var(E(G_{ij}|D_{ij}))$ will lead to an underestimate of this value at low depths.

$$Var(G_{ij}) = Var(E(G_{ij}|D_{ij})) + E(Var(G_{ij}|D_{ij})) \tag{2.4}$$

How this corresponds to the variance component of the test statistic depends on the relative depths and sample sizes of the two groups, as the group with the smallest sample size will contribute the most to the variance calculation, due to the inclusion of the average phenotype $\bar{Y}$ in the weights (see Equation 2.3). For

example, if we assume that the high coverage group has sufficient information to obtain reasonable variance estimates, while the lower coverage group does not, then when $N_{Low} >> N_{High}$ the variance component will be underestimated, while if $N_{High} >> N_{Low}$ the variance component may actually be overestimated. Underestimation of the variance component will lead to an overinflated test statistic, and vice versa.

Derkach et al. (2014) therefore proposed that, in the presence of systematic read depth differences between cases and controls, a more accurate test statistic could be obtained by calculating the variance components for the two groups separately (Equation 2.5).

$$\begin{aligned}\hat{V}ar\left(S_j\right) = N_{case} \left(\frac{N_{control}}{N}\right)^2 \hat{V}ar_{case} \left(E(G_{ij}|D_{ij})\right) \\ + N_{control}\left(\frac{N_{case}}{N}\right)^2 \hat{V}ar_{control}\big(E(G_{ij}|D_{ij})\big)\end{aligned} \tag{2.5}$$

This 'Robust Variance Score' (RVS) statistic can be extended to perform a burden test for multiple rare variants, using a similar approach as standard burden tests like CAST and CMC. The individual variant score statistics are simply summed together to given an overall score, while the variance component is calculated by combining the covariance matrices of the cases and controls, after estimating them separately. Unfortunately, however, the distribution of the resulting test statistic for the joint variant analysis is unknown. Instead, a permutation-style procedure needs to be used, whereby a p-value is generated by creating X bootstrap samples and counting up the number of times they generate a test statistic that is more significant than the original sample. Usually, evaluating significance using permutation would involved randomly permuting case and control status, but the different read depths between the groups precludes this. Instead, both the case and control groups are separately centred around their respective means, and then (still separately) sampled with replacement from these centred values, maintaining the same numbers of cases and controls as the original sample. In this way, the difference between the groups is reduced to one dimension (variance only), forming

an empirical null set from which bootstrap samples can be generated without swapping case and control status (Derkach et al., 2014).

2.4.3 Testing in a dataset with systematic read depth bias between cases and controls

In order to test the performance of the RVS in the presence of a known systematic bias in read depth between cases and controls, I considered a low coverage whole genome sequencing study of inflammatory bowel disease. The sample collection, sequencing and quality control procedures used to generate this dataset will be described in more detail in Chapter 3. However, briefly, it consists of 1,767 patients with ulcerative colitis (median coverage of 2x), 2,513 patients with Crohn's disease (4x), and 3,652 population controls (7x).

Implementing the RVS statistic in C++

Testing a dataset of this size using the original R implementation of the RVS statistic as provided by Derkach et al. (2014) would lead to extensive computer memory demands and excessive run times, such that it was not possible even given the sizeable computational resources available at the Wellcome Trust Sanger Institute. I therefore had to first implement the RVS statistic as an extension to the software ANGSD (Korneliussen et al., 2014), which makes use of the compiled language C++ and multi-threading to generate much more efficient run times. My implementation can be found at https://github.com/katiedelange/angsd.

I developed the algorithm described in Box 2.1 to perform the RVS association test within the framework defined by ANGSD. I optimised this solution to minimise memory requirements (currently the most limiting resource within the cluster computer framework to be used for association testing) and made use of multi-threading in order to parallelise steps wherever possible.

Box 2.1: Algorithm used to implement the RVS statistic within the ANGSD framework.

```
// Request the following inputs from the user
    - The number of burn-in bootstrap resampling permutations to
      perform before significance is evaluated
    - The number of bootstrap resampling permutations to perform
      (-1 specifies that adaptive permutation should be used)

// Extract the relevant summary data from the genotype probabilities
For each site j
    For each individual i
        Compute and store the expected genotype
          E(G_ij|D_ij) = sum_g P(G_ij = g|D_ij), for g=0,1,2
        Compute and store the expected variance
          Var(G_ij = g|D_ij) = E(G_i^2|D_ij) - E(G_ij|D_ij)^2
    Determine the population allele frequency estimate
      (G_ij|D_ij)/2N across both samples at this site.

// Compute the score statistic components for the unpermuted sample
Append the burden score S to the list of scores
    S = sum_{j=0}^{N}(S_j), where S_j = sum(Y_i - Ybar)E(G_ij|D_ij)
Append the burden variance Var(S) to the list of variances
    Var(S) = sum_i sum_j sum_k cov(E(G_ij|D_ij), E(G_ik|D_ik)

// Centre the stored genotype dosages around their respective means
Separately for cases and controls
    For each site j
        Compute the mean expected genotype
        Subtract this from each individual using a matrix transform

// Run permutation testing to evaluate the significance of the test
For the requested number of permutations
    Separately for N_0 controls and N_1 cases
        Randomly sample N_(0,1) times (with replacement)
    Append the permuted sample score S to the list of scores
    Append the permuted sample variance Var(S) to the list of variances

// Return the fraction of times that a permuted sample is more
// significant than the original sample
```

Performance of the RVS in systematically biased data

I tested the performance of the RVS burden test on rare (0.0001<MAF<0.01) functional coding variation within genes. I define functional coding variants to be those with one of the following Variant Effect Predictor (McLaren et al., 2010) annotations: frameshift_variant, stop_gained, initiator_codon_variant, splice_donor_variant, splice_acceptor_variant, missense_variant, stop_lost, inframe_deletion, or inframe_insertion. The MAF range used is also defined so as to exclude singletons, due to the lack of specificity at this frequency for very low coverage data (Figure 2.6). Despite these restrictions, I observe a very large excess of apparently significant associations after 10^6 permutations (Figure 2.7), and systematic over-inflation of the test statistic ($\lambda = 1.34$).

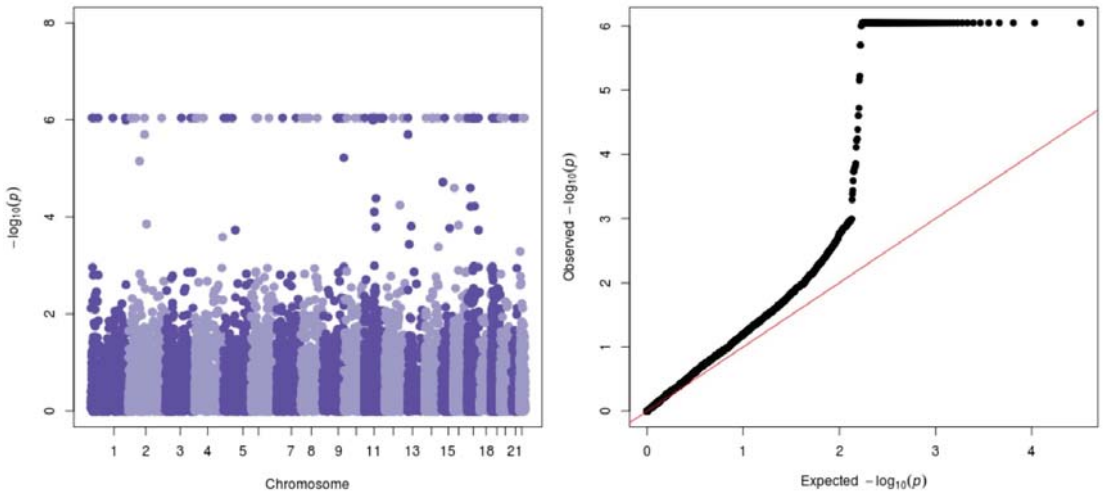

Figure 2.7: Burden testing using the RVS statistic (up to 1,000,000 permutations) on rare (0.0001<MAF<0.01) functional coding variation within genes.

When trying to determine why this statistic does not appear to be adjusting for the systematic coverage bias in this dataset as well as the authors suggest it will (Derkach et al., 2014), I note that there are a few crucial assumptions that must be met. In particular, the method assumes that 1) the variants being tested are true sites of variation; and 2) a variant has been successfully detected if it is present. However, particularly when dealing with rare variants in very low coverage datasets, it is likely that these assumptions will be violated at a number of tested sites. This

includes both errors that have been mistakenly included in the lower coverage group due to reduced specificity, and rare variants that have failed to be detected in the lower coverage group due to reduced sensitivity. I therefore looked to modify the standard sequencing quality control procedure that was applied to this test dataset (see Chapter 3 for details) to include additional filters tailored to rare variants, in order to both better remove potential errors and try to identify sites that, whilst true sites of variation, failed to be identified in one group due to low coverage (rather than disease association).

2.4.4 Adjusting the quality control procedures

Identifying variants sites that were missed at lower coverage

I first focus on trying to deal with rare variant calls that are likely to be true sites of variation, but were missed in the lower coverage group due to a lack of sensitivity. Hu et al. (2016) show that this particular problem can sometimes be overcome by modelling the error rate and using it to predict loci that are likely to be true variants. In particular, they aim to include the maximal set of possible variants in the test, applying only minimal filtering to try and remove sites that are predicted to be truly monomorphic in both datasets. This is done by screening out sites that are predicted to be uninformative, in that they have a score $S = 0$ and therefore do not contribute to the burden test. However, because this minimal screening step is unlikely to capture all problematic sites, they then adjust the permutation procedure to try and generate bootstrap datasets that have identical allele frequencies between cases and controls, but match the read depths, error rates, and the number of true variants and monomorphic loci that are seen in the original dataset. Unfortunately, this method relies on a sufficiently strong signal-to-noise ratio at very rare sites in at least one of the groups being tested, in order to properly model errors for the initial screening step. For situations where both cases and controls are of low coverage, this method is not expected to offer any significant advantages over Derkach et al's RVS model.

I therefore looked to capture these sites as part of the filtering process instead, by trying to measure how accurately a given site is likely to have been captured across all the individuals in each cohort. To do this, I calculate the INFO score α, which can be interpreted as describing the amount of 'missing' information such that the observed data at a site is equivalent to a set of perfectly observed genotypes in a sample of size αN (Marchini and Howie, 2010), separately for each cohort. It is computed using the likelihood of the true population allele frequency θ_j at a given site j if we had observed genotypes G_{ij}, as shown in Equation 2.6.

$$\mathcal{L}(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}} \tag{2.6}$$

The score (first derivative) and information (second derivative) for this likelihood are shown in Equations 2.7 and 2.8, where N is the sample size, and $X = \sum_{i=1}^{N} G_{ij}$. The score reflects how sensitively $\mathcal{L}(\theta_j)$ depends on θ_j, while the information describes how much information the observable variable G_{ij} carries about θ_j.

$$U(\theta_j) = \frac{d \log \mathcal{L}(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1-\theta_j)} \tag{2.7}$$

$$I(\theta_j) = \frac{d^2 \log \mathcal{L}(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1-\theta_j)^2} \tag{2.8}$$

If we then consider that the genotypes G_{ij} are not perfectly observable, but are instead approximated through the data D_{ij}, we can compute a similar likelihood for the allele frequency parameter θ_j that is integrated over the missing data that comes from estimating G_{ij} using D_{ij} (Equation 2.9). In order to do this, the data is partitioned into the observed data Y_O and the missing data Y_M.

$$\mathcal{L}^*(\theta_j, Y_O) = \log(P(Y_O|\theta)) = \log \int P(Y_O, Y_M|\theta)\, \mathrm{d}Y_M. \tag{2.9}$$

The score and information of this observed data likelihood is heavily related to that of the full likelihood, as shown in Equations 2.10 and 2.11 (Louis, 1982).

$$U^*(\theta) = \frac{d\mathcal{L}^*(\theta_j)}{d\theta_j} = E_{Y_M|Y_O.G_{ij}}[U(\theta_j)] \tag{2.10}$$

$$I^*(\theta_j) = \frac{d^2\mathcal{L}^*(\theta_j)}{d\theta_j^2} = E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] - V_{Y_M|Y_O.G_{ij}}[U(\theta_j)] \tag{2.11}$$

Of particular interest here is the information statistic, which we can use to describe the amount of missing information about the true allele frequency due to estimation using observed data as opposed to true genotypes. If we consider $I^*(\theta_j)$ to represent the observed information, and $E_{Y_M|Y_O.G_{ij}}[I(\theta_j)]$ the complete information, it follows that $V_{Y_M|Y_O.G_{ij}}[U(\theta_j)]$ is the missing information. These components can be calculated using Equations 2.12 and 2.13. Importantly, we can see that the top line of Equation 2.13 is actually calculating $Var(G_{ij}|D_{ij})$: as mentioned earlier, this converges to 0 as the read depth improves. Therefore, we expect more missing data in lower coverage samples.

$$E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \tag{2.12}$$

$$V_{Y_M|Y_O.G_{ij}}[U(\theta_j)] = \frac{\sum_{i=1}^{N} E(G_{ij}|D_{ij}) - E(G_{ij}^2|D_{ij})}{\hat{\theta}^2(1-\hat{\theta})^2} \tag{2.13}$$

Using these two terms, we can compute the ratio of observed data to complete data (Equation 2.14), giving the INFO score α that can then be used to generate an effective sample size αN for the amount of informative data in the sample set at site j.

$$\alpha = \frac{E_{Y_M|Y_O.G_{ij}}[I(\theta_j)] - V_{Y_M|Y_O.G_{ij}}[U(\theta_j)]}{E_{Y_M|Y_O.G_{ij}}[I(\theta_j)]} \tag{2.14}$$

This INFO score provides an estimate of how well a variant has been captured across all the individuals in each cohort, and (as can be seen in Equations 2.12 and 2.13) is also closely related to the terms being tested by the RVS statistic. I

therefore computed this statistic for each site separately in each of the test cohorts, and plotted the distributions as shown in Figure 2.8.

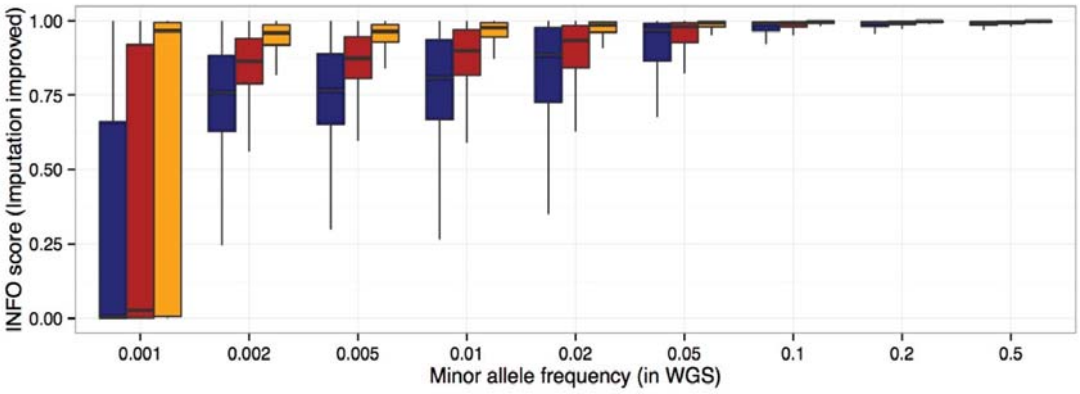

Figure 2.8: The distribution of the INFO score for 2x (blue), 4x (red) and 7x (yellow) data across a range of minor allele frequencies.

Immediately apparent are the large differences in median INFO scores between each of the cohorts below a minor allele frequency of $\sim 2\%$. This is particularly pronounced for very rare variants, where the datasets sequenced to 2-4x average coverage retain almost no information about sites with a MAF $< 0.2\%$. Given these observations, it is unsurprising that a score statistic calculated using datasets that are so distinct in their ability to capture the true genotypes resulted in such an excess of false positive associations. However, the extent to which each cohort differed on their median INFO measure, and how this changed between rare and common sites, was more unexpected.

One possibility is that this effect may be related to the use of genotype refinement via imputation, which is the major MAF-dependent factor affecting the genotype probabilities from which both the INFO score and RVS statistic are calculated. This process aims to remove noise and improve confidence in genotype calls made: in essence, producing a set of 'smoothed' genotype probabilities through the incorporation of population-level information. However when the true signal is low, such as for sites of rare variation, it may be that this refinement step is overzealous. To evaluate if this is the case, I investigated the use of genotype probabilities generated directly from the samtools Genotype Quality (GQ) field, without any genotype refinement.

The GQ value represents the phred-scaled genotype probability of the most likely genotype, as calculated by GQ$= -10 \log_{10} \max \left(P(G_{ij} = g | D_{ij}) \right.$, for $g \in \{0, 1, 2\}\big)$. Unfortunately, this does not provide enough information to resolve all three possible genotype probabilities (homozygous reference, RR; heterozygous, RA; and homozygous alternate, AA). Therefore, in order to produce a set of genotype probabilities I assign the probability reflected in the GQ score to the genotype called in the VCF file, and all the remaining probability to the most likely alternate call:

$$P(\text{Call}) = 1 - 10^{\frac{-\text{GQ}}{10}}$$

$$P(\text{Alt}) = 1 - P(\text{Call})$$

$$P(\text{Remainder}) = 0$$

When the called genotype is homozygous, the next most likely genotype is assumed to be the heterozygous genotype (i.e. if Call=RR or AA, then Alt=RA). If the genotype call was heterozygous I assume, given the low MAF (≤ 0.01) of the variants being considered for burden testing, that the rare homozygote is not likely to be observed and thus I define the next most likely genotype as being homozygous reference (i.e. Call=RA, Alt=RR).

I compute these unrefined genotype probabilities across the complete dataset, and recalculate the INFO score separately for each of the three cohorts, across all sites. As can be seen in Figure 2.9, using unrefined genotype data leads to a dramatic improvement in the amount of information obtained at sites of rarer variation (MAF $\leq 2\%$). The utility of performing genotype refinement at common sites is also apparent, with improved INFO score distributions for higher MAFs (particularly MAF $\geq 10\%$, Figure 2.8).

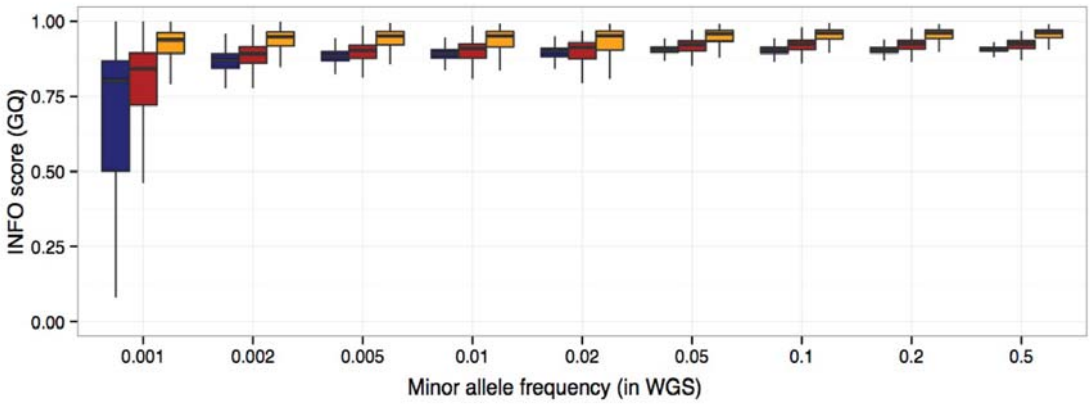

Figure 2.9: The distribution of the INFO score for 2x (blue), 4x (red) and 7x (yellow) data across a range of minor allele frequencies, using raw genotype probabilities estimated directly from the samtools genotype quality score.

In order to minimise the possible differences in INFO score between the case and control cohorts during association testing, and thus attempt to reduce the inclusion of rare variants that have been detected in the high coverage group but missed in the low coverage group due to reduced sensitivity, I filter out any sites with INFO < 0.6 in either of the relevant cohorts for each test. In general, this allows more sites to be retained when comparing the 4x cases (as opposed to the 2x cases) to the 7x controls.

Additional error filtering

I then applied the following additional quality control filters, to try and reduce the number of erroneous sites included (particularly from the lower coverage group, which has poorer specificity during variant calling):

– Sites with a missingness rate > 0.9. When using unrefined genotype probabilities, the missingness rate across all sites is greatly increased, compared to the refined set that has attempted to infer a number of missing genotypes. I remove any sites with a high number of samples where a genotype could not be called.

– Sites with low confidence observations comprising $\geq 1\%$ of non-missing data. I define a low confidence observation as one with a maximum genotype probability ≤ 0.9. This filter helps to capture sites where it is particularly difficult to confidently call variants, or where a large number of samples happen to have particularly low coverage.

– 'Uncertain' sites. These are sites that I first identified by analysing some of the most significant associations originally produced by the RVS, that did not lie in known IBD loci. In general, I noted a number of sites with low quality scores and a high proportion of individuals with a maximum genotype probability less than one (although not sufficiently low so as to be captured by the low-confidence filter described above). As can be seen in Figure 2.10, these sites have quite different distributions of genotype probabilities compared to high-quality sites. In order to systematically detect such variants, I used the output of five independent Support Vector Machines (SVMs) that were trained on 1,000 high-quality sites that overlapped with the HapMap3 dataset (Altshuler et al., 2010), and 1,000 poor-quality sites with a quality score < 10 in the raw VCF files. Any site with an SVM score < 0.1 in any of the five runs was removed.

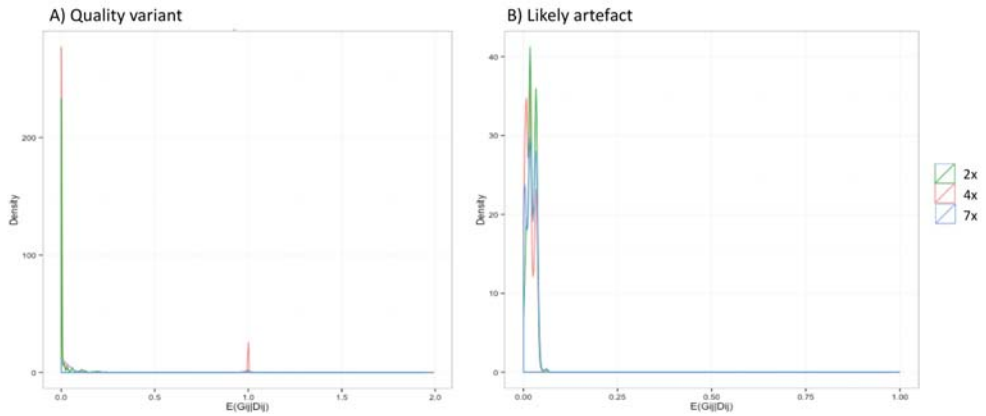

Figure 2.10: An example of a site captured to high quality (panel A), compared to a site with mostly low confidence genotype probabilities (panel B).

Using these additional quality control filters, and unrefined genotype probabilities, I repeated the RVS burden test on rare (0.0001<MAF<0.01) functional coding variation within genes. As can be seen in Figure 2.11, the Type I (false positive) error rate is now properly controlled and no systematic over-inflation of the test statistic is observed (λ=1.06).

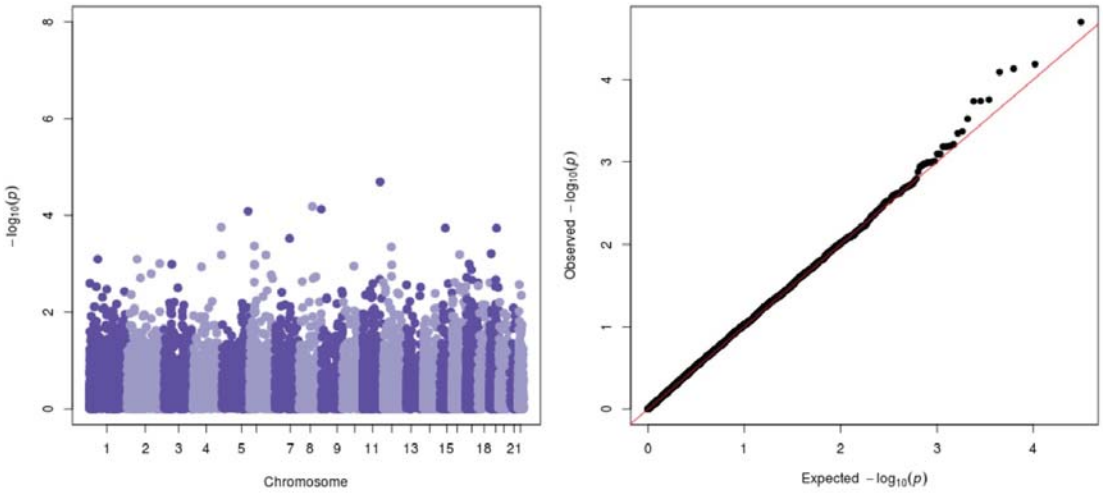

Figure 2.11: The performance of the RVS statistic in a dataset with systematic read depth bias between the cases (4,280 samples at 2-4x coverage) and controls (3,652 samples at 7x).

2.4.5 Increasing the size of the burden test

The logical extension of these gene-based rare variant burden tests is to combine individual tests together into larger, more powerful, gene set tests. However, the RVS statistic is a simple burden test, and does not account for potential differences in the direction of effect of its constituent variants. Within individual genes, one possible way to try and overcome this is to select for variation that is predicted to have a damaging effect on the protein, in the hope that all variation affecting a given gene will therefore act in the same direction. However, for larger gene set tests this is unlikely to help, particularly as previous research has already shown that loss of some genes will lead to an increase in risk, while loss of others will be protective. For example, if we consider just the two most strongly associated genes in IBD, variation in *NOD2* is risk-increasing, while variation in *IL23R* is risk-decreasing.

I therefore extended the RVS statistic to perform larger burden set tests using an enrichment procedure that allows for opposing directions of effect. For each gene (or other form of primary aggregation set, such as enhancers or promoters), the absolute scores are summed together to form an overall score statistic that is independent of effect direction. Overall variances are also summed together, meaning that whilst covariance is included when computing the variance component for an individual gene, the inter-gene covariance is not accounted for. This decision was made in order to greatly reduce the number of between-variant comparisons that were required, which generated massive improvements in the computational efficiency of this method. However, overall I expect the loss of inter-gene covariances to be of minimal consequence. In general, covariance is used to capture the effects of linkage disequilibrium between variants in the test, increasing the overall variance component of the test statistic when highly-correlated variants are present, in order to avoid over-estimating the significance of an association. It is therefore retained for individual gene tests, where all included variation is in very close proximity, but overall it is expected to be relatively small given the rarity of the variants being tested (and therefore their low correlation with other variation in the region). For gene set tests in particular, where many of the contributing genes are not even on the same chromosome, linkage disequilibrium between variants from different genes should be very low.

The resulting set statistic is then divided by the equivalent statistic produced using the set consisting of all genes, in an approach based on the SMP method devised by Purcell et al. (2014). Accounting for the exome-wide statistic in this way helps to remove any residual case-control coverage bias that may accumulate over the large numbers of variants contributing to these gene set tests. Significance is evaluated using permutation testing, where individual gene statistics are re-computed in bootstrapped samples (with the exact same samples drawn for every gene during each permutation round) and summed to produce both set and exome-wide permutation statistics.

2.5 Discussion

Large-scale sequencing studies such as the Exome Aggregation Consortium (Lek et al., 2016), the 1000 Genomes project (1000 Genomes Project Consortium et al., 2015), and the UK10K project (Walter et al., 2015) have revealed important insights into human population biology, and provided vital resources for interpreting the clinical relevance of variation. However, they have also highlighted the practical difficulties associated with combining multi-source sequencing data at scale, as systematic biases in exome capture technology and sequencing depth lead to crucial sensitivity and specificity differences when performing variant calling. As researchers now look to extend the success of these cohort studies to investigate genetic disease risk using large case-control comparisons, the effects of these systematic biases can be observed as a slew of false associations.

In this chapter, I have described various methods that can be used to overcome systematic biases in read depth in a case-control setting, in order to prevent over-inflation of the test statistic and tightly control the Type I error rate. While the effects of sequencing coverage can be largely overcome at sites of low frequency variation, through joint calling of variants followed by genotype refinement, ultimately disease associations for such variants are best tested by imputing them into the wealth of existing GWAS cohorts currently available. Not only does this increase sample size, and therefore power to detect association, but the resulting imputed sequences will not be affected by any of the systematic sequencing biases present in the original cohorts.

For rare variation, which is poorly correlated with nearby variation and therefore cannot be accurately imputed, studies must be performed in the directly sequenced data. As the rare allele for these sites is observed so infrequently in the population, joint calling and genotype refinement offer little power to alleviate the effects of sequencing depth on the sensitivity and specificity of variant calling. Rare variant association testing in the presence of systematic read depth bias between cases and controls therefore required the development of a novel approach that accounts for this bias directly in the association test.

To this end, I implemented the RVS statistic described by Derkach et al. (2014), which adjusts for read depth bias by using genotype dosages (as opposed to hard genotype calls) and calculating the variance component of the test statistic (which is read depth dependent) separately for cases and controls. I then test the performance of this statistic in real data, using cases that had been sequenced at 2-4x average coverage, and controls that were sequenced to 7x. Unfortunately, when using a standard sequencing processing and quality control pipeline, this statistic failed to control the Type I error rate. However, I overcame this problem by reverting to the use of unrefined genotype probabilities, as the genotype refinement process is overzealous when acting upon sites of rare variation, and applying additional quality control filters. Using these adjustments, the number of false positive associations when performing rare variant burden testing across genes can be well controlled, and no systematic over-inflation of the test statistic is observed.

This process has emphasised the difficulties associated with performing large-scale sequencing studies, particularly in a case-control setting. However, I have also shown that, through the use of carefully chosen methods and very stringent quality control, it is possible to perform association testing on this scale even in the presence of systematic read depth bias between cases and controls. This analysis proves that it is feasible for researchers to cost-effectively investigate the role of low frequency and rare variation in genetic disease risk by combining their own sequenced cases with large, publicly-available control datasets.

Chapter 3

The role of rare and low frequency variation in IBD risk

3.1 Introduction

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

Recent reductions in the cost of DNA sequencing means that rare variants may now be captured at scale. In order to maximise sample size, early IBD sequencing studies concentrated on genes in GWAS-implicated loci (Rivas et al., 2011; Beaudoin et al., 2013; Hunt et al., 2013; Prescott et al., 2015), which can logically be extended to

study the entire exome. However, coding variation has been shown to explain at most 20% of the IBD associations uncovered using GWAS (Huang et al., 2015), with the remaining variants lying in non-coding, presumed regulatory, regions of the genome. Low coverage whole genome sequencing has therefore been suggested as a cost-effective approach to capture both coding and non-coding variation in large numbers of samples (Li et al., 2011). This approach is well suited to explore rarer variants than are accessible using GWAS (Cai et al., 2015; Danjou et al., 2015), although the low individual sequencing depth precludes the capture of extremely rare and private mutations.

3.1.1 Chapter overview

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

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

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

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

3.1.2 Contributions

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

3.2 Data preparation

3.2.1 Low coverage whole genome sequencing

Sample ascertainment

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

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

Sequencing and data processing

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

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

3.2.2 Variant calling and imputation improvement

Generating a SNP and indel call set

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

Initial SNP filtering

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

– DP: Raw read depth

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

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

Initial indel filtering

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

- DP: Approximate read depth, after reads with MQ= 255 or bad mates are removed
- FS: Phred-scaled p-value using Fisher's exact test to detect strand bias
- ReadPosRankSum: Z-score from Wilcoxon rank sum test of alternate vs. reference read position bias
- MQRankSum: Z-score from Wilcoxon rank sum test of alternate vs. reference read mapping qualities

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

Genotype refinement

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

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

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

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

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

Copy number variants (CNVs) are usually detected via the identification of localised changes in read depth, an individual read that spans a deletion or insertion breakpoint, or read pairs that map unexpectedly far apart. However, the low average read depth of this particular dataset means that this form of variant detection is not particularly sensitive for individual samples. I therefore called CNVs using GenomeSTRiP 2.0 (Handsaker et al., 2015), which was designed to discover and genotype shared deletions, duplications and multiallelic copy number variants (mCNVs) across whole-genome sequences from multiple individuals. As this study uses low coverage sequences, power to detect variation is limited to larger CNVs. Thus GenomeSTRiP 1.0, which is more sensitive to smaller deletions

and therefore usually recommended as a complementary CNV analysis, was not used for this project.

