Chapter 1. Introduction and historical perspective

Members of the human race are fascinatingly diverse. No two individuals – not even identical twins – are exactly alike in height, body weight, skin colour, blood type, personality, or football club allegiance. Yet it is no coincidence that for most traits, people who are related to each other are, on average, more similar than those who are not. Part of this reflects the shared environments of closely related individuals. Families live under one roof, eat the same food, with children going to the same schools and playing with the same toys. Then there are genetics factors. Individuals who are related to each other also share more stretches of identical DNA.

Of all the traits that vary between individuals, understanding the causes of disease susceptibility is perhaps the most pertinent. Identifying the specific genetic factors that are associated with disease risk will offer insights into understanding disease biology with a goal for better treatment outcomes for patients. Much of this dissertation describes the identification of genetic loci associated with risk for four poorly understood autoimmune and autoinflammatory disorders: primary biliary cirrhosis, primary sclerosing cholangitis, Crohn's disease and ulcerative colitis. For the remainder of this chapter, I provide a rationale for studying these diseases, as well as a historical perspective on how our understanding of the genetic contribution to complex traits has been shaped.

1.1 Immune-mediated diseases

1.1.1 The immune system

The human immune system encompasses three broad layers of protection against infectious agents such as bacteria and viruses. Firstly, physical barriers such as the skin prevent pathogens from entering the body in the first place. When these are breached, the innate immune system, consisting of ever-present cells ready at the site of infection, provides an immediate and generic response to the pathogen. If the agent is able to overcome these innate defences, the adaptive immune system may become activated. Here, pathogen recognition is specific and becomes part of immunological memory, allowing for a more potent response to infection and acquisition of immunity.

How the human immune system discriminates between its own cells and that of a pathogen is one of the central questions of immunology. To be effective, the immune system needs to strike a balance between its ability to recognise and destroy a pathogen while leaving endogenous cells alone. A weak immune response can lead to immunodeficiency and a greater risk of infection, while an overactive response, whereby the host's own cells are targeted, can result in autoimmune and autoinflammatory diseases.

Over 100 such immune-mediated diseases (IMDs) have been described, and together represent a diverse array of clinical features, epidemiological profiles and risk factors (Ricard Cervera and Munther, 2009). Such disorders can affect either a single tissue type or organ, such as inflammatory bowel disease or type 1 diabetes, or can affect multiple parts of the body, such as systemic lupus erythematosus. For the majority of these diseases, symptoms are chronic there are no known cures or preventive measures, and are thought to be triggered by combinations of environmental factors (e.g. an infection from a pathogen or a microbiome imbalance) in a genetically susceptible host. Treatments to control symptoms generally begin with medication to suppress the immune response, though for some disorders, an organ transplant may ultimately be required.

1.1.2 Epidemiology

Individually, IMDs are quite rare, though they collectively affect 3-7% of the population and represent a large and growing public health issue (Cooper *et al.,* 2009; Parkes *et al.,* 2013). It has been estimated that the direct annual medical cost of IMDs in the United States is over $125 billion (Blumberg *et al.,* 2012), with further economic costs incurred through loss in productivity and working days from these chronic conditions. Indeed, the prevalence of many IMDs has increased over the past 50 years, and is thought to be a reflection of greater awareness and better disease diagnoses, as well as changing environmental factors (Cooper *et al.,* 2009). One often-cited explanation for the rising prevalence is the "hygiene hypothesis", whereby the decreasing incidence of infections in developed countries inhibits proper development of the immune system, which in turn increases risk to allergies and IMDs in later life (Okada *et al.,* 2010).

Epidemiological studies have also shown significant comorbidity between several IMDs, where an individual with one IMD is at significantly increased risk to develop a second IMD (Cooper *et al.,* 2009). For instance, patients with inflammatory bowel disease are at higher risk of also developing primary sclerosing cholangitis and primary biliary cirrhosis (Roman and Munoz, 2011; Saich and Chapman, 2008). It is also possible having one IMD can offer protection against others. For instance, it has been suggested sufferers of multiple sclerosis have reduced risk of rheumatoid arthritis (Somers *et al.,* 2009). Increased risks for IMDs also extends to family members of affected individuals, both for the same disease and increased risk for other IMDs (Cooper *et al.,* 2009). In Crohn's disease, for instance, familial clustering showed that 2-14% of patients have a family history of Crohn's (Halme *et al.,* 2006), while estimates of the sibling recurrence risk ratio (the ratio of disease risk among siblings of patients compared with that in the general population, i.e. the population prevalence) ranged from 15-42 (Halme *et al.,* 2006). The variation in these estimates highlights the difficulty in obtaining accurate prevalence and comorbidity measures for relatively rare disorders. Confounders also include inconsistent study design (e.g. only counting first degree relatives rather than all relatives), sample selection bias (e.g. hospitalised cases that are likely to have a more severe form of the disease than those sent home), and variation in disease prevalence, both between different populations and over time (Farrokhyar *et al.,* 2001; Halme *et al.,* 2006; Hiatt and Kaufman, 1988; Mathew and Lewis, 2004; Shivananda *et al.,* 1996). Nevertheless, this "kaleidoscope of autoimmunity" (Anaya *et al.,* 2007) suggests shared biological mechanisms present in many of these disorders, for which genetic factors are likely to play a role. Identifying the genes that underlie disease risk allows for a greater understanding of disease biology, and potentially, better treatment options for patients.

