Chapter 1

Introduction and Background

1.1 Justifications of phenotype and approach

Since Man first climbed down from the trees and stared out at the world around him, he has wondered why both he and his sister suffer from acute inflammation of their digestive tracts. This thesis describes a series of statistical analyses designed to uncover and understand the genetics of complex disease, with particular application to the discovery of loci associated with inflammatory bowel disease (IBD).

Why should we dedicate time and effort to the study of IBD? And, given this, why should we study it through the medium of complex disease genetics?

1.1.1 Why study inflammatory bowel disease?

Much of this thesis will be concerned with inflammatory bowel disease (IBD), and in particular its two major forms: Crohn's disease (CD) and ulcerative colitis (UC). IBD is characterised by an inappropriate inflammatory response in the gastrointestinal tract, and symptoms include abdominal pain, diarrhoea, weight loss and damage to the intestinal wall (often requiring surgery to correct). Its incidence varies geographically, with a mean of around 7 new cases per 100,000 people per year in Europe, and has been increasing for at least the last 30 years throughout the world (Vatn, 2011).

IBD, and CD in particular, has been a "model" disease in complex disease genetics, with many linkage, candidate gene and genome-wide association studies carried out over the last 20 years. There are two aspects of IBD that make it an ideal complex trait to study.

It is a poorly understood disease with a high burden

While IBD is not a fatal disease, it does lead to a significant decrease in life expectancy. The standardised mortality ratio for Crohn's disease is 1.39 (95% CI 1.30 - 1.49) (Duricova et al., 2010), corresponding to a decreased life expectancy of approximately 5 years (95% CI 3.8-6.1, using the method of Tsai et al. (1992)). Ulcerative colitis does not show the same decrease in life expectancy, though approximately 17% of UC patients die from UC-related complications (Jess et al., 2007). Most deaths occur due to gastrointestinal disease, though a significant minority of deaths come from respiratory and genitourinary complications (Duricova et al., 2010).

As well as increased mortality, IBD is a life-long disease that is diagnosed early in life (mean age of diagnosis is 27). 40-50% of patients will require surgery within 10 years of diagnosis, and most will require drug therapy

throughout their lives (Bernstein, 2011). As well as the costs in human suffering, it is estimated that in Europe each patient with Crohn's disease costs €2898-6960 in direct health costs, and a total of up to €16.7 billion per year in economic costs (Yu et al., 2008).

The aetiology of IBD is still poorly understood (Zhang et al., 2008), with treatment focusing mostly on dietary changes to maintain remission, and interventions to reduce acute inflammation. The most effective treatment of acute inflammation is anti-TNF therapy (Bernstein, 2011), which is widely used in a range of inflammatory diseases, but often has negative side effects (Keane et al., 2001). Studies into environmental risk factors have had mixed results (Vatn, 2011), making genetics a good candidate to shed light on the biology of the disease. A better understanding of the aetiology of IBD could lead to treatments that target the underlying disease pathways, significantly lowering the costs of the disease.

It is highly heritable, and well characterised via twin studies

IBD is a highly heritable and genetically complex trait. Brant (2011) reviewed data from 6 twin studies of inflammatory bowel disease over the past 14 years, consisting of 657 sets of twins. Given these data, and the liability threshold methods described in Chapter 2, we can make inferences about the genetic architecture of disease. I analysed these data using two different liability models: one where siblings have some degree of shared environmental risk (C) but where genetic risk is purely additive (the ACE model), and one where genetic risk has additive and dominant components (A and D), but no shared environmental risk (Table 1.1). Both of these models are approximations, and should be viewed as such, but both can shed light on the genetic basis of IBD.

Phenotype	h^2 (=A) (95% CI)	C (95% CI)	D (95% CI)	H^2 (=A+D) (95% CI)
CD (ACE)	0.77 (0.70 - 0.84)	0 (0 - 0.06)	0 (NA)	0.77 (0.70 - 0.84)
CD (ADE)	0.48 (0.41 - 0.54)	0 (NA)	0.31 (0.24 - 0.38)	0.78 (0.69 - 0.88)
UC (ACE)	0.53 (0.46 - 0.61)	0.11 (0.05 - 0.18)	0 (NA)	0.53 (0.46 - 0.61)
UC (ADE)	0.66 (0.58 - 0.74)	0 (NA)	0 (0 - 0.08)	0.66 (0.58 - 0.77)

Table 1.1: The inferred liability components of CD and UC, using two different liability threshold models.

We can draw a number of conclusions from the twin study data. Firstly, both CD and UC are highly heritable: 70-85% and 45-70% respectively. Secondly, Crohn's disease has a significantly higher heritability than ulcerative colitis ($p = 1.04 \times 10^{-5}$). Thirdly, there is strong evidence of shared environment in UC, and strong evidence of non-additivity in CD, showing that IBD is both environmentally and genetically complex. The high heritability makes IBD a good candidate for genetic study.

It has been well studied by linkage and GWAS

Since the rise of genome-wide genetic studies IBD has been at the forefront of locus discovery. The discovery of the *NOD2* locus via genome-wide linkage (Hampe et al., 1999), and its subsequent fine-mapping to multiple causal variants (Hugot et al., 2001), was a notable success of linkage studies. The discovery of the *IL23R* locus was one of the first successes during the early days of GWAS (Duerr et al., 2006).

The genetic basis of IBD has also been well studied through large, collaborative meta-analyses. The largest linkage meta-analyses in IBD, though unsuccessful in mapping new loci, were successful in bringing together nearly

2000 families (van Heel et al., 2004). The largest international GWAS meta-analyses of Crohn's disease (Franke et al., 2010) and ulcerative colitis (Anderson et al., 2011) discovered nearly a hundred IBD loci in total. Notably, they also collected together over 13,000 total cases with genome-wide data, and over 25,000 other cases for the purposes of replication.

As a result of these studies, the IBD genetics community has a great deal of experience in successful genetic research, a series of long-standing collaborations with a history of data sharing, and a very large shared pool of patient samples for study. Together, these contribute to the highly productive research community that makes IBD a model disease for genetic studies.

1.1.2 Why study complex disease genetics?

I justified the study of inflammatory bowel disease by saying that the disease was costly to society, not well understood, and a heritable and genetically complex trait. However, the discovery of genetic risk factors is not in itself of use to society. To justify the approach, one must show how the discovery of these risk factors will positively impact science or medicine.

In this section I will discuss some of the ways risk loci can be used to the benefit of scientists and patients. I will start with two uncontroversial uses (helping to understand disease biology, and aiding further studies of disease), and move on to the more hotly debated topic of genetic risk prediction.

To directly understand biology

The dominant reason for discovering loci associated with disease is to allow us to understand disease biology. A better understanding of the aetiology of human diseases can allow the development of improved options for treatment, diagnosis and prevention, and ultimately reduce the incidence of, and

suffering from, disease.

The identification of loci has improved the understanding of many complex diseases. GWAS of type 2 diabetes have played an important role in shifting focus away from insulin resistance and towards insulin production (McCarthy and Zeggini, 2009), in particular towards defects in β-cell development, and have identified many new drug targets (Wolfs et al., 2009). New disease loci have uncovered previously unexpected pathways in inflammatory bowel disease including, notably, the role of autophagy in Crohn's disease (Zhang et al., 2008), and barrier defence in ulcerative colitis (Lees et al., 2011). Another notable success for GWAS was the discovery of the *BCL11A* locus as a major modifier of disease severity in haemoglobinopathies (Akinsheye et al., 2011), which has "reinvigorated the field of globin gene regulation" and is leading to the development of new treatment options for sickle cell disease and beta-thalassemia (Bauer and Orkin, 2011).