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

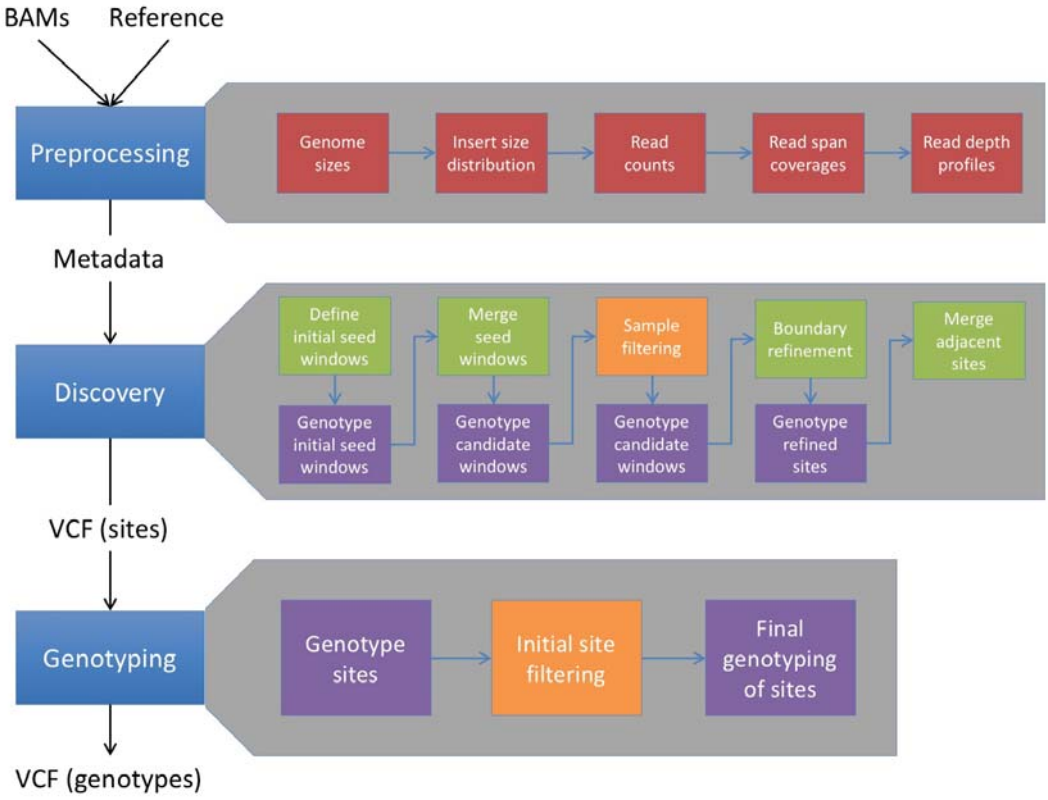

Figure 3.1: Overview of the modular structure employed by GenomeSTRiP 2.0 to discover and genotype CNVs across a number of low coverage whole genome sequences.

Default GenomeSTRiP configurations were used, as per the example configuration files provided within the software releases. Window sizing parameters, which define the size of CNVs that can be detected, matched those used for the 1,000 Genomes Project's low coverage (6-8x) dataset:

```
tilingWindowSize 5000
tilingWindowOverlap 2500
maximumReferenceGapLength 2500
boundaryPrecision 200
minimumRefinedLength 2500
```

Because reads realigned from GrCH37 to hs37d5 using BridgeBuilder did not contain appropriate metadata information for use by GenomeSTRiP 2.0, these reads were excluded from discovery and genotyping.

3.2.3 Quality control

Sample filtering

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

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

Site filtering for SNPs and indels

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

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

Site filtering for copy number variants

Initial CNV filtering was performed in accordance with the default thresholds set in the GenomeSTRiP 2.0 CNVDiscoveryPipeline workflow. These thresholds are generous, and many poor-quality sites are expected to remain: nevertheless, this process removed $86,379$ variants (out of $179,774$) variants from the discovery set, and made manual quality control more manageable. The filters applied at this step include:

- Deletion or mixed CNV length $> 1,000$. Given the search windows used, this still allows variants slightly smaller than those we expect to confidently detect to be included.
- Duplication length $> 2,000$. This follows the recommendations of Handsaker et al. (2015), who note that small duplications appear to have a higher false discovery rate than equivalently sized deletions or mixed CNVs.
- Call rate > 0.9, to remove those variants with excessive missingness.
- Density > 0.5, with density calculated by dividing GSELENGTH (the effective CNV length) by GCLENGTH (the denominator of GC content).
- Cluster separation > 5. This measure checks that appropriate cluster separation was achieved by the Gaussian mixture model used in read depth genotyping.

– GSVDJFRACTION > 0. Remove variants with any evidence of V(D)J recombination, based on the vdjregions.bed file provided with the GenomeSTRiP metadata.

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

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

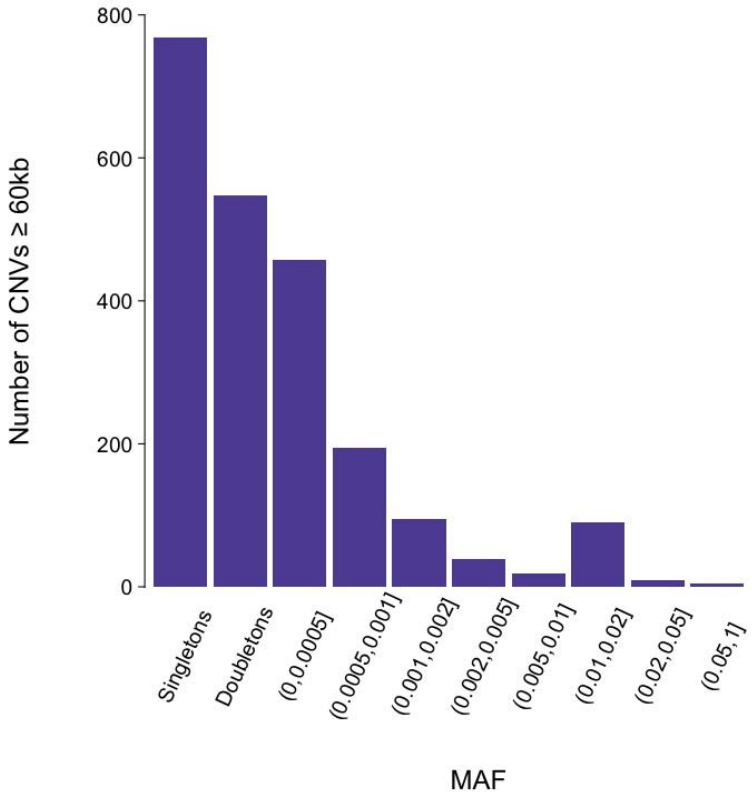

Figure 3.2: Due to a stretch of missing data on chromosome 6 for 95 control samples, there is an apparent excess of large copy number variants with a MAF of 1-2%.

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

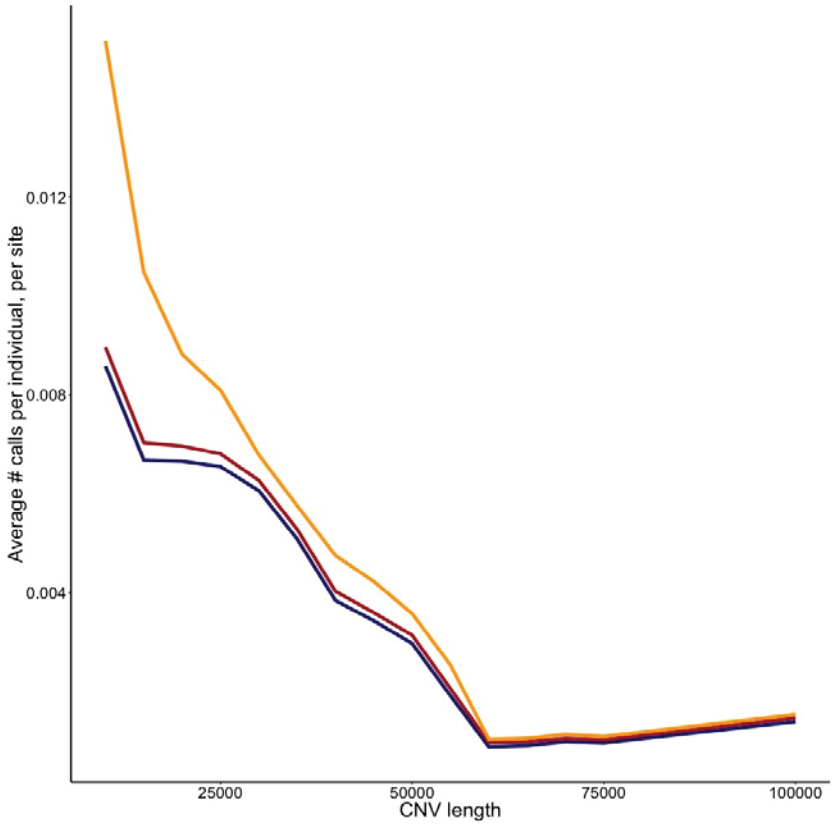

Figure 3.3: The average number of calls per individual per site, across different copy number variant (CNV) lengths. UK10K controls (7x) in yellow, Crohn's disease cases (4x) in red, and ulcerative colitis cases (2x) in blue.

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

3.3 Structural variation

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

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

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

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

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

3.4 Rare variation

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

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

3.4.1 Additional quality control

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

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

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

3.4.2 Burden testing across coding regions

Gene-based burden tests

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

Table 3.2: Variant annotations used to define each of the gene-based burden test subsets.

Annotation	Functional coding	Predicted damaging
frameshift_variant	✓	✓
stop_gained	✓	CADD≥21
initiator_codon_variant	✓	CADD≥21
splice_donor_variant	✓	CADD≥21
splice_acceptor_variant	✓	CADD≥21
missense_variant	✓	CADD≥21
stop_lost	✓	CADD≥21
inframe_deletion	✓	X
inframe_insertion	✓	X

Every test was repeated to independently check for association with CD, UC and IBD at every gene containing one or more relevant variants. This resulted in a total of 100,335 tests, with an average of 5.84 variants contributing to each test (Table 3.3). To correct for this multiple testing, I used a Bonferroni-adjusted threshold for significance of 5×10^{-7}, reflecting an overall alpha value of 0.05. This does not take into account the correlation between the different tests (as the predicted damaging variant set is a direct subset of the functional coding set, and the CD and UC individuals are a subset of the IBD set) and therefore may be too stringent a threshold.

Table 3.3: The number of gene-based burden tests performed for each combination of annotation set and phenotype, with the average number of variants contributing to each of those tests given in parentheses.

Test	Functional coding	Predicted damaging	Total
UC	18,149 (6.83)	14,850 (4.25)	32,999 (5.67)
CD	18,670 (7.42)	15,406 (4.56)	34,076 (6.13)
IBD	18,293 (6.88)	14,967 (4.26)	33,260 (5.70)

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

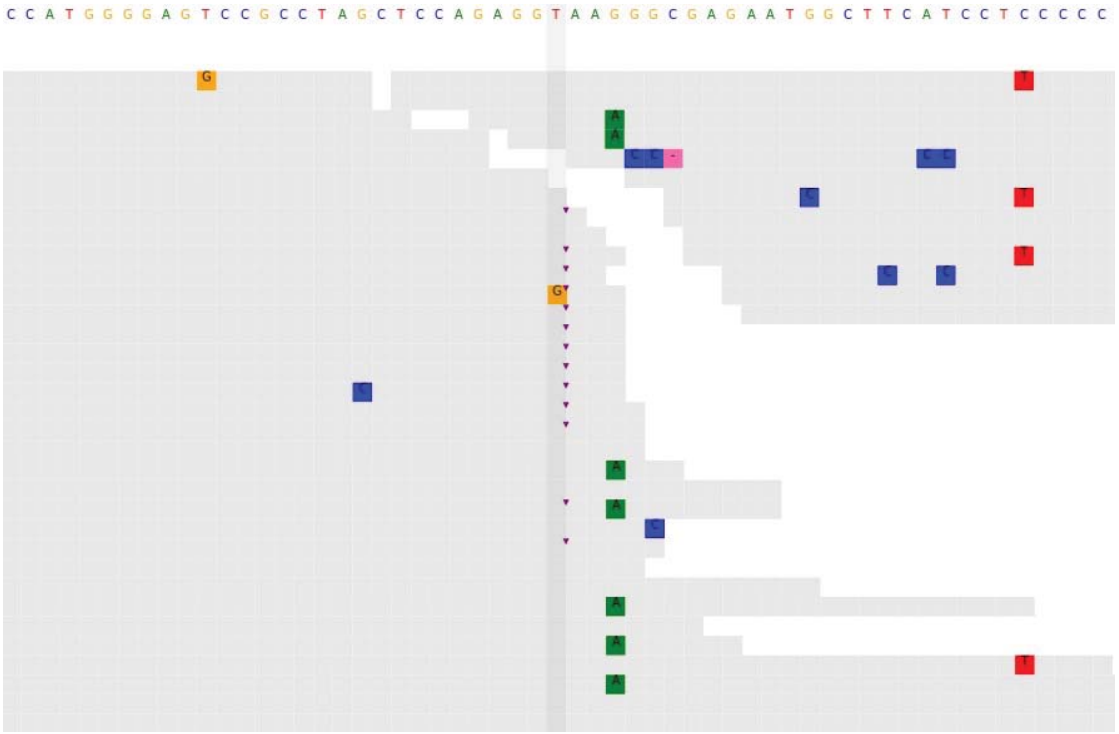

Figure 3.4: Manual inspection of variant calling at nominally associated sites, to identify low quality sites that may have passed the broad quality control thresholds.

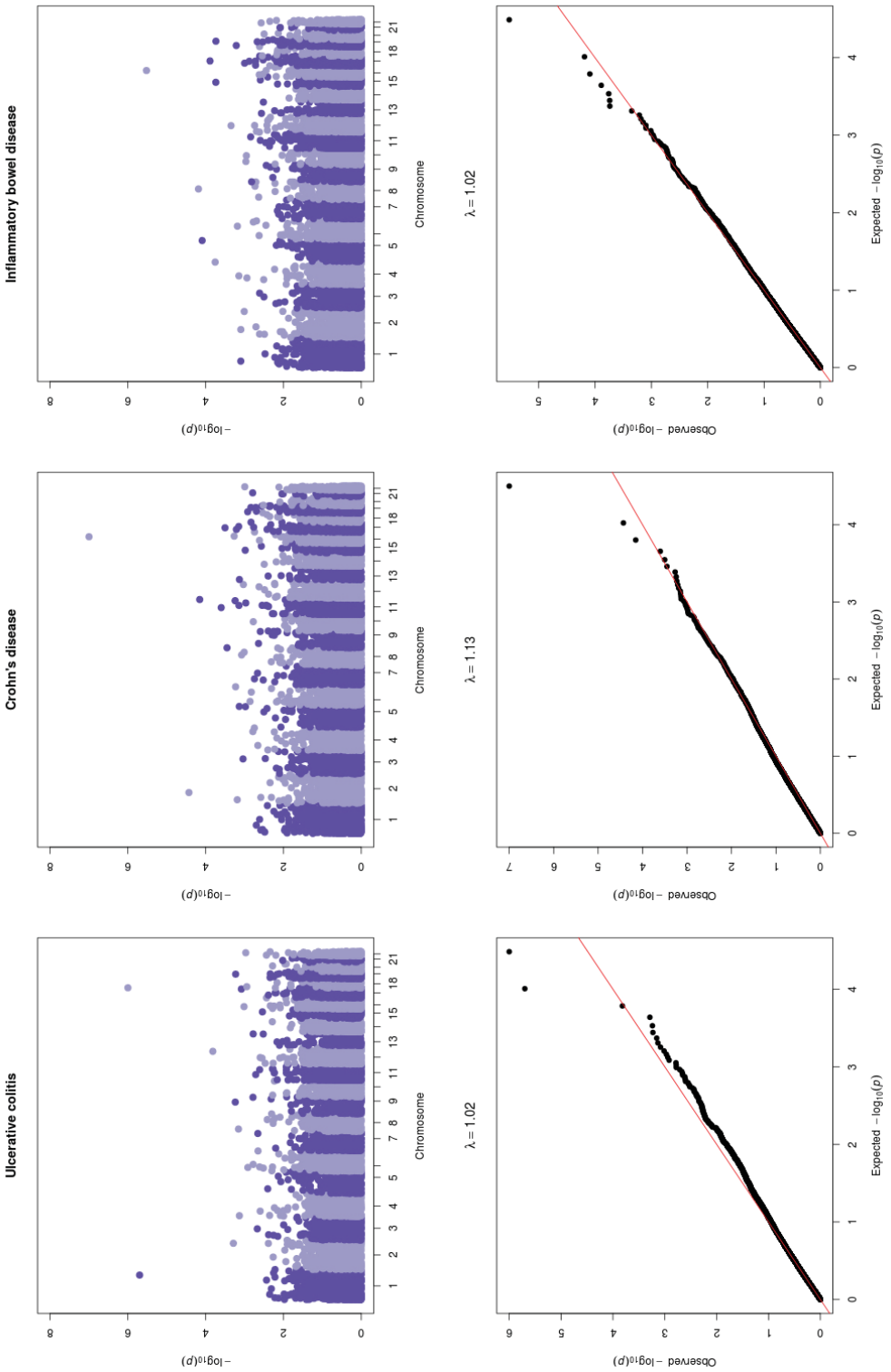

Figure 3.5: Manhattan and QQ plots showing the results of gene-based burden tests using rare, functional coding variation.

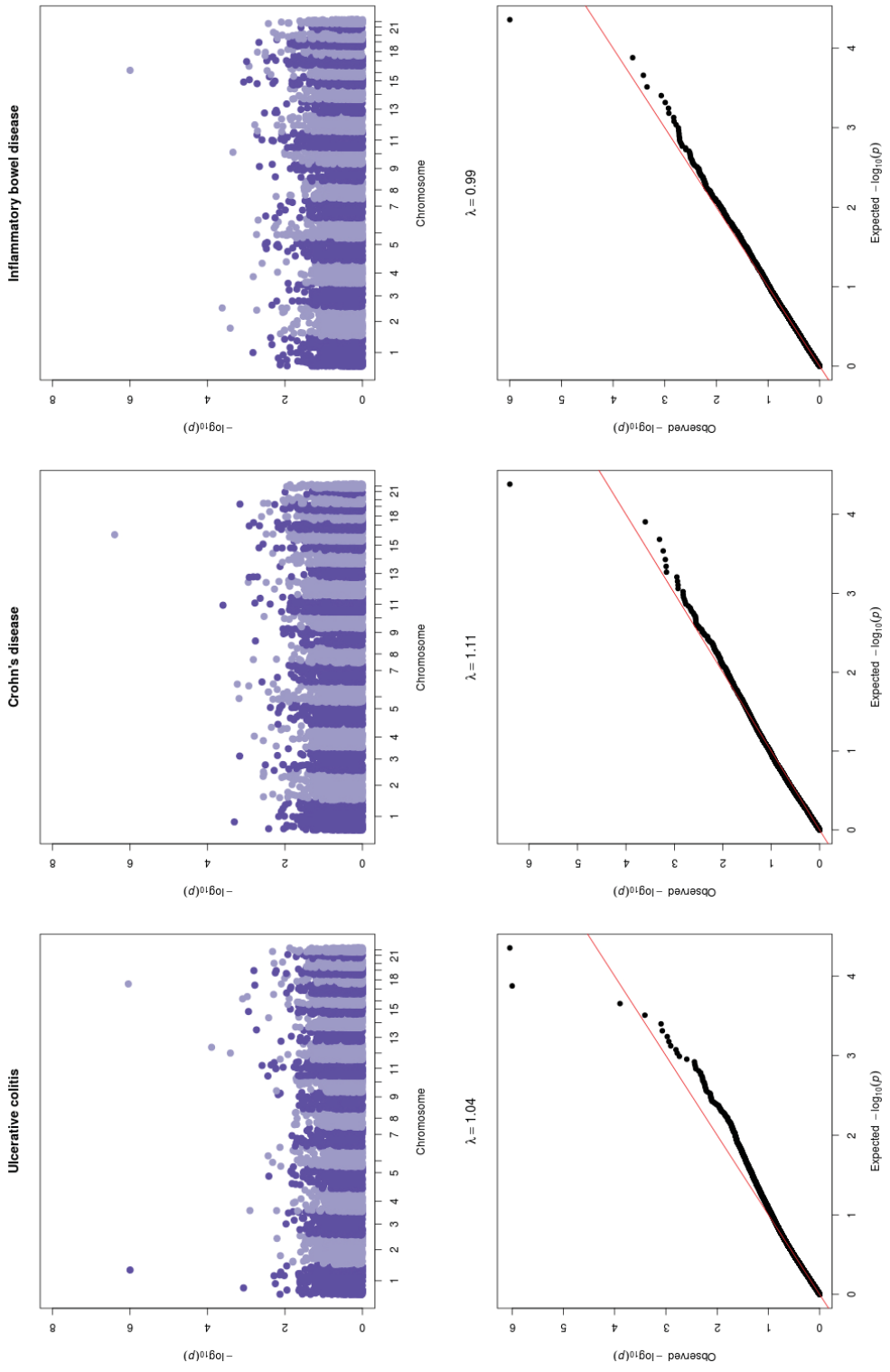

Figure 3.6: Manhattan and QQ plots showing the results of gene-based burden tests using rare, functional coding variation that is predicted to be damaging.

Table 3.4: Genes with a p-value $< 5 \times 10^{-4}$ in the gene-based burden tests. For each gene exceeding this threshold, the BAM files for the three variants with the largest contribution to the overall gene signal were inspected, and any with questionable variant calls were excluded from this table.

Gene Name	Ensembl ID	P value	Phenotype	Annotation set	Effect
NOD2	ENSG00000167207	0.0000001	CD	Functional coding	Risk
NOD2	ENSG00000167207	0.0000004	CD	Predicted damaging	Risk
NOD2	ENSG00000167207	0.000001	IBD	Predicted damaging	Risk
NOD2	ENSG00000167207	0.000003	IBD	Functional coding	Risk
IGKC	ENSG00000211592	0.000037	CD	Functional coding	Risk
WWP1	ENSG00000123124	0.000065	IBD	Functional coding	Protective
VWA5A	ENSG00000110002	0.00007	CD	Functional coding	Risk
CTB-78H18.1	ENSG00000253110	0.000081	IBD	Functional coding	Risk
KRT16	ENSG00000186832	0.000129	IBD	Functional coding	Protective
DCTD	ENSG00000129187	0.000175	IBD	Functional coding	Protective
CADM4	ENSG00000105767	0.000183	IBD	Functional coding	Risk
UGT1A3	ENSG00000243135	0.000239	IBD	Predicted damaging	Risk
LRRC55	ENSG00000183908	0.00025	CD	Functional coding	Risk
LRRC55	ENSG00000183908	0.00025	CD	Predicted damaging	Risk
MYO19	ENSG00000141140	0.000314	CD	Functional coding	Protective
DOCK8	ENSG00000107099	0.000353	CD	Functional coding	Risk
ERBB3	ENSG00000065361	0.000388	UC	Predicted damaging	Protective
SOAT2	ENSG00000167780	0.000448	IBD	Functional coding	Protective
ARHGAP19-SLIT1	ENSG00000269891	0.000453	IBD	Predicted damaging	Risk
IL23R	ENSG00000162594	0.000492	CD	Predicted damaging	Protective

The only gene for which I detected a significant burden of rare variants was *NOD2* ($P_{functional} = 1 \times 10^{-7}$), the well-known Crohn's disease risk gene. To ensure this association was not due to the known low frequency *NOD2* risk variants, I evaluated the independence of the rare variant signal against the common IBD-associated coding variants rs2066844, rs2066845, and rs2066847. Individuals with a minor allele at any of these sites were assigned to one group, and those with reference genotypes to another. Burden testing for this new phenotype produced $P_{functional} = 0.0117$ and $P_{damaging} = 0.7311$. On average, contributing rare variants were at an elevated frequency in non-*NOD2* canonical mutation carriers, compared to those individuals with a minor allele at any of these three sites.

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

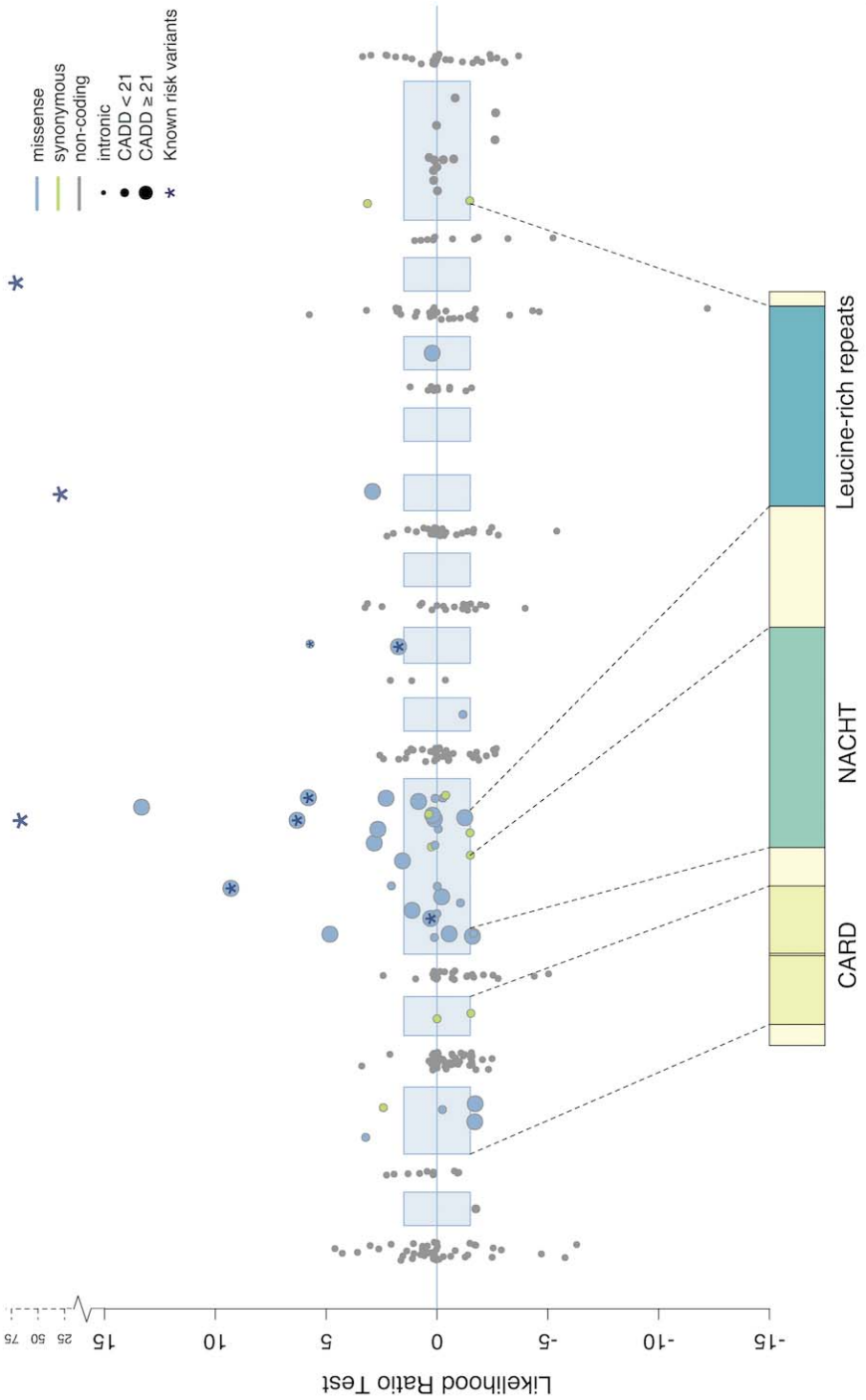

Figure 3.7: Each point represents the contribution of an individual variant to my *NOD2* burden test. Three common variants (rs2066844, rs2066845, rs2066847) are shown, and the six rare variants identified by Rivas et al. (2011) using targeted sequencing are starred. Exonic regions (not to scale) are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

Gene set tests

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

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

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

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

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

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

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

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

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

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

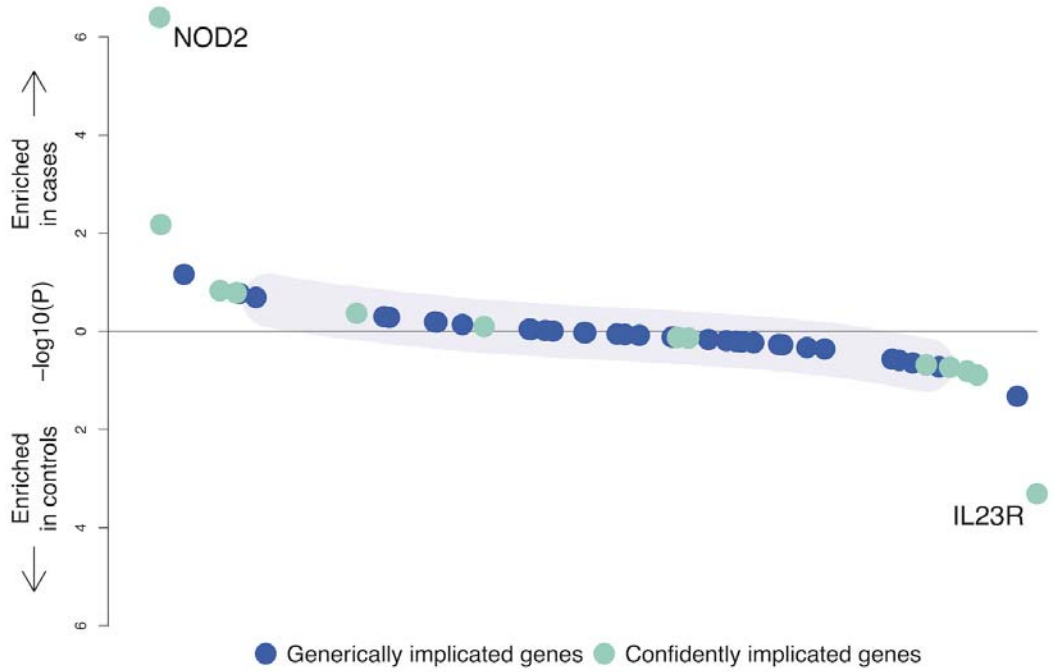

Figure 3.8: The burden of rare damaging variants in Crohn's disease. Each point represents a gene in my confidently implicated (green) or generically implicated (blue) gene sets. Genes are ranked on the x-axis from most enriched in cases to most enriched in controls, and position on the y-axis represents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

3.4.3 Burden testing across non-coding regions

Enhancer-based burden tests

Using the same approach outlined above for individual genes, I evaluated the role of rare (MAF $\leq$ 0.5% in controls) regulatory variation using burden tests across enhancer regions. I consider enhancer regions as defined by the FANTOM5 project (Andersson et al., 2014), which used cap analysis of gene expression (CAGE) to identify enhancer activity through the presence of balanced bidirectional capped transcripts. In particular, I focus my testing on those enhancers that were transcribed at a significant expression level in at least one of the 432 primary cell

or 135 tissue samples tested by the FANTOM5 consortium, which are referred to as 'robust enhancers' by Andersson et al. (2014). The locations of these robust enhancers were downloaded using the `robust_enhancers.bed` track available at http://enhancer.binf.ku.dk/presets/.

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

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

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

I considered all ENCODE transcription factor ChIP-seq motifs (Kheradpour and Kellis, 2014) that had an overall information content ≥ 14 bits (which is equivalent to 7 perfectly conserved positions), and checked if a given variant created or disrupted that motif at a high-information site (IC ≥ 1.8).

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

Table 3.8: The number of enhancer-based burden tests performed for each combination of annotation set and phenotype, with the average number of variants contributing to each of those tests given in parentheses.

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

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

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

However, regardless of the method used, identifying the role of a given enhancer requires testing in the correct cell type and under the correct conditions. For example, Fairfax et al. (2014) discover a number of important immune eQTLs that only occur in monocytes after application of specific stimuli. To try and capture some of this cell-specific expression, studies such as the GTEx consortium are mapping eQTLs across a range of tissues in multiple individuals (GTEx Consortium,

2015), while others are undertaking similar endeavours using Hi-C (Mifsud et al., 2015). As these resources continue to grow, refining of enhancer variant sets to test and interpretation of individual enhancer results may be improved in the future.