1.2 Genetic studies of complex autoimmune disorders

1.2.1 Mendelian inheritance, multifactorial traits and heritability

First laid out by Gregor Mendel in the 1860s and rediscovered in the 1900s, the Mendelian laws of inheritance describe how heredity factors (genes), of which an offspring acquires two versions (alleles – one from each parent) can affect variation in phenotypes (Bateson and Mendel, 1902). Mendel observed through the crossing of pea plants how a phenotype, in his case the colour of the flower, is passed through to subsequent generations in a discrete manner (rather then being a blend of the colour of the parents) via certain principles of segregation. For a given gene, which of the two parental alleles an offspring receives is random, and by performing a large number of crosses, Mendel was able to infer the two alleles (genotype) of each individual plant depending on whether the phenotype displayed dominance or recessive characteristics (Figure 1.1). Traits that adhere to this mode of inheritance are known as Mendelian traits, and include diseases such as sickle-cell anaemia and cystic fibrosis, where a single recessive allele is responsible for disease.

While Mendel's laws could adequately describe the observed discrete inheritance patterns of some traits, they did not appear to apply to the majority of traits where variation appeared to be continuous, nor to discrete traits that did not follow any obvious patterns of Mendelian inheritance. Moreover, Mendel's laws appeared to be inconsistent with natural selection, where evolution occurs via the accumulation of small, gradual changes. These apparent conflicting observations were reconciled in the 1930s in what became known as the modern evolutionary synthesis. Ronald Fisher and others showed that quantitative traits such as height can be described by multiple genes, each with small, additive effects acting according Mendel's laws of inheritance (Fisher, 1930). Together, these small independent effects, along with the environment give rise to a phenotype that approximates the normal distribution (Figure 1.2).

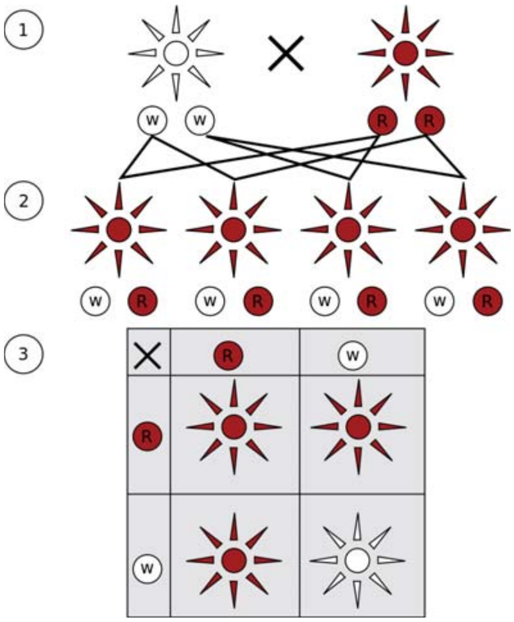

Figure 1.1. Mendel's laws of inheritance. In this example, there are two alleles: W and R which give rise to either a white or red phenotype respectively when both copies are present. Red is dominant and white is recessive. In (1) the parental generation, the parents are homozygotes for each of the alleles. In (2) the first generation, all offspring are heterozygotes and will show the red phenotype. When heterozygotes cross, (3) the offspring will show a 3:1 red:white ratio depending on which of the two alleles they inherit. (Image source: Magnus Manske, Wikimedia Commons)

Binary phenotypes such as disease status are also often the result of multiple genes, each with small effects, and the environment. These complex (or multifactorial/polygenic) disorders can be modelled quantitatively with a liability threshold model in a similar manner to that proposed by Fisher (Falconer and Mackay, 1996). Each individual of a population will have a disease liability – a quantitative measure that incorporates all genetic and environment factors in disease risk. Disease liability itself is rarely observed directly, but can be described in a population as a normally distributed continuous trait. When an individual's liability exceeds a given threshold, they are said to be affected by the disease (Figure 1.3).

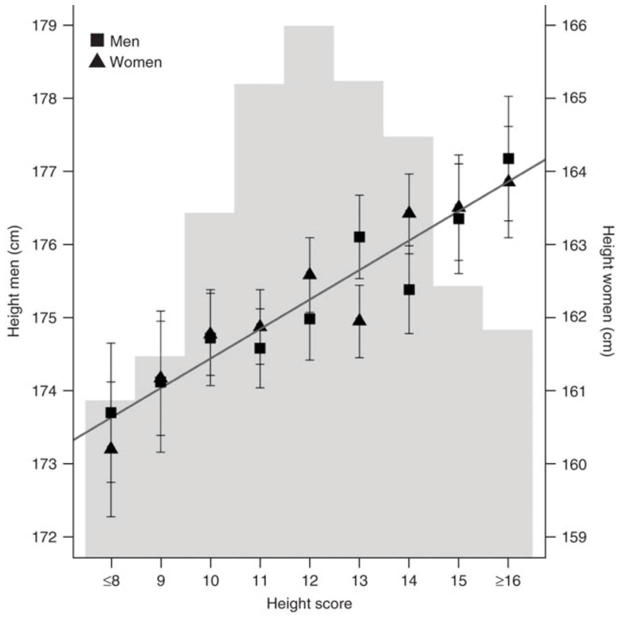

Figure 1.2. Polygenic inheritance in a normally distributed trait: height. Using 12 SNPs associated with height, 7,566 individuals were grouped according to the number of height-increasing alleles they carried (height score on x-axis). The gray bars represent the fraction of individuals in each height score group. For each height score, the average heights in men and women are plotted. The diagonal regression line indicates that each height-increasing allele increases height by 0.4 cm. Figure sourced from Lettre et al. (2008)

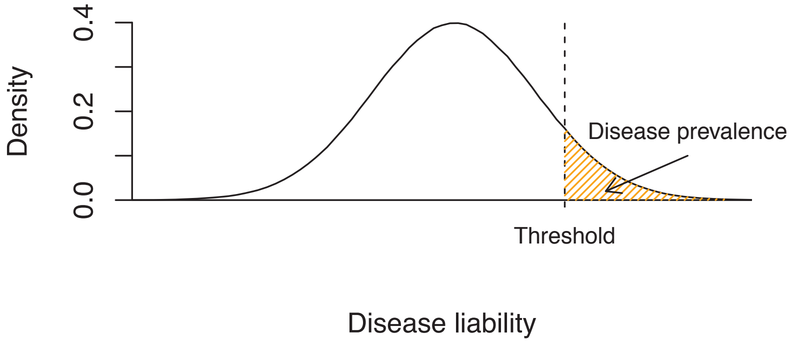

Figure 1.3. The liability threshold model. Disease liability can be thought of as a continuous trait that incorporates all environmental and genetic risk factors of a disease and is normally distributed in the population. Individuals who exceed a given threshold (dashed vertical line) will be affected by the disease (shaded orange area).