Locus identification can also give us information about biological factors that are shared across diseases. GWAS of different diseases will often implicate overlapping loci (Hindorff et al., 2009), and these loci can be informative about the shared aetiologies of these diseases. For instance, cross-phenotype comparisons of disease loci allow us to understand the relationships between Crohn's disease and both autoimmune and infectious disease (Lees et al., 2011). More generally, GWAS have highlighted the remarkable degree of genetic overlap between immune-mediated diseases (Cotsapas et al., 2011), and is starting to drive the creation of new classifications of immune disease based on shared pathways rather than affected tissue (McGonagle et al., 2009).

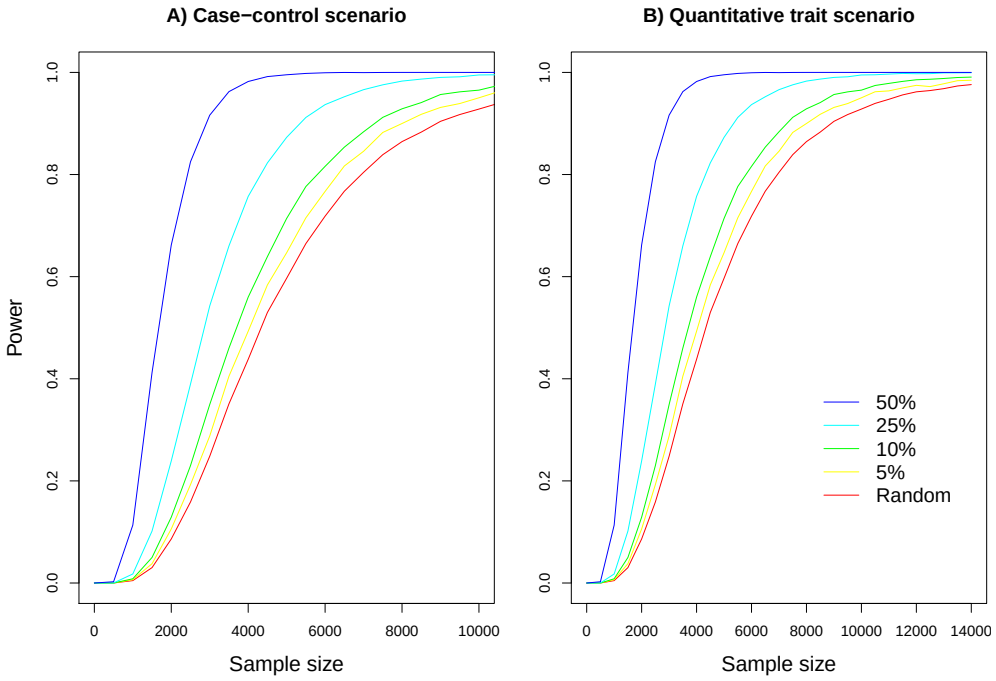

Figure 1.1: Improvement in power curves gained by prioritising samples based on genetic risk scores with different predictive powers. The colour of the line represents the proportion of total variance captured by the risk score, with the red line representing a random (i.e. non-prioritised) selection of samples. A) A case-control scenario for a disease with 1% prevalence. The total cohort size for prioritisation is 10,000 cases and an equal number of controls, and we measured power to detect a risk allele with an odds ratio of 2 and a frequency of 1% at genome-wide significance. B) A quantitative trait scenario. The total cohort size is 100,000, and we measured power to detect an allele with 1% frequency that increased a normally-distributed quantitative trait by 0.2 standard deviations.

To facilitate further research

Beyond the direct biological information that they can give us, disease loci can also be used as tools to aid future experiments. One obvious example is the use of genes in disease loci as candidates for functional studies, such as gene knock-out studies in mice (Kitsios et al., 2010), in much the same way as any candidate gene would be studied. However, there are also a number of uses of disease loci that utilise the unique properties of risk loci.

One such property of risk alleles is that an affected patient who carries a large number of protective alleles is more likely to have been subject to another, non-observed risk factor. We can therefore use known disease loci to select cases that have a low risk allele count (or low genetic prediction score), and test these for the presence of other risk factors. This is particularly relevant to the detection of low-frequency causal variants by sequencing, where often only a small subset of a larger cohort can be sequenced cost effectively. Figure 1.1 shows how this approach can increase the power of sequencing experiments for an example disease trait and an example quantitative trait. This approach is particularly well powered when selecting from large population cohorts of healthy individuals.

Another property of risk alleles is that they are acquired from birth (through Mendelian segregation), and remain constant through an individual's lifetime. As a result they cannot be caused by other risk factors, helping to resolve epidemiological problems of causality (this is called " Mendelian randomisation"). This approach has allowed some previously difficult-to-answer questions to be settled. For example high LDL cholesterol has been shown to be causally related to heart disease (Linsel-Nitschke et al., 2008), but high HDL is not (Voight et al., 2012). The same approach can be used to perform "retrospective" drug trials, for instance using Mendelian randomisation to establish *IL6R* as a drug target for heart disease (Hingorani et al., 2012).

To predict disease genetically

In his 1999 Shattuck lecture on the impact of the Human Genome Project (Collins, 1999), Francis Collins predicted the GWAS era, the rise of pharmacogenomics and the revolution in Mendelian disease genetics. However, the

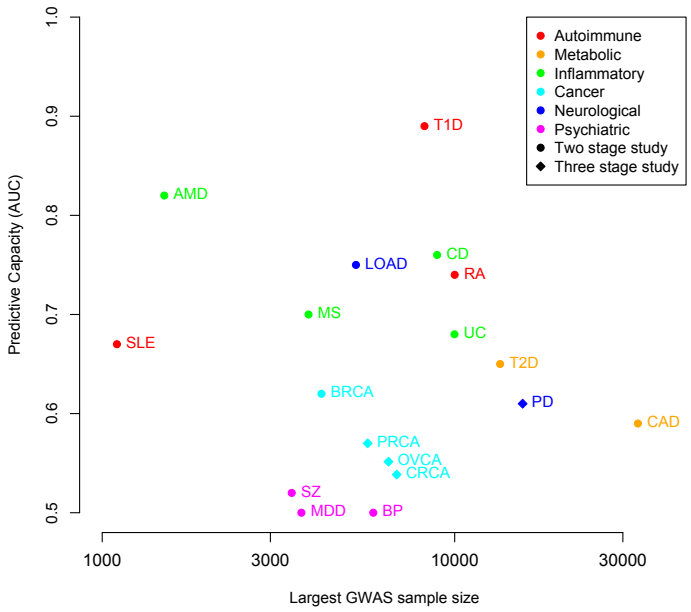

Figure 1.2: The predictive accuracy of variants discovered by genome-wide association studies, as a function of the effective sample size ($= \frac{2}{1/N_{case}+1/N_{control}}$), adjusted for the number of stages in the study (three stage studies have a smaller fraction of samples with GWAS data, and thus have lower power). Risk prediction is performed using logistic regression evaluated on datasets simulated from allele frequencies and odds ratios taken from replication data. PD: Parkinson's Disease (International Parkinson's Disease Genomics Consortium and Wellcome Trust Case Control Consortium 2, 2011; Nalls et al., 2011), AMD: Age-related Macular Degeneration (Chen et al., 2010), T1D: Type 1 Diabetes (Clayton, 2009), T2D: Type 2 Diabetes (Voight et al., 2010), UC: Ulcerative Colitis (Anderson et al., 2011), CD: Crohn's Disease (Franke et al., 2010; Yazdanyar et al., 2009), RA: Rheumatoid Arthritis (Stahl et al., 2010), CAD: Coronary Artery Disease (Schunkert et al., 2011), BRCA: Breast Cancer (Turnbull et al., 2010), LOAD: Late-Onset Alzheimer's Disease (Harold et al., 2009; Corneveaux et al., 2010), MS: Multiple Sclerosis (De Jager et al., 2009), MDD: Major Depressive Disorder (Shyn et al., 2009), BP: Bipolar Disorder (Scott et al., 2009), SLE: Systemic Lupus Erythematosus (Harley et al., 2008), SZ: Schizophrenia (Purcell et al., 2009), CRCA: Colorectal Cancer (Houlston et al., 2008), PRCA: Prostate Cancer (Eeles et al., 2009), OVCA: Ovarian Cancer (Goode et al., 2010; Song et al., 2009).