Cell- and tissue-specific enhancer set tests

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

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

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

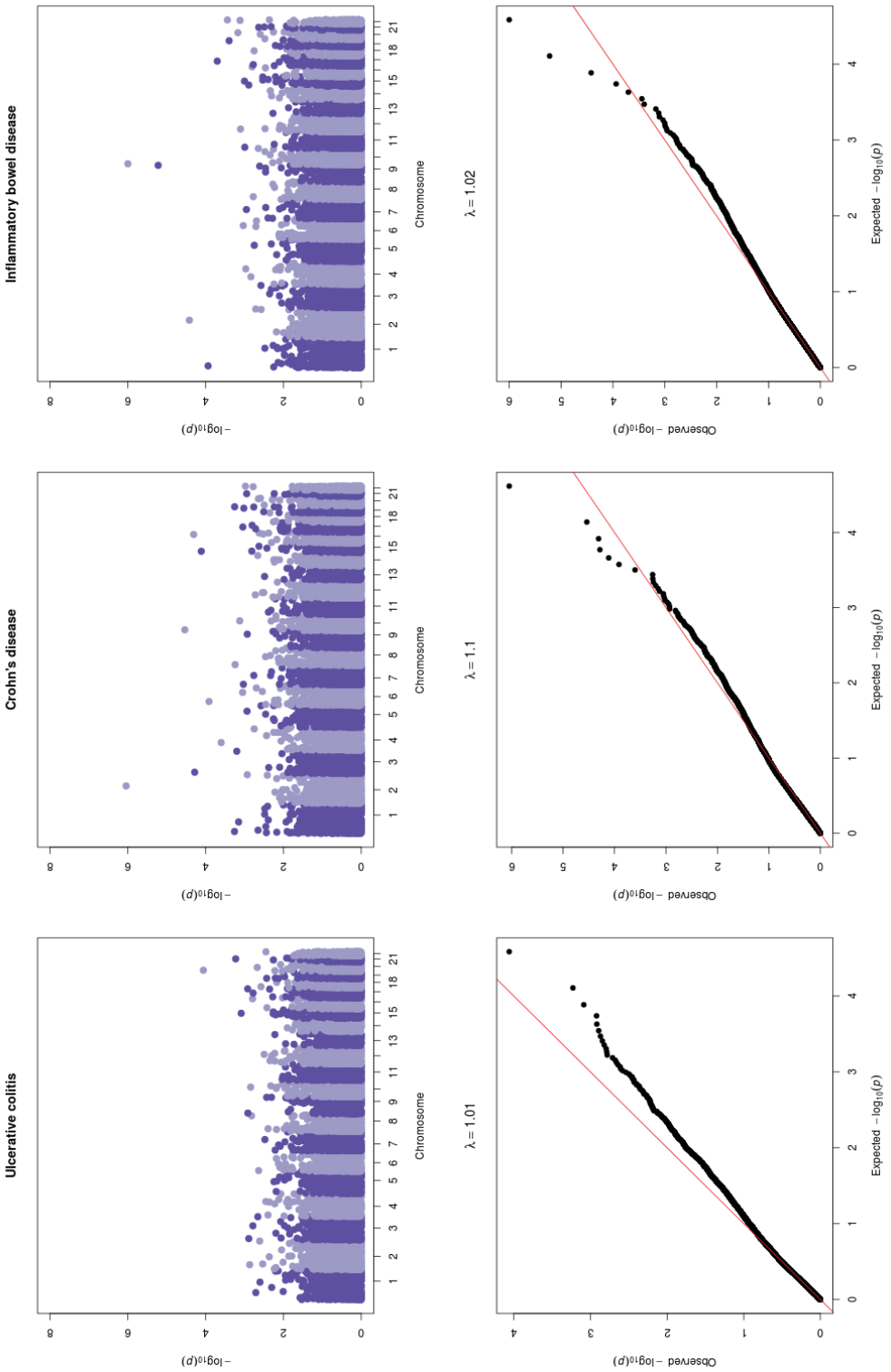

Figure 3.9: Manhattan and QQ plots showing the results of enhancer-based burden tests using all rare variation.

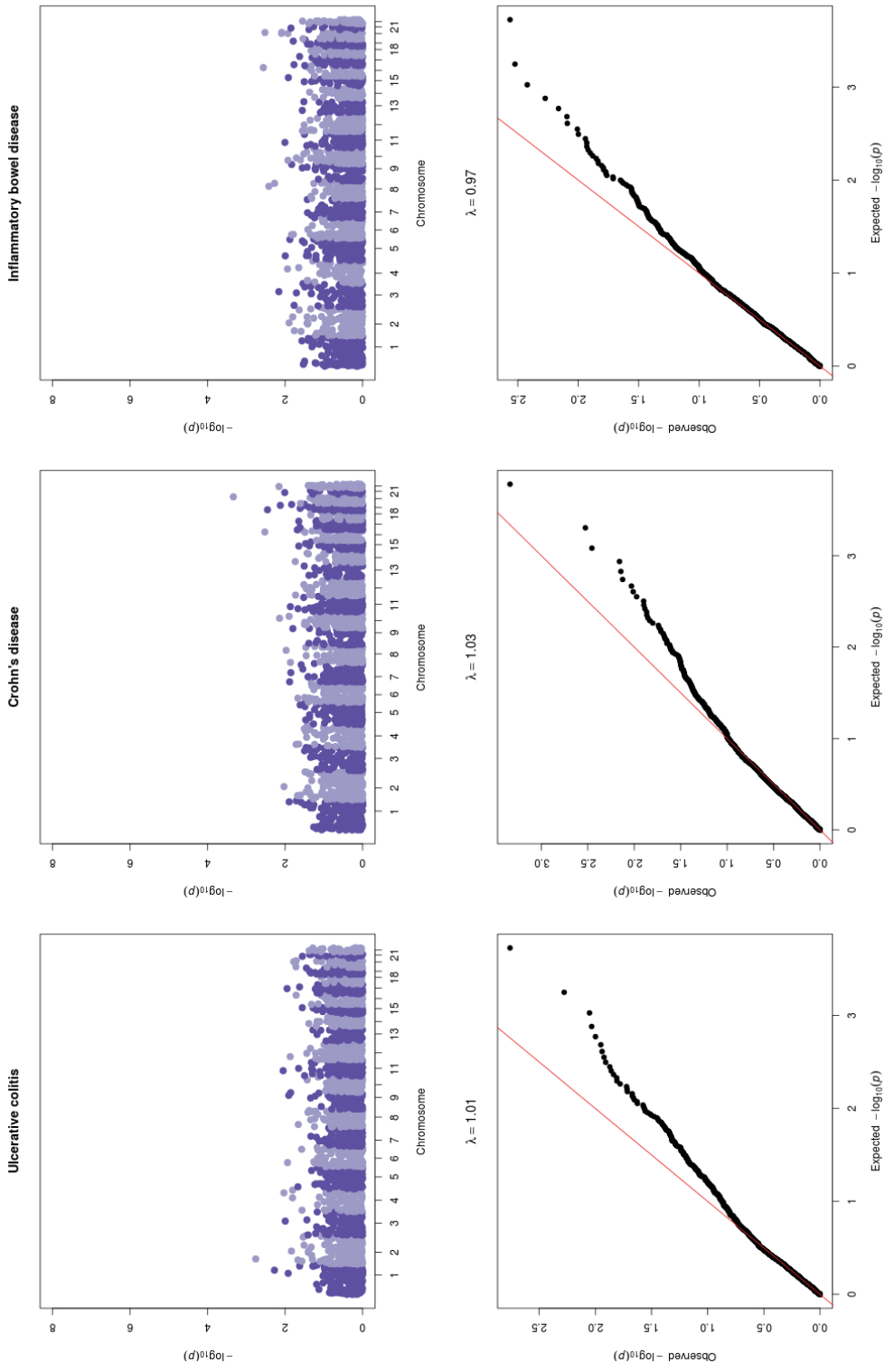

Figure 3.10: Manhattan and QQ plots showing the results of enhancer-based burden tests using rare variation predicted to disrupt or create a transcription factor binding motif. A depletion of very small p-values is observed, possibly due to the low average number of variants contributing to these burden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

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

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

Continued on next page

Table 3.9 – *Continued from previous page*

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

Table 3.10: Enhancer set-based tests with $P < 0.005$. 'TFBM' refers to set tests performed only using rare variants predicted to create or disrupt a transcription factor binding motif, while 'All' includes all rare variants within the relevant enhancer region. No set test reaches significance after multiple correction testing for the 660 tests performed.

Cell/tissue type	*P*-value	Disease	Annotation	# enhancers	# variants
skeletal muscle tissue	0.00058	CD	All	67	222
skeletal muscle tissue	0.00068	IBD	All	61	188
skeletal muscle cell	0.00253	IBD	TFBM	293	397
melanocyte	0.0039	CD	All	379	1,241
stromal cell	0.00398	IBD	TFBM	272	401
cardiac fibroblast	0.00425	UC	TFBM	192	278

3.5 Low frequency variation

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

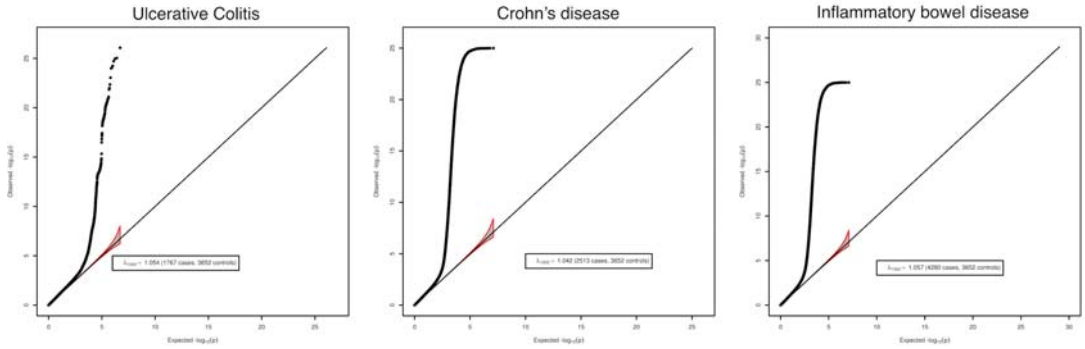

Figure 3.11: QQ plots of genome-wide association studies for variants with MAF $\geq 0.1\%$ in the sequencing dataset. λ_{1000} values are reported for the ulcerative colitis, Crohn's disease and inflammatory bowel disease analyses. Grey shapes show 95% confidence intervals. Figures produced by Yang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

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

3.5.1 Imputation into GWAS

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

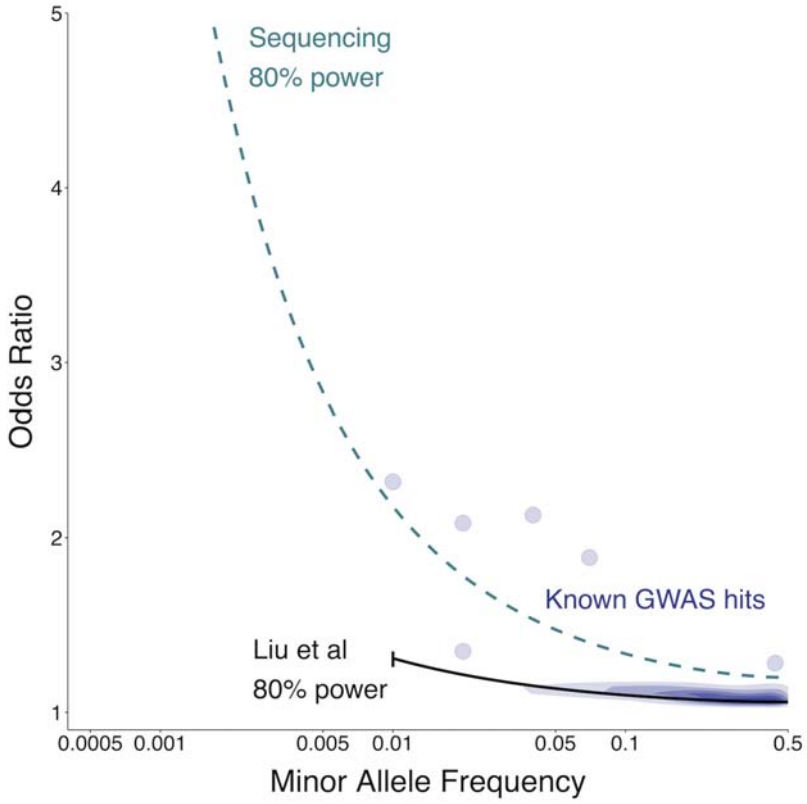

Figure 3.12: Relative power of this study compared to previous GWAS. The black line shows the path through frequency-odds ratio space where the latest International IBD Genetics Consortium (IIBDGC) meta-analysis (Liu et al., 2015) had 80% power, and the green line shows the same for this study. The earlier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

We therefore sought to increase power by using imputation to leverage both new and existing IBD GWAS. As previous data has shown that expanded reference panels can significantly improve the imputation accuracy of low frequency variants (McCarthy et al., 2016), we built a phased reference panel of 10,971 individuals

by combining our low coverage whole genomes with the 1000 Genomes Phase 3 haplotypes (1000 Genomes Project Consortium et al., 2015).

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

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

Table 3.11: Sample counts of the imputed GWAS cohorts.

Cohort	Case	Control	Total
WTCCC1	1,206	2,918	4,124
WTCCC2	1,921	2,776	4,697
GWAS3_CD	4,264	9,495	13,759
GWAS3_UC	4,072	9,495	13,567
GWAS3_IBD	8,860	9,495	18,355
Sequencing_CD	2,513	3,652	6,165
Sequencing_UC	1,767	3,652	5,419
Sequencing_IBD	4,280	3,652	7,932
Total	16,267	18,841	35,108

3.5.2 Quality control and association testing

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

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

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

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

these had MAF > 10%, so were carried forward to a meta-analysis of our data and published IBD GWAS summary statistics as will be discussed in Chapter 4.

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

The fourth new association ($P = 9\times10^{-12}$) was a 0.6% missense variant (p.Asp439Glu, rs78534766) in *ADCY7* that doubles risk of ulcerative colitis (OR=2.19, 95% CI =1.75-2.74), and is strongly predicted to alter protein function (SIFT=0, PolyPhen=1, MutationTaster=1). This variant was associated ($P = 1 \times 10^{-6}$) in a subset of directly genotyped individuals, suggesting the signal was unlikely to be driven by imputation errors. However, to further validate this finding we obtained two replication cohorts:

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

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

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

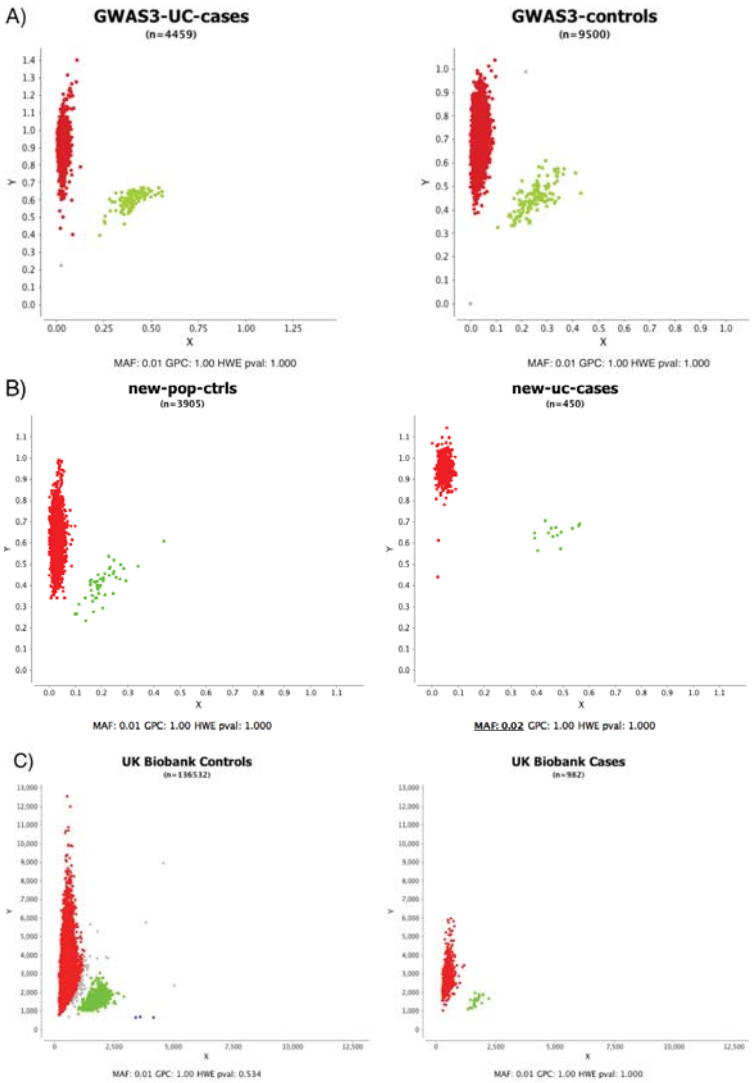

Figure 3.13: Cluster plots are shown for rs78534766 (chr16:50335074, *ADCY7* p.Asp439Glu) for the A) new UK IBD GWAS, B) replication and C) UK Biobank samples that passed quality control. The SNP genotypes have been assigned based on cluster formation in scatter plots of normalized allele intensities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

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

Cohort	Cases	Controls	OR [95% CI]	*P*-value	MAF (controls)	Method	Info	Missingness (cases/controls)	P_{het}
WTCCC2	1,921	2,918	2.62 [1.63-4.22]	7.03×10^{-05}	0.0061	Imputed	0.82	N/A	
GWAS3	4,072	9,495	2.05 [1.53-2.75]	1.43×10^{-06}	0.0065	Genotyped	N/A	0.00025/0.0024	
Sequencing	1,767	3,652	2.14 [1.27-3.60]	0.0042	0.0060	Sequenced	0.88	N/A	
All discovery	7,760	16,065	2.19 [1.75-2.74]	9.20×10^{-12}		(Meta-analysis)			0.69
UK Biobank	982	136,464	1.70 [1.18-2.44]	0.0189	0.0061	Genotyped	N/A	0.0000/0.0004	
Replication	450	3,905	4.10 [1.76-9.51]	0.0009	0.0069	Genotyped	N/A	0.0000/0.0044	
All directly genotyped	5,504	149,864	2.06 [1.63-2.60]	1.62×10^{-09}		(Meta-analysis)			0.19
All cohorts	13,264	165,929	2.16 [1.77-2.62]	1.17×10^{-14}		(Meta-analysis)			0.39

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

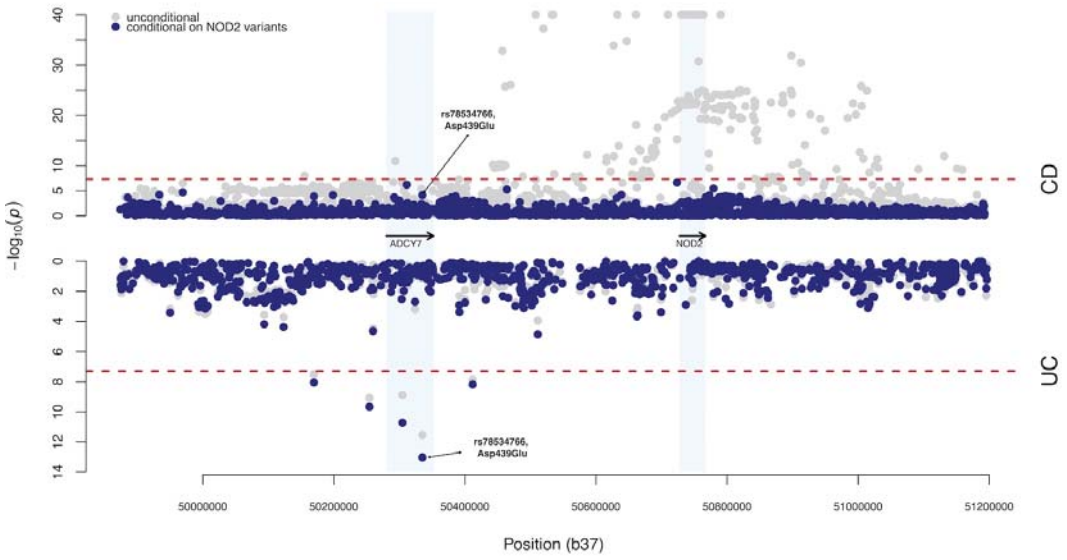

Figure 3.14: Association analysis for the *NOD2*/*ADCY7* region of chromosome 16. Results from the single variant association analysis are presented in gray, and results after conditioning on seven known *NOD2* risk variants in blue. Results for Crohn's disease (CD) are shown in the top half, and ulcerative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

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

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

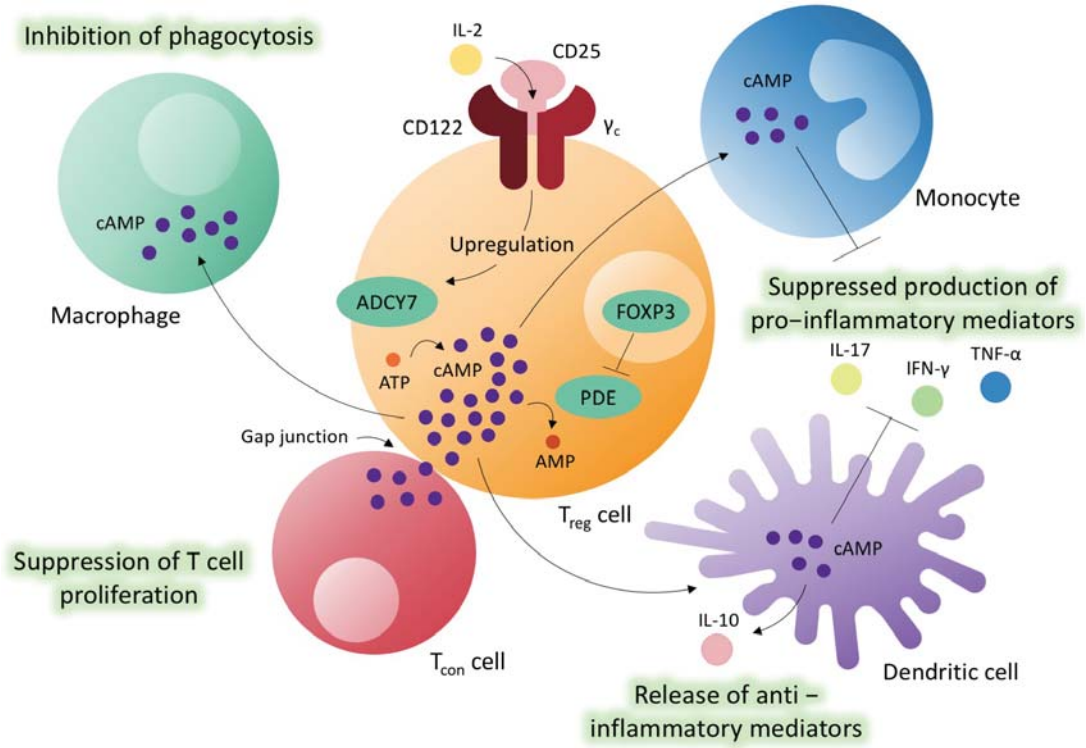

Figure 3.15: An overview of the role of ADCY7 in the inflammatory response, where it is responsible for the conversion of ATP to cAMP in haemopoietic cells. A subset of the immune-related functions performed by the secondary signalling molecule cAMP are depicted here (Rossi et al., 1998; Tiemessen et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

3.6 Discussion

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

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

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

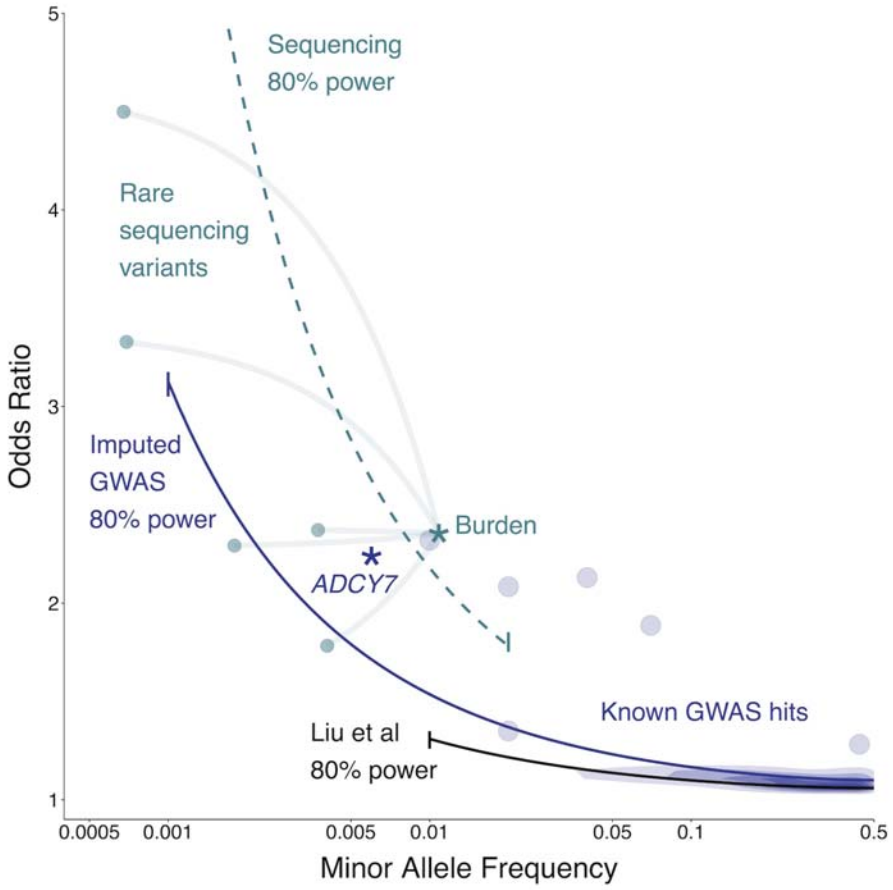

Figure 3.16: The frequency-odds ratio space investigated by this study, comparing the latest IIBDGC meta-analysis (black line) to the sequencing (green) and imputed GWAS (purple) used in this study. The earlier study had more samples but restricted their analysis to MAF > 1%. Purple density and points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

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

The challenge of detecting a burden of variation in regulatory regions is further compounded by our current inability to clearly distinguish likely functional variation

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

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

Therefore, if sequence-based rare variant association studies are to be as successful as common variant GWAS, computationally efficient methods and accepted standards for combining these novel datasets need to be developed. An example of one such effort is the Haplotype Reference Consortium (HRC), which has collected whole genome sequences from more than 32,000 individuals (including the IBD samples discussed here) in order to create a reference panel that can be used for imputation of low frequency and common variants (McCarthy et al., 2016). Imputation into GWAS using this large HRC panel is as accurate as low-coverage sequencing down to MAF $\sim$0.05% (McCarthy et al., 2016), suggesting that in the future the most effective way to discover low frequency variants associated with complex disease will be to impute the huge resources of existing GWAS data with large new reference

panels. Thus, while projects such as this one provide valuable resources in the form of publicly available reference panels, it is unlikely that there will be much need for low coverage whole genome sequencing in the future. Together, our results suggest that a combination of continued GWAS imputed using substantial new reference panels and large scale deep sequencing projects will be required in order to fully understand the genetic basis of complex diseases like IBD.

Chapter 4

Uncovering the biological mechanisms driving association

4.1 Introduction

Next-generation sequencing represents a powerful tool for analysing the contribution of rare variation to a range of disorders, and is currently enjoying rapid growth in popularity as we usher in the so-called 'sequencing era'. But does this advance in technology mean the end of genotyping?

For low frequency and common variation, new discoveries are more likely to arise from continuing to increase sample sizes using cost-effective genotyping arrays. Indeed, this approach has proven very successful at identifying genetic risk loci for IBD. To date, 215 associated loci have been uncovered using genome-wide association studies (GWAS) and targeted follow-up using the Immunochip. However, the utility of performing these ever-larger genome-wide association studies in order to identify common variation of relatively small effect sizes has been questioned. In particular, it is notable that just 20 of these 215 IBD-associated loci have been narrowed down to a causal gene, and to date the increased biological understanding from genetic studies has not yet had a substantial impact on disease therapies.

However, recent methodological and technological advances offer the opportunity to derive more therapeutically-relevant information from these genome-wide association studies. This includes novel fine-mapping techniques that can better resolve a given association signal down to a likely causal variant, and improved statistical co-localization methods that can associate a GWAS signal with an expression quantitative trait locus (eQTL) from a variety of cell types and conditions. Such improvements, coupled with rapidly expanding databases of eQTLs and other functional annotations, may prove to be the important missing links required in order to unravel the biological mechanisms underlying many GWAS associations.

4.1.1 Chapter overview

In this chapter, I conduct a new genome-wide association study of inflammatory bowel disease in 18,355 individuals from the United Kingdom. I then meta-analyse these data with the whole genome sequences described in Chapter 3 and published GWAS summary statistics, yielding a total sample size of 59,957 subjects. This leads to the identification of 25 new IBD susceptibility loci, which are then evaluated to try to resolve the potential biological mechanisms underlying each association.

Likely causal missense variants are identified in the genes *SLAMF8*, a negative regulator of inflammation, and *PLCG2*, a gene that has been implicated in primary immune deficiency. A potentially causal variant is also observed in an intron of *NCF4*, which is another gene associated with an immune-related Mendelian disorder. In general, a significant enrichment of genes associated with Mendelian disorders of inflammation and immunity is observed for all 241 IBD-associated loci.

In addition, three novel loci lie proximal to integrin genes, which encode proteins in pathways that have been identified as important therapeutic targets in IBD. Co-localization with eQTL signals confirm that the associated IBD risk-increasing variants are also correlated with expression changes in monocytes in response to immune stimulus at two of these genes (*ITGA4* and *ITGB8*), and at two previously implicated loci (*ITGAL* and *ICAM1*). Overall, we note that new associations at common variants continue to identify genes that are relevant to therapeutic target identification and prioritization.

4.1.2 Contributions

This study was conceived and designed by the UK IBD Genetics Consortium (UKIBDGC), with case ascertainment, phenotyping and sample collection performed by the numerous clinics that contribute to this effort: please see Appendix A for a full list of contributors. DNA sample preparation and genotyping was performed by the Wellcome Trust Sanger Institute pipelines facility. Imputation of GWAS datasets using an IBD-specific reference panel was performed in collaboration with Shane McCarthy; quality control, LD score regression and conditional analysis of the resulting meta-analysis was performed by Loukas Moutsianas. Principal components were generated by Carl Anderson. Overlap with existing eQTL datasets was evaluated by Sun-Gou Ji. Fine-mapping and eQTL co-localization testing was run by Luke Jostins-Dean, but I analysed the output. Disease localisation analysis of variation in *NCF4* was performed by Jeffrey Barrett. Identification of therapeutically-relevant genes and pathways, and evaluation of the biological significance of novel findings was done in discussion with James Lee, Christopher Lamb and Nick Kennedy. Unless stated, I carried out all other analyses.

4.2 Data preparation

4.2.1 A new UK IBD genome wide association study

Sample ascertainment and genotyping

Following ethical approval by Cambridge MREC (reference: 03/5/012), 11,768 British IBD cases, diagnosed using accepted endoscopic, histopathological and radiological criteria, were consented into a new study by the UK IBD Genetics Consortium. These samples consisted of 5,695 Crohn's disease cases, 5,299 ulcerative colitis cases, and 764 inflammatory bowel disease cases of indeterminate type. In parallel, 10,484 controls were obtained by the UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. Both cases and controls were genotyped at the Wellcome Trust Sanger Institute; controls on the Human Core Exome v12.0 chip, and cases on the Human Core Exome v12.1 chip.

Genotype calling

I called genotypes for this dataset using the software optiCall (Shah et al., 2012), run in five separate batches (four case batches, and a single control batch) to reflect the groupings by which samples were processed in the laboratory. Called genotypes were then strand aligned using files provided by William Rayner (http://www.well.ox.ac.uk/~wrayner/strand/). I removed any sites not included on both versions of the chip, leaving a total of 535,434 genotyped sites.

Sample filtering

Prior to sample quality control, sites were pruned to remove those with a missingness rate in excess of 5%. Individuals failing on one or more of the following filtering criteria were then removed from the dataset:

– Mismatching gender between that listed in the manifest, and that determined genetically. Genders were determined using PLINK v1.9 (Chang et al., 2015), which computes the inbreeding coefficient F based on data from the X chromosome. Under Hardy-Weinberg equilibrium, females should have an X-chromosome F coefficient close to zero, while for males it should be close to one.

– Heterozygosity rate ±3 standard deviations from the mean (Figure 4.1).

– Missingness rate > 1% (Figure 4.1).

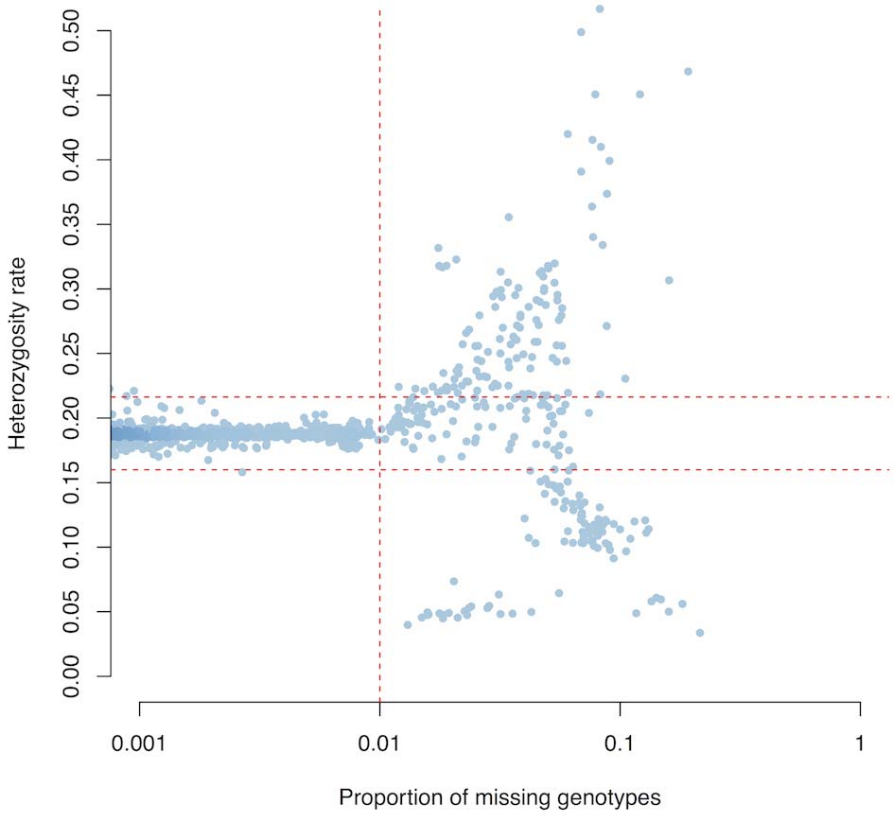

Figure 4.1: Missingness versus heterozygosity rate for samples in the new UK IBD GWAS. Samples falling outside of the dotted lines (missingness > 1% and heterozygosity rate ±3 standard deviations from the mean) were removed from the analysis. Script for figure generation available from Anderson et al. (2010).