The concept of heritability is often used when describing the genetic contribution to variation in a trait or disease. The variation of a continuous trait seen in the population can be partitioned into genetic (heritable) and non-genetic (environmental) components. The heritable component can also be further partitioned into additive and non-additive components. Additive genetic variation, or narrow sense heritability, describes the extent to which an individual's phenotype can be determined by that of their parents. In the context of a gene affecting a quantitative trait, this means that each additional copy of an allele increases (or decreases) the value of the trait by the same amount. Non-additive components include dominance and gene-gene interaction effects, and together with the additive effects, make up broad sense heritability. In the context of complex diseases and for the remainder of this dissertation, I will refer to the narrow-sense heritability of disease liability as "heritability" (Falconer and Mackay, 1996). These components of phenotypic variation have typically been estimated based on expected genetic relatedness across families, the most useful of which is the twin study (described below). In recent years, heritability can also be estimated from directly observed genotypes (e.g. SNP microarrays) across both related (Visscher *et al.,* 2006) and unrelated individuals (Yang *et al.,* 2010).

1.2.2 Twin studies

Familial recurrence and disease comorbidity do not always themselves suggest a role for genetics in disease, as these observations can also be a consequence of shared environment. Twin studies, however, can provide compelling evidence for a significant genetic component to disease risk. Identical (monozygotic) twins are genetically identical, while non-identical twins (dizygotic) share half their polymorphic alleles. The twin design assumes that the environmental component to phenotypic variation is the same between monozygotic and dizygotic twins, and thus the difference in disease concordance rates between sets of monozygotic and dizygotic twin pairs can be used to estimate the additive genetic, shared environmental and unique environmental components of disease risk.

The assumptions that underlie the twin study have often been the subject of scrutiny. For instance, the assumption of shared environment does not hold when considering the pre-natal intrauterine environment. Monozygotic twins, for example, often share a single placenta, whereas dizygotic twins have separate placentas. Moreover, it may be the case that monozygotic twins tend to copy each other more or are treated differently by those around them than dizygotic twins throughout their lives. These assumptions are often difficult to test and violations may lead to inflated heritability estimates (Devlin *et al.,* 1997). Nevertheless, studies that use twins reared apart, which do not rely on the equal environment assumption, consistently show higher concordance between monozygotic twins than dizygotic twins for a range of traits and diseases (Bouchard *et al.,* 1990; Hanson *et al.,* 1991). In addition, recent assumption-free methods of estimating heritability from directly genotyped genetic markers in related (Visscher *et al.,* 2006) and unrelated (Lee *et al.,* 2011; Yang *et al.,* 2010) individuals are consistent with those estimated from twin studies.

Twin studies have demonstrated that most IMDs do have a significant genetic component. In Crohn's disease, the largest meta-analysis of 112 monozygotic and 196 dizygotic twins reported concordance rates of 30.3% and

3.6% respectively (Brant, 2011). Significant differences in monozygotic/dizygotic concordance rates have also been found for multiple sclerosis (25.4% and 5.4%) (Willer *et al.,* 2003), coeliac disease (75% and 11%) (Greco *et al.,* 2002) and type 1 diabetes (27.3% and 3.8%) (Hyttinen *et al.,* 2003). These results implied that that given sufficient sample sizes and genetic markers, it is theoretically possible to identify the genetic variants that contribute to disease risk.

1.2.3 The major histocompatibility complex

The first robust associations between a genetic locus and IMDs were identified in the major histocompatibility complex (MHC) in the 1970s, many decades before the genes and genetic variants in question were mapped. The human MHC is located on chromosome 6 and contains many genes that are collectively known as the human leucocyte antigen (HLA). These genes encode cell surface molecules that are responsible for a range of immune-related functions, including the establishment of adaptive immunity and the destruction of infected cells. As part of the immune system's self/non-self recognition processes, genes in the MHC were first discovered as being crucial for whether an organ transplant was successful (Sheldon and Poulton, 2006). Throughout the 70s and 80s, HLA variants were found to be associated with almost all IMDs, albeit with larger effects in some than others. These early studies took a molecular rather than genetic approach to identifying disease associations. That is, associations were inferred via serological typing in affected and unaffected individuals rather than later genetic studies that sought to capture genetic variation directly. These later approaches, starting with linkage mapping and then moving on to association, would become the prevailing methods by which genetic risk factors for complex disease are discovered.

1.2.4 Linkage

A linkage study identifies regions of the human genome underlying disease susceptibility by testing a series of marker alleles for cosegregation (linkage) with disease status across a family or number of families. Technological advances in the 1970s and 1980s lead to the easy genotyping of restriction fragment length polymorphisms (RFLPs) (Botstein *et al.,* 1980) spread throughout the genome, and later, denser maps of repeat regions (microsatellites) (Weber and May, 1989). Owing to the large size of chromosomal segments segregating within a typical family, around 300-400 evenly distributed around one every 10 cM microsatellite markers are usually sufficient to capture the majority of recombination events (Evans and Cardon, 2004). The evidence for linkage in a region is evaluated by metrics such as a LOD (logarithm of odds) score, which compares the probability that the genotyped marker and the hypothetical disease locus are inherited together in the observed data versus the probability of observing the cosegregation pattern purely by chance. A typical linkage study will report all loci with LOD scores greater than three, which corresponds to the data being 1000 times more likely to arise due to cosegregation with disease than by chance (Lander and Kruglyak, 1995). By the mid-1990s, linkage studies had proven to be a robust means of identifying highly penetrant loci underlying monogenic disease such as cystic fibrosis (Tsui *et al.,* 1985) and Huntington's disease (Gusella *et al.,* 1983) and the utility of the method for mapping complex disease loci was increasingly being explored.