most controversial forecast was about the advent of prediction for complex disease, and its role in medical practice. He told a hypothetical story about a patient (named John), visiting his doctor in 2010:

> After working through an interactive computer program that explains the benefits and risks of such tests, John agrees (and signs informed consent) to undergo 15 genetic tests that provide risk information for illnesses for which preventive strategies are available. [...]
>
> Confronted with the reality of his own genetic data, he arrives at that crucial "teachable moment" when a lifelong change in health-related behaviour, focused on reducing specific risks, is possible. And there is much to offer. By 2010, the field of pharmacogenomics has blossomed, and a prophylactic drug regimen based on the knowledge of John's personal genetic data can be precisely prescribed to reduce his cholesterol level and the risk of coronary artery disease to normal levels.

While this exact scenario was not common by 2010, personal genetic testing for disease risk has become available to those who want it (and are willing to pay). Many companies now carry out such tests, using genome-wide data, for a range of diseases (Ng et al., 2009). The largest such companies, such as 23andMe and deCODEme, provide testing for tens of thousands of customers a year (Wright and Gregory-Jones, 2010). The potential utility of such genetic risk prediction has been widely debated (Gulcher and Stefansson, 2010; Kraft and Hunter, 2009; Hall et al., 2010).

Hundreds of GWAS and ever-larger meta-analyses have discovered a lengthening list of variants associated with complex disease, which can in

turn be used to construct disease predictors. Figure 1.2 shows the Area Under the ROC Curve (AUC) of predictors based on the current genetic knowledge of 18 diseases. In this context, the AUC can be interpreted as the probability that a genetic test could correctly identify the affected individual in a pair of individuals of which exactly one is affected. Many diseases cannot be well predicted (including virtually all psychiatric diseases and cancers), but others have relatively good predictive power (including type 1 diabetes, Crohn's disease and age-related macular degeneration). Note that, while the AUC is a useful indicator of predictive power, it needs to be considered in the context of the prevalence of the disease. For example, the low prevalence of Crohn's disease makes prediction difficult, even given the high predictive power of Crohn's GWAS loci.

The range of genetic AUCs for these diseases is very similar to the range found in classical (non-genetic) risk prediction based on epidemiological predictors (Lloyd-Jones et al., 2006; Cassidy et al., 2008; Seddon et al., 2009; Wacholder et al., 2010; Buijsse et al., 2011). There are additional advantages to genetic risk prediction compared to classical risk prediction, due to the fact that genetics do not change over an individual's lifetime. This means that risk models can be fitted with retrospective genotype data without fear of confounding, and that risk prediction can be carried out much further in advance. For instance, genetics is better than classical risk factors in predicting type 2 diabetes more than 30 years in the future (Lyssenko et al., 2008). This may be important for cases where prevention is most effective if started long before disease onset, or carried out over a long period. However, when both genetic and non-genetic predictors are available, prospective studies are required to determine how much power genetic testing adds: common variants increase the AUC of risk prediction from 0.76 to 0.83 in age-related

macular degeneration (Seddon et al., 2009), but add negligible improvement for prediction of metabolic diseases (Companioni et al., 2011; Buijsse et al., 2011).

Of course, these numbers only tell part of the story. To properly assess the utility of genetic risk prediction, it must be considered in the context of the cost of testing, the actionability of the results, and the framework in which these results will be used. Deciding the optimal way to use genetic risk prediction, and its potential utility in such an optimal framework, will be a significant challenge for medical practice in the future.

1.2 A brief history of human disease genetics

1.2.1 The age of molecular disease: 1940 to 1980

The concept of disease as influenced by hereditary factors originates at the turn of the 19th century, with the rise of family studies (discussed in Chapter 5). However, the modern formulation of disease genetics, characterised by the search for inherited polymorphisms in disease loci that increase or decrease disease risk, is a product of the mid-20th century.

The adoption of Mendelian laws of inheritance (Mendel, 1866) in the early 20th century led to the discovery that many diseases follow a Mendelian pattern of inheritance within families (Garrod, 1902; Punnett, 1908). While these early studies were before the discovery of DNA, and were thus unable to establish the genetic cause of these diseases, they nonetheless established that they were caused by the presence (or absence) of a specific molecular factor. It was the search for these molecular factors that led to rise of the molecular disease paradigm, and the discoveries of the first true disease loci.

In the 1940s a series of landmark experiments established the central dogma of heredity (Beadle and Tatum, 1941; Avery et al., 1944): DNA is the agent of heredity, and it acts via the production of proteins. The coming decades would see the structure of DNA solved (Watson and Crick, 1953) and the genetic code for proteins described (Crick et al., 1961). These discoveries gave us the modern framework of disease genetics: mutations in DNA lead to changes in the functioning of proteins, which in turn lead to defects in body function that cause disease.

The age of molecular disease lasted from the establishment of the central dogma to the rise of recombinant DNA techniques in the 1970s. It was characterised by an increasing understanding of the action of proteins in

disease, and the resulting discovery of inherited functional polymorphisms that underlie them. The first disease to be explained in molecular terms was sickle cell disease, which in 1949 Linus Pauling and colleagues showed to be caused by differences in the activity and amino acid composition of the haemoglobin protein (Pauling et al., 1949). Remarkably, a single amino acid sequence difference underlying this disease was discovered only 8 years later (Ingram, 1957), though the gene itself was not cloned and mapped until the late 1970s (Lawn et al., 1978; Deisseroth et al., 1978). Other successes rapidly followed, such as the discovery of the enzymatic cause of phenylketonuria in 1953 (Jervis, 1953).

One group of proteins that were first understood in this period were the proteins of human leukocyte antigen (HLA) system. First identified as important in matching donor and host tissue for transplant, in the course of the 1960s and 70s the HLA came to be recognised as having a centrally important role in diseases of immunity (Dick, 1978). Many associations between HLA alleles and immune-mediated diseases were discovered at this time, including relatively simple associations with a single HLA allele, and more complex associations with multiple HLA alleles (such as those in type 1 diabetes (Cudworth and Festenstein, 1978)). The HLA has been under almost constant study as a source of risk alleles for the last 50 years.

The above disease loci were identified in an essentially "backward" manner. The disease biology led to the investigation of a candidate protein, which in turn led to the discovery of pathogenic variation and, eventually, mapping of disease genes. While this process was "molecular", it was not truly "genetic" in the modern sense, in that it did not proceed from DNA. The first truly genetic programme for the study of disease came with the development of recombinant DNA technology, and the sequential rises of linkage,

candidate gene and genome-wide association studies.