– Duplicated or related individuals with a kinship coefficient > 0.177 (indicating first-degree relatives or closer). Kinship coefficients were calculated for samples passing the heterozygosity and missingness checks, using markers with a MAF > 0.05 and the software KING (Manichaikul et al., 2010). The sample with the lowest call rate (or mismatching gender, if applicable) of each related pair was removed.

– Non-European samples, as determined using a principal component analysis (Figure 4.2) incorporating samples from the HapMap3 project (Altshuler et al., 2010).

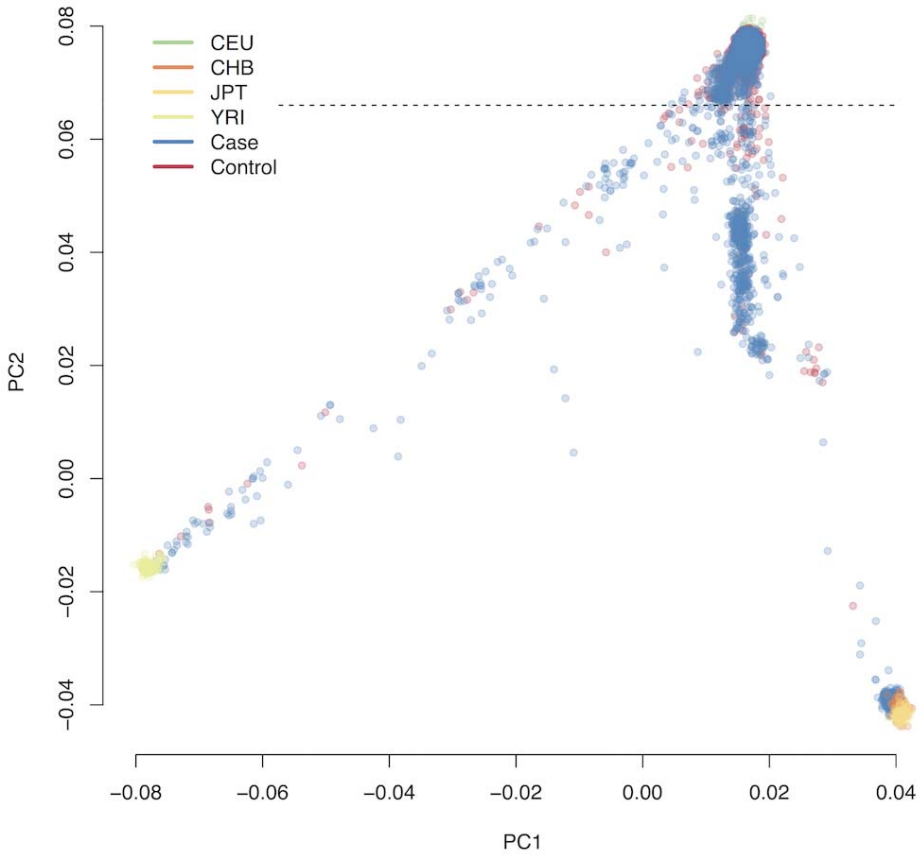

Figure 4.2: Principal component analysis of samples in the new UK IBD GWAS, analysed jointly with samples from the HapMap3 project (Altshuler et al., 2010). Samples with PC2 ≤ 0.066 (dotted line) were considered to be of non-European ancestry.

Site filtering

A final set of quality control filters were then used to remove markers still performing poorly amongst the high-quality samples, as determined by:

- Significant difference ($P < 1 \times 10^{-5}$) in call rate between cases and controls
- Evidence for a deviation from Hardy-Weinberg equilibrium in controls, where the p-value $< 1 \times 10^{-5}$
- One of 429 markers affected by a genotyping batch effect. These sites were identified by Yang Luo by computing within-sample principal components (PCs) using common variants (MAF $> 1\%$), which highlighted a clear outlier group of case samples all belonging to one genotyping batch (Figure 4.3a). PC1 was used to split cases into outliers and non-outliers, and an association test between these groups identified significant sites ($P < 1 \times 10^{-5}$). Once these sites were removed, the within-sample PCs no longer produced any outlier groups (Figure 4.3b).

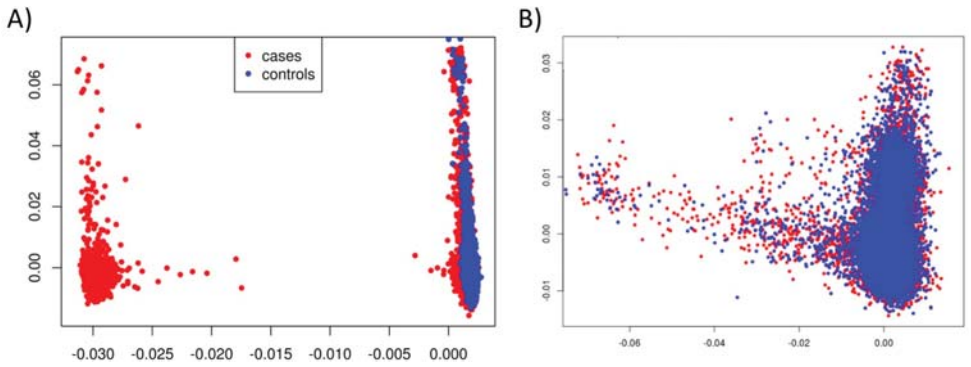

Figure 4.3: Panel A) depicts a genotyping batch effect identified in the new UK IBD GWAS using a principal component analysis, while panel B) shows the improvement after the removal of 429 sites that were significant when comparing the outlier samples from A) against the remaining samples. Figures generated by Yang Luo.

This left a high-quality dataset consisting of 510,520 genotyped sites in 9,239 cases (4,474 CD, 4,173 UC, 592 indeterminate IBD), and 9,500 controls. Before imputation, these sites were further pruned to those with a MAF $> 0.1\%$, leaving a total of 296,203 markers.

4.2.2 Imputation using an IBD-specific reference panel

Previous data has suggested that increasing the size of the reference panel used during imputation can significantly improve the accuracy of imputed low frequency variants (McCarthy et al., 2016), Therefore, as mentioned in Chapter 3, we created an expanded imputation reference panel, consisting of 4,686 low coverage IBD sequences collected by the UKIBDGC (retaining those individuals that were excluded from association analyses due to non-European ancestry), combined with 3,781 UK10K and 2,504 1000 Genomes Phase 3 control sequences. The inclusion of IBD samples helps to enrich the resulting reference panel with IBD-associated variants.

Prior to imputation, I remove any genotyped samples that were already included in the UKIBDGC low coverage sequencing study, as these would be present in the reference panel. I also remove any samples also included in the Wellcome Trust Case Control Consortium datasets (The Wellcome Trust Case Control Consortium, 2007; Barrett et al., 2009), as these samples contributed to the latest International IBD Genetics Consortium (IIBDGC) study that I shall be meta-analysing with this dataset. This left a total of 18,355 samples (4,264 Crohn's disease, 4,072 ulcerative colitis, 524 indeterminate inflammatory bowel disease, and 9,495 controls).

We then imputed whole genome sequences, down to a MAF $\sim 0.1\%$. Given the large size of both the reference and genotype panel, the computationally efficient software PBWT (Durbin, 2014) was used in order to obtain results in a tractable amount of time.

4.2.3 Meta-analysis of sequencing and imputed genomes with existing summary statistics

I tested these imputed sequences separately for association to ulcerative colitis, Crohn's disease and IBD using SNPTEST v2.5 (Marchini and Howie, 2010), performing an additive frequentist association test conditioned on the first ten principal components for each cohort. I then filtered out variants with MAF $< 0.1\%$,

INFO < 0.4, or strong evidence for deviations from Hardy-Weinberg equilibrium in controls ($P < 1 \times 10^{-7}$).

In order to increase power for the analysis of common variation, I obtained the publicly available summary statistics from the latest IIBDGC meta-analysis (Liu et al., 2015), and applied the same MAF $\geq 0.1\%$ and INFO ≥ 0.4 filters. I then used METAL (Willer et al., 2010) to perform a standard error weighted meta-analysis of the summary statistics from the UKIBDGC sequencing and imputed GWAS datasets together with the IIBDGC GWAS data.

4.2.4 Quality control

We filtered the output of this meta-analysis, removing sites with high evidence for heterogeneity ($I^2 > 0.90$) in any of the cohorts, and a meta-analysis p-value higher than all of the cohort-specific p-values. After this quality control, overall inflation of the summary statistics was still observed ($\lambda_{GC} = 1.23$ and 1.29 for Crohn's disease and ulcerative colitis, respectively). To determine if this was due to confounding population substructure that had not been properly accounted for, LD score regression was applied using LDSC v1 (Bulik-Sullivan et al., 2015) and European linkage disequilibrium (LD) scores from the 1000 Genomes Project (downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2) on all sites with INFO > 0.95. Both intercepts were 1.09, indicating that the observed inflation is more likely to be due to broad polygenic signal.

In total, we tested 9.7 million high-quality sites across 25,042 IBD cases and 34,915 controls (Table 4.1), the largest genome-wide association test performed in inflammatory bowel disease to date. This dataset therefore offers us the opportunity to not only uncover further common variant IBD associations of small effect size, but also gives us reasonable power to perform causal variant fine-mapping in IBD-associated loci that were not covered by the Immunochip genotyping array used by Huang et al. (2015).

Table 4.1: Sample numbers and variant counts are described for each contributing dataset, at each stage of the analysis (Total = raw numbers, QC+ = post quality control, and N-o = after removing overlapping samples). Numbers are described separately for the ulcerative colitis, Crohn's disease, and inflammatory bowel disease analyses.

Study	Data	CD			UC			IBD			Controls			Grand Total		
		Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o
UK sequences	Samples	2697	2513	1974	1817	1767	1326	4514	4280	3300	3910	3652	3650	8424	7932	6950
	Sites	95.1M	13.2M		95.1M	13.2M		95.1M	13.2M		95.1M	13.2M		95.1M	13.2M	
New UK GWAS	Samples	5695	4474	4264	5299	4173	4072	11758	9239	8860	10484	9500	9495	22242	18739	18355
	Genotyped	535k	296k		535k	296k		535k	296k		535k	296k		535k	296k	
	Imputed	109.5M	19.0M		109.5M	18.9M		109.5M	19.0M		109.5M	19.0M		109.5M	19.0M	
IIBDGC CD GWAS	Samples		5956	5956								14927	14927		20883	20883
	Sites	12.3M	12.0M								12.3M	12.0M		12.3M	12.0M	
IIBDGC UC GWAS	Samples					6968	6968					20464	20464		27432	27432
	Sites				12.3M	12.1M					12.3M	12.1M		12.3M	12.1M	
IIBDGC IBD GWAS	Samples								12882	12882		21770	21770		34652	34652
	Sites							12.7M	12.5M		12.7M	12.5M		12.7M	12.5M	
Meta-analysis (CD/UC/IBD)	Samples	12194			12366			25042			28072/ 33609/ 34915			40266/ 45975/ 59957		
	Sites	20.8M	9.6M		20.8M	9.6M		20.9M	9.7M		9.6M/ 9.6M/ 9.7M			9.6M/ 9.6M/ 9.7M		

4.3 Unravelling common variant associations

Overall, we identified 25 new IBD-associated loci at genome-wide significance (Table 4.2), including a number of associations of very small effect (OR < 1.1). In order to uncover causal variants, genes and mechanisms amongst these new associations, we performed a range of fine-mapping, eQTL co-localization, and gene enrichment tests as discussed in the following sections.

4.3.1 Fine-mapping and functional annotation of new and known loci

We performed a summary statistics fine-mapping analysis on the 25 novel IBD-associated loci, together with 40 previously discovered loci that reached genome-wide significance in this dataset but where fine-mapping had not previously been attempted. To do this, approximate Bayes factors were calculated from the meta-analysis effect sizes and standard errors, assuming the SNPTEST default prior variance on the log odds ratio of 0.04. These Bayes factors were then fine-mapped using the method outlined by the Wellcome Trust Case Control Consortium et al. (2012), to generate a posterior probability for each variant that reflects its likelihood of being causal in a given locus. The credible set for an association signal is defined as the smallest set of variants with posteriors that sum to at least 95%.

In order to properly resolve a GWAS signal down to the causal variant(s), it is important that all common SNPs in the locus have been directly genotyped or imputed to high quality (Spain and Barrett, 2015). This is to ensure that the truly causal SNPs are actually included in the fine-mapping comparison, when determining the relative evidence for causality of each associated SNP in the region. Therefore, to be confident about the conclusions drawn from this fine-mapping procedure, I only considered loci which had high quality imputed data for all relevant variants. This is defined as having no variants in the Phase 3 v5 release of the 1000 Genomes project (2013-05-02 sequence freeze) that are in high LD ($r^2 \geq 0.6$) with our hit SNP, but missing from our dataset, and no variants in our data within high LD ($r^2 > 0.8$) that fail during our QC procedure.

Table 4.2: Twenty five novel IBD-associated loci identified via a meta-analysis of 25,042 cases and 34,915 controls. The locus boundaries are defined by the left- and right-most variants that have an r^2 of 0.6 or more with the main variant.'RAF' refers to the risk allele frequency in the 1000 Genomes CEU and GBR populations.

Rsid	Chr	Position	Locus (Mb)	Risk Allele	Non-risk Allele	RAF	P_{Meta}	OR	95% CI	Phenotype	Implicated gene
rs34687326	1	159799910	159.80-159.80	G	A	0.900	1.06×10^{-08}	1.18	1.12-1.24	CD	*SLAMF8*
rs59043219	1	209970610	209.97-210.02	A	G	0.379	1.09×10^{-08}	1.08	1.05-1.10	IBD	-
rs6740847	2	182308352	182.31-182.33	A	G	0.508	1.22×10^{-13}	1.10	1.07-1.12	IBD	*ITGA4*
rs144344067	2	187576378	187.50-187.68	A	AT	0.895	1.29×10^{-08}	1.12	1.08-1.16	IBD	-
rs1811711	2	228670476	228.67-228.67	C	G	0.826	6.09×10^{-09}	1.14	1.10-1.18	UC	-
rs76527535	2	242484701	242.47-242.49	C	T	0.745	2.87×10^{-08}	1.09	1.06-1.12	IBD	-
rs2581828	3	53133149	53.10-53.17	C	G	0.597	6.46×10^{-09}	1.10	1.07-1.13	CD	-
rs2593855	3	71175495	71.16-71.19	C	T	0.663	2.54×10^{-09}	1.09	1.06-1.11	IBD	-
rs503734	3	101023748	100.91-101.27	A	G	0.513	2.67×10^{-08}	1.07	1.05-1.10	IBD	-
rs56116661	3	188401160	188.40-188.40	C	T	0.795	5.67×10^{-10}	1.14	1.10-1.18	CD	-
rs11734570	4	38588453	38.58-38.59	A	G	0.368	4.80×10^{-08}	1.07	1.05-1.10	IBD	-
rs17656349	5	149605994	149.59-149.63	T	C	0.466	1.54×10^{-08}	1.09	1.06-1.13	UC	-
rs113986290	6	19781009	19.72-19.83	C	T	0.989	7.59×10^{-09}	1.36	1.25-1.46	UC	-
rs67289879	6	42007403	42.00-42.01	T	C	0.179	3.04×10^{-08}	1.09	1.06-1.13	IBD	-

Continued on next page

Table 4.2 – *Continued from previous page*

Rsid	Chr	Position	Locus (Mb)	Risk Allele	Non-risk Allele	RAF	P_{Meta}	OR	95% CI	Phenotype	Implicated gene
rs11768365	7	6545188	6.50-6.55	A	G	0.816	3.88×10^{-08}	1.09	1.06-1.12	IBD	-
rs149169037	7	20577298	20.58-20.58	G	A	0.895	3.26×10^{-08}	1.14	1.10-1.19	IBD	*ITGB8*
rs243505	7	148435339	148.40-148.58	A	G	0.624	3.04×10^{-10}	1.08	1.06-1.11	IBD	-
rs7911117	10	27179596	27.16-27.18	T	G	0.871	1.84×10^{-08}	1.14	1.10-1.19	UC	-
rs111456533	10	126439381	126.32-126.55	G	A	0.829	1.18×10^{-09}	1.11	1.08-1.14	IBD	-
rs80244186	13	42917861	42.84-42.94	C	T	0.111	3.66×10^{-08}	1.13	1.09-1.18	CD	-
rs11548656	16	81916912	81.91-81.92	A	G	0.961	5.18×10^{-11}	1.27	1.20-1.34	IBD	*PLCG2*
rs10492862	16	82867456	82.87-82.92	A	C	0.308	1.26×10^{-09}	1.11	1.08-1.15	CD	-
rs4256018	20	6093889	6.08-6.10	G	T	0.250	1.23×10^{-08}	1.08	1.05-1.11	IBD	-
rs138788	22	35729721	35.72-35.74	A	G	0.418	2.95×10^{-08}	1.09	1.06-1.13	UC	-
rs4821544	22	37258503	37.26-37.26	C	T	0.321	1.76×10^{-08}	1.10	1.07-1.13	CD	-

Because of the relative sparsity with which the genome-wide microarrays cover each region (as opposed to dense genotyping arrays, such as the Immunochip), only 12 loci pass this filtering step. For 6 of these, there exists a single variant with $> 50\%$ probability of being causal (Table 4.3). For those implicated variants that were directly genotyped in the new UKIBDGC GWAS dataset, the cluster plots were manually checked to confirm quality data (Figure 4.4).

Of particular interest are two loci where a single variant had >99% probability of being causal. The first causally implicated variant, rs34687326, is a missense change predicted to affect protein function in *SLAMF8* (p.Gly99Ser, Figure 4.5a). As can be seen in Figure 4.5a, this signal was relatively easy to resolve given the low linkage disequilibrium between this lead SNP and the surrounding variation. While this sparse Manhattan plot was initially concerning, we were reassured by the very clean cluster plots produced by direct genotyping of the variant rs34687326 in the new UKIBDGC GWAS dataset (Figure 4.4a), increasing our confidence that this is a true association.

SLAMF8 is a cell surface receptor expressed by various myeloid cells (including neutrophils, macrophages and dendritic cells) after exposure to gram- or gram+ bacteria, lipopolysaccaride (LPS) or interferon (IFN)-γ, where it has been reported to inhibit the migration of these cells to sites of inflammation (Wang et al., 2015). In addition, SLAMF8 has been shown to play a role in repressing the production of reactive oxygen species (ROS) by these cells, further negatively regulating inflammatory responses (Wang et al., 2012). The risk-decreasing allele in our dataset (MAF=0.1, Table 4.2) is predicted to strongly affect protein function (CADD=32, 92nd percentile of missense variants, Kircher et al. (2014)), suggesting that further experimental follow up to evaluate a possible gain-of-function mechanism may be worthwhile.

Table 4.3: Variants fine-mapped to > 50% probability of being causal in their given signal.

Rsid	Chr	Position	P_{Causal}	Effect	Credible set size	Phenotype	P_{Meta}	Locus type
rs34687326	1	159799910	1.000	*SLAMF8* p.Gly99Ser (missense)	1	CD	1.06×10^{-08}	Novel
rs4845604	1	151801680	0.999	*RORC* (intronic)	1	IBD	7.09×10^{-14}	Known
rs1811711	2	228670476	0.914		2	UC	6.09×10^{-09}	Novel
rs56116661	3	188401160	0.561	*LPP* (intronic)	11	CD	5.67×10^{-10}	Novel
rs11548656	16	81916912	0.502	*PLCG2* p.His244Arg (missense)	3	IBD	5.18×10^{-11}	Novel
rs1143687	16	81922813	0.746	*PLCG2* p.Arg268Trp (missense)	5	IBD	3.83×10^{-08}	Novel
rs4821544	22	37258503	0.804	*NCF4* (intronic)	2	CD	1.76×10^{-08}	Novel

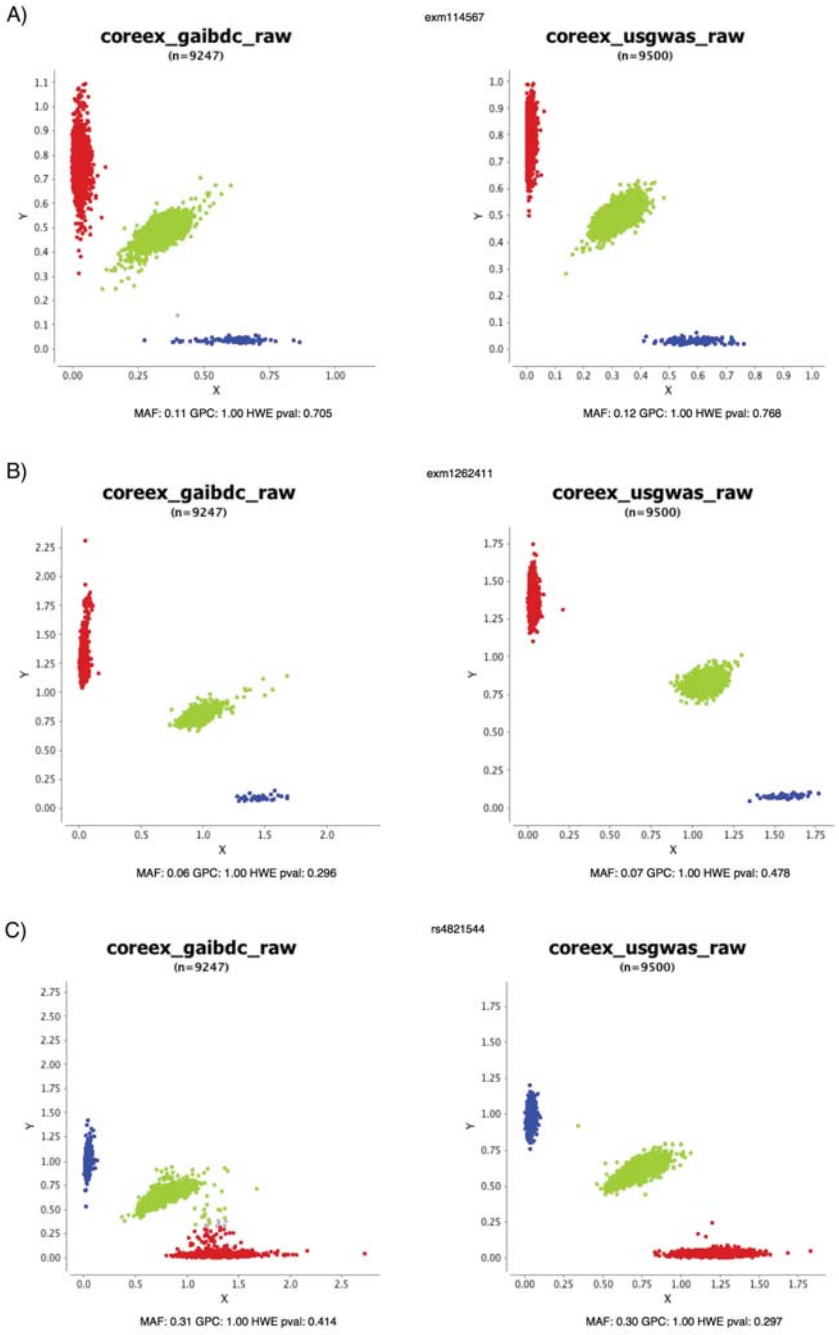

Figure 4.4: Cluster plots for A) rs34687326, B) rs1143687 and C) rs4821544 for the new UK IBD GWAS samples that passed quality control. The SNP genotypes have been assigned based on cluster formation in scatter plots of normalized allele intensities X and Y. Each circle represents one individual's genotype. Blue and red clouds indicate homozygote genotypes for the SNP (CC/AA), green heterozygote (CA) and grey undetermined. Figures generated by Daniel Rice.

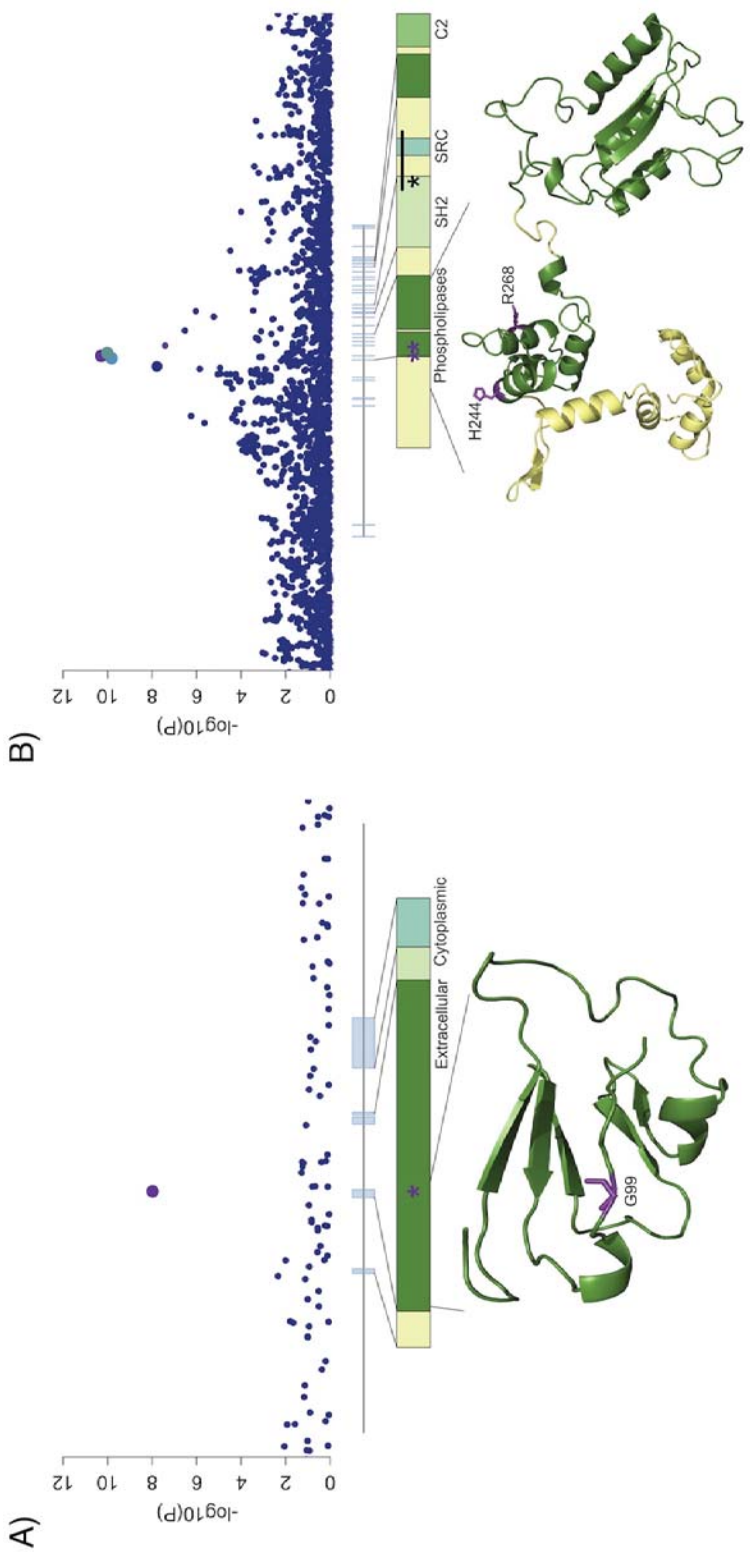

Figure 4.5: Likely causal missense variants in *SLAMF8* and *PLCG2*. For A) *SLAMF8* and B) *PLCG2*, local association results are plotted, with point size corresponding to LD to our lead variant and color to fine-mapping probability (purple $> 50\%$, intermediate blue $10 - 50\%$, navy blue $< 10\%$). Gene body diagrams and protein domain annotations are taken from ENSEMBL, and partial predicted crystal structures for both proteins are obtained from the SWISS-MODEL repository. Our fine-mapped variants (purple) and known variants leading to autosomal dominant immune disorders (black) are marked on the gene body diagrams.

The second variant with >99% probability of being causal is rs4845604, an intronic variant in the key regulator of T_H17 cell differentiation, *RORC* (Ivanov et al., 2006). *RORC* encodes RORγt, which is the master transcriptional regulator of both T_H17 cells (Ivanov et al., 2006) and group 3 innate lymphoid cells (Luci et al., 2009). These cell types both play important roles in defense at mucosal surfaces: in particular, they have been shown to help maintain homeostasis between the intestinal immune system and the gut microbiota (Yang et al., 2014c; Sawa et al., 2011). Loss of this equilibrium is often seen in inflammatory bowel disease (Gevers et al., 2014). Furthermore, pharmacologic inhibition of RORγt has been shown to be of therapeutic benefit in mouse models of intestinal inflammation, and reduces the frequency of T_H17 (but not innate lymphoid) cells isolated from primary intestinal samples of patients with inflammatory bowel disease (Withers et al., 2016).

Also of note is another likely functional variant amongst the remaining, less clearly resolved, fine-mapped loci (Table 4.3). This missense variant (CADD=16.5, 50.2% probability of causality) affects the gene *PLCG2* (Figure 4.5b). Interestingly, after conditioning on this variant we observe a second, independent missense variant in the same gene ($P = 2 \times 10^{-8}$), that is highly likely to affect protein function (CADD=34.0, 74.6% probability of causality). *PLCG2* encodes a phospholipase enzyme that plays an important role in regulating immune pathway signalling and T cell selection (Fu et al., 2012). It has also been implicated in two autosomal dominant immune disorders: intragenic deletions in the autoinhibitory domain of *PLCG2* cause antibody deficiency and immune dysregulation (familial cold autoinflammatory syndrome 3, MIM 614468), while heterozygous missense variants (e.g. p.Ser707Tyr) lead to a phenotype that includes intestinal inflammation (Ombrello et al., 2012; Zhou et al., 2012).

4.3.2 Enrichment amongst IBD loci for genes associated with Mendelian disorders of inflammation and immunity

An association is also observed between Crohn's disease and an intronic variant in *NCF4* (P=1.76×10^{-8}, 80.4% probability of causality), a gene which has also been associated with a Mendelian disorder of inflammation and immunity. In particular, *NCF4* encodes p40phox, part of the NADPH-oxidase system that destroys phagocytosed bacteria via an oxidative burst in innate immune cells (Tarazona-Santos et al., 2013). Rare pathogenic variants in *NCF4* cause autosomal recessive chronic granulomatous disease, which is characterized by intestinal inflammation and defective ROS production in neutrophils (Matute et al., 2009). Interestingly, the variant associated in our dataset, rs4821544, had previously been suggestively associated with small bowel Crohn's disease (Rioux et al., 2007; Roberts et al., 2008). When we stratified patients by disease location we found that the effect was consistently stronger for ileal disease (affecting the small bowel) compared to colonic (affecting the large bowel), as shown in Figure 4.6. This is consistent with growing genetic evidence that Crohn's disease may in fact be better defined as two distinct subtypes, ileal Crohn's disease and colonic Crohn's disease (Cleynen et al., 2016).

In order to test whether these observations in *PLCG2* and *NCF4* reflected a more general overlap between candidate IBD GWAS genes and Mendelian disorders of inflammation and immunity, I performed a gene set enrichment analysis. I defined the set of Mendelian disorder genes of interest as being those associated with primary immune deficiences according to the latest curated release by the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency (Picard et al., 2015), as well as a secondary list of genes associated with rare disorders in OMIM that include inflammatory bowel disease as a clinical diagnostic. The secondary genes were obtained using a clinical synopsis search in OMIM (https://www.omim.org/search/advanced/clinicalSynopsis, as accessed on Sep 08, 2016) for the terms "Inflammatory bowel disease", "Crohn's disease" and "Ulcerative colitis", restricting the output to results where the molecular basis

has been identified. This list was then manually curated to exclude those entries corresponding to complex disorders.

