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.

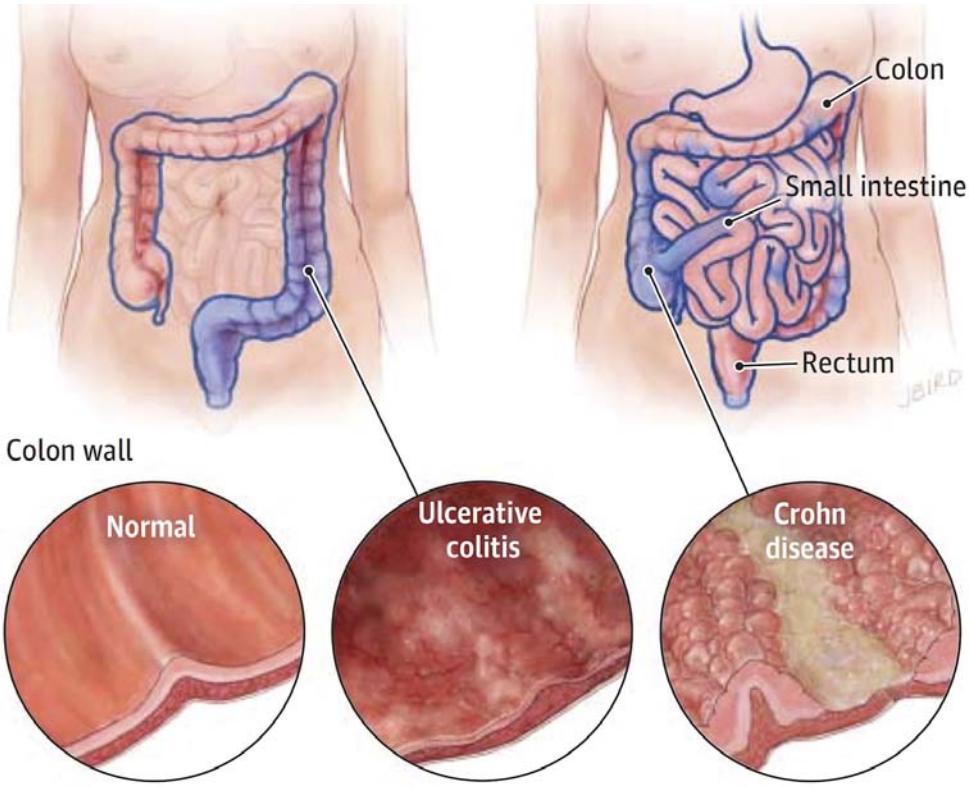

Figure 1.1: Disease localisation and appearance of Crohn's disease and ulcerative colitis, the two major forms of inflammatory bowel disease. Image sourced from Jin (2014)

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).