In addition to confirming many of the known associations with the HLA, an early success for linkage studies in complex traits was the identification of the *NOD2* locus associated with Crohn's disease in 1996 (Hugot *et al.,* 1996). This result was confirmed in subsequent studies (Brant *et al.,* 1998; Cavanaugh, 2001; Cavanaugh *et al.,* 1998; Cho *et al.,* 1998; Curran *et al.,* 1998; Mirza *et al.,* 1998; Ohmen *et al.,* 1996) and in 2001 the specific causal mutations that underlie risk were localised to three low frequency coding variants (R702W, G908R and L1007fs) within the *NOD2* gene (at that time, also known as *CARD15*) (Cuthbert *et al.,* 2002; Hampe *et al.,* 2001; Hugot *et al.,* 2001; Ogura *et al.,* 2001; Vermeire *et al.,* 2002). These three variants individually had odds ratios (ORs) of 2-4 in heterozygotes and 20-40 for homozygotes, and at least one mutation was present in 30-40% of Crohn's disease cases compared with 6-7% in European controls (Mathew and Lewis, 2004). Other notable well-replicated linkage findings in IMDs during this time include *INS* and *CTLA4* in type 1 diabetes (Bain *et al.,* 1992; Bennett *et al.,* 1997; Nisticò *et al.,* 1996) and *PTPN22* in rheumatoid arthritis (Begovich *et al.,* 2004; Jawaheer *et al.,* 2003).

It soon became apparent that strong linkage signals for complex disorders were the exception rather than the rule. Overall, the results of linkage studies were largely disappointing, with few loci being consistently replicated across different studies. This lack of reproducibility suggested that complex diseases, in contrast to Mendelian diseases, were unlikely to be driven by the highly penetrant risk loci that linkage is well powered to detect. In 1996 a seminal paper was published in Science proposing that complex diseases are underpinned by common variants of modest effect (Risch and Merikangas, 1996). The authors demonstrated that, for a risk allele of 50% frequency and OR of 1.5, around 18,000 affected sib-pairs would be needed to detect the locus via linkage. In contrast, they reported that less than 1000 trios would be needed to detect such a locus adopting the transmission/disequilibrium association test of Spielman *et al.* (1993). Technological limitations at the time restricted the immediate uptake of the association study design; such studies require that a causal variant (or another variant in high linkage disequilibrium to the causal variant) is directly genotyped in order to detect a significant signal of association.

1.2.5 Candidate genes

While it was infeasible to test for association at markers across the entire genome, technological improvements during the late 1990s and through the 2000s made it possible to genotype markers within individual genes to then test for association. Genes were selected based on *a priori* knowledge of biological function or because they reside within a region implicated through linkage analysis. These candidate gene studies typically involved genotyping a set of markers within a gene of interest in a sample of disease cases and controls, and testing for statistically significant differences in allele frequencies between the two groups. Other study designs such as transmission disequilibrium tests in parent-offspring trios were also often used.

Results from the majority of candidate gene studies for complex traits were disappointing, with initial findings often failing to replicate in subsequent experiments. A combination of small sample sizes, false-positive association, publication bias and failure to account for multiple comparisons meant that as many as 95% of findings from candidate gene studies of complex traits during this era were false (Colhoun *et al.,* 2003; Ioannidis *et al.,* 2001). In some cases, the lack of power in these studies meant that variants in genes that later became established risk loci were missed altogether (for instance, *IL10* in Crohn's disease) (Parkes *et al.,* 1998; Castro-Santos *et al.,* 2006; Franke *et al.,* 2010). Ultimately however, it would take a combination of technological advances and a greater appreciation of the need for much larger sample sizes to make the identification of bona fide risk loci routine.

1.2.6 Genome-wide association studies

In the early 2000s, along with the closing phases of Human Genome Project, concurrent efforts were underway to gauge the extent of human genetic variation at the population level. Projects such as the SNP Consortium and dbSNP had catalogued over 1.4 million single nucleotide polymorphisms (SNPs) by 2001 (Sachidanandam *et al.,* 2001; Sherry *et al.,* 1999). It was found that common SNPs in physical proximity formed LD blocks punctuated by hotspots of recombination (McVean *et al.,* 2004). These correlation patterns were further characterised through the International Hapmap Project, which by 2007 had identified a further 3.1 million SNPs across 270 individuals from three distinct ancestry groups (International HapMap Consortium *et al.,* 2007). At the same time, technological advances in microarray technologies made possible the cost-effective genotyping of hundreds of thousands of SNPs spread throughout the genome (Syvanen, 2005). The patterns of LD meant that these arrays could effectively survey the majority of common genetic variation in a population by directly genotyping only a fraction of the total number of variants in the genome.

In Europeans and East Asians, around 5 million common SNPs (those with minor allele frequency greater than 5%) can be almost entirely tagged by a selection of approximately 500,000 SNPs (Barrett and Cardon, 2006; International HapMap Consortium *et al.,* 2007). Together, these advances paved the way for researchers to perform genome-wide association studies (GWAS) in order to identify loci associated with complex traits or disease risk.