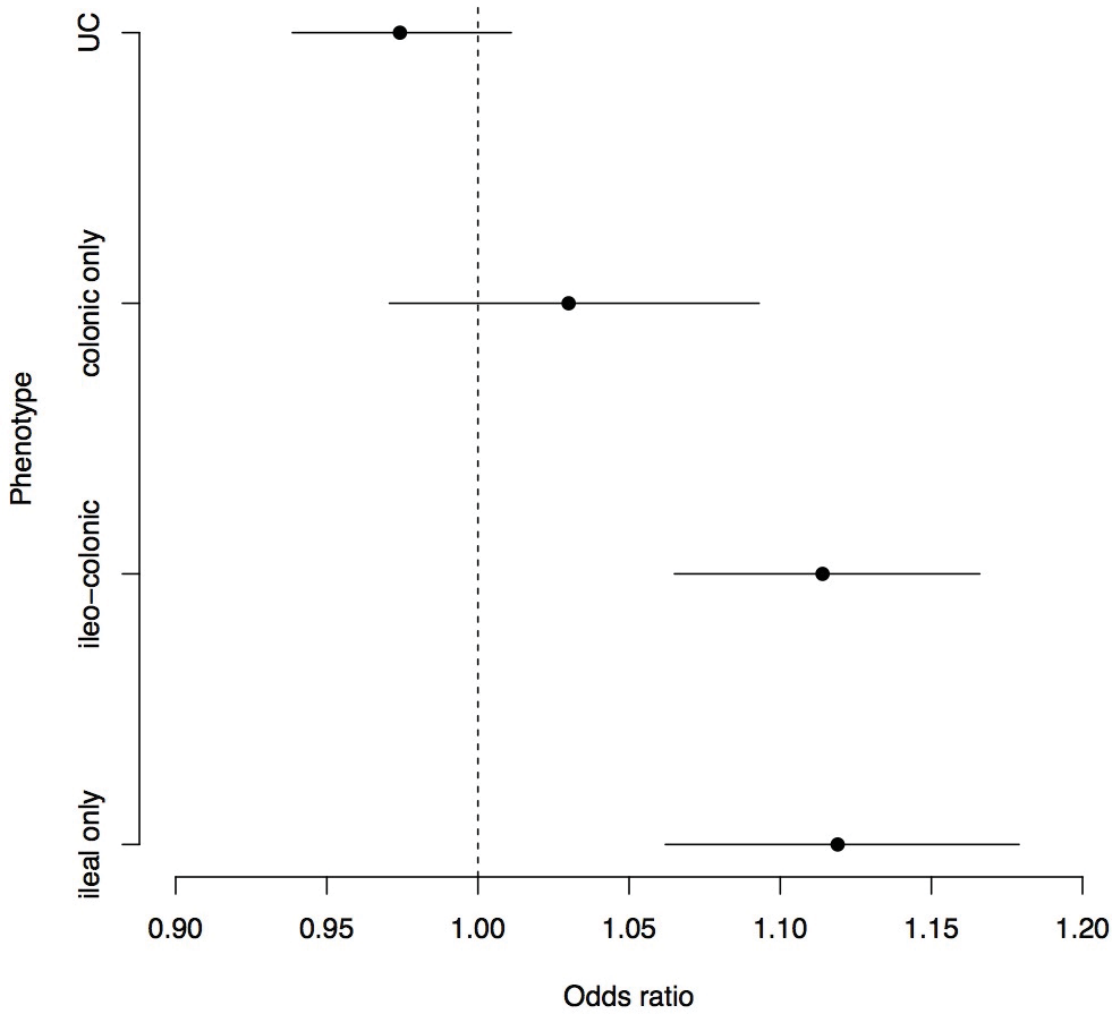

Figure 4.6: The effect of the well fine-mapped variant rs4821544, which is intronic in *NCF4*, is consistently stronger in small bowel compared to large bowel disease. Figure produced by Jeffrey Barrett.

Using the software INRICH (Lee et al., 2012a), I observe a significant enrichment ($P < 1 \times 10^{-6}$) of these genes within all 241 loci now associated with IBD (Appendix B). I then refine this test to just the 26 loci where a gene can be confidently implicated by fine-mapping to a coding variant or co-localization with an eQTL (Huang et al., 2015; Fairfax et al., 2014; Wright et al., 2014), as described in Table 4.4 and Chapter 3. Within the set of loci with a high-confidence gene assignment, the enrichment for genes associated with Mendelian disorders of inflammation and immunity is even stronger (27% vs 3%, $P = 2 \times 10^{-5}$ using a Fisher's exact test).

Table 4.4: Association of known IBD genes with Mendelian disorders of inflammation and immunity. These disorders include Primary Immune Deficiencies as defined by Picard et al. (2015), and Mendelian disorders which include IBD as a symptom, according to OMIM.

Gene	Phenotype	Primary Immune Deficiency	Additional rare disorders
CARD9	IBD	CARD9 deficiency	-
IFIH1	UC	Aicardi-Goutieres syndrome 7	Singleton-Merten syndrome 1
IL2RA	CD	CD25 deficiency	-
NOD2	CD	Blau syndrome	Early-onset sarcoidosis
PLCG2	IBD	PLAID (PLCγ2 associated antibody deficiency and immune dysregulation); Familial cold autoinflammatory syndrome 3; APLAID (autoinflammation and PLAID)	-
SMAD3	IBD	-	Loeys-Dietz syndrome 3

Remaining known IBD genes without an associated Mendelian disorder:
ADCY7 (UC), *ATG16L1* (CD), *CD6* (CD), *ERAP2* (CD), *FCGR2A* (IBD), *FUT2* (CD), *ICAM1* (IBD), *IL18RAP* (IBD), *IL23R* (IBD), *ITGA4* (IBD), *ITGAL* (UC), *ITGB8* (IBD), *MST1* (IBD), *NXPE1* (UC), *PTPN22* (CD), *SLAMF8* (CD), *SP140* (CD) and *TNFSF8* (IBD)

4.3.3 Co-localization of GWAS and eQTL associations

Among the remaining 21 novel loci, it was interesting to observe that three associations were within 150kb of integrin genes (*ITGA4*, *ITGAV* and *ITGB8*), while a previously associated locus also overlaps with a fourth integrin gene, *ITGAL*. In addition, a recent study has demonstrated that there is an IBD specific association that affects expression of *ICAM1*, which encodes the binding partner of *ITGAL* (Dendrou et al., 2016). The integrins encoded by these genes act as cell adhesion mediators that are capable of signalling across the plasma membrane in both directions, and have been shown to play a crucial role in leukocyte homing and cell differentiation in inflammation (Hynes, 2002). An overview of how integrins are involved in leukocyte homing to different tissues is given in Figure 4.7.

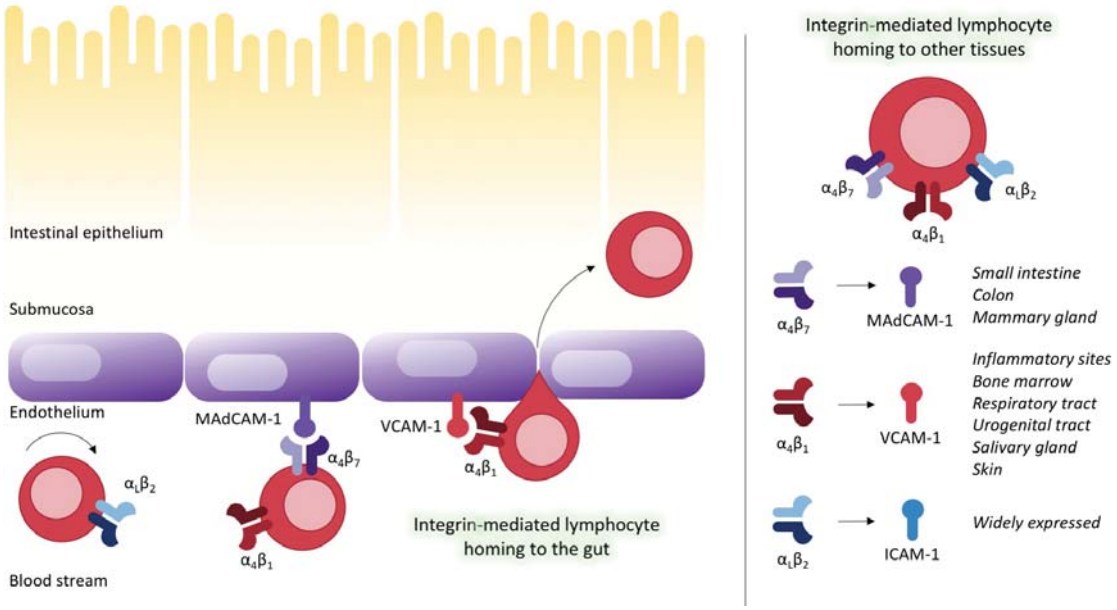

Figure 4.7: The role of integrins in leukocyte homing. Integrin complexes are expressed on the surface of leukocytes, and will bind to corresponding intercellular adhesion molecules on the surface of endothelial cells, prompting infiltration of the leukocytes into the tissue. Some of these binding partners exhibit tissue-specific expression patterns, allowing for tight control of leukocyte homing to specific sites (Kunkel and Butcher, 2003; Pals et al., 2007).

These integrin genes reflect very strong candidates for involvement in inflammatory bowel disease. However, both the gene set enrichment result discussed above, as well as the rare variant burden tests across known IBD genes described in Chapter 3, highlight the importance of using methods such as fine-mapping and eQTL association studies in order to confidently assign GWAS signals to their causal genes. As the fine-mapping analysis had already excluded the possibility that these associations could be caused by protein-coding changes, we next searched for effects of IBD risk SNPs on integrin gene expression in immune cells using a number of publicly available eQTL datasets (Zeller et al., 2010; Fairfax et al., 2012; Westra et al., 2013; Battle et al., 2014; Fairfax et al., 2014; Lee et al., 2014a; Raj et al., 2014; Ye et al., 2014; GTEx Consortium, 2015; Zhernakova et al., 2015).

While many eQTL and GWAS signals show some degree of correlation, inferences about causality require more robust statistical co-localization of the two signals. One means of obtaining this statistical support is to directly test for co-localization between IBD association signals and eQTLs using the coloc2 method (Giambartolomei et al., 2014), implemented in the R package coloc. We ran this method across our dataset, using a window size of 250kb on each side of the IBD association and default settings. Each test was repeated using two different p_{12} values ($p_{12} = 1 \times 10^{-5}$ and $p_{12} = 1 \times 10^{-6}$), which represents the prior probability of co-localization. For each gene, we test for co-localization with eQTLs in unstimulated monocytes, as well as monocytes stimulated with lipopolysaccaride (LPS) after 2 and 24 hours, monocytes stimulated with IFN-γ, and in unstimulated B cells, as described by Fairfax et al. (2014). The results of this analysis are summarised in Table 4.5.

Table 4.5: Co-localization between meta-analysis association statistics and monocyte stimulus response eQTLs. The co-localization of the meta-analysis and eQTL signals is tested with two different priors (1×10^{-5} and 1×10^{-6}) across the genes *ITGA4*, *ITGB8*, *ITGAL* and *ICAM1*. For each gene, we test co-localization with eQTLs in unstimulated monocytes, as well as monocytes stimulated with LPS after 2 and 24 hours, monocytes stimulated with IFN-γ, and in unstimulated B cells.

	Prior (p12)	Posterior probability of co-localization between GWAS association and monocyte eQTLs (after the application of stimuli)				
		Naive	LPS2HR	LPS24HR	IFN-γ	BCELL
ITGAL	1×10^{-05}	0.089	0.045	0.980	0.989	0.045
	1×10^{-06}	0.010	0.005	0.833	0.896	0.005
ITGB8	1×10^{-05}	0.061	0.057	0.712	0.051	0.178
	1×10^{-06}	0.006	0.006	0.198	0.005	0.021
ITGA4	1×10^{-05}	0.979	0.736	0.984	0.992	0.228
	1×10^{-06}	0.823	0.218	0.864	0.923	0.029
ICAM1	1×10^{-05}	0.050	0.961	0.093	0.162	0.064
	1×10^{-06}	0.005	0.713	0.010	0.019	0.007

Remarkably, three of the associations near integrin genes had $> 90\%$ probability of being driven by the same variants as monocyte-specific stimulus response eQTLs (*ITGA4*, $P_{\text{LPS\_24hr}} = 0.984$; *ITGAL*, $P_{\text{LPS\_24hr}} = 0.980$; *ICAM1*, $P_{\text{LPS\_2hr}} = 0.961$). A fourth association, *ITGB8*, is difficult to map due to extended linkage disequilibrium in the locus, but shows intermediate evidence of co-localization ($P_{\text{LPS\_24hr}} = 0.712$) in response to the same stimulus (Figure 4.8). All four of the IBD risk increasing alleles are associated with upregulated expression of their respective genes, suggesting that an increased level of pro-inflammatory cell surface markers in response to stimulus may be a consistent mechanism of action for these associations. Determining if this is indeed the case, however, would require functional follow up to prove that these IBD risk alleles causally change gene expression in response to stimulus, and indeed that changes in integrin gene expression are relevant to the inflammatory bowel disease phenotype.

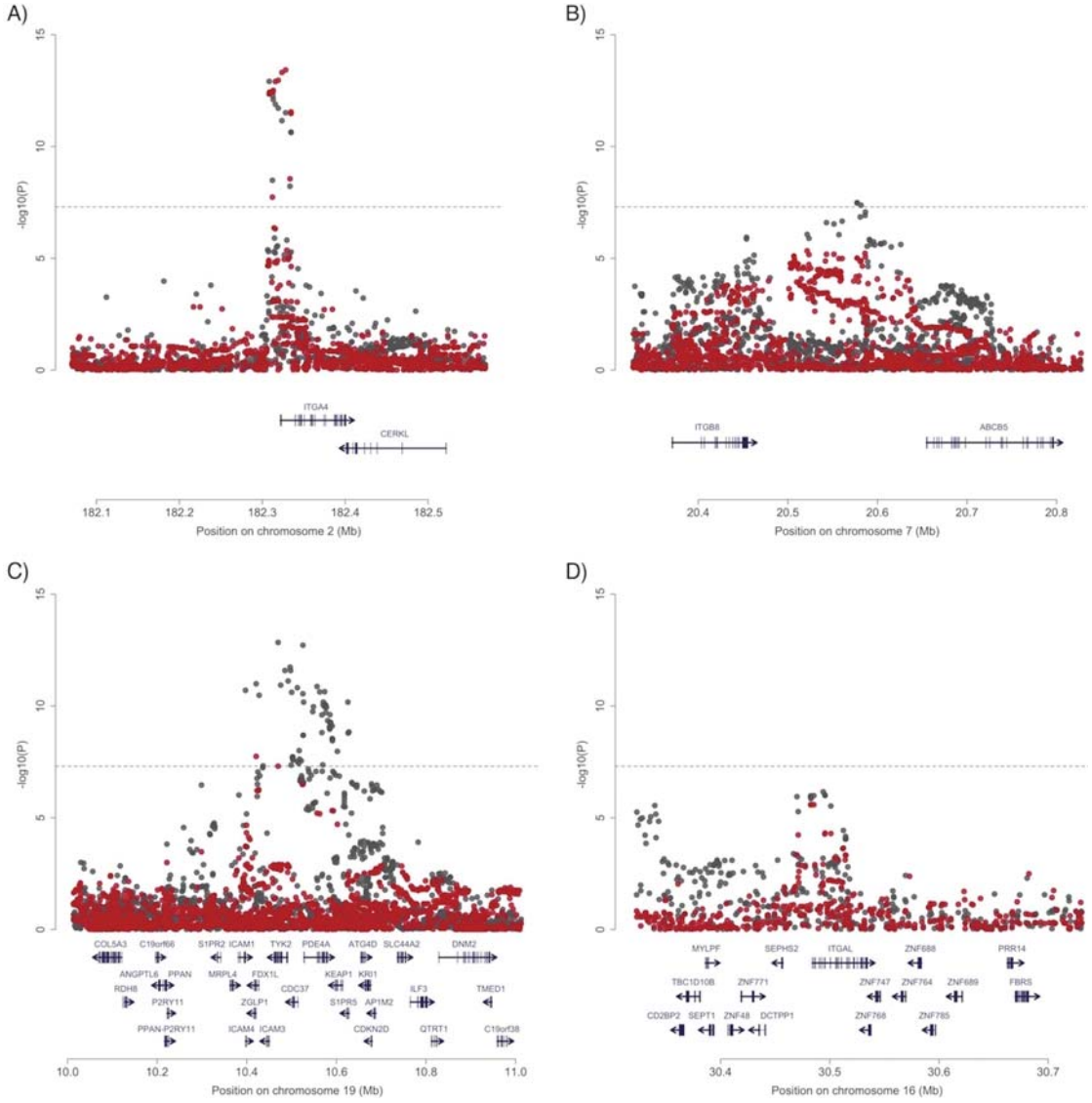

Figure 4.8: Co-localization of disease association and stimulus response eQTLs in monocytes. The local pattern of disease association (IBD: (A) *ITGA4*, (B) *ITGB8*, (C) *ICAM1*; (D) UC: *ITGAL*) in grey, and the association of that variant with response to LPS stimulation in red. Evidence of co-localization (probability > 70%) is observed for all for signals.

This second point is supported by the recent emergence of integrins and their counter-receptors as important therapeutic targets in IBD. In particular, the monoclonal antibodies vedolizumab and etrolizumab, which target the components of the $\alpha 4\beta 7$ dimer (encoded by *ITGA4* and *ITGB7*, and responsible for the gut-homing specificity of certain leukocytes), have demonstrated efficacy in both CD

and UC (Feagan et al., 2013; Sandborn et al., 2013; Vermeire et al., 2014). In addition, an antisense oligonucleotide that targets *ICAM1* has recently shown promise in the treatment of ulcerative colitis and pouchitis (Hosten et al., 2014).

Therapeutics targeting integrin αL (efalizumab) and α4 (natalizumab) have also demonstrated potential in the treatment of Crohn's disease (Sandborn et al., 2005; James et al., 2011), but have subsequently been associated with progressive multifocal leukoencephalopathy, or PML (Carson et al., 2009). This association highlights the importance of gut-selectivity in therapeutic approaches, with the potentially fatal PML condition likely to be mediated by binding to integrin dimers that are not gut-specific (leading to deficiencies in leukocyte migration to the central nervous system, and allowing for JC virus infection in the brain). Because of the risk of PML, efalizumab has been withdrawn from the market and natalizumab is not licensed for Crohn's disease in Europe.

Integrins are not only important in cell trafficking, but can also contribute to cellular signalling. For example, the αVβ8 heterodimer - both subunits of which are encoded by genes which are now within confirmed IBD loci (*ITGAV* and *ITGB8*, respectively) - is a potent activator of TGFβ. Notably, mice with dendritic-cell specific deletion of this complex had impaired regulatory T cell function and severe colitis (Travis et al., 2007), while deleting it in regulatory T cells themselves prevented the suppression of pathogenic T cell responses during active inflammation (Worthington et al., 2015). Although no therapeutics directly target αVβ8, there have been promising early results from an oral antisense oligonucleotide to the inhibitory TGFβ-signalling protein SMAD7 (Monteleone et al., 2015), itself encoded by a locus identified by genetic association studies (Jostins et al., 2012), that emphasises the therapeutic potential of modifying TGFβ in inflammatory bowel disease.

4.3.4 Therapeutic relevance of genetic associations

The associations to anti-integrin and anti-TGFβ therapies described above are just a few examples of therapeutically relevant genes that have been implicated using genetic studies of inflammatory bowel disease. To investigate these connections on a broader scale, we identified the following immune pathways as relevant to classes of approved IBD therapeutics: the IL12 and IL23 signalling pathways (ustekinumab, Sandborn et al. (2012)), the TNFα signalling pathway (infliximab, Hanauer et al. (2002); adalimumab, Colombel et al. (2007)), and the integrin signalling pathway (vedolizumab, Feagan et al. (2013) and Sandborn et al. (2013)). Genes involved in these pathways were then identified using the Molecular Signatures Database canonical pathways gene sets (C2; available at http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=CP), which have been curated by the Pathway Interaction Database (Schaefer et al., 2009). The integrin signalling gene list was comprised of all unique genes from the following gene sets: the integrin β1 pathway, integrin β7 pathway and integrin cell surface interactions. The list of TNFα signalling genes was obtained from the TNF pathway, and the list of IL-23/IL-12 p40 signalling genes was comprised of all unique genes from the IL12 and IL23 pathways.

Based on these gene lists, I identified genes in known IBD loci of therapeutic relevance (Table 4.6). As Figure 4.9 highlights, the importance of the biological pathways underlying associations, and their potential therapeutic significance, are not necessarily reflected in their GWAS effects sizes, with many relevant associations requiring tens of thousands of samples to identify.

Table 4.6: IBD-associated loci containing genes in immune pathways related to classes of approved therapeutics. We highlight loci that contain a gene in one of four signalling pathways related to targets of three classes of approved IBD therapeutics. In each case the relevant gene, signalling pathway, and therapeutic is marked. Genes marked with a ⋆ have been confidently implicated as the causal IBD gene.

Chr	Locus (Mb)	Relevant Gene	Pathway	Therapeutic(s)
1	67.2-68.1	*IL23R*⋆,*IL12RB2*	IL12, IL23	Ustekinumab
4	123-123.6	*IL2*	IL12, IL23	Ustekinumab
7	107.4-107.6	*LAMB1*	Integrinβ1	Vedolizumab
3	46.2-46.5	*CCR5*	IL12	Ustekinumab
14	75.7-75.7	*FOS*	IL12	Ustekinumab
16	11.3-11.7	*SOCS1*	IL12	Ustekinumab
6	149.6-149.6	*TAB2*	TNF	Infliximab, Adalimumab
4	102.7-103.5	*NFKB1*	IL12, IL23, TNF	Ustekinumab, Infliximab, Adalimumab
2	191.9-192	*STAT4*	IL12, IL23	Ustekinumab
10	75.5-75.7	*PLAU*	Integrinβ1, Integrinβ5-8	Vedolizumab
16	30.5-30.5	*ITGAL*⋆	Integrin cell interactions	Vedolizumab
17	32.6-32.6	*CCL2*	IL23	Ustekinumab
2	102.6-103.2	*IL18RAP*⋆,*IL18R1*,*IL1R1*	IL12, IL23	Ustekinumab
10	6.0-6.5	*IL2RA*⋆	IL12	Ustekinumab
5	158.7-158.9	*IL12B*	IL12, IL23	Ustekinumab
17	40.4-40.7	*STAT5A*,*STAT3*	IL12, IL23	Ustekinumab
19	10.4-10.6	*TYK2*⋆	IL12, IL23	Ustekinumab
2	182.3-182.3	*ITGA4*⋆	Integrinβ1, Integrinβ5-8, Integrin cell interactions	Vedolizumab
2	187.5-187.7	*ITGAV*	Integrinβ1, Integrinβ5-8, Integrin cell interactions	Vedolizumab
7	20.6-20.6	*ITGB8*⋆	Integrinβ5-8, Integrin cell interactions	Vedolizumab

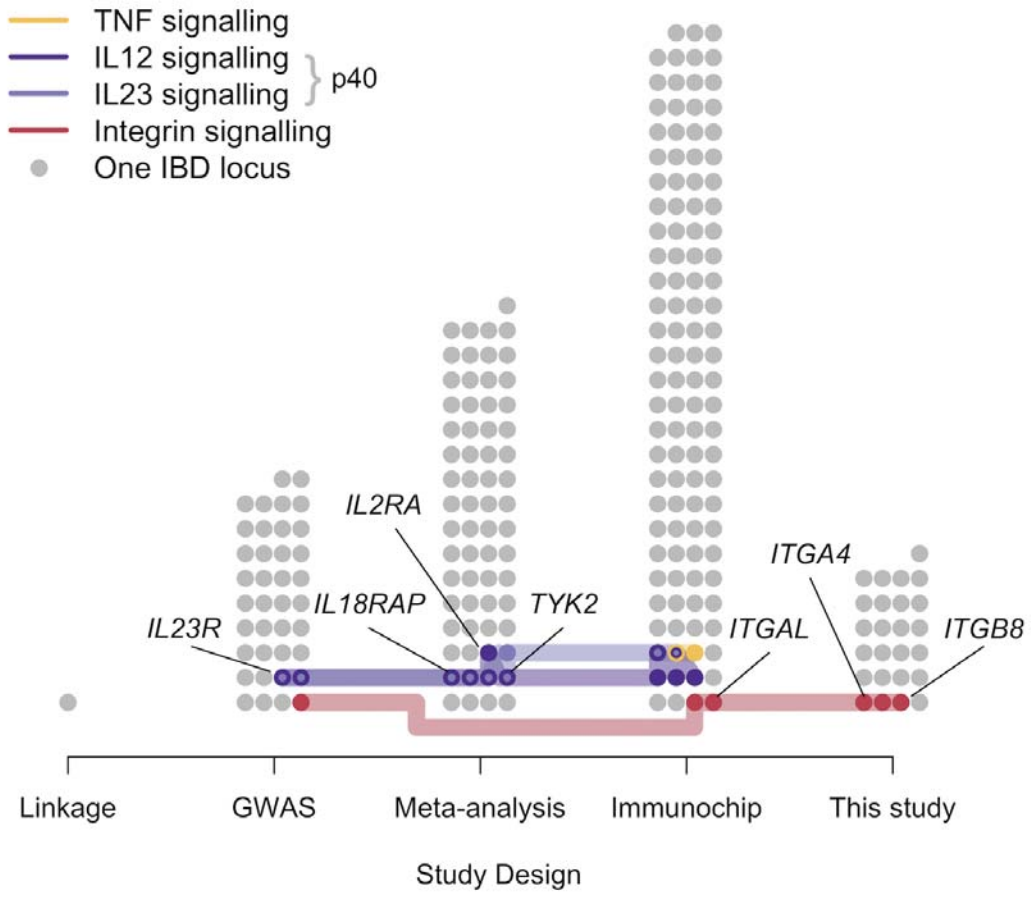

Figure 4.9: IBD-associated loci containing genes in immune pathways related to classes of approved therapeutics. All IBD loci are divided into the studies where they were first identified. Loci that contain a gene in one of four signalling pathways related to targets of three classes of approved IBD therapeutics are highlighted, with those where the pathway gene has been confidently identified as the causal IBD gene labelled. Despite the general pattern that effect size decreases from left to right, therapeutically relevant associations continue to be found.

4.4 Discussion

In this chapter I have described the discovery of 25 novel IBD-associated loci via the imputation and meta-analysis of nearly 60,000 samples, bringing the total number of loci associated with IBD to 241. Summary-statistic fine-mapping on these loci identified likely causal missense variants in the genes *SLAMF8*, a negative regulator of inflammation, and *PLCG2*, a gene implicated in primary immune deficiency. A potentially causal variant is also observed in an intron of *NCF4*, which is another gene associated with an immune-related Mendelian disorder.

A potential relationship between genes associated with Mendelian disorders of inflammation and immunity and those implicated in IBD has long been recognised, with the first Crohn's disease risk gene discovered, *NOD2*, also linked to the autosomal dominant granulomatous disorder Blau syndrome (Miceli-Richard et al., 2001). I confirm this link more generally, showing a strong enrichment for such genes amongst all known IBD loci. Furthermore, this enrichment is significantly stronger when considering just those IBD-associated loci for which a gene can be causally assigned with high confidence, either through fine-mapping or eQTL co-localization, highlighting the importance of using such methods when trying to draw conclusions about the biological mechanisms underlying an association.

Amongst the novel loci that could not be fine-mapped to a likely causal variant, three are proximal to integrin genes, which encode proteins in pathways that have been identified as important therapeutic targets in inflammatory bowel disease. Co-localization with eQTL signals confirm that the associated IBD risk-increasing variants are also correlated with expression changes in monocytes in response to immune stimulus at two of these genes (*ITGA4* and *ITGB8*), and at two previously implicated loci (*ITGAL* and *ICAM1*). This suggests that an increased level of pro-inflammatory cell surface markers in response to stimulus may be a consistent mechanism of action for these particular associations, although further functional follow up would be required to confirm this.

The discovery of this association between integrin genes and inflammatory bowel disease was particularly exciting for two key reasons. Firstly, assigning the signal

detected using GWAS to the likely causal genes would not have been possible without the ability to test for co-localization in an eQTL dataset that had analysed both the relevant cell type, and used the correct stimulus. No co-localization was observed between our data and eQTLs for integrin expression in B cells; similarly, a number of the GWAS associations failed to co-localize with integrin eQTLs from naive monocytes, or even monocytes stimulated with interferon-γ (Table 4.5). As studies that aim to uncover the specific cellular contexts in which different genes are active continue to grow in number and coverage, there is an exciting opportunity to potentially resolve the biological mechanisms underlying a number of other GWAS loci that can not be assigned to causal coding variation. Secondly, despite the relatively modest effect size of the signals near integrin genes (OR 1.10-1.12), they are of high therapeutic relevance. If we extend the idea of therapeutic relevance to other IBD-associated loci, it is clear that the importance of the biological pathways underlying genetic associations, and their potential use as drug targets, do not necessarily correlate with their GWAS effect sizes (Figure 4.9).

Overall, our findings suggest that there are still a number of potential benefits to be obtained by continuing to pursue genome-wide association studies, even in a well-studied complex disease like IBD, as valuable complementary analyses to large-scale sequencing endeavours.

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.

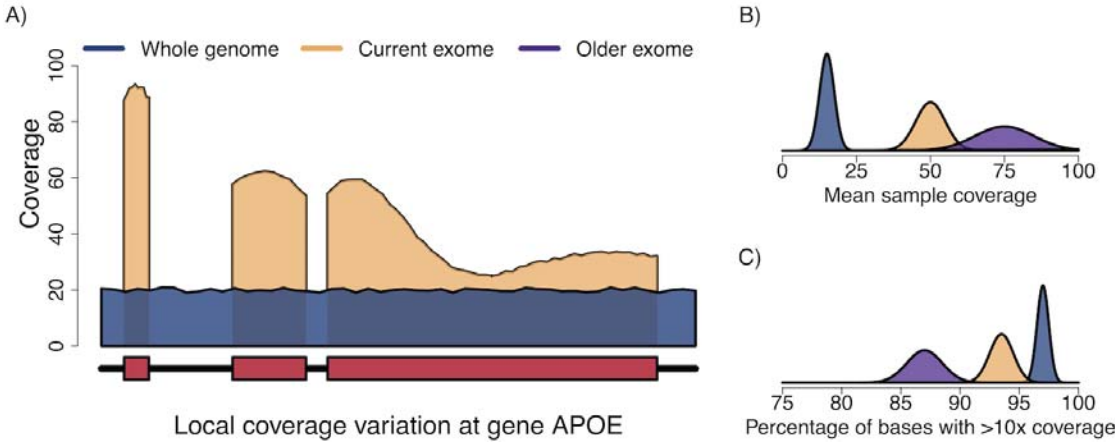

Figure 5.1: An illustration of the relative coverage that can be obtained using current exome (WES) and whole genome (WGS) sequencing techniques, and the improvements that have already been seen compared to initial WES protocols. WGS is able to produce much more even local coverage (panel A), that allows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

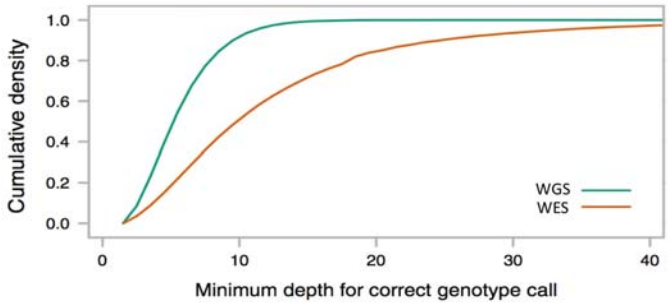

Figure 5.2: The minimum depth required to make a correct heterozygous genotype call in matched whole exome (WES) and whole genome (WGS) sequencing samples. Figure sourced from Meynert et al. (2014).

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.

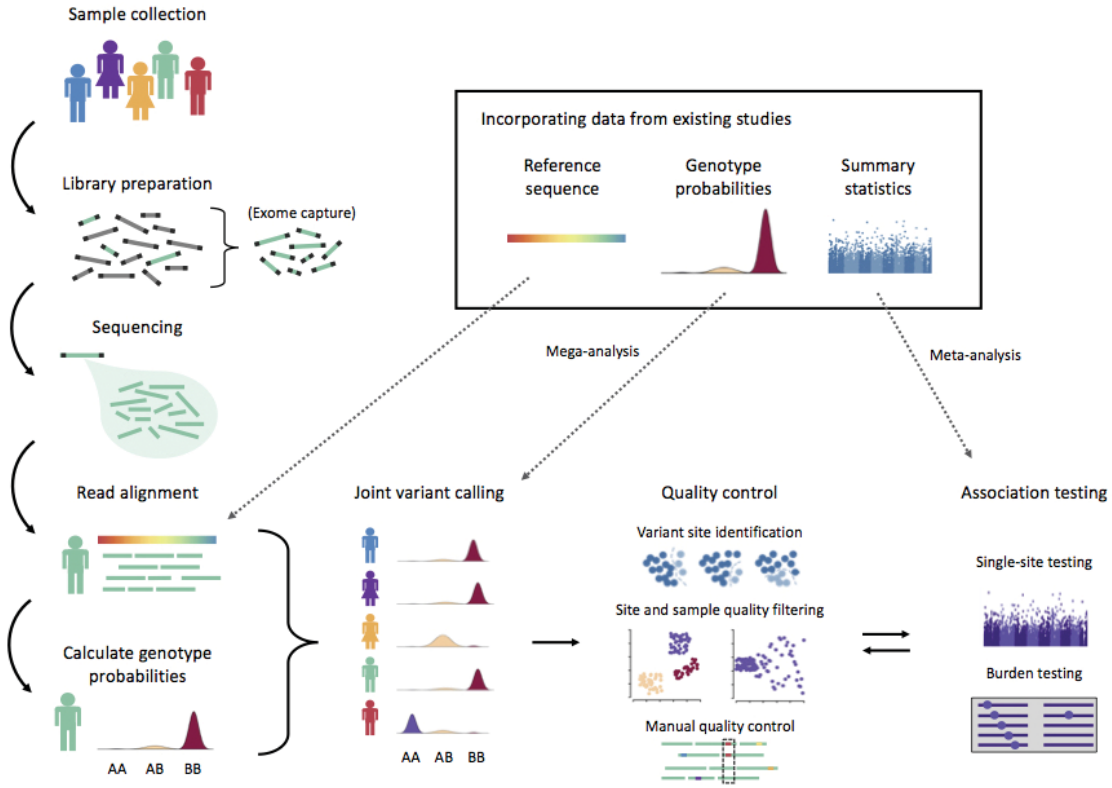

Figure 5.3: An overview of the key features of a sequence-based study, from initial sample collection and sequencing through to the final association testing and joint analysis with external datasets.