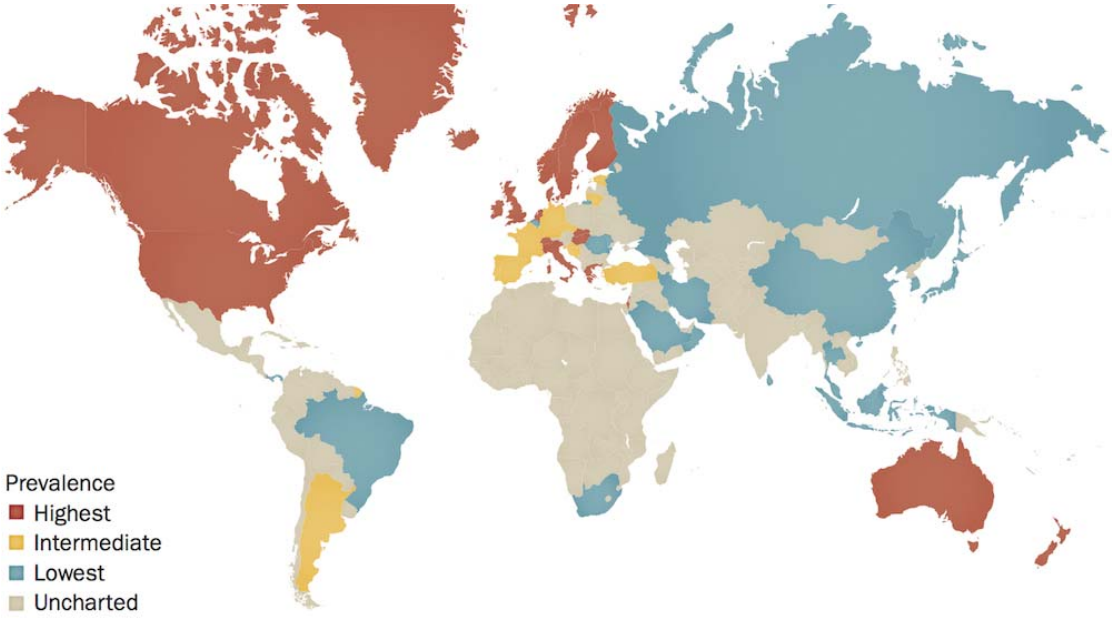

Figure 1.2: Global prevalence of inflammatory bowel disease in 2015. Image sourced from Kaplan (2015).

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.