Genome-wide association studies typically look for statistically significant differences in allele (or genotype) frequencies between a large number of diseased individuals and population controls across hundreds of thousands of SNPs spread throughout the genome. The SNPs that show significant association with disease status point to regions of the genome likely to harbour disease relevant genes. Unlike linkage studies, GWAS are not restricted to sibling pairs and families, and also have generally greater statistical power to detect associated loci of small to moderate effect sizes (Figure 1.4) (Risch and Merikangas, 1996). Due to patterns of LD, there is no reason to conclude that an associated SNP is the causal variant, but rather it is correlated with ("tags") the true causal variant. In addition, genotypes at SNPs that were not directed assayed can be inferred through imputation algorithms (Li *et al.,* 2009; Marchini and Howie, 2010) based on the genotypes from a representative reference set of haplotypes (International HapMap Consortium *et al.,* 2007; 1000 Genomes Project Consortium *et al.,* 2012; International HapMap Consortium *et al.,* 2010), allowing for individual studies using different genotyping platforms to be effectively combined into meta-analyses.

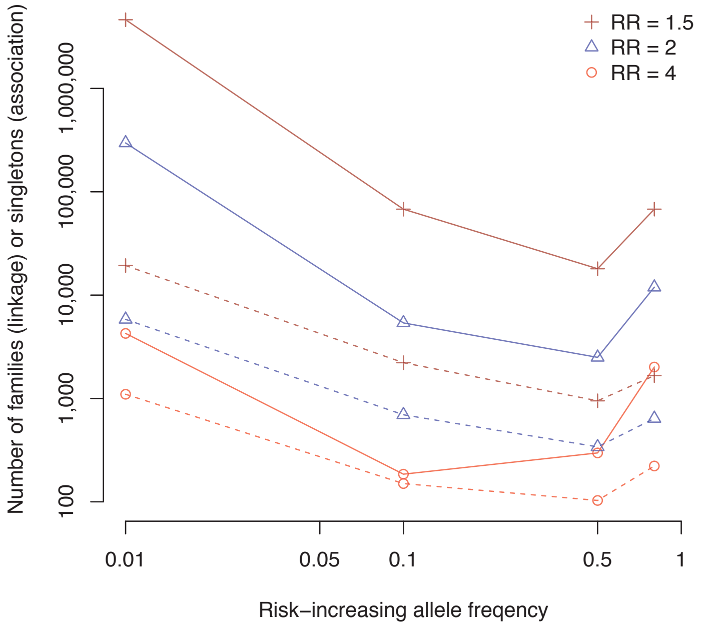

Figure 1.4. Power of linkage vs. association outlined in Risch and Merikengas (1996). The minimum number of samples required to detect a genetic variant with genotypic relative risks of 1.5, 2 and 4 at 80% power (at genome-wide significance) are plotted for linkage studies using related individuals (solid lines) and association studies using unrelated individuals (dashed lines). At all effect sizes and allele frequencies, association designs have greater power than linkage.

The first successful GWAS was published in 2005 for age-related macular degeneration (AMD) (Klein *et al.,* 2005), where the authors genotyped ~100,000 SNPs and identified a variant in the *CFH* gene that increased the risk of AMD by a factor of ~7.4. Some of the first GWAS for autoimmune disorders such as Crohn's disease and ulcerative colitis also appeared during this period (Duerr *et al.,* 2006; Yamazaki *et al.,* 2005). These early studies typically used small sample sizes compared to modern studies (usually a few hundred) and often differed in terms of association methods, the strength of statistical evidence used to declare significance, and quality control procedures. Standard protocols for GWAS became established following the seminal publication from the Wellcome Trust Case Control Consortium in 2007 of 14,000 cases across seven diseases and 3000 common controls (Wellcome Trust Case Control Consortium, 2007). Methods to deal with population stratification, HapMap imputation, manual inspection of intensity cluster plots, large sample sizes, stringent statistical criteria for declaring association and the requirement for independent replication were some of the many protocols in this paper that became standard in subsequent GWAS. The genome-wide significance threshold for association of $p < 5\times10^{-8}$ was also established around this time. This figure roughly corresponds to a 5% type-I error rate when considering the number of independent regions tagged by common variants in the genome in individuals of European descent (~1-2 million) (Hoggart *et al.,* 2008; International HapMap, 2005). Unlike linkage studies, these standardised protocols and strict statistical criteria meant that the vast majority SNPs that exceeded genome-wide significance were true positives.

These early GWAS showed that, with the exception of the HLA, the typical effect size of a susceptibility locus for complex traits was modest (OR < 1.3), such that the loci identified only explain a fraction of the estimated genetic component of disease risk (often referred to as the “missing heritability” (Maher, 2008; Manolio *et al.,* 2009)). While it is likely that a proportion of this missing heritability is due to rare (minor allele frequency less than 1%) and structural variants that are not well-captured on the current generation of GWAS microarrays, a substantial number of common variants will have even smaller effects than those identified, requiring much larger sample sizes to detect (Yang *et al.,* 2010). Indeed, for Crohn’s disease, it has been estimated that 22% of the variance in disease liability can be explained by common variants tagged on microarrays (Lee *et al.,* 2011) – more than double that explained by known risk loci at the time (Barrett *et al.,* 2008). Heritability is not missing, but rather resides at common variants with small effects that cannot be confidently associated with disease risk.

After the first wave of GWAS, an appreciation of the need for larger sample sizes lead to many studies being combined to perform meta-analyses. Again, taking the example from Crohn’s disease, three GWAS meta-analyses were published from 2008 to 2012. The first of these combined data for ~13,000 individuals from three previously published GWAS and identified 21 new Crohn’s susceptibility loci (Barrett *et al.,* 2008). This was followed two years later by a meta-analysis of six GWAS with a total sample size of ~50,000 individuals where 30 new loci were identified, bringing the total count to 71 (Franke *et al.,* 2010). The most recent meta-analysis in 2012 included 75,000 individuals, including both Crohn's disease and ulcerative colitis, and in total identified 163 inflammatory bowel disease loci, the most for any complex disease to date (Jostins *et al.,* 2012). One hundred and ten of these loci were associated with both Crohn's disease and ulcerative colitis. Similar large-scale meta-analyses have also been performed for other IMDs such as type 1 diabetes (30,000 individuals and 40 loci) (Barrett *et al.,* 2009), multiple sclerosis (80,000 individuals and 110 loci) (International Multiple Sclerosis Genetics, 2013), rheumatoid arthritis (48,000 individuals and 46 loci) (Eyre *et al.,* 2012) and celiac disease (24,000 individuals and 40 loci) (Trynka *et al.,* 2011a).