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.

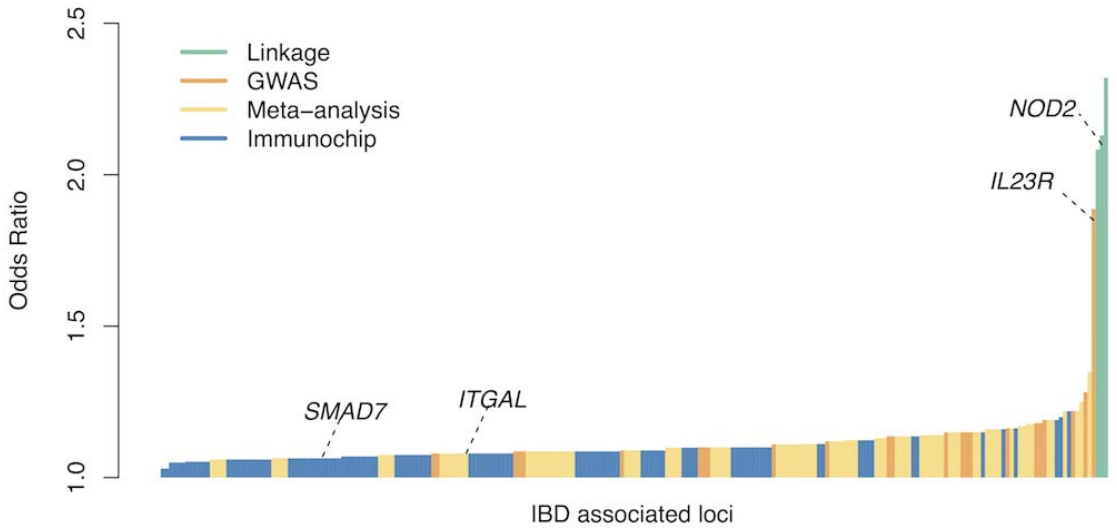

Figure 5.4: Effect sizes of IBD-associated loci identified using various study designs. The largest effect sizes are seen for the first two genes associated with IBD, *NOD2* and *IL23R*. Nevertheless, the genes *SMAD7* and *ITGAL*, which have relatively small effect sizes, are both confirmed drug targets.

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.

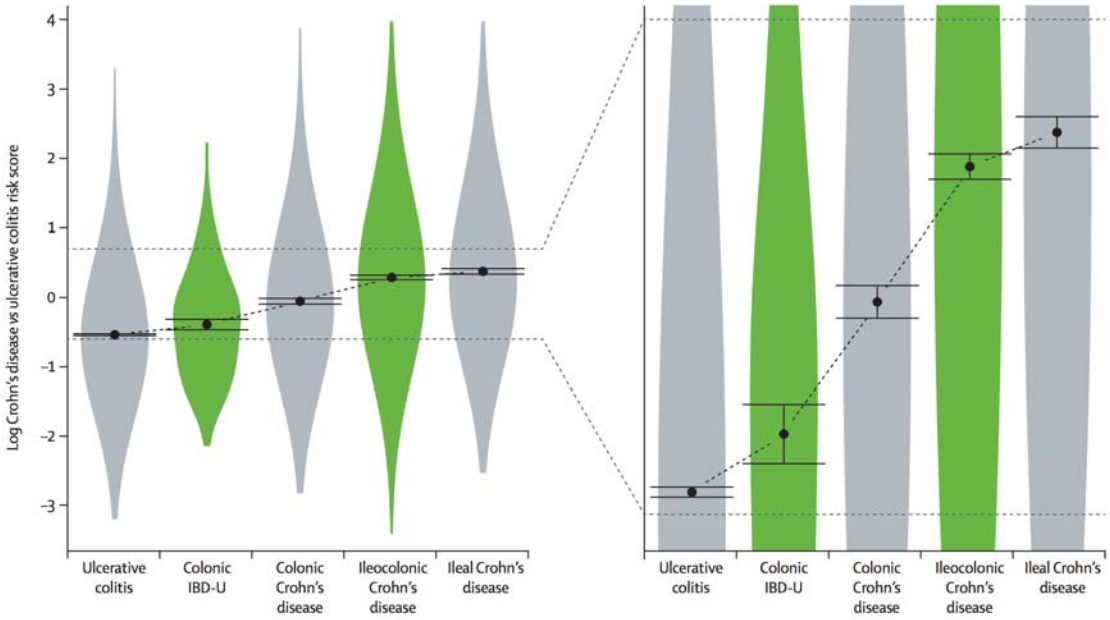

Figure 5.5: The genetic substructure of inflammatory bowel disease location, identified using genetic risk scores. A continuum of disorders based on disease location can clearly be seen, from those largely affecting the colon (UC, and colonic CD) to those largely affecting the ileum (ileal CD). Figure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

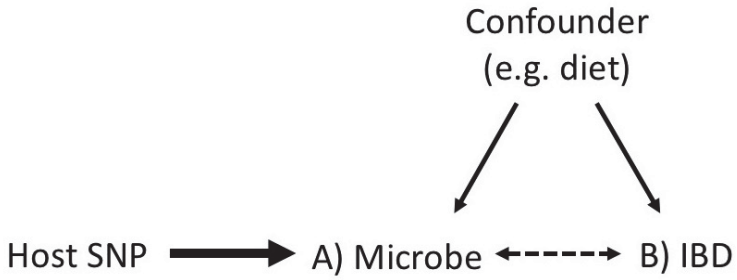

Figure 5.6: Mendelian randomization can be used to infer a causal relationship between two correlated traits, A and B (in this case the microbiome and IBD). If this correlation has arisen because A causes B, then it follows that any variable that affects trait A should also affect trait B (but not vice 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.

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, et al. (2015). "A global reference for human genetic variation". *Nature* 526.7571, pp. 68–74.

Abecasis, Gonçalo R, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467.7319, pp. 1061–1073.

Albert, Frank W and Leonid Kruglyak (2015). "The role of regulatory variation in complex traits and disease". *Nat. Rev. Genet.* 16.4, pp. 197–212.

Altshuler, David M, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Penelope E Bonnen, Paul I W de Bakker, Panos Deloukas, et al. (2010). "Integrating common and rare genetic variation in diverse human populations". *Nature* 467.7311, pp. 52–58.

Anderson, Carl A, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan (2010). "Data quality control in genetic case-control association studies". *Nat. Protoc.* 5.9, pp. 1564–1573.

Anderson, Carl A, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D'Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, et al. (2011). "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47". *Nat. Genet.* 43.3, pp. 246–252.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, et al. (2014). "An atlas of active enhancers across human cell types and tissues". *Nature* 507, pp. 455–461.

Antonopoulos, Dionysios A, Susan M Huse, Hilary G Morrison, Thomas M Schmidt, Mitchell L Sogin, and Vincent B Young (2009). "Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation". *Infect. Immun.* 77.6, pp. 2367–2375.

Asano, Kouichi, Tomonaga Matsushita, Junji Umeno, Naoya Hosono, Atsushi Takahashi, Takahisa Kawaguchi, Takayuki Matsumoto, Toshiyuki Matsui, Yoichi Kakuta, et al. (2009). "A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population". *Nat. Genet.* 41.12, pp. 1325–1329.

Asimit, Jennifer L, Aaron G Day-Williams, Andrew P Morris, and Eleftheria Zeggini (2012). "ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data". *Hum. Hered.* 73.2, pp. 84–94.

Barbieri, Christopher E, Sylvan C Baca, Michael S Lawrence, Francesca Demichelis, Mirjam Blattner, Jean-Philippe Theurillat, Thomas A White, Petar Stojanov, Eliezer Van Allen, et al. (2012). "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer". *Nat. Genet.* 44.6, pp. 685–689.

Barrett, Jeffrey C and Lon R Cardon (2006). "Evaluating coverage of genome-wide association studies". *Nat. Genet.* 38.6, pp. 659–662.

Barrett, Jeffrey C, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease". *Nat. Genet.* 40.8, pp. 955–962.

Barrett, Jeffrey C, James C Lee, Charles W Lees, Natalie J Prescott, Carl A Anderson, Anne Phillips, Emma Wesley, Kirstie Parnell, Hu Zhang, et al. (2009). "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region". *Nat. Genet.* 41.12, pp. 1330–1334.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, et al. (2014). "Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals". *Genome Res.* 24.1, pp. 14–24.

Baumgart, D C and W J Sandborn (2007). "Inflammatory bowel disease: clinical aspects and established and evolving therapies". *Lancet* 369.5, pp. 1641–1657.

Beaudoin, Mélissa, Philippe Goyette, Gabrielle Boucher, Ken Sin Lo, Manuel A Rivas, Christine Stevens, Azadeh Alikashani, Martin Ladouceur, David Ellinghaus, et al. (2013). "Deep Resequencing of GWAS Loci Identifies Rare Variants in CARD9, IL23R and RNF186 That Are Associated with Ulcerative Colitis". *PLoS Genet.* 9.9.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker (2012). "Hi-C: a comprehensive technique to capture the conformation of genomes". *Methods* 58.3, pp. 268–276.

Bernstein, Charles N, Michael Fried, J H Krabshuis, Henry Cohen, R Eliakim, Suleiman Fedail, Richard Gearry, K L Goh, Saheed Hamid, et al. (2010). "World gastroenterology organization practice guidelines for the diagnosis and management of IBD in 2010". *Inflamm. Bowel Dis.* 16.1, pp. 112–124.

Blaydon, Diana C, Paolo Biancheri, Wei-Li Di, Vincent Plagnol, Rita M Cabral, Matthew A Brooke, David A van Heel, Franz Ruschendorf, Mark Toynbee, et al. (2011). "Inflammatory skin and bowel disease linked to ADAM17 deletion". *N. Engl. J. Med.* 365.16, pp. 1502–1508.

Botstein, D, R L White, M Skolnick, and R W Davis (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms". *Am. J. Hum. Genet.* 32.3, pp. 314–331.

Boyd, Andy, Jean Golding, John Macleod, Debbie A Lawlor, Abigail Fraser, John Henderson, Lynn Molloy, Andy Ness, Susan Ring, and George Davey Smith (2013). "Cohort Profile: the 'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children". *Int. J. Epidemiol.* 42.1, pp. 111–127.

Boyman, Onur and Jonathan Sprent (2012). "The role of interleukin-2 during homeostasis and activation of the immune system". *Nat. Rev. Immunol.* 12.3, pp. 180–190.

Brant, Steven R (2011). "Update on the heritability of inflammatory bowel disease: the importance of twin studies". *Inflamm. Bowel Dis.* 17.1, pp. 1–5.

Brant, Steven R, Yifan Fu, Carter T Fields, Romulo Baltazar, Geoffrey Ravenhill, Michael R Pickles, Patrick M Rohal, Jasdeep Mann, Barbara S Kirschner, et al. (1998). "American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12". *Gastroenterology* 115, pp. 1056–1061.

Brown, Eric M, Manish Sadarangani, and B Brett Finlay (2013). "The role of the immune system in governing host-microbe interactions in the intestine". *Nat. Immunol.* 14.7, pp. 660–667.

Browning, Brian L and Sharon R Browning (2016). "Genotype Imputation with Millions of Reference Samples". *Am. J. Hum. Genet.* 98.1, pp. 116–126.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". *Nat. Genet.* 47.3, pp. 291–295.

Cai, Na, Tim B Bigdeli, Warren Kretzschmar, Yihan Li, Jieqin Liang, Li Song, Jingchu Hu, Qibin Li, Wei Jin, et al. (2015). "Sparse whole-genome sequencing identifies two loci for major depressive disorder". *Nature* 12415800.

Carson, Kenneth R, Daniele Focosi, Eugene O Major, Mario Petrini, Elizabeth A Richey, Dennis P West, and Charles L Bennett (2009). "Monoclonal antibody-associated progressive multifocal leucoencephalopathy in patients treated with rituximab, natalizumab, and efalizumab: a Review from the Research on Adverse Drug Events and Reports (RADAR) Project". *Lancet Oncol.* 10.8, pp. 816–824.

Cavanaugh, J A, D F Callen, S R Wilson, P M Stanford, M E Sraml, M Gorska, J Crawford, S A Whitmore, C Shlegel, et al. (1998). "Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16". *Ann. Hum. Genet.* 62.4, pp. 291–298.

Cavanaugh, J and The International IBD Genetics Consortium (2001). "International Collaboration Provides Convincing Linkage Replication in Complex Disease through Analysis of a Large Pooled Data Set : Crohn Disease and Chromosome 16". *Am. J. Hum. Genet.* 68, pp. 1165–1171.

Chang, Christopher C, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* 4.1, p. 7.

Chang, Diana, Feng Gao, Andrea Slavney, Li Ma, Yedael Y Waldman, Aaron J Sams, Paul Billing-Ross, Aviv Madar, Richard Spritz, and Alon Keinan (2014). "Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases". *PLoS One* 9.12, e113684.

Chen, Gui-Bo, Sang Hong Lee, Marie-Jo A Brion, Grant W Montgomery, Naomi R Wray, Graham L Radford-Smith, Peter M Visscher, and International IBD Genetics Consortium (2014). "Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data". *Hum. Mol. Genet.* 23.17, pp. 4710–4720.

Choi, Murim, Ute I Scholl, Weizhen Ji, Tiewen Liu, Irina R Tikhonova, Paul Zumbo, Ahmet Nayir, Ayin Bakkaloğlu, Seza Özen, et al. (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing". *Proceedings of the National Academy of Sciences* 106.45, pp. 19096–19101.

Cleynen, Isabelle, Juan R González, Carolina Figueroa, Andre Franke, Dermot McGovern, Martin Bortlík, Bart J A Crusius, Maurizio Vecchi, Marta Artieda, et al. (2013). "Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project". *Gut* 62.11, pp. 1556–1565.

Cleynen, Isabelle, Gabrielle Boucher, Luke Jostins, L Philip Schumm, Sebastian Zeissig, Tariq Ahmad, Vibeke Andersen, Jane M Andrews, Vito Annese, et al. (2016). "Inherited determinants of Crohn's disease and

ulcerative colitis phenotypes: a genetic association study". *Lancet* 387.10014, pp. 156–167.

Colombel, Jean-Frédéric, William J Sandborn, Paul Rutgeerts, Robert Enns, Stephen B Hanauer, Remo Panaccione, Stefan Schreiber, Dan Byczkowski, Ju Li, et al. (2007). "Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial". *Gastroenterology* 132.1, pp. 52–65.

Curran, Mark E, K I T F Lau, Jochen Hampe, Stefan Schreiber, Steven Bridger, Andrew J S Macpherson, L O N R Cardon, Hakan Sakul, Timothy J R Harris, et al. (1998). "Genetic Analysis of Inflammatory Bowel Disease in a Large European Cohort Supports Linkage to Chromosomes 12 and 16". *Gastroenterology* 115, pp. 1066–1071.

Dahle, Maria K, Anders E Myhre, Ansgar O Aasen, and Jacob E Wang (2005). "Effects of forskolin on Kupffer cell production of interleukin-10 and tumor necrosis factor alpha differ from those of endogenous adenylyl cyclase activators: possible role for adenylyl cyclase 9". *Infect. Immun.* 73.11, pp. 7290–7296.

Danjou, Fabrice, Magdalena Zoledziewska, Carlo Sidore, Maristella Steri, Fabio Busonero, Andrea Maschio, Antonella Mulas, Lucia Perseu, Susanna Barella, et al. (2015). "Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels". *Nat. Genet.* 47.11, pp. 1264–1271.

Davey Smith, George and Gibran Hemani (2014). "Mendelian randomization: genetic anchors for causal inference in epidemiological studies". *Hum. Mol. Genet.* 23.R1, R89–98.

Daye, Z John, Hongzhe Li, and Zhi Wei (2012). "A powerful test for multiple rare variants association studies that incorporates sequencing qualities". *Nucleic Acids Res.* 40.8, e60.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, et al. (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data". *Nat. Genet.* 43.5, pp. 491–498.

Dendrou, Calliope A, Adrian Cortes, Lydia Shipman, Hayley G Evans, Kathrine E Attfield, Luke Jostins, Thomas Barber, Gurman Kaur, Subita Balaram Kuttikkatte, et al. (2016). "Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity". *Sci. Transl. Med.* 8.363, 363ra149–363ra149.

Derkach, Andriy, Jerry F Lawless, and Lei Sun (2013). "Robust and Powerful Tests for Rare Variants Using Fisher's Method to Combine Evidence of Association From Two or More Complementary Tests". *Genet. Epidemiol.* 37.1, pp. 110–121.

Derkach, Andriy, Theodore Chiang, Jiafen Gong, Laura Addis, Sara Dobbins, Ian Tomlinson, Richard Houlston, Deb K Pal, and Lisa J Strug (2014). "Association analysis using next-generation sequence data from publicly available control groups: The robust variance score statistic". *Bioinformatics* 30.15, pp. 2179–2188.

Dethlefsen, Les, Sue Huse, Mitchell L Sogin, and David A Relman (2008). "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing". *PLoS Biol.* 6.11, pp. 2383–2400.

de Vries, A Boudewijn, Marcel Janse, Hans Blokzijl, and Rinse K Weersma (2015). "Distinctive inflammatory bowel disease phenotype in primary sclerosing cholangitis". *World J. Gastroenterol.* 21.6, pp. 1956–1971.

D'haeseleer, Patrik (2006). "What are DNA sequence motifs?" *Nat. Biotechnol.* 24.4, pp. 423–425.

Dilthey, Alexander T, Pierre-Antoine Gourraud, Alexander J Mentzer, Nezih Cereb, Zamin Iqbal, and Gil McVean (2016). "High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs". *PLoS Comput. Biol.* 12.10, e1005151.

Dilthey, Alexander, Charles Cox, Zamin Iqbal, Matthew R Nelson, and Gil McVean (2015). "Improved genome inference in the MHC using a population reference graph". *Nat. Genet.* 47.6, pp. 682–688.

Duan, Biyan, Richard Davis, Eva L Sadat, Julie Collins, Paul C Sternweis, Dorothy Yuan, and Lily I Jiang (2010). "Distinct roles of adenylyl cyclase VII in regulating the immune responses in mice". *J. Immunol.* 185.1, pp. 335–344.

Duerr, R H, K D Taylor, S R Brant, J D Rioux, M S Silverberg, M J Daly, a H Steinhart, C Abraham, M Regueiro, et al. (2006). "A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene". *Science* 314.5804, pp. 1461–1463.

Durbin, Richard (2014). "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)". *Bioinformatics* 30.9, pp. 1266–1272.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489.7414, pp. 57–74.

Eiseman, B, W Silen, G S Bascom, and A J Kauvar (1958). "Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis". *Surgery* 44.5, pp. 854–859.

Ellinghaus, David, Luke Jostins, Sarah L Spain, Adrian Cortes, Jörn Bethune, Buhm Han, Yu Rang Park, Soumya Raychaudhuri, Jennie G Pouget, et al. (2016). "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". *Nat. Genet.* 48.5, pp. 510–518.

Evans, David M and Lon R Cardon (2004). "Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps". *Am. J. Hum. Genet.* 75.4, pp. 687–692.

Fairfax, Benjamin P, Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O Vannberg, and Julian C Knight (2012). "Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles". *Nat. Genet.* 44.5, pp. 502–510.

Fairfax, Benjamin P, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, et al. (2014). "Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression". *Science* 343.6175, p. 1246949.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J H Ryan, et al. (2015). "Genetic and epigenetic fine mapping of causal autoimmune disease variants". *Nature* 518.7539, pp. 337–343.

Feagan, Brian G, Paul Rutgeerts, Bruce E Sands, Stephen Hanauer, Jean-Frédéric Colombel, William J Sandborn, Gert Van Assche, Jeffrey Axler, Hyo-Jong Kim, et al. (2013). "Vedolizumab as induction and maintenance therapy for ulcerative colitis". *N. Engl. J. Med.* 369.8, pp. 699–710.

Fisher, Sheila A, Mark Tremelling, Carl A Anderson, Rhian Gwilliam, Suzannah Bumpstead, Natalie J Prescott, Elaine R Nimmo, Dunecan Massey, Carlo Berzuini, et al. (2008). "Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease". *Nat. Genet.* 40.6, pp. 710–712.

Ford, D, D F Easton, M Stratton, S Narod, D Goldgar, P Devilee, D T Bishop, B Weber, G Lenoir, et al. (1998). "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families". *Am. J. Hum. Genet.* 62.3, pp. 676–689.

Franke, Andre, Tobias Balschun, Tom H Karlsen, Jurgita Sventoraityte, Susanna Nikolaus, Gabriele Mayr, Francisco S Domingues, Mario Albrecht, Michael Nothnagel, et al. (2008). "Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility". *Nat. Genet.* 40.11, pp. 1319–1323.

Franke, Andre, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". *Nat. Genet.* 42.12, pp. 1118–1125.

Fu, Guoping, Yuhong Chen, James Schuman, Demin Wang, and Renren Wen (2012). "Phospholipase Cγ2 plays a role in TCR signal transduction and T cell selection". *J. Immunol.* 189.5, pp. 2326–2332.

Fuchsberger, Christian, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, et al. (2016). "The genetic architecture of type 2 diabetes". *Nature* 536.7614, pp. 41–47.

GTEx Consortium (2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans". *Science* 348.6235, pp. 648–660.

Ganna, Andrea, Giulio Genovese, Daniel P Howrigan, Andrea Byrnes, Mitja I Kurki, Seyedeh M Zekavat, Christopher W Whelan, Mart Kals, Michel G Nivard, et al. (2016). "Ultra-rare disruptive and damaging mutations influence educational attainment in the general population". *Nat. Neurosci.* 19.12, pp. 1563–1565.

Garner, Chad (2011). "Confounded by sequencing depth in association studies of rare alleles". *Genet. Epidemiol.* 35.4, pp. 261–268.

Genovese, Giulio, Menachem Fromer, Eli A Stahl, Douglas M Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L Moran, Shaun M Purcell, Pamela Sklar, et al. (2016). "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia". *Nat. Neurosci.* 19.11, pp. 1433–1441.

Gevers, Dirk, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, et al. (2014). "The treatment-naive microbiome in new-onset Crohn's disease". *Cell Host Microbe* 15.3, pp. 382–392.

Giambartolomei, Claudia, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol (2014). "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics". *PLoS Genet.* 10.5, e1004383.

Gibson, Greg (2011). "Rare and common variants: twenty arguments". *Nat. Rev. Genet.* 13.2, pp. 135–145.

Glocker, E O, D Kotlarz, Kaan Boztug, E Michael Gertz, Alejandro A Schäffer, Fatih Noyan, Mario Perro, Jana Diestelhorst, Anna Allroth, et al. (2009). "Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor". *N. Engl. J. Med.* 361.21, pp. 2033–2045.

Goodwin, Sara, John D McPherson, and W Richard McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies". *Nat. Rev. Genet.* 17.6, pp. 333–351.

Goyette, Philippe, Gabrielle Boucher, Dermot Mallon, Eva Ellinghaus, Luke Jostins, Hailiang Huang, Stephan Ripke, Elena S Gusareva, Vito Annese, et al. (2015). "High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis". *Nat. Genet.* 47.2, pp. 172–179.

Gusella, J F, N S Wexler, P M Conneally, S L Naylor, M a Anderson, R E Tanzi, P C Watkins, K Ottina, M R Wallace, et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease". *Nature* 306.17, pp. 234–238.

Hall, Stephen S (2013). "Genetics: a gene of rare effect". *Nature* 496.7444, pp. 152–155.

Halme, Leena, P Paavola-Sakki, Ulla Turunen, Maarit Lappalainen, Martti Farkkila, and Kimmo Kontula (2006). "Family and twin studies in inflammatory bowel disease". *World J. Gastroenterol.* 12.23, pp. 3668–3672.

Hampe, Jochen, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, Gabriele Mayr, Francisco M De La Vega, et al. (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1". *Nat. Genet.* 39.2, pp. 207–211.

Han, Jian-Wen, Hou-Feng Zheng, Yong Cui, Liang-Dan Sun, Dong-Qing Ye, Zhi Hu, Jin-Hua Xu, Zhi-Ming Cai, Wei Huang, et al. (2009). "Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus". *Nat. Genet.* 41.11, pp. 1234–1237.

Hanauer, Stephen B, Brian G Feagan, Gary R Lichtenstein, Lloyd F Mayer, S Schreiber, Jean Frederic Colombel, Daniel Rachmilewitz, Douglas C Wolf, Allan Olson, et al. (2002). "Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial". *Lancet* 359.9317, pp. 1541–1549.

Handsaker, Robert E, Vanessa Van Doren, Jennifer R Berman, Giulio Genovese, Seva Kashin, Linda M Boettger, and Steven A McCarroll (2015). "Large multiallelic copy number variations in humans". *Nat. Genet.* 47.3, pp. 296–303.

Haritunians, Talin, Kent D Taylor, Stephan R Targan, Maria Dubinsky, Andrew Ippoliti, Soonil Kwon, Xiuqing Guo, Gil Y Melmed, Dror Berel, et al. (2010). "Genetic Predictors of Medically Refractory Ulcerative Colitis". *Inflamm. Bowel Dis.* 16.11, pp. 1830–1840.

Hosten, Terron Anthony, Ke Zhao, Hong Qiu Han, Gang Liu, and Xiang Hui He (2014). "Alicaforsen: An Emerging Therapeutic Agent for Ulcerative Colitis and Refractory Pouchitis". *Gastroenterol. Res. Pract.* 7.2, pp. 51–55.

Hu, Yi-Juan, Peizhou Liao, H Richard Johnston, Andrew S Allen, and Glen A Satten (2016). "Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls". *PLoS Genet.* 12.5, e1006040.

Huang, Hailiang, Ming Fang, Luke Jostins, Masa U Mirkov, Gabrielle Boucher, Carl A Anderson, Vibeke Andersen, Isabelle Cleynen, Adrian Cortes, et al. (2015). "Association mapping of inflammatory bowel disease loci to single variant resolution". *bioRxiv*, p. 028688.

Hugot, J P, P Laurent-Puig, C Gower-Rousseau, J M Olson, J C Lee, L Beaugerie, I Naom, J L Dupas, A Van Gossum, et al. (1996). "Mapping of a susceptibility locus for Crohn's disease on chromosome 16". *Nature* 379.6568, pp. 821–823.

Hugot, Jean-Pierre, Mathias Chamaillard, Habib Zouali, Suzanne Lesage, Jean-Pierre Cézard, Jacques Belaiche, Sven Almer, Curt Tysk, Colm A O'Morain, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease". *Nature* 411.6837, pp. 599–603.

Hunt, Karen a, Vanisha Mistry, Nicholas a Bockett, Tariq Ahmad, Maria Ban, Jonathan N Barker, Jeffrey C Barrett, Hannah Blackburn, Oliver Brand, et al. (2013). "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability". *Nature* 498.7453, pp. 232–235.

Hynes, Richard O (2002). "Integrins: bidirectional, allosteric signaling machines". *Cell* 110.6, pp. 673–687.

Ioannidis, John P A (2003). *Genetic associations: False or true?*

Ivanov, Ivaylo I, Brent S McKenzie, Liang Zhou, Carlos E Tadokoro, Alice Lepelley, Juan J Lafaille, Daniel J Cua, and Dan R Littman (2006). "The Orphan Nuclear Receptor RORγt Directs the Differentiation Program of Proinflammatory IL-17+ T Helper Cells". *Cell* 126.6, pp. 1121–1133.

James, Dustin G, Da Hea Seo, Jiajing Chen, Caroline Vemulapalli, and Christian D Stone (2011). "Efalizumab, a human monoclonal anti-CD11a antibody, in the treatment of moderate to severe Crohn's disease: An open-label pilot study". *Dig. Dis. Sci.* 56.6, pp. 1806–1810.

Jiang, Lily I, Paul C Sternweis, and Jennifer E Wang (2013). "Zymosan activates protein kinase A via adenylyl cyclase VII to modulate innate immune responses during inflammation". *Mol. Immunol.* 54.1, pp. 14–22.

Jin, Jill (2014). "JAMA patient page. Inflammatory bowel disease". *JAMA* 311.19, p. 2034.

Jostins, Luke, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, et al. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease". *Nature* 491.7422, pp. 119–124.

Julià, Antonio, Eugeni Domènech, María Chaparro, Valle García-Sánchez, Fernando Gomollón, Julián Panés, Míriam Mañosa, Manuel Barreiro-De Acosta, Ana Gutiérrez, et al. (2014). "A genome-wide association study identifies a novel

locus at 6q22.1 associated with ulcerative colitis". *Hum. Mol. Genet.* 23.25, pp. 6927–6934.

Juyal, Garima, Sapna Negi, Ajit Sood, Aditi Gupta, Pushplata Prasad, Sabyasachi Senapati, Jacques Zaneveld, Shalini Singh, Vandana Midha, et al. (2015). "Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis". *Gut* 64, pp. 571–579.

Kandt, R S, J L Haines, M Smith, H Northrup, R J Gardner, M P Short, K Dumars, E S Roach, S Steingold, and S Wall (1992). "Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease". *Nat. Genet.* 2.1, pp. 37–41.

Kaplan, Gilaad G (2015). "The global burden of IBD: from 2015 to 2025". *Nat. Rev. Gastroenterol. Hepatol.* 12.12, pp. 720–727.

Kenny, Eimear E, Itsik Pe'er, Amir Karban, Laurie Ozelius, Adele A Mitchell, Sok Meng Ng, Monica Erazo, Harry Ostrer, Clara Abraham, et al. (2012). "A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci". *PLoS Genet.* 8.3, e1002559.

Kheradpour, Pouya and Manolis Kellis (2014). "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments". *Nucleic Acids Res.* 42.5, pp. 2976–2987.

Khor, Bernard, Agnès Gardet, and Ramnik J Xavier (2011). "Genetics and pathogenesis of inflammatory bowel disease". *Nature* 474, pp. 308–317.

Kichaev, Gleb, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc (2014). "Integrating functional data to prioritize causal variants in statistical fine-mapping studies". *PLoS Genet.* 10.10, e1004722.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure (2014). "A general framework for estimating the relative pathogenicity of human genetic variants". *Nat. Genet.* 46.3, pp. 310–315.

Korneliussen, Thorfinn, Anders Albrechtsen, and Rasmus Nielsen (2014). "ANGSD: Analysis of Next Generation Sequencing Data". *BMC Bioinformatics* 15, p. 356.

Kostic, Aleksandar D, Ramnik J Xavier, and Dirk Gevers (2014). "The microbiome in inflammatory bowel disease: current status and the future ahead". *Gastroenterology* 146.6, pp. 1489–1499.

Kunkel, Eric J and Eugene C Butcher (2003). "Plasma-cell homing". *Nat. Rev. Immunol.* 3.10, pp. 822–829.

Lee, James C, Daniele Biasci, Rebecca Roberts, Richard B Gearry, John C Mansfield, Tariq Ahmad, Natalie J Prescott, Jack Satsangi, David C Wilson, et al. (2017). "Genome-wide association study identifies

distinct genetic contributions to prognosis and susceptibility in Crohn's disease". *Nat. Genet.* 49.2, pp. 262–268.

Lee, Mark N, Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M Eisenhaure, Selina H Imboywa, Portia I Chipendo, F Ann Ran, et al. (2014a). "Common genetic variants modulate pathogen-sensing responses in human dendritic cells". *Science* 343.6175, p. 1246980.

Lee, Phil H, Colm O'Dushlaine, Brett Thomas, and Shaun M Purcell (2012a). "INRICH: interval-based enrichment analysis for genome-wide association studies". *Bioinformatics* 28.13, pp. 1797–1799.

Lee, Seunggeun, Michael C Wu, and Xihong Lin (2012b). "Optimal tests for rare variant effects in sequencing association studies". *Biostatistics* 13.4, pp. 762–775.

Lee, Seunggeung, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin (2014b). "Rare-variant association analysis: study designs and statistical tests". *Am. J. Hum. Genet.* 95.1, pp. 5–23.

Leek, Jeffrey T, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". *Nat. Rev. Genet.* 11.10, pp. 733–739.

Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H ODonnell-Luria, James S Ware, Andrew J Hill, et al. (2016). "Analysis of protein-coding genetic variation in 60,706 humans". *Nature* 536.7616, pp. 285–291.

Li, Bingshan and S M Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data". *Am. J. Hum. Genet.* 83, pp. 311–321.

Li, Bingshan, Dajiang J Liu, and Suzanne M Leal (2013a). "Identifying rare variants associated with complex traits via sequencing". *Curr. Protoc. Hum. Genet.* Pp. 1.26.1–1.26.22.