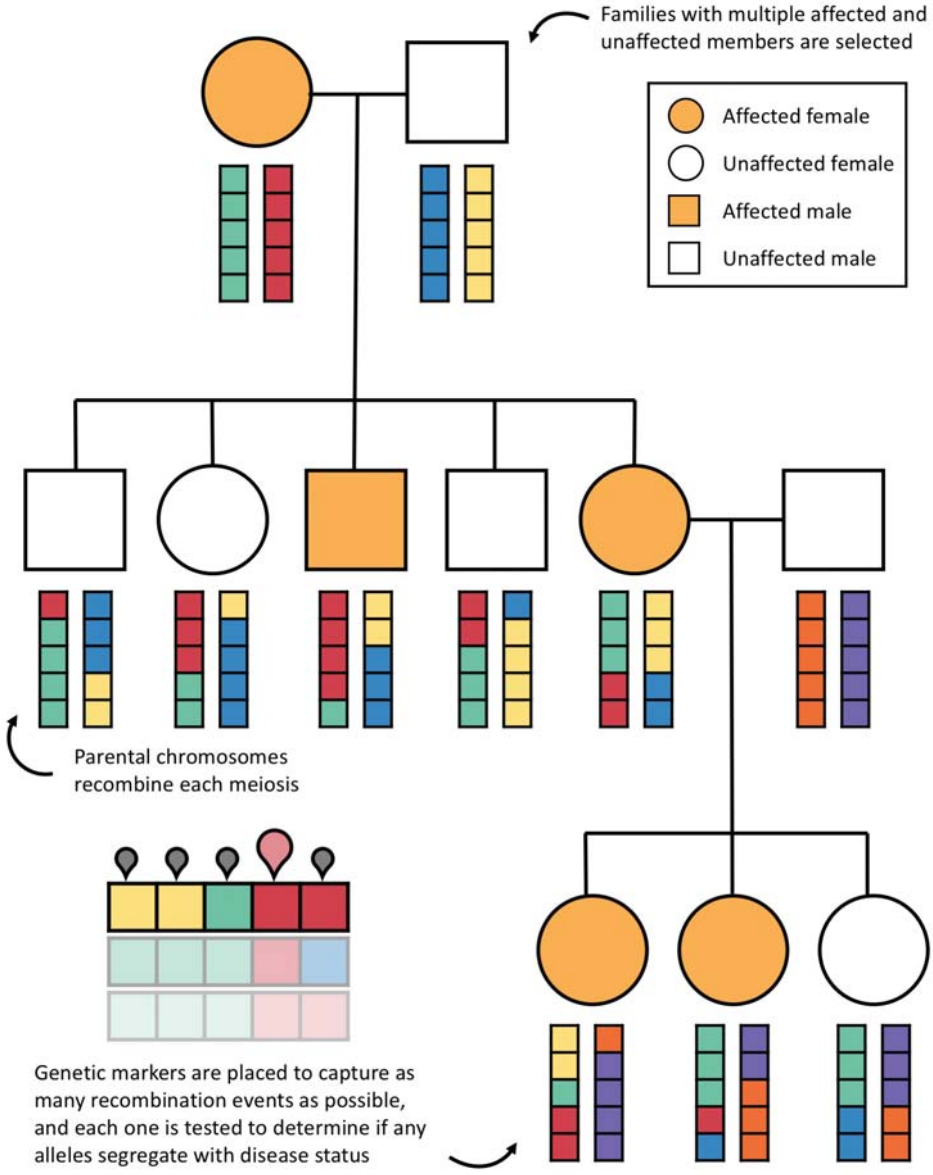

Figure 1.3: Overview of the linkage analysis study design for identifying disease-associated loci within a family containing multiple affected individuals.

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.

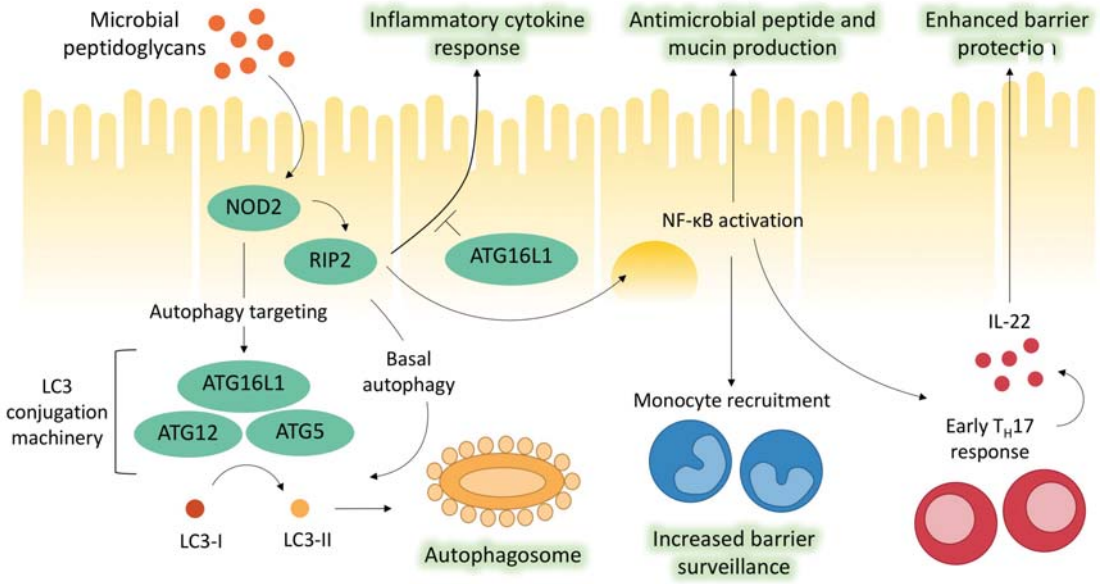

Figure 1.4: The signalling pathways through which *NOD2* responds to microbial peptidoglycan stimuli to promote innate mucosal defence and an autophagic response. Figure adapted from Philpott et al. (2014).

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).