For a whirlwind tour of the 40 years I am about to describe, one only needs to look at the study of the HLA regions in type 1 diabetes. The HLA association with diabetes was first identified via HLA typing in the 1970s (Cudworth and Festenstein, 1978). The strongest signal was localised to the *HLA-D* region in the early 1980s via linkage to restriction fragment length polymorphisms (RFLPs). Fine-mapping of this signal to the gene *HLA-DQB*, however, had to wait until the late 1980s and the rise of the polymerase chain reaction (PCR) (Todd et al., 1987). Even then, a full characterisation of all the different HLA associations in diabetes had to wait for the development of microarray genotyping at the turn of the 20th century (Nejentsev et al., 2007), forty years after the association was first reported. The same locus identified during the early days of molecular disease studies has taken four decades of technological advance to crack.

1.2.2 The age of linkage for Mendelian traits: 1980-1994

The concept of linkage is an old one. In essence, linkage involves discovering the relative positions of different genetic markers by measuring their coinheritance within families. Markers that are present on the same chromosome are more likely to be coinherited than would be expected by chance, and markers that are closer together on the genome are even more likely to be coinherited, as recombination is less likely to separate them. For a fully penetrant Mendelian disease, presence of a mutation is synonymous with disease status, and thus linkage can be used to determine the location of the mutated gene on a genetic map.

Linkage studies have a sophisticated statistical heritage. In the 1930s both Haldane (1934) and Fisher (1935) described statistical methods for

detecting genetic linkage between dominant traits. Linkage has undergone constant statistical refinement for over half a century, with the development of the parametric LOD score (Morton, 1955), pedigree likelihood modelling (Elston and Stewart, 1971), the multipoint Lander-Green algorithm (Lander and Green, 1987), Non-Parametric Linkage (Kruglyak et al., 1996) and the development of sparse gene flow trees (Abecasis et al., 2002). Each of these statistical developments has been in response to the development of linkage from small-scale breeding experiments to massive whole-genome meta-analyses with hundreds of markers and thousands of individuals.

The original linkage maps were based on physical characteristics, and were almost exclusively generated for model organisms via breeding experiments. For instance, in 1940 the chicken linkage map consisted of 6 chromosomes with a total of 21 genes, each defined by mutant phenotype (Hutt et al., 1940). This specified that, for instance, there were 10 centimorgans between the genes that produce the Silkie and Flightless phenotypes. While these maps allowed the first real understandings of genome structure, they were of limited use for human disease. Firstly, without selective breeding, multiple obviously Mendelian traits rarely segregated in the same family, so the maps were difficult to produce. Secondly, the information provided was of little direct relevance, since there existed no method of turning location on a linkage map into biological insight.

Technological revolutions during the 1970s provided a platform for linkage studies of human disease to come of age. This began with the development of amplification in DNA within viral or bacterial vectors (Jackson et al., 1972), and developed rapidly to sequencing of entire genes by dye termination (Sanger) sequencing (Sanger et al., 1977). These developments meant that, if the location of a gene could be identified, it could theoretically lead to

the gene being cloned and sequenced, its protein sequence determined and its tissue expression distribution characterised. The development of the Southern blot during the same period (Southern, 1975) allowed easy genotyping of RFLPs (variants in the DNA that interfered with the action of restriction enzymes). This was the first time that genotypes could be efficiently measured from DNA itself, and led to the development of human linkage maps without the need for mutation phenotypes (Botstein et al., 1980). Suddenly, discovering disease loci by linkage became both possible, and potentially highly biologically informative.

It did not take long for linkage results to arrive in multiple Mendelian diseases. The first disease locus to be identified purely by linkage was Huntington's disease (via a very fortuitous study of only 12 RFLPs), followed soon by a flurry of papers reporting linkage to chromosome 7 in cystic fibrosis (Tsui et al., 1985; Knowlton et al., 1985; Wainwright et al., 1985; White et al., 1985). However, while these loci were rapidly identified, the journey from linkage to a mapped, cloned gene was often difficult. For instance, a large international collaboration was required to discover the *CFTR* gene and ΔF508 mutation that underlies cystic fibrosis, using a laborious positional cloning approach (Rommens et al., 1989; Riordan et al., 1989; Kerem et al., 1989). For Huntington's, discovering the responsible mutations took 10 years from when linkage was first detected (The Huntington's Disease Collaborative Research Group, 1993).

The number of samples and variants typed in these early studies were counted in double digits, and the researchers only managed to discover mutations with extremely high penetrance in diseases with simple genetic architecture. The methods used to solve these diseases required monumental effort to use, and seem primitive and laborious by modern standards. However, in

other ways they contained many of the essential principles of modern genetics. They used direct typing of DNA, without requiring any prior knowledge of the disease biology, to uncover disease loci. They utilised state-of-the-art technology, combined with rigorous statistical analysis, and in many cases shared data, samples and expertise across large, international consortia.

The success of this approach in solving these diseases inspired similar projects aimed at solving more challenging diseases. These early forays into the genetics of common complex diseases were less immediately successful. It would require a series of technological revolutions, combined with a number of false starts, before complex disease genetics would come of age.

1.2.3 The beginning of complex disease genetics: 1994-2005

The diseases described in the previous section are all Mendelian diseases. These diseases are caused by a mutation in a single gene, and this mutation (and thus the disease itself) is passed on to offspring in a Mendelian fashion. However, many diseases, including virtually all diseases with prevalence greater than around 1 in 500, are complex diseases. These include most immune-mediated diseases, such as type 1 diabetes, Crohn's disease and rheumatoid arthritis, most metabolic diseases such as cardiovascular disease and type 2 diabetes, and most cancers. They do not appear to have a single cause (genetic or otherwise), but most have been known from families to have a genetic component since the early 20th century (see Chapter 5). In the 1990s, many geneticists turned their attention to the genetic underpinnings of these complex diseases.

The RFLP linkage approach had some ability to detect common alleles of unusually large effect in complex diseases, including the discoveries of

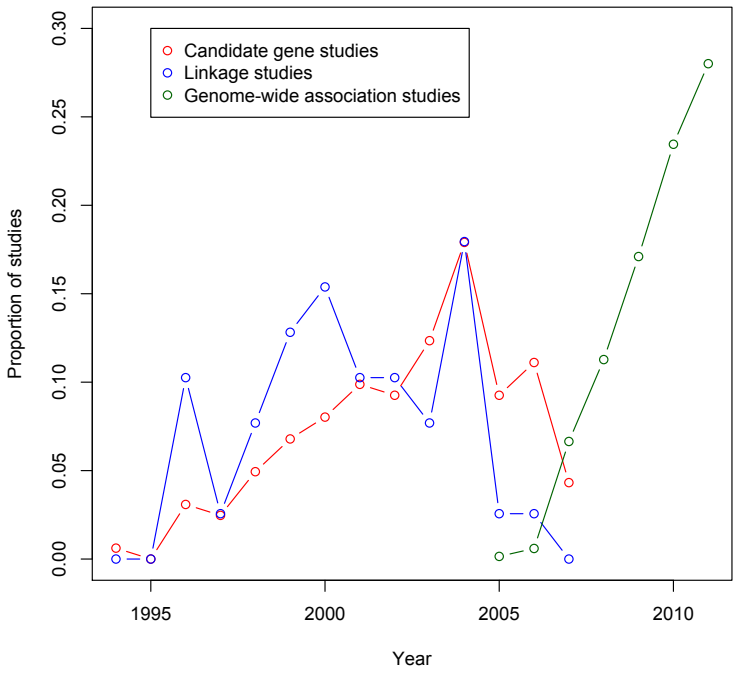

Figure 1.3: A timeline of complex disease genetics. Candidate gene studies (prior to 2007) are taken from reviews by Bosker et al. (2011) and Morgan et al. (2007). Linkage studies (prior to 2007) are taken from reviews by Guan et al. (2008) and Baumgart and Carding (2007). GWAS taken form the NHGRI GWAS catalogue (Hindorff et al., 2009).

the *INS* locus in type 1 diabetes (Bell et al., 1984) and the *ApoE* locus in early onset Alzheimer's disease (St George-Hyslop et al., 1987; Goate et al., 1991). However, these discoveries were the exception, not the rule, and the high genetic heterogeneity and low effect sizes in complex disease made it ill suited to study using the old techniques. Another wave of technological innovation in the late 1980s and early 1990s fundamentally changed the way complex disease genetics was done.