1.3 Insights from GWAS

1.3.1 Biology

The genes (and their corresponding pathways) implicated the variants identified through GWAS have provided invaluable insights into the biological processes underlying IMDs. In multiple sclerosis, most of the associated genes are involved in known immunological pathways (e.g. cytokine pathway, T-cell differentiation and signal transduction) rather than neurodegeneration (International Multiple Sclerosis Genetics Consortium, 2013; Sawcer *et al.,* 2011). Moreover, the *KIF21B* gene that may be involved in neurodegeneration is also associated with Crohn's disease and ankylosing spondylitis, suggesting that this gene may also have an immune-related function despite being exclusively expressed in the brain and spleen (Visscher *et al.,* 2012). Additionally, two of the genes identified were previously known targets for multiple sclerosis drugs (natalizumab for *VCAM1* and daclizumab for *IL2RA*) (Sawcer *et al.,* 2011), suggesting that there is great therapeutic potential among the list of associated genes.

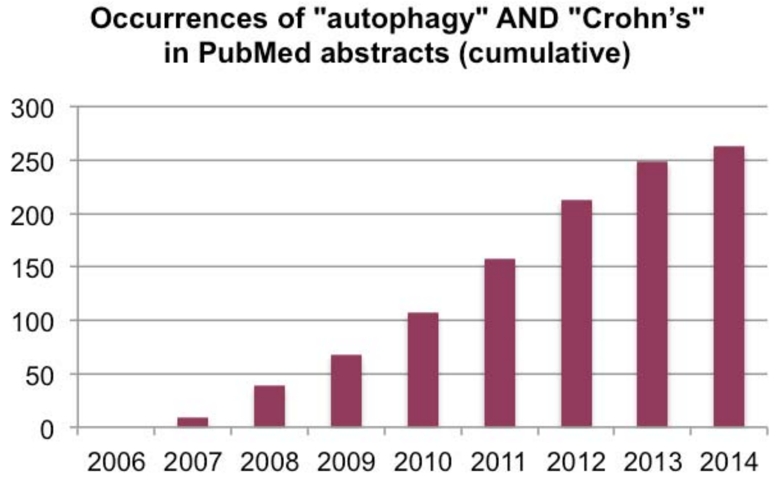

Figure 1.5. Number of publications indexed in PubMed with the terms "autophagy" and "Crohn's" in the abstract since 2006.

GWAS have also provided biological insights into inflammatory bowel disease. Perhaps most notably, early GWAS for Crohn's disease for suggested a role for autophagy via associations at *ATG16L1* and *IRGM*, in disease etiology (Hampe *et al.,* 2007; Khor *et al.,* 2011; Parkes *et al.,* 2007). Autophagy is the process by which a cell cleanses and recycles unnecessary components, including the elimination of pathogens. It has been suggested that the coding variant in *ATG16L1* associated with Crohn's disease degrades this protein, thus impairing autophagy function such that cells were unable to clear bacterial infections (Murthy *et al.,* 2014). Autophagy is now an active area of Crohn's disease research, perhaps best illustrated by the number of Pubmed abstracts containing "Crohn's" and "autophagy" that have appeared since 2007 (Figure 1.5). These and other examples of previously unsuspected pathways in inflammatory bowel disease (e.g. IL23R pathway, innate immunity) demonstrate the value of hypothesis-generating genetic associations studies in enabling a greater understanding of disease biology (Visscher *et al.,* 2012).

1.3.2 Genetic overlap between immune-mediated disorders

Insights into biology can also be gained from identifying shared and unique associations among a set of related disorders. While the role of the HLA in autoimmunity has been known since the 1970s, one of the major findings of early GWAS was the extent to which non-HLA risk loci are shared among IMDs. Perhaps surprisingly, where patterns of familial aggregation appeared to cluster into seropositive autoimmune (e.g. primary biliary cirrhosis, celiac disease and type 1 diabetes) and seronegative disorders (e.g. Crohn's disease, psoriasis and anklyosing spondylitis) the pattern of pleiotropic loci has been observed across all these diseases (Parkes *et al.,* 2013).

In a review of six IMDs where large GWAS have been undertaken (ankylosing spondylitis, celiac disease, inflammatory bowel disease, psoriasis, rheumatoid arthritis and type 1 diabetes) Parkes *et al.* (2013) found 71 loci that are associated with two or more diseases. Notably, of the 416 pairwise combinations of overlapping loci, 45% were concordant (same associated variant and same direction of effect), 14% discordant (same variant, but risk increasing in one disease and risk decreasing in the other) and 42% not correlated (same locus, but different associated variant).

Together, these observations support the observations that the increased occurrence of IMDs within individuals and family members may in part be driven by the shared genetic risk factors underlying these diseases. Identifying the genes and pathways that are shared between IMDs can provide insights into shared biology and potential drug targets across various disorders. Conversely, variants that are discordant between disorders may explain why some drugs may be effective for one disorder, but ineffective or even exacerbate the condition in another. Taking advantage of this genetic overlap was one of the driving motivations for the development of the Immunochip genotyping array.