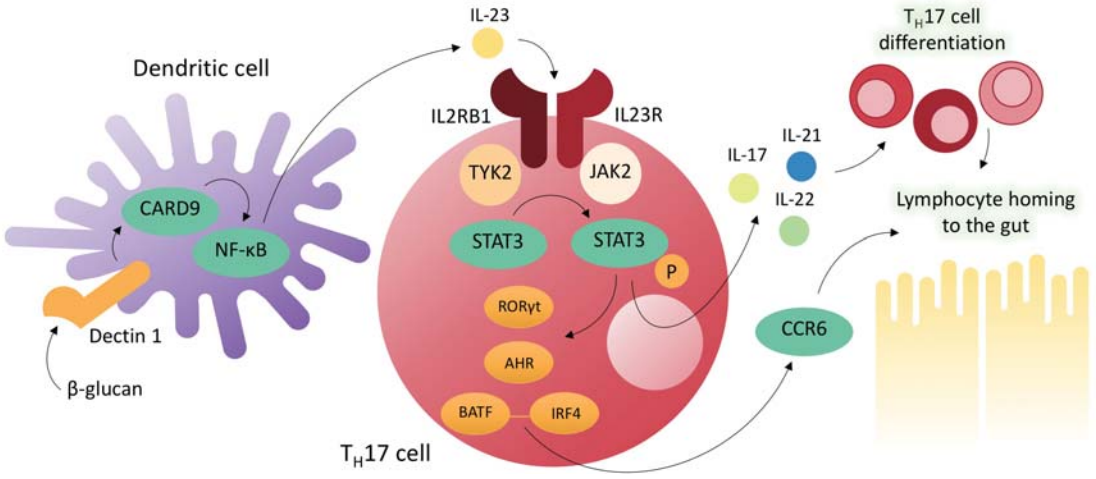

Figure 1.5: The *IL23R* signalling pathway used to activate the adaptive immune response, and the downstream T_H17 cell differentiation program (Weaver and Hatton, 2009; Zhernakova et al., 2009; Khor et al., 2011; Parkes et al., 2013).

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.

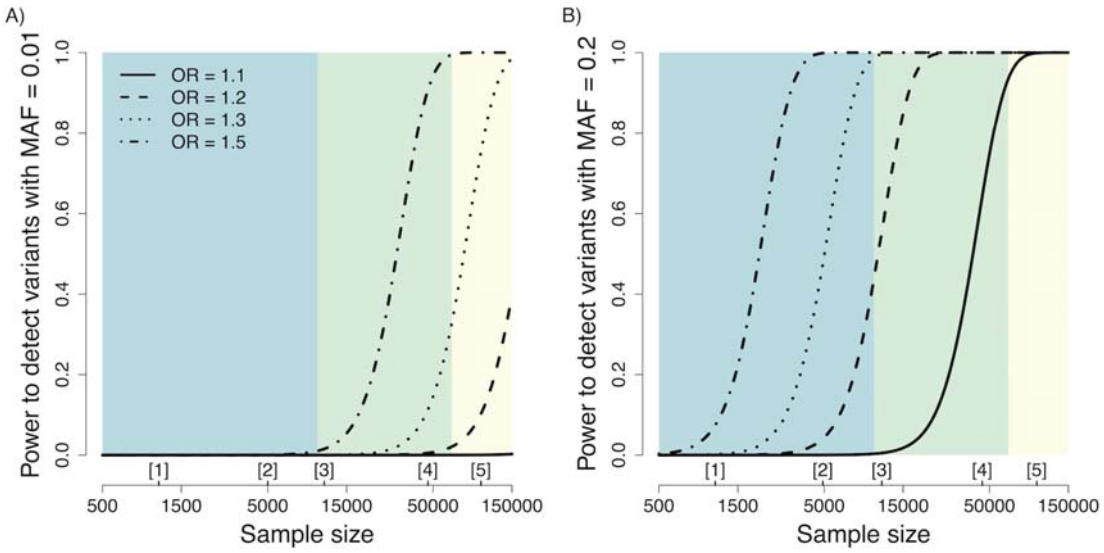

Figure 1.6: Power to detect associations of different effect size (odds ratio, OR) are compared for rare variants (MAF = 0.01, panel A) and common variants (MAF = 0.2, panel B). Effective sample sizes of several key studies are indicated along the x-axis, to reflect the power of the GWAS studies (blue), 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.