In 1986, Kary Mullis and colleagues published the polymerase chain reaction (PCR), a method for rapidly amplifying specific DNA sequences in vivo (Mullis et al., 1986). This revolutionised the study of DNA. In 1989, Variable Number Tandem Repeats (VNTRs) were described as a class of

variant easily genotyped by PCR (Weber and May, 1989), and linkage maps based on VNTRs appeared soon after. Additionally, by 1993 the TaqMan system was being used to genotype SNPs and small indels using PCR (Lee et al., 1993). These new techniques allowed genotyping of denser maps, in many more samples, at much lower cost than the old techniques. As well as making studies into many more Mendelian diseases affordable, this new technology also drove an explosion of studies into the genetics of complex disease, including both genome-wide linkage studies and association studies of candidate genes (see Figure 1.3).

The first success of the new linkage technology was the discovery in 1990 of strong linkage in early onset breast cancer (Miki et al., 1994) (soon generalised to all breast cancer (Margaritte et al., 1992)). The new techniques also allowed relatively rapid mapping of the causal gene (*BRCA1*) in less than four years (Miki et al., 1994). There were also notable early successes in type 1 diabetes, including replication of the *INS* association using linkage (Bain et al., 1992), along with the discovery that it was driven by VNTR variation in the gene itself (Bennett et al., 1995), and confirmation of a third linkage driven by a mutation in *CTLA4* (Nistico et al., 1996). Later successes include the discovery of linkage (Jawaheer et al., 2003) and then association (Begovich et al., 2004) to *PTPN22* in rheumatoid arthritis, and the detection of linkage (Hampe et al., 1999) and then causal variants (Hugot et al., 2001) in the gene *NOD2* in Crohn's disease.

Despite these successes, however, many of the linkage peaks discovered were sporadic, and could not be consistently replicated. Even more disappointing was the failure of linkage meta-analysis. The Genome Search Meta-analysis (GSMA) method (Wise et al., 1999) was introduced in 1999 to allow the results of linkage scans to be combined without sharing genotyp-

ing data, and theory created the possibility of very highly powered linkage studies. However, when the large linkage meta-analyses arrived, including thousands of affected families and representing millions of dollars of total investment, they produced almost no significant, novel results (van Heel et al., 2004; Guan et al., 2008; Concannon et al., 2009).

In retrospect, the relative failure of later linkage studies was a result of the high genetic heterogeneity and low effect sizes of complex disease associations (a fact later uncovered by GWAS). It has long been known that the power of across-family linkage falls off very rapidly with effect size and allele frequency (Risch and Merikangas, 1996), meaning that even the large linkage meta-analyses would not be well powered to detect true associations.

The history of candidate gene studies is an even more chequered. The advent of relatively inexpensive genotyping, combined with gene mapping and variant discovery efforts, made it possible to select at least one SNP in a candidate gene and test it for association to a disease of interest. A large number of associations were identified in this manner. There were some notable successes that have stood the test of time, such as the discovery of the *PPARG* association in type 2 diabetes (Altshuler et al., 2000)). However, in general less than 5% of associations identified in candidate gene studies were replicated in larger GWAS (Ioannidis et al., 2011), suggesting that, on the whole, candidate gene studies failed to reliably identify true associations. This failure is especially worrying given the fact that many candidate gene studies are still carried out today.

The reasons for this failure have been widely debated. The use of post-hoc adjustment to push p-values into nominal significance has been suggested (as has been demonstrated in other fields (Masicampo and Lalande, 2012)), often with an implication that this is a result of "hypothesis driven" investigators

pushing their pet gene. However, I believe that most of the failure of candidate gene studies follows naturally from the sample size, p-value thresholds and the (then unknown) distribution of effect sizes in truly associated loci.

Examining the candidate gene studies for major depression reviewed by Bosker et al. (2011), we find that half of the positive studies reported a p-value between 0.01 and 0.05, and that the median effective sample size was 170 cases and 170 controls. Even under optimistic assumptions that odds ratios are large (>2) and the SNP selection criteria is good (one in 20 is truly associated), this will produce false positives 49% of the time. However, from GWAS we now know that the typical odds ratio is closer to 1.25, which increases the rate of false positives to over 80%. In practice, a more appropriate set of criteria for candidate gene studies would be to use $p < 0.005$ and $N > 1500$, which would give a 60% true positive rate even given a 1 in 100 success rate in candidate SNP selection and an odds ratio of 1.25. These are approximately the criteria used by Altshuler et al. (2000) to successfully establish the true *PPARG* association in type 2 diabetes. The majority of candidate gene studies, however, fell well short of these criteria, and were thus doomed to failure from the start.

By 2005, a small number of important new disease associations had been identified. Many of these triggered new scientific investigations, such as the role of innate immunity in Crohn's disease inspired by the discovery of *NOD2*. Others led to new developments in patient care, such as the (soon routine) testing of *BRCA1* mutations in individuals with a family history of breast cancer. Others still generated significant social debate, notably the strong *ApoE* association in Alzheimer's disease. However, while the genes identified were important, they were not many of them, with no diseases having more than two or three loci identified. Ultimately, it would take the technological

developments accompanying the Human Genome Project to increase the pace of locus discovery.

1.2.4 The technological build-up to genome-wide association studies: 1986-2005

The idea of a genome-wide association study (GWAS) was established even in 1996, when Risch and Merikangas (1996) noted the greater power of association testing compared to linkage in almost all scenarios, but especially for lower effect sizes (OR < 2). They suggested that by mapping polymorphisms genome-wide, the Human Genome Project would allow the creation of high-density polymorphism maps that, when combined with advances in genotyping technology, would allow well-powered association testing across all genes. In this design, a large number of cases (probably the cases already collected as part of linkage studies) would be genotyped throughout the genome, along with a set of controls, and each variant could be tested for differences in frequency between cases and controls. Again, the concept and the statistics were well established, and waiting for the technology to catch up. In this case, the technology consisted of advances in DNA sequencing and SNP discovery, and the development of DNA microarrays for large-scale genotyping.

In 1986, a description of the first automated DNA sequencing machine was published (Smith et al., 1986). This machine used 4-colour dye termination, separated fragments through gel electrophoresis and imaged them digitally. It was commercialised as the ABI 370-series, and at its peak a single machine could produce 7200 bp (base pairs) of sequence per hour (Dovichi, 1997). In 1996 ABI released its first capillary sequencing machine, the ABI 310, followed two years later by the 96-capillary ABI 3700-series, capable of

producing approximately 80kbp of sequence per hour (Dovichi, 1997). This was the technology that drove the sequencing of the human genome, the first full drafts of which were published in 2001 (Lander et al., 2001; Venter et al., 2001).