1.4 Locus discovery beyond GWAS

1.4.1 Dense genotyping

A feature of many locus discovery projects in IMDs since 2011 has been the use of the Immunochip custom genotyping array. The Immunochip was designed after the first wave of GWAS meta-analyses to aid in the replication, fine-mapping and discovery of loci associated with inflammatory and IMDs (Cortes and Brown, 2011). To take advantage of the pervasive genetic overlap between many of these diseases, the Immunochip contains a dense panel of ~130,000 SNPs located in 186 regions with known association with one or more of 12 immune-related diseases. SNPs within the regions were ascertained via dbSNP, the 1000 Genomes Project (February 2010 release), and IMD resequencing projects. While not all SNPs passed the Illumina design process and made it onto the microarray, the Immunochip provides unprecedented coverage of common, low-frequency and rare variants across these 186 genomic regions. A further 50,000 SNPs that were suggestively significant in the original GWAS studies were also included. The cost-effectiveness of the Immunochip (at ~20% that of a GWAS microarray at the time) allows for studies with much larger sample sizes than GWAS and also enables powerful disease subphenotype and cross-disease comparisons (Parkes *et al.,* 2013).

1.4.2 Finemapping and inferring causality

The causal variants that underlie the majority of loci discovered through GWAS remain unidentified. An associated locus will often consist of dozens of correlated SNPs in high LD spanning across many genes, with very similar association signals. In the 140 loci associated with Crohn's risk, the number of SNPs that are tagged ($r^2 > 0.8$) by the reported GWAS SNP range from 1 to 306 per locus (median 13). The *IRGM* locus associated with Crohn's disease exemplifies some of the challenges in assigning causality to a particular variant. The initial reported associated SNP was later found to be in perfect LD with a 20kb deletion upstream of *IRGM* (McCarroll *et al.,* 2008; Parkes *et al.,* 2007). This deletion was thought to be causal because it affects the expression of *IRGM*, which in turn regulates the efficiency of autophagy. A later study showed, however, that this deletion is one of several highly correlated Crohn's disease associated variants in the region that affect *IRGM* expression, none of which can reasonably be ruled out as causal (Prescott *et al.,* 2010). Furthermore, the variants are also not associated with Crohn's disease in the Japanese population, suggesting either European-specific gene-environment interactions or the presence of an untyped causal variant that arose after the European-Asian population split (Prescott *et al.,* 2010).

Narrowing multiple correlated associations signals down to a single causal variant is difficult and will initially require a combination of many complementary approaches. Firstly, much larger sample sizes will be required to differentiate statistical signals at causal variants over their highly correlated neighbours. Secondly, as patterns of LD differ between different ancestral groups, obtaining samples from multiple populations can narrow the associated region for risk loci that are shared across populations. Thirdly, combining functional genetic information with association results allows variants with relevant annotations to be up-weighted in association analyses. Data from projects such as ENCODE (ENCODE Project Consortium *et al.,* 2012) and GTEx (Lonsdale *et al.,* 2013) provide rich functional genomic information that can potentially be integrated with GWAS results. Methods for integrating these various data sources are under active development. In addition to providing functional candidates, these functional annotations can also uncover potential biological mechanisms through which variants act, either through the specific cell type or functional element (Liu *et al.,* 2012; Schaub *et al.,* 2012; Trynka and Raychaudhuri, 2013), or can be used to weight genetic association signals in order to identify additional associations (Pickrell, 2014).

1.4.3 Sequencing and rare variant associations

The role of rare variants in complex diseases is currently an important area of focus in human genetics. High-throughput discovery and accurate genotyping of rare variants has recently been made feasible through large reductions in the cost of next-generation sequencing. Often cited as a possible explanation for missing heritability, rare variants are in theory likely to have much larger effect sizes than common variants due to purifying selection maintaining damaging alleles at low frequencies (Manolio *et al.,* 2009). Indeed, loci that are associated with complex disease are enriched for rare variants that cause known Mendelian disorders and it has been suggested that recessive variants confer risk to related complex diseases when the carrier is heterozygote (Blair *et al.,* 2013). Independent rare variant associations are also often found in genes with known common associated variants (Momozawa *et al.,* 2011; Nejentsev *et al.,* 2009; Sanna *et al.,* 2008).

Since the rare allele of individual rare variants are observed so infrequently, single variant tests of association will be underpowered for all but the most highly penetrant alleles. For instance, for an allele that doubles disease risk (OR=2) and has a frequency of 0.1%, nearly 60,000 cases and a similar number of controls will be required for the variant to reach genome-wide significance. To increase power to detect association, rare variants are often aggregated based on characteristics such as their position within genes, functional features (e.g. loss-of-function alleles) and allele frequencies (Bansal *et al.,* 2010). Dozens of these burden tests have been proposed (Asimit and Zeggini, 2010; Bansal *et al.,* 2010; Basu and Pan, 2011; Kiezun *et al.,* 2012) along with methods for meta-analysis and replication (Hu *et al.,* 2013; Lee *et al.,* 2013b; Liu *et al.,* 2014). These statistical tests typically differ in the way variants are weighted and whether they incorporate alleles with opposite directions of effects. Indeed, the most powerful method to use will differ from gene to gene and will depend on the specific genetic architecture, which is seldom known in advance.

Taking Crohn's disease as an example, the degree to which such variants contribute to disease heritability is unclear, and the results from early large scale sequencing studies targeted at known susceptibility genes have been disappointing (Momozawa *et al.,* 2011; Rivas *et al.,* 2011; Hunt *et al.,* 2013). These studies typically involved sequencing the coding regions of several candidate genes in a few hundred cases and controls followed by the direct genotyping of putatively associated variants in a much larger replication cohort. Coding regions are targeted because the functional consequences of variants in these regions are much better understood than those in noncoding parts of the genome. These variants are hypothesized to have larger effect sizes given their direct impact on protein product and are generally more evolutionarily conserved than noncoding variants (Chen *et al.,* 2007). Momozawa *et al.*