Li, Heng (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". *Bioinformatics* 27.21, pp. 2987–2993.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16, pp. 2078–2079.

Li, Heng, Bob Handsaker, Petr Danecek, Shane McCarthy, and John Marshall (2013b). *SAMtools and BCFtools*. URL: `https://sourceforge.net/projects/samtools/files/samtools/0.1.19/`.

Li, Yun R, Jin Li, Sihai D Zhao, Jonathan P Bradfield, Frank D Mentch, S Melkorka Maggadottir, Cuiping Hou, Debra J Abrams, Diana Chang, et al.

(2015). "Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases". *Nat. Med.* 21.9, pp. 1018–1027.

Li, Yun, Carlo Sidore, Hyun Min Kang, Micheal Boehnke, and Gonçalo R Abecasis (2011). "Low-coverage sequencing: implications for design of complex trait association studies". *Genome Res.* 21, pp. 940–951.

Libioulle, Cécile, Edouard Louis, Sarah Hansoul, Cynthia Sandor, Frédéric Farnir, Denis Franchimont, Séverine Vermeire, Olivier Dewit, Martine De Vos, et al. (2007). "Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4". *PLoS Genet.* 3.4, pp. 0538–0543.

Liu, Jimmy Z and Carl A Anderson (2014). "Genetic studies of Crohn's disease: past, present and future". *Best Pract. Res. Clin. Gastroenterol.* 28.3, pp. 373–386.

Liu, Jimmy Z, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, et al. (2015). "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations". *Nat. Genet.* 47.9, pp. 979–989.

Loftus, Edward V (2004). "Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences". *Gastroenterology* 126.6, pp. 1504–1517.

Louis, Thomas A (1982). "Finding the Observed Information Matrix when Using the EM Algorithm". *J. R. Stat. Soc. Series B Stat. Methodol.* 44.2, pp. 226–233.

Luci, Carmelo, Ana Reynders, Ivaylo I Ivanov, Celine Cognet, Laurent Chiche, Lionel Chasson, Jean Hardwigsen, Esperanza Anguiano, Jacques Banchereau, et al. (2009). "Influence of the transcription factor RORγt on the development of NKp46+ cell populations in gut and skin". *Nat. Immunol.* 10.1, pp. 75–82.

Luo, Yang, Katrina M de Lange, Luke Jostins, Loukas Moutsianas, Joshua Randall, Nicholas A Kennedy, Christopher A Lamb, Shane McCarthy, Tariq Ahmad, et al. (2017). "Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7". *Nat. Genet.* 49.2, pp. 186–192.

M'Koma, Amosy E (2013). "Inflammatory bowel disease: an expanding global health problem". *Clin. Med. Insights Gastroenterol.* 6, pp. 33–47.

Madsen, Bo Eskerod and Sharon R Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic". *PLoS Genet.* 5.2, e1000384.

Maher, Brendan (2008). "The case of the missing heritability". *Nature News Features* 456, pp. 18–21.

Manichaikul, Ani, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* 26.22, pp. 2867–2873.

Manolio, Teri A, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, et al. (2009). "Finding the missing heritability of complex diseases". *Nature* 461.7265, pp. 747–753.

Marceau, Kristine, Minni T B McMaster, Taylor F Smith, Joost G Daams, Catharina E M van Beijsterveldt, Dorret I Boomsma, and Valerie S Knopik (2016). "The Prenatal Environment in Twin Studies: A Review on Chorionicity". *Behav. Genet.* 46.3, pp. 286–303.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies". *Nat. Rev. Genet.* 11.7, pp. 499–511.

Matute, Juan D, Andres A Arias, Nicola A M Wright, Iwona Wrobel, Christopher C M Waterhouse, Xing Jun Li, Christophe C Marchal, Natalie D Stull, David B Lewis, et al. (2009). "A new genetic subgroup of chronic granulomatous disease with autosomal recessive mutations in p40 phox and selective defects in neutrophil NADPH oxidase activity". *Blood* 114.15, pp. 3309–3315.

Maurano, Matthew T, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, et al. (2012). "Systematic localization of common disease-associated variation in regulatory DNA". *Science* 337.6099, pp. 1190–1195.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". *Nat. Genet.* 48.10, pp. 1279–1283.

McLaren, William, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham (2010). "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor". *Bioinformatics* 26.16, pp. 2069–2070.

McVean, Gilean A T, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly (2004). "The fine-scale structure of recombination rate variation in the human genome". *Science* 304, pp. 581–584.

Meynert, Alison M, Morad Ansari, David R FitzPatrick, and Martin S Taylor (2014). "Variant detection sensitivity and biases in whole genome and exome sequencing". *BMC Bioinformatics* 15, p. 247.

Miceli-Richard, C, S Lesage, M Rybojad, A M Prieur, S Manouvrier-Hanu, R Häfner, M Chamaillard, H Zouali, G Thomas, and J P Hugot (2001). "CARD15 mutations in Blau syndrome". *Nat. Genet.* 29.1, pp. 19–20.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, et al. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". *Nat. Genet.* 47.6, pp. 598–606.

Mills, Ryan E, W Stephen Pittard, Julienne M Mullaney, Umar Farooq, Todd H Creasy, Anup A Mahurkar, David M Kemeza, Daniel S Strassler, Chris P Ponting, et al. (2011). "Natural genetic variation caused by small insertions and deletions in the human genome". *Genome Res.* 21.6, pp. 830–839.

Moayyeri, Alireza, Christopher J Hammond, Deborah J Hart, and Timothy D Spector (2013). "The UK Adult Twin Registry (TwinsUK Resource)". *Twin Res. Hum. Genet.* 16.1, pp. 144–149.

Molodecky, Natalie A, Ing Shian Soon, Doreen M Rabi, William A Ghali, Mollie Ferris, Greg Chernoff, Eric I Benchimol, Remo Panaccione, Subrata Ghosh, et al. (2012). "Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review". *Gastroenterology* 142.1, pp. 46–54.

Monteleone, Giovanni, Markus F Neurath, Sandro Ardizzone, Antonio Di Sabatino, Massimo C Fantini, Fabiana Castiglione, Maria L Scribano, Alessandro Armuzzi, Flavio Caprioli, et al. (2015). "Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease". *N. Engl. J. Med.* 372.12, pp. 1104–1113.

Morgenthaler, Stephan and William G Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)". *Mutat. Res.* 615.1-2, pp. 28–56.

Morris, Andrew P and Eleftheria Zeggini (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies". *Genet. Epidemiol.* 34.2, pp. 188–193.

Moutsianas, Loukas and Andrew P Morris (2014). "Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits". *Brief. Funct. Genomics* 13.5, pp. 362–370.

Neale, Benjamin M, Manuel a Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly (2011). "Testing for an unusual distribution of rare variants". *PLoS Genet.* 7.3, e1001322.

Negoro, K, D P B McGovern, Y Kinouchi, S Takahashi, N J Lench, T Shimosegawa, A Carey, L R Cardon, D P Jewell, and D A Van Heel (2003). "Analysis of the IBD5 locus and potential gene-gene interactions in Crohns disease". *Gut* 52, pp. 541–546.

Nejentsev, Sergey, Joanna M M Howson, Neil M Walker, Jeffrey Szeszko, Sarah F Field, Helen E Stevens, Pamela Reynolds, Matthew Hardy, Erna King,

et al. (2007). "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A". *Nature* 450.7171, pp. 887–892.

Ng, Sarah B, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, et al. (2009). "Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes". *Nature* 461.7261, pp. 272–276.

Ng, Sarah B, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, et al. (2010). "Exome sequencing identifies the cause of a mendelian disorder". *Nat. Genet.* 42.1, pp. 30–35.

Novak, Adam M, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, M A Saleh Elmohamed, Sally Guthrie, et al. (2017). "Genome Graphs".

Ogura, Yasunori, Denise K Bonen, Naohiro Inohara, Dan L Nicolae, Felicia F Chen, Richard Ramos, Heidi Britton, Thomas Moran, Reda Karaliuskas, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease". *Nature* 411.6837, pp. 603–606.

Ohmen, Jeffrey D, Hui-Ying Yang, Karen K Yamamoto, Hong-Yu Zhao, Yuanhong Ma, L Gordon Bentley, Zhihan Huang, Scott Gerwehr, Sheila Pressman, et al. (1996). "Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohns disease, but not in ulcerative colitis". *Hum. Mol. Genet.* 5.10, pp. 1679–1683.

Ombrello, Michael J, Elaine F Remmers, Guangping Sun, Alexandra F Freeman, Shrimati Datta, Parizad Torabi-Parizi, Naeha Subramanian, Tom D Bunney, Rhona W Baxendale, et al. (2012). "Cold urticaria, immunodeficiency, and autoimmunity related to PLCG2 deletions". *N. Engl. J. Med.* 366.4, pp. 330–338.

Pals, Steven T, David J J de Gorter, and Marcel Spaargaren (2007). "Lymphoma dissemination: the other face of lymphocyte homing". *Blood* 110.9, pp. 3102–3111.

Pan, Wei (2009). "Asymptotic tests of association with multiple SNPs in linkage disequilibrium". *Genet. Epidemiol.* 33.6, pp. 497–507.

Parkes, M, J Satsangi, G M Lathrop, J I Bell, and D P Jewell (1996). "Susceptibility loci in inflammatory bowel disease". *Lancet* 348.9041, p. 1588.

Parkes, Miles, Jeffrey C Barrett, Natalie J Prescott, Mark Tremelling, Carl A Anderson, Sheila A Fisher, Roland G Roberts, Elaine R Nimmo, Fraser R Cummings, et al. (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility". *Nat. Genet.* 39.7, pp. 830–832.

Parkes, Miles, Adrian Cortes, David A van Heel, and Matthew A Brown (2013). "Genetic insights into common pathways and complex relationships among immune-mediated diseases". *Nat. Rev. Genet.* 14.9, pp. 661–673.

Philpott, Dana J, Matthew T Sorbara, Susan J Robertson, and Kenneth Croitoru (2014). "NOD proteins: regulators of inflammation in health and disease". *Nat. Rev. Immunol.* 14, pp. 9–23.

Picard, Capucine, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Mary Ellen Conley, Charlotte Cunningham-Rundles, Amos Etzioni, Steven M Holland, et al. (2015). "Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015". *J. Clin. Immunol.* 35.8, pp. 696–726.

Pierre, Sandra, Thomas Eschenhagen, Gerd Geisslinger, and Klaus Scholich (2009). "Capturing adenylyl cyclases as potential drug targets". *Nat. Rev. Drug Discov.* 8.4, pp. 321–335.

Prescott, Natalie J, Benjamin Lehne, Kristina Stone, James C Lee, Kirstin Taylor, Jo Knight, Efterpi Papouli, Muddassar M Mirza, Michael A Simpson, et al. (2015). "Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes". *PLoS Genet.* 11.2, e1004955.

Purcell, Shaun M, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". *Nature* 506.7487, pp. 185–190.

Raj, Towfique, Katie Rothamel, Sara Mostafavi, Chun Ye, Mark N Lee, Joseph M Replogle, Ting Feng, Michelle Lee, Natasha Asinovski, et al. (2014). "Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes". *Science* 344.6183, pp. 519–523.

Raker, Verena Katharina, Christian Becker, and Kerstin Steinbrink (2016). "The cAMP Pathway as Therapeutic Target in Autoimmune and Inflammatory Diseases". *Front. Immunol.* 7, p. 123.

Rioux, John D, Ramnik J Xavier, Kent D Taylor, Mark S Silverberg, Philippe Goyette, Alan Huett, Todd Green, Petric Kuballa, M Michael Barmada, et al. (2007). "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis". *Nat. Genet.* 39.5, pp. 596–604.

Risch, Neil and Kathleen Merikangas (1996). "The Future of Genetic Studies of Complex Human Diseases". *Science* 273, pp. 1516–1517.

Risøe, Petter K, Arkady Rutkovskiy, Joanna Ågren, Ingrid B M Kolseth, Signe Flood Kjeldsen, Guro Valen, Jarle Vaage, and Maria K Dahle (2015).

"Higher TNFα responses in young males compared to females are associated with attenuation of monocyte adenylyl cyclase expression". *Hum. Immunol.* 76.6, pp. 427–430.

Rivas, Manuel A, Mélissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, et al. (2011). "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease". *Nat. Genet.* 43.11, pp. 1066–1073.

Rivas, Manuel A, Daniel Graham, Patrick Sulem, Christine Stevens, A Nicole Desch, Philippe Goyette, Daniel Gudbjartsson, Ingileif Jonsdottir, Unnur Thorsteinsdottir, et al. (2016). "A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis". *Nat. Commun.* 7.

Roberts, R L, J E Hollis-Moffatt, R B Gearry, M A Kennedy, M L Barclay, and T R Merriman (2008). "Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population-based cohort". *Genes Immun.* 9.6, pp. 561–565.

Rossi, Adriano G, Judith C Mc Cutcheon, Noemi Roy, Edwin R Chilvers, Christopher Haslett, and Ian Dransfield (1998). "Regulation of Macrophage Phagocytosis of Apoptotic Cells by cAMP1". *J. Immunol.* 160, pp. 3562–3568.

Rueda, Cesar M, Courtney M Jackson, and Claire A Chougnet (2016). "Regulatory T-Cell-Mediated Suppression of Conventional T-Cells and Dendritic Cells by Different cAMP Intracellular Pathways". *Front. Immunol.* 7, p. 216.

Sachidanandam, R, D Weissman, S C Schmidt, J M Kakol, L D Stein, G Marth, S Sherry, J C Mullikin, B J Mortimore, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms". *Nature* 409.6822, pp. 928–933.

Sandborn, William J, Jean Frédéric Colombel, Roberts Enns, Brian G Feagan, Stephen B Hanauer, Ian C Lawrance, Remo Panaccione, Martin Sanders, Stefan Schreiber, et al. (2005). "Natalizumab induction and maintenance therapy for Crohn's disease". *N. Engl. J. Med.* 353.18, pp. 1912–1925.

Sandborn, William J, Christopher Gasink, Long-Long Gao, Marion A Blank, Jewel Johanns, Cynthia Guzzo, Bruce E Sands, Stephen B Hanauer, Stephan Targan, et al. (2012). "Ustekinumab induction and maintenance therapy in refractory Crohn's disease". *N. Engl. J. Med.* 367.16, pp. 1519–1528.

Sandborn, William J, Brian G Feagan, Paul Rutgeerts, Stephen Hanauer, Jean-Frédéric Colombel, Bruce E Sands, Milan Lukas, Richard N Fedorak, Scott Lee, et al. (2013). "Vedolizumab as induction and maintenance therapy for Crohn's disease". *N. Engl. J. Med.* 369.8, pp. 711–721.

Sands, Bruce E and Stacey Grabert (2009). "Epidemiology of Inflammatory Bowel Disease and Overview of Pathogenesis". *Med. Health R. I.* 92.3, pp. 73–77.

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors". *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–5467.

Satsangi, J, Ken I Welsh, Mike Bunce, Cecile Julier, J Mark Farrant, John I Bell, and Derek P Jewell (1996). "Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease". *Lancet* 347, pp. 1212–1217.

Sawa, Shinichiro, Matthias Lochner, Naoko Satoh-Takayama, Sophie Dulauroy, Marion Bérard, Melanie Kleinschek, Daniel Cua, James P Di Santo, and Gérard Eberl (2011). "RORγt+ innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota". *Nat. Immunol.* 12.4, pp. 320–326.

Schaefer, Carl F, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow (2009). "PID: the Pathway Interaction Database". *Nucleic Acids Res.* 37.Database issue, pp. D674–9.

Schloss, Patrick D, Kathryn D Iverson, Joseph F Petrosino, and Sarah J Schloss (2014). "The dynamics of a family's gut microbiota reveal variations on a theme". *Microbiome* 2.1, p. 25.

Seed, Cotton, Alex Bloemendal, Jonathan M Bloom, Jacqueline I Goldstein, Daniel King, Timothy Poterba, and Benjamin M Neale (2017). *Hail: An Open-Source Framework for Scalable Genetic Data Analysis.* URL: `https://github.com/hail-is/hail`.

Seizinger, B R, G A Rouleau, L J Ozelius, A H Lane, A G Faryniarz, M V Chao, S Huson, B R Korf, D M Parry, et al. (1987). "Genetic linkage of von Recklinghausen neurofibromatosis to the nerve growth factor receptor gene". *Cell* 49.5, pp. 589–594.

Shah, T S, J Z Liu, J A B Floyd, J A Morris, N Wirth, J C Barrett, and C A Anderson (2012). "optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants". *Bioinformatics* 28.12, pp. 1598–1603.

Shih, David Q, Lola Y Kwan, Valerie Chavez, Offer Cohavy, Rivkah Gonsky, Elmer Y Chang, Christopher Chang, Charles O Elson, and Stephan R Targan (2009). "Microbial Induction of Inflammatory Bowel Disease Associated Gene TL1A (TNFSF15) in Antigen Presenting Cells". *Eur. J. Immunol.* 39.11, pp. 3239–3250.

Siddique, T, D A Figlewicz, M A Pericak-Vance, J L Haines, G Rouleau, A J Jeffers, P Sapp, W Y Hung, J Bebout, and D McKenna-Yasek (1991). "Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity". *N. Engl. J. Med.* 324.20, pp. 1381–1384.

Silverberg, Mark S, Lucia Mirea, Shelley B Bull, Janet E Murphy, A Hillary Steinhart, Gordon R Greenberg, Robin S McLeod, Zane Cohen, Judith A Wade, and Katherine A Siminovitch (2003). "A population- and family-based study of Canadian families reveals association of HLA DRB1*0103 with colonic involvement in inflammatory bowel disease". *Inflamm. Bowel Dis.* 9.1, pp. 1–9.

Silverberg, Mark S, Judy H Cho, John D Rioux, Dermot P B McGovern, Jing Wu, Vito Annese, Jean-Paul Achkar, Philippe Goyette, Regan Scott, et al. (2009). "Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study". *Nat. Genet.* 41.2, pp. 216–220.

Sims, David, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting (2014). "Sequencing depth and coverage: key considerations in genomic analyses". *Nat. Rev. Genet.* 15.2, pp. 121–132.

Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen (2012). "Association testing for next-generation sequencing data using score statistics". *Genet. Epidemiol.* 36.5, pp. 430–437.

Spain, Sarah L and Jeffrey C Barrett (2015). "Strategies for fine-mapping complex traits". *Hum. Mol. Genet.* 24.R1, R111–9.

Speer, M C, L H Yamaoka, J H Gilchrist, C P Gaskell, J M Stajich, J M Vance, A Kazantsev, A A Lastra, C S Haynes, and J S Beckmann (1992). "Confirmation of genetic heterogeneity in limb-girdle muscular dystrophy: linkage of an autosomal dominant form to chromosome 5q". *Am. J. Hum. Genet.* 50.6, pp. 1211–1217.

Stephens, Philip J, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, et al. (2012). "The landscape of cancer genes and mutational processes in breast cancer". *Nature* 486.7403, pp. 400–404.

Sun, Jianping, Yingye Zheng, and Li Hsu (2013). "A unified mixed-effects model for rare-variant association in sequencing studies". *Genet. Epidemiol.* 37.4, pp. 334–344.

Suskind, David L, Mitchell J Brittnacher, Ghassan Wahbeh, Michele L Shaffer, Hillary S Hayden, Xuan Qin, Namita Singh, Christopher J Damman, Kyle R Hager, et al. (2015). "Fecal Microbial Transplant Effect on Clinical Outcomes and Fecal Microbiome in Active Crohns Disease". *Inflamm. Bowel Dis.* 21.3, pp. 556–563.

Tarazona-Santos, Eduardo, Moara Machado, Wagner C S Magalhães, Renee Chen, Fernanda Lyon, Laurie Burdett, Andrew Crenshaw, Cristina Fabbri, Latife Pereira, et al. (2013). "Evolutionary dynamics of the human NADPH oxidase genes CYBB, CYBA, NCF2, and NCF4: functional implications". *Mol. Biol. Evol.* 30.9, pp. 2157–2167.

The 1000 Genomes Project Consortium (2010). *1000 Genomes Project Phase I.* URL: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz.

– (2011). *1000 Genomes Project Phase II.* URL: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz.

The International HapMap Consortium (2005). "A haplotype map of the human genome". *Nature* 437.7063, pp. 1299–1320.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". *Nature* 447.7145, pp. 661–678.

Tiemessen, Machteld M, Ann L Jagger, Hayley G Evans, Martijn J C van Herwijnen, Susan John, and Leonie S Taams (2007). "CD4+CD25+Foxp3+ regulatory T cells induce alternative activation of human monocytes/macrophages". *Proc. Natl. Acad. Sci. U. S. A.* 104.49, pp. 19446–19451.

Travis, Mark A, Boris Reizis, Andrew C Melton, Emma Masteller, Qizhi Tang, John M Proctor, Yanli Wang, Xin Bernstein, Xiaozhu Huang, et al. (2007). "Loss of integrin alpha(v)beta8 on dendritic cells causes autoimmunity and colitis in mice". *Nature* 449.7160, pp. 361–365.

Trynka, Gosia, Karen A Hunt, Nicholas A Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F Bakker, Maria Teresa Bardella, Leena Bhaw-rosun, et al. (2011). "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease". *Nat. Genet.* 43.12, pp. 1193–1201.

Tsui, Lap-Chee, Manuel Buchwald, David Barker, Jeffrey C Braman, Robert Knowlton, James W Schumm, Hans Eiberg, Jan Moher, Dara Kennedy, et al. (1985). "Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic Marker". *Science* 230, pp. 1054–1057.

Uhlig, Holm H (2013). "Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease". *Gut* 62.12, pp. 1795–1805.

Van Limbergen, Johan, David C Wilson, and Jack Satsangi (2009). "The genetics of Crohn's disease". *Annu. Rev. Genomics Hum. Genet.* 10, pp. 89–116.

Vance, J M, G A Nicholson, L H Yamaoka, J Stajich, C S Stewart, M C Speer, W Y Hung, A D Roses, D Barker, and M A Pericak-Vance (1989). "Linkage of Charcot-Marie-Tooth neuropathy type 1a to chromosome 17". *Exp. Neurol.* 104.2, pp. 186–189.

Vermeire, Séverine, Sharon O'Byrne, Mary Keir, Marna Williams, Timothy T Lu, John C Mansfield, Christopher A Lamb, Brian G Feagan, Julian Panes, et al.

(2014). "Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial". *Lancet* 384.9940, pp. 309–318.

Visscher, Peter M, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin (2006). "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings". *PLoS Genet.* 2.3, e41.

Walter, Klaudia, Josine L Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R B Perry, Changjiang Xu, Marta Futema, et al. (2015). "The UK10K project identifies rare variants in health and disease". *Nature* 526.7571, pp. 82–90.

Wang, Guoxing, Ana C Abadía-Molina, Scott B Berger, Xavier Romero, Michael S O'Keeffe, Domingo I Rojas-Barros, Marta Aleman, Gongxian Liao, Elena Maganto-García, et al. (2012). "Cutting edge: Slamf8 is a negative regulator of Nox2 activity in macrophages". *J. Immunol.* 188.12, pp. 5829–5832.

Wang, Guoxing, Boaz J van Driel, Gongxian Liao, Michael S O'Keeffe, Peter J Halibozek, Jacky Flipse, Burcu Yigit, Veronica Azcutia, Francis W Luscinskas, et al. (2015). "Migration of myeloid cells during inflammation is differentially regulated by the cell surface receptors Slamf1 and Slamf8". *PLoS One* 10.3, e0121968.

Weaver, Casey T and Robin D Hatton (2009). "Interplay between the TH17 and TReg cell lineages: a (co-)evolutionary perspective". *Nat. Rev. Immunol.* 9.12, pp. 883–889.

Weber, J L and P E May (1989). "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction". *Am. J. Hum. Genet.* 44.3, pp. 388–396.

Wellcome Trust Case Control Consortium, Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna M M Howson, Adam Auton, et al. (2012). "Bayesian refinement of association signals for 14 loci in 3 common diseases". *Nat. Genet.* 44.12, pp. 1294–1301.

Westra, Harm-Jan, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, et al. (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations". *Nat. Genet.* 45.10, pp. 1238–1243.

Willer, Cristen J, Yun Li, and Gonçalo R Abecasis (2010). "METAL: fast and efficient meta-analysis of genomewide association scans". *Bioinformatics* 26.17, pp. 2190–2191.

Withers, David R, Matthew R Hepworth, Xinxin Wang, Emma C Mackley, Emily E Halford, Emma E Dutton, Clare L Marriott, Verena Brucklacher-Waldert, Marc Veldhoen, et al. (2016). "Transient inhibition

of ROR-[gamma]t therapeutically limits intestinal inflammation by reducing TH17 cells and preserving group 3 innate lymphoid cells". *Nat. Med.* 22.3, pp. 319–323.

Worthington, John J, Aoife Kelly, Catherine Smedley, David Bauché, Simon Campbell, Julien C Marie, and Mark A Travis (2015). "Integrin αvβ8-Mediated TGF-β Activation by Effector Regulatory T Cells Is Essential for Suppression of T-Cell-Mediated Inflammation". *Immunity* 42.5, pp. 903–915.

Wright, Caroline F, Tomas W Fitzgerald, Wendy D Jones, Stephen Clayton, Jeremy F McRae, Margriet van Kogelenberg, Daniel A King, Kirsty Ambridge, Daniel M Barrett, et al. (2015). "Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data". *Lancet* 385.9975, pp. 1305–1314.

Wright, Fred A, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, et al. (2014). "Heritability and genomics of gene expression in peripheral blood". *Nat. Genet.* 46.5, pp. 430–437.

Wu, Michael C, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test". *Am. J. Hum. Genet.* 89.1, pp. 82–93.

Yamazaki, Keiko, Masakazu Takazoe, Torao Tanaka, Toshiki Kazumori, and Yusuke Nakamura (2002). "Absence of mutation in the NOD2 / CARD15 gene among 483 Japanese patients with Crohns disease". *J. Hum. Genet.* 47, pp. 469–472.

Yamazaki, Keiko, Masakazu Takazoe, Torao Tanaka, Toshiki Ichimori, Susumu Saito, Aritoshi Iida, Yoshihiro Onouchi, Akira Hata, and Yusuke Nakamura (2004). "Association analysis of SLC22A4 , SLC22A5 and DLG5 in Japanese patients with Crohn disease". *J. Hum. Genet.* 49, pp. 664–668.

Yamazaki, Keiko, Dermot McGovern, Jiannis Ragoussis, Marta Paolucci, Helen Butler, Derek Jewell, Lon Cardon, Masakazu Takazoe, Torao Tanaka, et al. (2005). "Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease". *Hum. Mol. Genet.* 14.22, pp. 3499–3506.

Yamazaki, Keiko, Junji Umeno, Atsushi Takahashi, Atsushi Hirano, Todd Andrew Johnson, Natsuhiko Kumasaka, Takashi Morizono, Naoya Hosono, Takaaki Kawaguchi, et al. (2013). "A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population". *Gastroenterology* 144.4, pp. 781–788.

Yang, Suk-Kyun, Myunghee Hong, Wanting Zhao, Yusun Jung, Naeimeh Tayebi, Byong Duk Ye, Kyung-Jo Kim, Sang Hyoung Park, Inchul Lee, et al. (2013). "Genome-Wide Association Study of Ulcerative Colitis in Koreans Suggests

Extensive Overlapping of Genetic Susceptibility With Caucasians". *Inflamm. Bowel Dis.* 19.5, pp. 954–966.

Yang, Suk-Kyun, Myunghee Hong, Jiwon Baek, Hyunchul Choi, Wanting Zhao, Yusun Jung, Talin Haritunians, Byong Duk Ye, Kyung-Jo Kim, et al. (2014a). "A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia". *Nat. Genet.* 46.9, pp. 1017–1020.

Yang, Suk-Kyun, Myunghee Hong, Wanting Zhao, Yusun Jung, Jiwon Baek, Naeimeh Tayebi, Kyung Mo Kim, Byong Duk Ye, Kyung-Jo Kim, et al. (2014b). "Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations". *Gut* 63.1, pp. 80–87.

Yang, Yi, Miriam B Torchinsky, Michael Gobert, Huizhong Xiong, Mo Xu, Jonathan L Linehan, Francis Alonzo, Charles Ng, Alessandra Chen, et al. (2014c). "Focused specificity of intestinal TH17 cells towards commensal bacterial antigens". *Nature* 510.7503, pp. 152–156.

Ye, Chun Jimmie, Ting Feng, Ho-Keun Kwon, Towfique Raj, Michael T Wilson, Natasha Asinovski, Cristin McCabe, Michelle H Lee, Irene Frohlich, et al. (2014). "Intersection of population variation and autoimmunity genetics in human T cell activation". *Science* 345.6202, p. 1254665.

Zeissig, Yvonne, Britt-Sabina Petersen, Snezana Milutinovic, Esther Bosse, Gabriele Mayr, Kenneth Peuker, Jelka Hartwig, Andreas Keller, Martina Kohl, et al. (2015). "XIAP variants in male Crohn's disease". *Gut* 64.1, pp. 66–76.

Zeller, Tanja, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, Marine Germain, Karl Lackner, et al. (2010). "Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility". *PLoS One* 5.5, e10693.

Zhernakova, Alexandra, Cleo C van Diemen, and Cisca Wijmenga (2009). "Detecting shared pathogenesis from the shared genetics of immune-related diseases". *Nat. Rev. Genet.* 10.1, pp. 43–55.

Zhernakova, Daria, Patrick Deelen, Martijn Vermaat, Maarten van Iterson, Michiel van Galen, Wibowo Arindrarto, Peter van t Hof, Hailiang Mei, Freerk van Dijk, et al. (2015). "Hypothesis-free identification of modulators of genetic risk factors".

Zhou, Qing, Geun-Shik Lee, Jillian Brady, Shrimati Datta, Matilda Katan, Afzal Sheikh, Marta S Martins, Tom D Bunney, Brian H Santich, et al. (2012). "A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency". *Am. J. Hum. Genet.* 91.4, pp. 713–720.

Zimmerman, Noah P, Suresh N Kumar, Jerrold R Turner, and Michael B Dwinell (2012). "Cyclic AMP dysregulates intestinal epithelial cell restitution through PKA and RhoA". *Inflamm. Bowel Dis.* 18.6, pp. 1081–1091.

Zuk, Or, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander (2014). "Searching for missing heritability: designing rare variant association studies". *Proc. Natl. Acad. Sci. U. S. A.* 111, E455–64.

Appendix A

Members of the UK IBD Genetics Consortium who contributed to these studies

Ailsa Hart
Department of Medicine, St Mark's Hospital, Harrow, Middlesex, UK

Alison Simmons
Translational Gastroenterology Unit, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DS, UK; Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK

Carl A. Anderson
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Cathryn Edwards
Department of Gastroenterology, Torbay Hospital, Torbay, Devon, UK

Charlie W. Lees
Gastrointestinal Unit, Wester General Hospital University of Edinburgh, Edinburgh, UK

Chris Hawkey
Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK

Christopher A. Lamb
Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne

Christopher G. Mathew
Department of Medical and Molecular Genetics, Faculty of Life Science and Medicine,

King's College London, Guy's Hospital, London, UK; Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of Witwatersrand, South Africa.

Craig Mowat
Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK

Daniel L. Rice
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

David C. Wilson
Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK; Child Life and Health, University of Edinburgh, Edinburgh, Scotland, UK

Elaine R. Nimmo
Gastrointestinal Unit, Wester General Hospital University of Edinburgh, Edinburgh, UK

Eva Goncalves Serra
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Graham Heap
Precision Medicine Exeter, University of Exeter, Exeter, UK; IBD Pharmacogenetics, Royal Devon and Exeter Foundation Trust, Exeter, UK

Holm Uhlig
Translational Gastroenterology Unit and the Department of Paediatrics, University of Oxford, Oxford, United Kingdom

Jack Satsangi
Gastrointestinal Unit, Wester General Hospital University of Edinburgh, Edinburgh, UK

James C. Lee
Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK

Javier Gutierrez-Achury
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Jeffrey C. Barrett
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Jeremy Sanderson
Guys and St Thomas NHS Foundation Trust, St Thomas Hospital, Department of Gastroenterology, London, UK

John C. Mansfield
Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK

Joshua Randall
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Katrina M. de Lange
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Loukas Moutsianas
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Luke Jostins
Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK; Christ Church, University of Oxford, St Aldates, UK

Mark Tremelling
Gastroenterology and General Medicine, Norfolk and Norwich University Hospital, Norwich, UK

Martin Pollard
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Miles Parkes
Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK

Natalie J. Prescott
Department of Medical and Molecular Genetics, Faculty of Life Science and Medicine, King's College London, Guy's Hospital, London, UK

Nicholas A. Kennedy
Precision Medicine Exeter, University of Exeter, Exeter, UK; IBD Pharmacogenetics, Royal Devon and Exeter Foundation Trust, Exeter, UK

Paul Henderson
Department of Child Life and Health, University of Edinburgh, Edinburgh, UK; De-

partment of Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children,Edinburgh, UK

Sam Nichols
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Shane McCarthy
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Sun-Gou Ji
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Tariq Ahmad
Precision Medicine Exeter, University of Exeter, Exeter, UK; IBD Pharmacogenetics, Royal Devon and Exeter Foundation Trust, Exeter, UK

William G. Newman
Genetic Medicine, Manchester Academic Health Science Centre, Manchester, UK; The Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK

Yang Luo
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; Division of Genetics and Rheumatology, Brigham and Womens Hospital, Harvard Medical School, Boston, MA, USA; Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Appendix B

Meta-analysis association statistics at all 241 known and novel loci

In the following table we report the top SNP as that with most significant association described in any of the listed studies (including this one), for all 241 known and novel IBD-associated loci. Association statistics for this study are included for each locus, based on the reported top SNP. A locus is defined around each reported SNP as the right- and left-most variants in linkage disequilibrium ≥ 0.6 to the top SNP, as calculated using the GBR and CEU samples from the 1000 Genomes Phase 3 dataset.