Simultaneously with the sequencing of the reference genome, many groups were discovering and cataloguing human genetic variation. dbSNP was founded in 1998, and by 1999 held 4713 unique variants (Sherry et al., 1999). This number did not stay this small for long: in 2001 the SNP Consortium published its list of 1.42M SNPs discovered during and alongside the Human Genome Project (Sachidanandam et al., 2001). In the same year, Mark Daly and colleagues published a study of linkage disequilibrium structure on chromosome 5 (Daly et al., 2001), and noted that SNPs tended to form LD blocks. This was soon confirmed independently on chromosome 21 (Patil et al., 2001). The importance of these LD blocks were reinforced by the discovery that a large proportion of recombination occurs in recombination hotspots (McVean et al., 2004). These observation made association studies based on a limited number of SNPs (so-called "tag SNPs") more plausible, and led to the founding of the HapMap project in 2002 (International HapMap Consortium, 2003). The HapMap Project set out to discover and characterise genetic variation within and across human populations, and by 2005 had brought the number of known SNPs up to 9.2M, 1M of which were genotyped in a reference panel of 270 individuals on a range of technologies (International HapMap Consortium, 2005). The project went on to genotype far more SNPs (3.1M) in the same samples using Perlegen technology (Hinds et al., 2005), and genotype 1.6M SNPs on an extended panel of 1184 individuals using Affymetrix and Illumina technology (Altshuler et al., 2010). The dataset generated by the HapMap project provided a backbone for

genome-wide association studies, locating hotspots and providing a resource for designing tag SNP sets across different populations.

Meanwhile, technology was advancing to allow these newly discovered variants to be genotyped efficiently. During the 1980s, many groups were working on parallelising Southern blotting. While a Southern blot allows the detection of a specific DNA sequence via binding to an oligonucleotide, it could only be performed one oligo at a time, making it costly and slow. A better solution would be a system where binding to a large number of oligos could be tested simultaneously. The publication of massively parallel light-directed synthesis in 1991 (Fodor et al., 1991) allowed sequences of DNA to be "printed" onto a chip, which could in turn be hybridised to a sample of DNA and digitally imaged. This technology was commercialised as the Affymetrix microarrays, with the first chip containing 64 kbp of sequence to assay the HIV genome for mutations (Lipshutz et al., 1995). The same approach was soon applied to human SNP variation, with a prototype chip being used to genotype 500 SNPs simultaneously in 1998 (Wang et al., 1998).

Throughout the early 2000s, a flurry of companies commercialised methods for genome-wide SNP genotyping, using a variety of methods and technologies (Syvanen, 2005). In retrospect, the most significant were Affymetrix and Illumina, whose chips went on to underlie most of the GWAS to date. Each used a slightly different form of microarray, but they also differed in their selection of SNPs: Affymetrix used a random selection of SNPs, whereas Illumina used a set of tag SNPs designed to maximise coverage in Europeans (Barrett and Cardon, 2006). Affymetrix released its 10K Mapping Array in 2003 (Matsuzaki et al., 2004b), which it quickly expanded to 100K SNPs in 2004 (Matsuzaki et al., 2004a) and 500K in 2006. Illumina released its GoldenGate BeadChip system for genotyping approximately 1200 SNPs in 2002

(Fan et al., 2003), followed by the Infinium chips, which in 2005 could genotype 100K SNPs, moving rapidly up to 650K SNPs in 2006. Higher density chips, capable of genotyping a million SNPs, followed from both companies, with the Illumina Human1M chip in 2007 and the Affymetrix SNP 6.0 array in 2008.

1.2.5 The age of genome-wide association studies: 2005-Present

By 2005, the technology for GWAS was in place. Genome-wide SNP sets that tagged the majority of common variation were on the market, with the possibility of performing statistical imputation (see Chapter 3) via the HapMap data to assay millions of SNPs. DNA microarrays were commercially available to genotype these SNPs in thousands of individuals. Additionally, many sample collections, originally collected for large linkage analyses, were already sitting in freezers ready for study.

The first published GWAS, a study of age-related macular degeneration (AMD), involved only 96 cases and 50 controls genotyped on the Affymetrix 100K chip. Despite the small sample size, they identified a strong, common association with a coding variant in the *CFH* gene (Klein et al., 2005). Other early successes include the discovery of the important Crohn's disease gene *IL23R* in 2006 (Duerr et al., 2006), and a second association for AMD in the same year (Dewan et al., 2006).

However, while the early days of GWAS were characterised by dramatic successes, they also suffered some teething troubles, driven mostly by a lack of a standardised GWAS protocol. For instance, in 2006 a genome-wide study of 649 individuals reported an association between a variant in the gene *INSIG2* (Herbert et al., 2006) and childhood obesity. This association did not meet

the modern definition of "genome-wide significant" (GWS) ($p < 5 \times 10^{-8}$), and reports soon came in that the association did not replicate in independent cohorts (Dina et al., 2007; Loos et al., 2007; Rosskopf et al., 2007). Another early GWAS reported an association between memory performance and a variant in the gene *KIBRA* that did not meet genome-wide significance (Papassotiropoulos et al., 2006), which itself spawned a series of contradictory and inconclusive candidate gene studies (Schaper et al., 2008; Need et al., 2008; Bates et al., 2009) (exactly the situation GWAS was designed to prevent). Other early genome-wide association studies employed statistical techniques that seem somewhat unusual by modern standards (e.g. Liu et al. (2006)).

The watershed moment in genome-wide association studies was the publication of the first study from the Wellcome Trust Case Control Consortium (WTCCC) in 2007 (Wellcome Trust Case Control Consortium, 2007). The WTCCC was the largest set of GWAS of its time by a wide margin, including 3000 shared controls and 7 different phenotypes, each with 2000 samples. It cost a total of £9 million. The study identified 21 loci, of which 14 were novel. All but one of these associations have been confirmed in later meta-analyses.

The first WTCCC study applied a number of techniques and protocols for the first time, many of which became standards in genome-wide association studies. The study gave a detailed treatment to population stratification, ensuring that associations were not driven by systematic differences between cases and controls. It was the first GWAS to use the HapMap data to perform genotype imputation (using the newly developed IMPUTE algorithm (Marchini et al., 2007)), allowing testing of variants that hadn't been directly genotyped. It also gave significant attention to genotype calling, developing a new calling algorithm, and ensuring that all associated SNPs were manually

inspected. Not all of these were novel techniques, but the WTCCC cemented these steps into a protocol that later GWAS followed.

Another aspect of the WTCCC was the extensive replication efforts that followed it. Both SNPs that passed genome-wide significance, and (importantly) SNPs that showed suggestive but not conclusive evidence in the original scan, were taken forward for replication in extensive cohorts. These studies, which included type 2 diabetes (Zeggini et al., 2007), rheumatoid arthritis (Thomson et al., 2007; Barton et al., 2008), Crohn's disease (Parkes et al., 2007) and type 1 diabetes (Todd et al., 2007), led to the establishment of many new associations. It also established the importance of performing replication in independent samples, using independent technologies, in order to provide additional robustness to existing associations, and to cost-effectively identify new loci. This replication paradigm has become an important part of modern GWAS.

Over the last five years the number of GWAS per year has increased linearly (Figure 1.3). As the number of association studies increased, the next logical step was to combine studies together into meta-analyses (as was done during the linkage era). Early GWAS meta-analyses often consisted of pairwise collaborations, such as Samani et al. (2007), and often did not produce many more significant hits than the original GWAS. However, meta-analyses soon started producing startling results. The first Crohn's disease meta-analysis, consisting of three studies, discovered 21 new loci, bringing the total to 30 (Barrett et al., 2008) (more than the entire WTCCC), and the type 2 diabetes meta-analysis discovered six new loci for the previously very hard to crack disease (Zeggini et al., 2008). In 2009 the type 1 diabetes meta-analysis broke the record for the disease with the largest number of associations, with 40 loci (Barrett et al., 2009a), topped by the 71 Crohn's

disease loci in 2010 (Franke et al., 2010). For almost all diseases studies, the majority of associations now came from large consortium meta-analyses.