(Momozawa *et al.,* 2011) initially sequenced 63 candidate genes in 112 Crohn's disease cases and 112 controls with replication in an additional 288 to 928 cases and 288 to 1216 controls, and identified four independent associations in *IL23R*, although only one of these exceeded genome-wide significance. Similarly, Rivas *et al*. (Rivas *et al.,* 2011) sequenced 56 genes in 350 cases and 350 controls with follow-up genotyping in 16,054 cases and 17,575 controls, and identified 12 independent rare variant associations across seven genes, of which two (coding variants in *NOD2* and *CARD9*) exceeded genome-wide significance. These three genome-wide significant variants were included on the Immunochip and subsequently confirmed in Jostins *et al*. (2012) using around 75,000 samples. However, a recent sequencing study of 25 candidate genes across 41,911 individuals in seven IMDs, failed to identify any novel associations (Hunt *et al.,* 2013). A natural extension for candidate gene sequencing studies is to sequence the entire exome of cases and controls. A recent exome sequencing study in 42 Crohn's cases with follow up genotyping in 9348 cases and 14,567 controls found suggestive rare variant associations in *PRDM1*(Ellinghaus *et al.,* 2013b). Again, the variant failed to reach genome-wide significance and other whole exome studies with much larger sample sizes are currently underway.

The sobering results from these studies highlight the challenges in rare variant association studies. As it is currently not economically feasible to perform high coverage whole-genome sequencing in a large number of cases and controls, compromises often need to be made in terms of the number of genomic regions covered and the number of individuals. Around 93% of SNPs reported in GWAS reside in noncoding regions (Maurano *et al.,* 2012), which have been overlooked by the current generation of sequencing studies. A large number of rare noncoding variants will play a role in gene regulation, though it remains to be seen whether their effects are large enough to be a major contributor to disease. Performing burden tests across rare variants in regulatory regions such as promoters and enhancers may show promise. Most importantly, the sample sizes used in these sequencing studies have thus far simply been insufficient to robustly identify rare variant associations. Under certain assumptions about the effect size distribution of rare variants and selection pressures, cohorts of more than 25,000 cases may be required in order to find these signals, along with an equally large number for replication (Zuk *et al.,* 2014).

1.5 Conclusions

Putting together the results from linkage, genome-wide association and sequencing studies, the genetic architecture of IMDs such as inflammatory bowel disease, multiple sclerosis and type 1 diabetes represents those of a typical multifactorial complex trait where a combination of multiple genes, along with the environment, lead to disease. With few exceptions, individual risk loci for these disorders confer only a modest effect on disease susceptibility and together, the known loci explain ~5-20% of variation in disease liability. The majority of the genetic contribution to disease risk remains to be explained, and will likely come from a combination of both common variants with ever smaller effects and rare variants.

1.6 Outline of dissertation

In the previous sections, I outlined the rationale for studying the genetics of IMDs, and provided a brief historical background to our understanding of how genetic variation contributes to phenotypic variation. I described the history of locus discovery experiments in complex traits, with specific examples from successful (and sometimes not so successful) efforts in IMDs. The remainder of this dissertation describes experiments to better understand the genetic basis of four IMDs: primary biliary cirrhosis, primary sclerosing cholangitis, and the two major forms in inflammatory bowel disease, Crohn's disease and ulcerative colitis.

In chapter 2, I describe a locus discovery experiment in primary biliary cirrhosis in 2,861 cases and 8,514 controls from the UK genotyped on the Immunochip. Three novel disease risk loci were identified, and, taking advantage of the much denser SNP coverage, we identified multiple novel independent signals within known loci. We highlight one of these regions (3q25) as an interesting example of where testing variants independently when there are multiple risk variants in LD can lead to both an over- and underestimation of effect sizes and significance levels. I explore methods by which combining risk loci with functional genomic information can provide insights into the functional elements and cell types that are specific to a disease.

In chapter 3, I describe a locus discovery experiment in primary sclerosing cholangitis (PSC) in 3,789 cases and 25,079 controls of European descent. Nine novel risk loci were identified, and associations in the HLA complex were refined via imputing the classic HLA haplotypes. A feature of PSC is the high degree of overlap with inflammatory bowel disease (IBD). Over 70% of PSC cases also suffer from ulcerative colitis, and the extent of genetic overlap between the disorders is yet to be determined. I show that around half the loci associated with PSC risk appear to be unique to PSC, and that there is little difference in the effects of PSC risk loci in PSC/IBD subphenotypes, suggesting distinct biological mechanisms behind PSC verses IBD.

In chapter 4, I describe a locus discovery and trans-ethnic association study of Crohn's disease and ulcerative colitis in ~75,000 European and ~11,000 non-European samples. The non-European dataset includes individuals of East Asian (Japan, South Korea, China), Indian and Iranian descent. By combining Immunochip and GWAS datasets and performing a trans-ethnic meta-analysis, we were able to identify 40 novel loci associated with Crohn's disease, ulcerative colitis or both. I showed that there is pervasive sharing of IBD risk loci between European and non-European populations, while also noting loci that appear to be specific to only Europeans, as well those with differences in effect sizes between various populations. The study demonstrates the utility of performing large-scale GWAS meta-analyses across different populations to identify novel susceptibility loci.

In chapter 5, I move beyond locus discovery and describe a simple method of integrating differential gene expression datasets with associated loci. I applied this method to two differential expression datasets: the first involves genes that are differentially expressed in the gut T cells vs. blood T cells in healthy humans, and the second consisting of murine cells from the cecum before and after infection by the nematode *Trichuris muris*. Differentially expressed genes between T cells in the gut are likely to be involved in maintaining intestinal homeostatsis, while those that are differentially expressed in infected and uninfected cells serve as a model for response to infection. I find that in both cases, genes that are differentially expressed between these conditions are significantly overrepresented among risk loci for a range of IMDs, allowing for the identification of additional candidate genes at these loci and the generation of hypotheses about the mechanism through which they mediate disease.

Finally, in chapter 6, I discuss the major themes that one can draw from the preceding chapters, and then look to the types of studies that will shape the field over the coming years.