I use the following definitions in this table:

- "FM" in place of a PMID refers to the Huang et al. (2015) bioRxiv preprint.
- "P_{Meta}" refers to the association p-value obtained in Chapter 4.
- "P_{Het}" refers to the meta-analysis heterogeneity p-value.
- "Trait" describes the phenotype reported by Liu et al. (2015), where applicable, and the one reported by the original paper (cited) if the locus was not replicated. In cases where two (or more) loci from Liu et al. (2015) are merged in this table, and for which different traits are reported, we keep both traits (e.g. the 1:20171860 locus)
- "Implicated gene" indicates those loci where a specific gene has been confidently implicated by fine-mapping, eQTL, or targeted sequencing studies.

Table B.1: Meta-analysis association statistics at all 241 known and novel loci. Table compiled in collaboration with Loukas Moutsianas.

Chr	Position	Rsid	LD_left	LD_right	P-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
1	1247494	rs12103	1193586	1346703	8.00×10^{13}	23128233	IBD	2.09×10^{07}	0.451	
1	2501338	rs10797432	2470671	2552217	3.00×10^{12}	23128233	UC	1.20×10^{10}	0.220	
1	8022197	rs3766606	7809383	8189936	8.84×10^{18}	26192919	IBD	1.35×10^{12}	0.092	
1	20171860	rs6426833	20133810	20238343	8.56×10^{69}	FM	UC, IBD	3.04×10^{42}, 1.01×10^{19}	0.049, 0.057	
1	22702231	rs12568930	22652911	22731392	1.00×10^{17}	23128233	IBD	7.53×10^{15}	0.074	
1	63049593	rs1748195	62900811	63202258	7.13×10^{08}	26192919	CD	3.75×10^{06}	0.377	
1	67707690	rs11581607	67206042	68100675	8.76×10^{175}	FM	IBD	4.59×10^{111}	0.078	*IL23R*
1	70995562	rs2651244	70991829	71035970	2.00×10^{08}	23128233	IBD	7.91×10^{01}	0.444	
1	78623626	rs17391694	78349467	78623626	2.62×10^{09}	26192919	CD	9.90×10^{06}	0.281	
1	92554283	rs34856868	92362377	93035991	9.80×10^{09}	26192919	IBD	8.86×10^{03}	0.771	
1	101466054	rs11583043	101293753	101587145	1.92×10^{06}	26192919	IBD	8.91×10^{04}	0.139	
1	114165757	rs2636006	114165757	114447914	2.34×10^{42}	FM	IBD	2.89×10^{03}	0.018	*PTPN22*
1	120451190	rs3897478	120437718	120638503	2.00×10^{11}	23128233	CD	1.66×10^{04}	0.202	
1	151801680	rs4845604	151801680	151801680	1.21×10^{17}	26192919	IBD	7.09×10^{14}	0.276	
1	155230131	rs3180018	155053719	156011444	2.00×10^{13}	21102463	CD	1.87×10^{07}	0.007	
1	159799910	rs34687326	159799910	159799910	1.06×10^{08}	NOVEL	CD	1.06×10^{08}	0.420	*SLAMF8*
1	160856964	rs4656958	160837622	160919496	2.05×10^{11}	26192919	IBD	1.68×10^{08}	0.197	
1	161479745	rs1801274	161460211	161638410	2.00×10^{38}	23128233	IBD	9.34×10^{14}	0.008	*FCGR2A*
1	169519049	rs6025	169090748	169733566	2.51×10^{08}	26192919	IBD	8.92×10^{05}	0.443	
1	172853460	rs7517810	172803958	172871681	4.53×10^{22}	26192919	CD	1.55×10^{21}	0.319	
1	186875459	rs10798069	186862512	186967702	2.49×10^{07}	26192919	IBD	2.29×10^{02}	0.965	
1	191559356	rs10801047	191438042	191795390	3.00×10^{08}	17554261	CD	1.75×10^{01}	0.940	
1	197701279	rs2488397	197342380	197822994	5.64×10^{18}	FM	IBD	4.05×10^{11}	0.033	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	P-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
1	198598663	rs7555082	198598389	198730097	2.54×10^{09}	26192919	IBD	3.82×10^{03}	0.760	
1	200101920	rs2816958	200074018	200105746	2.00×10^{17}	23128233	IBD	8.71×10^{08}	0.580	
1	200877562	rs7554511	200874229	201025850	3.20×10^{44}	26192919	IBD	1.00×10^{21}	0.332	
1	206939904	rs3024505	206939904	206969339	2.99×10^{50}	26192919	IBD	6.04×10^{31}	0.271	*IL10*
1	209970610	rs59043219	209965283	210019701	1.09×10^{08}	NOVEL	IBD	1.09×10^{08}	0.351	
2	5664008	rs11894081	5651813	5693857	4.00×10^{09}	23266558	CD	2.27×10^{01}	0.744	
2	25097644	rs13407913	25075281	25506107	1.69×10^{20}	26192919	IBD	3.39×10^{07}	0.898	
2	27730940	rs1260326	27598097	27752871	1.74×10^{21}	26192919	CD	6.32×10^{11}	0.186	
2	28614794	rs925255	28602911	28647084	3.00×10^{15}	23128233	IBD	5.83×10^{11}	0.229	
2	43806918	rs10495903	43451957	43864089	2.00×10^{14}	21102463	IBD	5.53×10^{09}	0.497	
2	61204856	rs7608910	61186829	61231014	2.60×10^{36}	26192919	IBD	2.72×10^{28}	0.981	
2	62552321	rs11679753	62549284	62723474	2.65×10^{12}	FM	CD	2.83×10^{04}	0.898	
2	65667272	rs6740462	65557287	65693513	5.59×10^{12}	26192919	IBD	6.05×10^{04}	0.284	
2	102939036	rs13001325	102610642	103176797	2.51×10^{23}	FM	IBD	4.43×10^{11}	0.732	*IL18RAP*
2	145492382	rs11681525	145417530	145627269	4.08×10^{11}	26192919	CD	5.19×10^{07}	0.640	
2	160794008	rs4664304	160707866	160879971	2.61×10^{08}	26192919	IBD	8.38×10^{02}	0.314	
2	163110536	rs2111485	162992004	163237390	4.60×10^{09}	26192919	IBD	1.16×10^{05}	0.426	*IFIH1*
2	182308352	rs6740847	182308352	182334753	1.22×10^{13}	NOVEL	IBD	1.22×10^{13}	0.003	*ITGA4*
2	187576378	rs144344067	187497988	187684426	1.29×10^{08}	NOVEL	IBD	1.29×10^{08}	0.118	
2	191931464	rs1517352	191907655	191972789	3.87×10^{14}	26192919	IBD	1.66×10^{04}	0.049	
2	198896895	rs6738825	198386175	198954831	4.00×10^{09}	21102463	UC	1.32×10^{01}	0.639	
2	199560757	rs17229679	199362905	200152198	7.05×10^{20}	FM	UC	3.18×10^{04}	0.969	
2	204592021	rs3116494	204585502	204649276	5.11×10^{07}	26192919	IBD	9.98×10^{05}	0.610	
2	219151218	rs2382817	218941916	219192755	1.13×10^{13}	26192919	IBD	1.90×10^{07}	0.158	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
2	228660112	rs111781203	228646855	228664568	2.16×10^{10}	26192919	IBD	5.09×10^{05}	0.820	
2	228670476	rs1811711	228670387	228671736	6.09×10^{09}	NOVEL	UC	6.09×10^{09}	0.870	
2	231097129	rs6716753	231083171	231187167	1.98×10^{17}	26192919	CD	1.46×10^{13}	0.822	*SP140*
2	234161448	rs6752107	234143048	234208258	1.42×10^{73}	FM	IBD	3.13×10^{28}	0.607	*ATG16L1*
2	241574401	rs4676408	241563739	241608453	2.59×10^{30}	FM	IBD, UC	7.63×10^{15}, 1.19×10^{17}	0.323, 0.597	
2	242484701	rs76527535	242471730	242489453	2.87×10^{08}	NOVEL	IBD	2.87×10^{08}	0.310	
2	242737341	rs35320439	242724543	242740537	1.73×10^{05}	26192919	IBD	2.38×10^{01}	0.051	
3	18767404	rs4256159	18610175	18825669	9.00×10^{15}	23128233	CD	2.90×10^{05}	0.032	
3	46457412	rs113010081	46183180	46461783	4.21×10^{08}	26192919	IBD	1.08×10^{03}	0.677	
3	49721532	rs3197999	48446237	49731861	1.55×10^{52}	26192919	IBD	1.43×10^{33}	0.661	*MST1*
3	53062661	rs9847710	52962681	53174638	1.00×10^{08}	23128233	UC	6.57×10^{05}	0.033	
3	53133149	rs2581828	53103155	53174638	6.46×10^{09}	NOVEL	CD	6.46×10^{09}	0.620	
3	71175495	rs2593855	71163596	71191350	2.54×10^{09}	NOVEL	IBD	2.54×10^{09}	0.432	
3	101023748	rs503734	100910925	101270494	2.67×10^{08}	NOVEL	IBD	2.67×10^{08}	0.718	
3	101569726	rs616597	101560223	101576029	9.34×10^{06}	26192919	UC	4.50×10^{02}	0.640	
3	141105570	rs724016	141072289	141154542	3.36×10^{06}	26192919	CD	7.60×10^{08}	0.969	
3	188401160	rs56116661	188398642	188404669	5.67×10^{10}	NOVEL	CD	5.67×10^{10}	0.592	
4	3444503	rs2073505	3444503	3446754	1.46×10^{07}	26192919	IBD	2.16×10^{04}	0.040	
4	26132361	rs4692386	26132361	26132361	1.21×10^{08}	26192919	IBD	1.40×10^{04}	0.136	
4	38325036	rs6856616	38325036	38345898	4.00×10^{14}	23850713	IBD	1.42×10^{04}	0.751	
4	38588453	rs11734570	38581581	38588453	4.80×10^{08}	NOVEL	IBD	4.80×10^{08}	0.053	
4	48363245	rs7438704	48344930	48430354	3.42×10^{11}	26192919	CD	1.26×10^{05}	0.012	
4	74857708	rs2472649	74778235	74875713	3.00×10^{08}	23128233	IBD	1.86×10^{03}	0.247	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
4	102865304	rs13126505	102702364	103547967	1.57×10^{14}	26192919	CD, UC	3.79×10^{08}, 1.11×10^{02}	0.029, 0.419	
4	106075498	rs2189234	106048360	106271522	1.95×10^{10}	26192919	UC	1.47×10^{05}	0.156	
4	123351431	rs1479918	123023206	123558828	1.07×10^{13}	FM	IBD	1.10×10^{07}	0.800	
5	583442	rs4957048	523995	697372	1.00×10^{09}	20228799	UC	1.64×10^{07}	0.682	
5	10695526	rs2930047	10670197	10759514	9.78×10^{14}	26192919	IBD	5.08×10^{06}	0.764	
5	35876274	rs3194051	35815846	35924748	4.00×10^{08}	21297633	UC	1.78×10^{03}	0.073	
5	38867732	rs395157	38861514	38867732	2.22×10^{20}	26192919	IBD	4.63×10^{10}	0.445	
5	40410584	rs11742570	40219972	40623346	2.00×10^{82}	23128233	IBD	3.64×10^{40}	0.285	
5	55438851	rs10065637	55436851	55444683	4.00×10^{12}	23128233	IBD	1.09×10^{05}	0.472	
5	71693899	rs4703855	71683885	71694246	7.16×10^{11}	26192919	IBD	1.41×10^{05}	0.587	
5	72539850	rs34804116	72502029	72559339	1.21×10^{12}	FM	CD	6.09×10^{06}	0.160	
5	96252803	rs1363907	96200770	96373750	4.87×10^{15}	26192919	IBD	1.10×10^{10}	0.767	*ERAP2*
5	101946798	rs7705924	101629599	101978963	2.00×10^{08}	22412388	CD	5.96×10^{01}	0.256	
5	130017287	rs4836519	129723552	130197645	4.00×10^{10}	23128233	IBD	1.49×10^{03}	0.266	
5	131770805	rs2188962	130370970	131833599	9.53×10^{53}	FM	IBD	3.67×10^{27}	0.805	
5	134443606	rs254560	134422204	134453390	8.55×10^{10}	26192919	IBD	1.63×10^{06}	0.003	
5	141513204	rs6863411	141435466	141543989	2.41×10^{14}	26192919	IBD	5.02×10^{10}	0.867	
5	149605994	rs17656349	149592334	149626547	1.54×10^{08}	NOVEL	UC	1.54×10^{08}	0.125	
5	150277909	rs11741861	149737015	150723626	3.00×10^{37}	23128233	IBD	3.28×10^{15}	0.305	
5	158827769	rs56167332	158687404	158856513	7.17×10^{50}	26192919	IBD	2.52×10^{38}	0.415	
5	172324978	rs564349	172324978	172324978	1.54×10^{07}	26192919	IBD	2.35×10^{06}	0.749	
5	173279842	rs359457	173269669	173393823	3.00×10^{12}	21102463	IBD	6.42×10^{04}	0.513	
5	176788570	rs4976646	176781209	176806636	3.23×10^{12}	26192919	IBD	3.99×10^{09}	0.229	
6	382559	rs7773324	382537	391623	5.84×10^{09}	26192919	IBD	1.69×10^{01}	0.291	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	P-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
6	3433318	rs17309827	3416922	3445671	7.00×10^{09}	21102463	IBD	3.94×10^{02}	0.131	
6	14719496	rs17119	14711961	14734463	2.18×10^{11}	26192919	IBD	9.35×10^{11}	0.320	
6	19781009	rs113986290	19716349	19825913	7.59×10^{09}	NOVEL	UC	7.59×10^{09}	0.150	
6	20812588	rs9358372	20640424	20891791	9.00×10^{14}	23128233	CD	1.25×10^{08}	0.113	
6	21430730	rs71559680	21426743	21444899	1.78×10^{15}	26192919	CD	2.08×10^{08}	0.607	
6	32612397	rs6927022	31011373	32778656	5.00×10^{133}	23128233	IBD	2.28×10^{31}	0.039	
6	42007403	rs67289879	41999358	42008203	3.04×10^{08}	NOVEL	IBD	3.04×10^{08}	0.232	
6	43795968	rs943072	43784803	43801582	2.00×10^{10}	21297633	UC	1.81×10^{04}	0.127	
6	90973159	rs1847472	90809560	91012867	2.00×10^{10}	23128233	IBD	1.79×10^{09}	0.401	
6	106435269	rs7746082	106435025	106530330	1.29×10^{20}	26192919	IBD	9.03×10^{13}	0.428	
6	111848191	rs3851228	111426965	112213778	1.50×10^{15}	26192919	IBD	3.91×10^{10}	0.501	
6	116768917	rs2858829	116768917	116818887	8.00×10^{10}	25082827	UC	8.69×10^{06}	0.044	
6	127456122	rs9491697	127419737	127532381	4.00×10^{10}	23128233	CD	2.24×10^{09}	0.765	
6	128277151	rs9491891	128215237	128339699	9.39×10^{17}	FM	CD	6.84×10^{06}	0.732	
6	138006504	rs6920220	137959235	138166068	1.00×10^{21}	23128233	IBD	1.00×10^{08}	0.122	
6	143898894	rs12199775	143865221	143924048	2.00×10^{08}	23128233	IBD	2.91×10^{05}	0.193	
6	149577079	rs7758080	149577079	149595505	1.15×10^{07}	26192919	IBD	1.09×10^{04}	0.220	
6	159490436	rs212388	159472149	159515309	1.80×10^{16}	26192919	CD	9.52×10^{11}	0.575	
6	167373547	rs1819333	167360389	167544278	7.00×10^{21}	23128233	IBD	8.81×10^{15}	0.093	
7	2789880	rs798502	2752152	2912928	6.00×10^{17}	23128233	IBD	1.60×10^{07}	0.221	
7	6545188	rs11768365	6497501	6545335	3.88×10^{08}	NOVEL	IBD	3.88×10^{08}	0.935	
7	17442679	rs1077773	17430004	17450531	5.96×10^{09}	26192919	UC	6.25×10^{06}	0.773	
7	20577298	rs14916903	20577298	20581696	3.26×10^{08}	NOVEL	IBD	3.26×10^{08}	0.259	*ITGB8*
7	26892440	rs10486483	26694926	27248891	1.55×10^{11}	26192919	CD, UC	4.32×10^{04}, 9.60×10^{01}	0.197, 0.391	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
7	28190730	rs56691573	28141159	28231103	2.22×10^{12}	FM	CD	7.01×10^{01}	1.000	
7	50304461	rs1456896	50096251	50323456	7.00×10^{15}	23128233	IBD	4.50×10^{11}	0.266	
7	98759117	rs9297145	98717625	98797924	8.00×10^{12}	23128233	IBD	8.37×10^{09}	0.414	
7	100315517	rs1734907	100219167	100433794	2.00×10^{13}	23128233	IBD	1.61×10^{05}	0.121	
7	107480315	rs4380874	107435187	107584780	2.00×10^{26}	23128233	UC	9.07×10^{21}	0.320	
7	116892846	rs38904	116889718	116917118	1.00×10^{08}	23128233	UC	6.28×10^{06}	0.695	
7	128573967	rs4728142	128567032	128581835	1.92×10^{14}	26192919	UC	3.23×10^{10}	0.491	
7	148220448	rs2538470	148211140	148253738	3.00×10^{11}	26192919	IBD	3.77×10^{05}	0.329	
7	148435339	rs243505	148395484	148576042	3.04×10^{10}	NOVEL	IBD	3.04×10^{10}	0.671	
8	27227554	rs17057051	27195121	27238052	5.50×10^{08}	26192919	IBD	9.90×10^{04}	0.235	
8	49129242	rs7011507	49047317	49206289	2.03×10^{08}	26192919	IBD	5.84×10^{06}	0.061	
8	74007347	rs12677663	73995603	74013420	2.00×10^{08}	22412388	CD	6.38×10^{02}	0.306	
8	90875918	rs7015630	90854846	90877546	1.00×10^{08}	23128233	IBD	9.98×10^{04}	0.459	
8	126534671	rs921720	126527765	126541090	8.00×10^{20}	23128233	IBD	2.73×10^{12}	0.458	
8	129567181	rs6651252	129503666	129571140	4.00×10^{18}	21102463	IBD	5.63×10^{07}	0.235	
8	130624105	rs1991866	130589676	130624661	2.00×10^{09}	23128233	IBD	1.22×10^{06}	0.043	
9	4981601	rs75900472	4980756	4984530	4.70×10^{48}	26192919	IBD	1.24×10^{28}	0.273	
9	34736158	rs9408254	34707814	34768894	1.95×10^{08}	26974007	CD	1.47×10^{04}	0.038	
9	93928416	rs4743820	93904561	94124414	3.80×10^{09}	26192919	IBD	3.63×10^{06}	0.189	
9	117545666	rs6478106	117535341	117698507	5.00×10^{46}	23266558	IBD	1.89×10^{01}	0.638	*TNFSF8*
9	120475302	rs4986790	120467576	120475602	6.99×10^{09}	26974007	CD	2.77×10^{05}	0.211	
9	139266405	rs10781499	138824899	139405093	4.00×10^{56}	23128233	IBD	5.06×10^{36}	0.121	*CARD9*
10	6094697	rs61839660	6038478	6527143	1.24×10^{14}	FM	IBD	3.81×10^{03}	0.013	*IL2RA*
10	27179596	rs7911117	27159385	27182592	1.84×10^{08}	NOVEL	UC	1.84×10^{08}	0.723	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
10	30728101	rs1042058	30689316	30808324	6.00×10^{11}	23128233	IBD	2.81×10^{12}	0.221	
10	35466185	rs34779708	35256960	35554054	2.39×10^{26}	FM	IBD	5.74×10^{12}	0.034	
10	59913151	rs1819658	59901559	60065351	9.00×10^{17}	21102463	IBD	1.44×10^{06}	0.191	
10	64445564	rs10761659	64349667	64566258	4.97×10^{53}	26192919	IBD	2.30×10^{36}	0.239	
10	75673101	rs2227564	75462199	75695724	7.00×10^{10}	23128233	IBD	1.35×10^{04}	0.049	
10	81060317	rs1250550	81032532	81067480	1.00×10^{30}	21102463	IBD	3.56×10^{12}	0.552	
10	82254047	rs6586030	82214586	82301536	9.00×10^{16}	23128233	IBD	5.27×10^{08}	0.600	
10	94436851	rs7911264	94248310	94495241	3.00×10^{08}	23128233	IBD	5.37×10^{07}	0.491	
10	101284237	rs4409764	101271789	101320612	1.16×10^{61}	26192919	IBD	1.90×10^{34}	0.117	
10	104232716	rs3740415	104222963	104403310	1.03×10^{07}	26192919	IBD	8.47×10^{04}	0.004	
10	112186148	rs11195128	112182708	112186148	2.00×10^{10}	23850713	CD	5.41×10^{11}	0.012	
10	126439381	rs111456533	126316604	126551228	1.18×10^{09}	NOVEL	IBD	1.18×10^{09}	0.568	
10	133172119	rs10734105	133170322	133172119	3.00×10^{08}	22412388	CD	3.45×10^{01}	0.932	
11	1874072	rs907611	1873232	1880596	1.00×10^{10}	21297633	IBD	1.06×10^{06}	0.100	
11	57203009	rs11229030	57184964	57209488	8.00×10^{09}	22412388	CD	2.98×10^{01}	0.906	
11	58408687	rs11229555	58174653	58434545	5.23×10^{12}	26192919	IBD	2.59×10^{05}	0.489	
11	60776209	rs11230563	60770426	60776781	1.71×10^{14}	26192919	IBD	1.95×10^{06}	0.138	*CD6*
11	61564299	rs4246215	61542006	61624181	2.00×10^{15}	23128233	IBD	4.37×10^{05}	0.332	
11	64150370	rs559928	63956102	64164833	3.33×10^{13}	26192919	IBD	1.75×10^{05}	0.235	
11	65656564	rs2231884	65575263	65663547	3.00×10^{10}	23128233	CD	1.40×10^{04}	0.647	
11	72863697	rs11235667	72863697	72863697	7.00×10^{09}	23850713	CD	NA	NA	
11	76299649	rs11236797	76270683	76302073	2.25×10^{50}	FM	IBD	7.19×10^{33}	0.837	
11	87125438	rs6592362	87123977	87130298	2.00×10^{08}	23128233	IBD	1.44×10^{03}	0.839	
11	96023427	rs483905	96015426	96045998	3.16×10^{10}	26192919	UC	6.81×10^{07}	0.013	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
11	114397162	rs200349593	114275412	114447782	3.56×10^{20}	FM	UC	2.19×10^{09}	1.000	*NXPE1*
11	118754353	rs630923	118754353	118754353	7.00×10^{09}	23128233	CD	6.55×10^{06}	0.495	
11	128380974	rs11221332	128380974	128396738	2.44×10^{09}	26974007	IBD	2.44×10^{08}	0.909	
12	6491125	rs7954567	6487161	6514969	1.30×10^{09}	26192919	CD	4.69×10^{09}	0.974	
12	12657513	rs11612508	12560673	12711368	1.00×10^{08}	23128233	UC	5.01×10^{04}	0.226	
12	40740223	rs148319899	40388109	40828306	5.54×10^{29}	FM	IBD	2.55×10^{15}	0.100	
12	48208368	rs11168249	48195939	48208368	8.00×10^{09}	23128233	UC	3.18×10^{07}	0.711	
12	68508122	rs11614178	68469642	68508276	2.22×10^{32}	FM	IBD	3.82×10^{23}	0.976	
12	112007756	rs653178	111826477	112486818	1.11×10^{08}	26192919	IBD	2.14×10^{09}	0.451	
12	120146925	rs11064881	120109284	120146925	5.95×10^{08}	26192919	IBD	5.34×10^{05}	0.911	
13	27531267	rs17085007	27531267	27543781	3.00×10^{19}	23128233	UC	1.21×10^{14}	0.942	
13	41013977	rs941823	40678443	41032853	2.00×10^{14}	23128233	IBD	3.59×10^{07}	0.240	
13	41558110	rs7329174	41552738	41771476	8.00×10^{09}	23266558	CD	3.90×10^{01}	0.164	
13	42917861	rs80244186	42838908	42938329	3.66×10^{08}	NOVEL	CD	3.66×10^{08}	0.362	
13	43052880	rs2062305	42951449	43055002	5.00×10^{10}	21102463	CD	3.60×10^{07}	0.139	
13	44457925	rs3764147	44406102	44490181	2.00×10^{21}	23128233	IBD	2.74×10^{08}	0.954	
13	49595331	rs2026029	49538512	49821244	9.73×10^{09}	26974007	CD	2.58×10^{03}	0.889	
13	99956622	rs9557195	99778655	100064765	2.00×10^{14}	23128233	IBD	1.38×10^{06}	0.366	
14	69210199	rs4902642	69201003	69307621	2.00×10^{10}	21102463	IBD	1.96×10^{04}	0.225	
14	75741751	rs1569328	75700675	75747118	3.21×10^{09}	26192919	IBD	1.98×10^{05}	0.819	
14	88472595	rs8005161	88398949	88555206	4.00×10^{18}	21102463	IBD	2.71×10^{11}	0.303	
15	38899190	rs16967103	38836777	38925195	4.00×10^{09}	23128233	IBD	6.35×10^{03}	0.864	
15	41563950	rs28374715	41367036	41779260	2.00×10^{08}	23128233	UC	3.27×10^{07}	0.750	
15	67442596	rs17293632	67441750	67468285	2.71×10^{20}	26192919	IBD	3.01×10^{21}	0.538	*SMAD3*

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	P-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
15	91172901	rs7495132	91172901	91192408	9.00×10^{11}	23128233	IBD	2.62×10^{07}	0.854	
16	11373320	rs529866	11344903	11718433	2.00×10^{16}	23128233	CD	1.05×10^{07}	0.244	
16	23864590	rs7404095	23841051	23867776	1.00×10^{09}	23128233	IBD	1.63×10^{07}	0.145	
16	28528781	rs28449958	28338039	28955702	7.50×10^{24}	FM	IBD	2.18×10^{12}	0.631	
16	30482494	rs11150589	30469919	30512443	1.63×10^{11}	FM	UC	1.21×10^{06}	0.936	*ITGAL*
16	50335074	rs78534766	50304685	50335074	1.17×10^{14}	NOVEL	UC	3.32×10^{13}	0.838	*ADCY7*
16	50745926	rs2066844	50511169	51009351	2.27×10^{217}	FM	IBD	1.42×10^{38}	0.003	*NOD2*
16	68591230	rs1728785	68554754	68680902	3.00×10^{08}	19915572	UC	3.76×10^{08}	0.702	
16	81916912	rs11548656	81910840	81922813	5.18×10^{11}	NOVEL	IBD	5.18×10^{11}	0.208	*PLCG2*
16	82867456	rs10492862	82866188	82916401	1.26×10^{09}	NOVEL	CD	1.26×10^{09}	0.840	
16	86014241	rs16940202	85989464	86019761	6.00×10^{19}	21297633	IBD	2.51×10^{11}	0.606	
17	25869033	rs10775412	25788058	25869033	2.51×10^{25}	FM	CD	3.32×10^{08}	0.974	
17	32593974	rs3091316	32567679	32625383	1.00×10^{26}	23128233	IBD	1.47×10^{12}	0.578	
17	37912377	rs12946510	37902887	38119638	2.16×10^{39}	26192919	IBD	1.69×10^{26}	0.916	
17	40527544	rs12942547	40412165	40690118	6.00×10^{22}	23128233	IBD	1.90×10^{17}	0.717	
17	54880993	rs3853824	54868990	54949047	7.70×10^{10}	26192919	IBD	9.79×10^{06}	0.807	
17	57963537	rs1292053	57801597	58046076	9.00×10^{13}	23128233	IBD	2.04×10^{05}	0.086	
17	70642923	rs17780256	70611194	70642923	3.19×10^{11}	26192919	IBD	3.74×10^{11}	0.914	
17	76737118	rs17736589	76667271	76833429	4.34×10^{08}	26192919	UC	1.28×10^{05}	0.079	
18	12818922	rs80262450	12745889	12886441	3.97×10^{28}	FM	IBD	1.04×10^{16}	0.696	
18	46395022	rs7240004	46393601	46395315	1.01×10^{10}	26192919	IBD	4.30×10^{08}	0.209	
18	56879827	rs9319943	56876228	56883319	9.05×10^{07}	26192919	CD	6.60×10^{05}	0.346	
18	67530439	rs727088	67511645	67562657	5.00×10^{09}	23128233	IBD	2.21×10^{03}	0.863	
18	77220616	rs7236492	77183529	77237142	1.45×10^{08}	26192919	IBD	9.99×10^{04}	0.156	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	*P*-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
19	1124031	rs2024092	1106477	1191682	8.00×10^{22}	23128233	IBD	5.00×10^{17}	0.620	
19	10512911	rs11879191	10408439	10600418	5.27×10^{20}	26192919	IBD	1.53×10^{11}	0.566	*ICAM1, TYK2*
19	33731551	rs17694108	33728381	33757062	6.00×10^{15}	23128233	IBD	1.36×10^{12}	0.985	
19	34656406	rs587259	34653364	34727202	2.98×10^{08}	26974007	CD	1.41×10^{03}	0.043	
19	46849806	rs4802307	46847901	47147226	9.04×10^{12}	26192919	CD, UC	3.59×10^{07}, 9.92×10^{02}	0.598, 0.698	
19	49206172	rs516246	49168942	49248730	1.33×10^{20}	26192919	CD	3.58×10^{11}	0.266	*FUT2*
19	55383051	rs11672983	55368865	55386920	7.00×10^{11}	23128233	IBD	2.74×10^{04}	0.994	
20	6093889	rs4256018	6076126	6095344	1.23×10^{08}	NOVEL	IBD	1.23×10^{08}	0.781	
20	30725648	rs6142618	30696392	31099311	6.00×10^{10}	23128233	IBD	6.14×10^{06}	0.334	
20	31376282	rs4911259	31329704	31471012	1.00×10^{09}	23128233	IBD	2.46×10^{03}	0.061	
20	33799280	rs6088765	33799280	33882720	2.00×10^{08}	23128233	IBD	8.47×10^{03}	0.079	
20	43065028	rs6017342	43065028	43258079	5.17×10^{50}	FM	UC	3.95×10^{30}	0.150	
20	44742064	rs1569723	44680853	44749251	1.00×10^{13}	23128233	IBD	1.98×10^{06}	0.195	
20	48955424	rs913678	48955424	48968438	5.35×10^{11}	26192919	IBD	4.21×10^{07}	0.411	
20	57824309	rs259964	57809343	57829821	1.00×10^{12}	23128233	IBD	3.37×10^{09}	0.936	
20	62329099	rs6062496	62270637	62378954	6.76×10^{26}	FM	IBD	2.83×10^{26}	0.075	
21	16817938	rs2823286	16781136	16841303	9.00×10^{30}	23128233	IBD	4.43×10^{29}	0.073	
21	34776695	rs2284553	34752334	34777409	5.63×10^{17}	26192919	CD	1.14×10^{14}	0.624	
21	40465534	rs2836878	40458508	40468838	5.00×10^{48}	23128233	IBD	2.30×10^{29}	0.093	
21	45615741	rs7282490	45611686	45634148	9.85×10^{30}	26192919	IBD	1.85×10^{23}	0.214	
22	21922904	rs2266959	21910280	21998833	1.00×10^{16}	23128233	IBD	4.12×10^{15}	0.642	
22	30493882	rs5763767	30130115	30592487	8.82×10^{15}	26192919	IBD	4.17×10^{08}	0.109	
22	35729721	rs138788	35724659	35737461	2.95×10^{08}	NOVEL	UC	2.95×10^{08}	0.347	
22	37258503	rs4821544	37258503	37258986	1.76×10^{08}	NOVEL	CD	1.76×10^{08}	0.574	

Continued on next page

Table B.1 – *Continued from previous page*

Chr	Position	Rsid	LD_left	LD_right	P-value	Study (PMID)	Trait	P_{Meta}	P_{Het}	Implicated gene
22	39659773	rs2413583	39659773	39756650	5.06×10^{38}	26192919	IBD	4.60×10^{24}	0.822	
22	41867377	rs727563	41215672	42216326	1.88×10^{10}	26192919	CD	8.20×10^{04}	0.016	
22	50435480	rs5771069	50353919	50457041	4.00×10^{08}	20228798	UC	2.47×10^{10}	0.373	

List of Tables

List of Figures