1.2.6 Technological advances post-GWAS: 2004-Present

Technological development did not halt with the advent of GWAS, and many new experimental techniques have been introduced in the last 5 years that are again dramatically altering the landscape of complex disease genetics.

The greatest leaps forwards have come in sequencing, with the advent of "next-generation" (sometimes called "second generation") sequencing. In 2004 the 454 pyrosequencing method was introduced, which allowed hundreds of thousands of sequencing reactions to be carried out in parallel (Langaee and Ronaghi, 2005). In 2006 Illumina commercialised the Solexa reversible termination sequencing method, and in 2007 ABI (now Life Tech) introduced the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) technology. By the end of 2007 it was possible to sequence over 500Mb a day on a single machine (Mardis, 2008). In the last few years other sequencing technologies have been introduced, including the small, low-cost "desktop sequencers" such as Illumina's MiSeq and Life Tech's Ion Torrent (Quail et al., 2012), and even more advanced technologies, such as nanopore sequencing (Eisenstein, 2012), are on the horizon. The rate of improvement in throughput has continued to climb, and at the time of writing the state of the art machines (e.g. Illumina's HiSeq 2500) can produce over 50Gb per day per machine. The cost of a high-quality fully sequenced human genome is now less than £5000 (Wetterstrand, 2012).

This technology spawned a new breed of systematic resequencing studies of human reference populations. In 2007 the 1000 Genomes Project was founded, to perform low-coverage (2-4X) sequencing on thousands of human

genomes. The project started with a pilot that detected 16M SNPs, indels and structural variants in 180 HapMap samples (Project, 2010). The full project will eventually sequence 2500 individuals from 25 populations, with the first phase producing calls for nearly 40M variants across 1092 individuals (Project, 2012). Unlike the HapMap, this dataset is a near-complete map of genetic variation in these samples, including all common SNPs and indels genotyped in all individuals, as well as an extensive catalogue of low frequency variation.

These results also underlie the development of a new generation of high-density genotyping chips, including the release of the Illumina Omni2.5, with 2.5 million SNPs, in 2010. Another result of this technology was the falling cost of designing custom genotyping chips, with the introduction of the Illumina iSelect high-density custom chips in 2006, and Affymetrix's Axiom system in 2010.

Other technological advances in sequencing followed these developments. In 2007 NimbleGen published their sequence capture technology (Albert et al., 2007), which used microarrays to pull down a specified subset of the genome, allowing low cost sequencing of a subset of the genome. This birthed the field of "whole exome sequencing", in which only the 1% of the genome coding that codes for proteins is sequenced. Interestingly, the benefits of this technology were first seen in the field of Mendelian diseases, where exome sequencing can identify all coding mutations in an individual's genome, and public databases (such as the 1000 Genomes Project) can exclude all polymorphic markers, leaving a small number of candidate causal mutations. The discovery of the causal mutation for Miller syndrome by exome sequencing (Ng et al., 2010) was rapidly followed by other successes, and this method is now the dominant method for solving Mendelian diseases (Bamshad et al.,

2011).

1.2.7 Next-generation GWAS and post-GWAS studies

The advent of GWAS has changed the landscape of complex disease genetics. In 2005 only a few dozen loci were known to be associated to complex diseases, across a handful of diseases. By the end of 2011, the NHGRI GWAS catalogue reported that GWAS have discovered over 2000 genome-wide significant associations for over 200 complex traits. But GWAS have their limits as a tool for locus discovery, and new methodologies are appearing the fill the gaps left by GWAS.

The tag SNP approach, the greatest strength of GWAS, is also its biggest limitation: a GWAS is only well powered to detect associations that are well covered by common tag SNPs. Populations with different LD to the HapMap populations, or meta-analyses across populations with different patterns of LD, can confound the tag SNP approach (Teo et al., 2010). This is especially problematic as many important diseases, including many infectious diseases, are more common in areas of the world with greater genetic diversity (e.g. Africa) or from areas that have been less well represented in reference panels (e.g. South Asia). Additionally, low frequency variants are not well tagged by common SNPs (Altshuler et al., 2010), making first generation GWAS ill-suited to discovering associations to such variants. This is an important limitation, as it has long been hypothesised that rare variants are likely to play an important role in complex disease (Pritchard, 2001). Finally, GWAS arrays are still relatively expensive, yet to discover loci with low-frequency or low-effect size risk variants we require tens or even hundreds of thousands of samples to be genotyped.

One potential method for overcoming problems of poor tagging is to use

a technique called genotype imputation, which can allow us to infer these poorly tagged sites statistically using the new sequence reference sets described above. As an example, the study of malaria in Africa has generally suffered from low LD and high diversity (Teo et al., 2010). However, a MalariaGEN study showed that genotype imputation using a well-matched reference set could overcome issues of low LD (The MalariaGEN Consortium, 2009). Similarly, imputation may allow us to assay associations at low frequency variation that is not well tagged by any one common SNP. Genotype imputation, combined with datasets such as that generated by the 1000 Genomes Project, may allow us to perform high-powered meta-analyses in African populations, and uncover new associations with low-frequency variants, without requiring more experimental genotyping.

The advent of low cost, high-density custom genotyping has allowed a many-fold expansion of genetic datasets of complex disease. By joining together in large meta-consortia, disease genetics consortia can club together to design genotyping chips. Because orders are large (>100,000 samples), chips can be purchased at very low cost, allowing very large sample sizes. The first example of such a chip was the Metabochip, designed to genotype 200,000 variants for deep replication and fine-mapping of metabolic and anthropometric traits (Cortes and Brown, 2011). The Metabochip has already expanded the number of known loci for both type 2 diabetes (Cortes and Brown, 2011) and glycemic traits (Scott et al., 2012). Other consortia have constructed similar platforms, including the Immunochip (for immune-mediated disease) and the Exome chip (to study coding variation).

The falling cost of sequencing has allowed the direct assaying of low-frequency variants via resequencing studies. Early studies involve the sequencing of sets of candidate regions using capture technology. A striking

early success came with the discovery of multiple rare variants in the gene *IFIH1* that protect again type 1 diabetes (Nejentsev et al., 2009). This study used 454 sequencing to sequence the exons of 10 candidate genes in 480 individuals, and marked the first major success of next-generation sequencing in complex disease genetics. A similar sequencing project in Crohn's disease identified a number of low frequency associated variants within existing GWAS loci, including a highly significant splice variant in the gene *CARD9* (Rivas et al., 2011).

Newer sequencing projects in complex diseases are focusing on whole-exome or whole-genome sequencing of case and control collections. Exome sequencing is relatively low cost, and can allow large sample sizes to be collected, but only allows us to study coding variation. A notable alternative approach is low-coverage, whole-genome sequencing, which is made plausible using the imputation-based genotype refinement techniques developed for the 1000 Genomes Project (Li et al., 2011). These techniques can allow us to infer genotypes in enough samples to test low-frequency variants genome-wide, at approximately the same cost of exome sequencing.

The success of whole-exome sequencing in solving Mendelian diseases has led people to ask whether family-based sequencing studies of complex disease may be able to identify low-frequency coding mutations that contribute to complex disease (Bamshad et al., 2011). While GWAS (and, indeed, the failure of linkage meta-analyses) ruled out the existence of high-frequency, high penetrance mutations (i.e. mutations likely to be shared between families), they do not rule out the possibility of rare variants of intermediate penetrance segregating with disease in a single family. The sequencing of multiply affected (or "multiplex") families, combined with new functional and genetic reference datasets, may allow us to identify such rare variants.