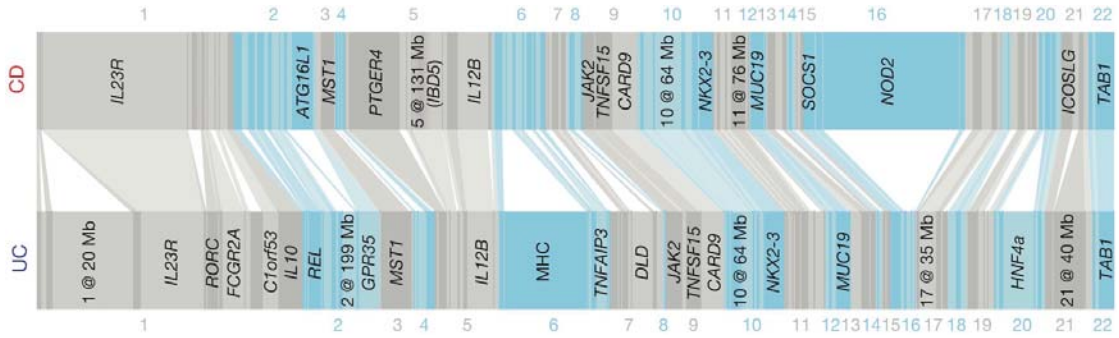

Figure 1.7: Belgravia plot of the 163 loci identified by Jostins et al. (2012), showing the shared genetic overlap between Crohn's disease and ulcerative colitis. The width of each bar is proportional to the variance explained by a given locus for the disease indicated, and bars are linked if they are 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).

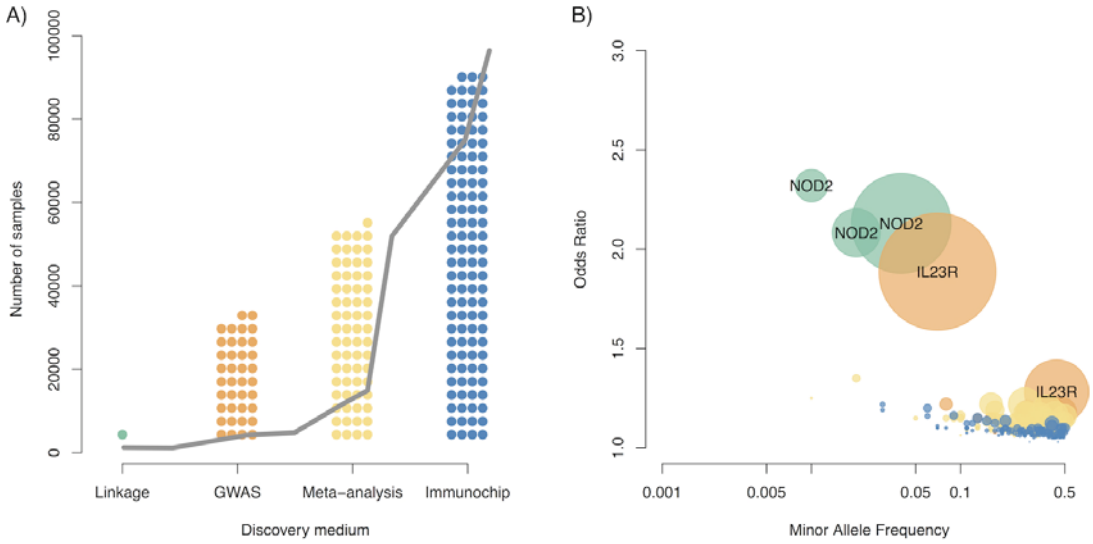

Figure 1.8: Panel A) describes the number of loci identified with different study designs (linkage=green; GWAS=orange; meta-analysis=yellow; Immunochip=blue), with each dot representing a novel locus. The number of samples required to identify these loci are shown in grey. Panel B) plots the odds ratio 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.