1.2.8 Conclusions

The history of locus discovery in human disease genetics has largely been a history of technology. The Southern blot and Sanger sequencing allowed the first disease genes to be mapped and cloned. PCR sparked the age of complex disease linkage and candidate gene studies, and microarrays and capillary sequencing led to GWAS. In each case, the general form of the studies were anticipated decades in advance, and the concepts underlying them were thus decades old by the time they came to be applied.

This is not a general property of genetics. For instance, sequence analysis has undergone a statistical renaissance in response to next-generation sequencing, with methodological advances in short read alignment (Ruffalo et al., 2011), de-novo assembly (Pop, 2009) and variant calling (Nielsen et al., 2011). It also has clear exceptions around chip design and processing, such as the development of tag SNP approaches (Li and Wang, 2010), of genotype calling algorithms (Shah et al., 2012) and of genotype imputation and methods to handle the resulting uncertainty (Marchini and Howie, 2010). But when it comes to locus discovery per-se, this conceptual preempting is the rule. Likewise, we are all aware that the ultimate locus discovery experiments will come within a few decades, via low-cost, high-quality whole-genome sequencing of hundreds of thousands of samples.

One effect of this technological drive is a tendency for statistical arguments to be raised, settled and often forgotten decades before the technology catches up. This can lead to a certain amount of historical blindness. Discussions of rare variants and genetic heterogeneity, for instance, seem to wax and then wane away every 10 years or so (with early family studies, with RFLP studies, with the failure of complex disease linkage, and in the GWAS era). Another effect is that methods can become ingrained, and used without

proper thought to what they mean. This was one of the reasons behind the failure of candidate gene studies, where a rule-of-thumb (a p-value threshold of 0.05) became a blindly applied law even in cases where it was not appropriate.

A more positive result of the established statistical methodologies is that far more attention is paid to downstream analysis of results. A good example of this is the development of gene prioritisation techniques, such as GRAIL (Raychaudhuri et al., 2009a) and DAPPLE (Rossin et al., 2011). A solid statistical framework is a platform that can easily be built upon to go beyond simple locus identification (e.g. see Chapter 4). This is especially important given that one of the main challenges of the next decade will be to turn the windfall of loci discovered by GWAS into detailed biological knowledge of disease.

1.3 Outline of this thesis

In this chapter, I have laid out the reasons for studying complex disease genetics in general, and the genetics of IBD in particular. I have shown how the process of locus discovery has proceeded over the last 70 years, and in particular how new technologies have continually opened up new avenues of research. We have seen that the greatest successes have come with the rise of genome-wide association studies, and in particular with large, collaborative GWAS meta-analyses. However, we have seen that there are still many loci to discover, as there are many classes of allele that the first generation of GWAS were unable to effectively study. I discussed how new technological advances are expanding our ability to study the gaps that GWAS left, and some of the strategies we can use to utilise these technologies to discover associations to rare and low-frequency variants, variants of small effect size and variants in diverse populations. The following chapters will lay out a series of investigations into the methods required, challenges faced and results generated by this next generation of studies.

However, before I describe these specific experiments, I will start by laying down a statistical framework to understand the methods and models that I am going to use. The twin studies used to infer heritability, the case-control studies used to discover risk variants, and the epidemiological studies that construct predictive models all use a related but distinct series of statistical methods. Likewise, many statements about genetic risk, such as the amount of heritability explained by GWAS, or the power of genetic risk prediction, are themselves built upon models of genetic risk. Throughout this thesis I make use of many of these different methods and models in the analysis of various datasets, and so before I report these analyses it is necessary to review this range of techniques, and unify them into a single rational framework.

To this end, Chapter 2 describes a family of models of genetic risk, built upon a normally distributed genetic risk score, with different models specified by different link functions connecting this risk score to disease probability. I show how the assumptions of most major statistical techniques correspond to a choice of one out of three link functions, and investigate the behaviour of these three models. I demonstrate that these models produce drastically different predictions about the distribution of observable quantities, and discuss how these differences can lead to inaccuracy or ambiguous results in studies of complex disease.

Once I have placed locus discovery efforts into both historical and statistical frameworks, I will proceed to describe a series of three projects designed to discover genetic risk factors in complex disease. Each of these projects is designed to extend, and overcome the limitations of, first-generation GWAS using a combination of new genetic data from patients, new publicly available genetic and functional datasets and new statistical techniques.

In Chapter 3, I investigate the use of genotype imputation algorithms in genome-wide association studies. As we saw above, genotype imputation can allow disease association to be tested with far more SNPs than have been genotyped in a GWAS, facilitating meta-analysis and increasing power. I begin by investigating the impact of reference set size and diversity on imputation in Europeans, using the HapMap data, with particular focus on the imputation of low frequency variants. I then investigate how effective the same reference sets are at performing imputation in African populations. Next, I expand this analysis to new datasets, looking at how well 1000 Genomes project data can impute low-frequency variation in a diverse African population. Finally, I show how imputation of variants from the 1000 Genomes pilot can be used to draw conclusions about disease biology,

by estimating the influence of loss-of-function variants on 7 complex diseases.

It is now clear from GWAS that a large proportion of disease risk is due to so-called polygenic risk. This consists of a large number of common variants, each with a small effect size, of which only those with the largest odds ratios have so far been identified. As we have seen, custom genotyping can allow us gather enough samples to identify loci in this long tail of low effect size polygenic risk. In Chapter 4, I discuss how a custom genotyping platform (the Immunochip) has been used to expand the IIBDGC GWAS meta-analyses collection to include over 40,000 cases of inflammatory bowel disease (IBD). This chapter details the analysis of this genotype data, including genotype calling, quality control, and association analysis. 71 new loci for IBD are described, bringing the total to 163 loci, with 193 genome-wide significant independent signals.

In order to biologically interpret this large list of associated loci, I present a number of bioinformatic analyses. This includes comparing genetic overlaps between the two forms of IBD (CD and UC), and the overlap between IBD and other complex and Mendelian diseases of immunity. It also includes gene prioritisation, functional enrichment and gene expression analyses. Finally, I outline two other projects that make use of the Immunochip data. The first is the use of Y chromosome markers to test relationships between Y chromosome haplogroups and IBD. The second is the use of densely genotyped fine-mapping regions on the Immunochip, combined with functional information, to draw conclusions about the nature and action of causal variants.

In contrast to the study of common variants of small effect, Chapter 5 describes a set of approaches to discover rare variants of large effect by using large, multiplex families. I begin by producing a joint model of common

polygenic and rare dominant penetrant genetic risk in families, and exploring how the probability of observing multiplex families of a certain size varies depending on heritability and penetrance. I then lay out a method of performing genetic risk prediction in families, and show that this method can effectively distinguish between multiplex families that do or do not harbour a penetrant mutation.

I go on to introduce a set of multiplex families with an abnormally high prevalence of IBD, including one extended family with over 40 affected individuals. I describe and apply an approach to studying such families using a combination of genotyping, whole-genome and/or whole-exome sequencing and functional annotation to detect candidate causal variants. I also discuss various methods by which these candidate variants can be validated and followed up.

In the final chapter I will highlight consistent themes and topics that tie together this thesis, including the importance of external datasets, the interplay between statistical and biological theory, and the nature of experimental design in the post-GWAS world. Next, I will look forward to locus discovery efforts in the near future and beyond. This will involve the description of a currently ongoing experiment involving low-coverage whole-genome sequencing of 5000 IBD patients and 4000 healthy controls, in order to identify low-frequency associations. Finally, I will consider the "ideal" locus discovery experiments of the coming decades, and the potential for an increased integration of genetic and functional biology.