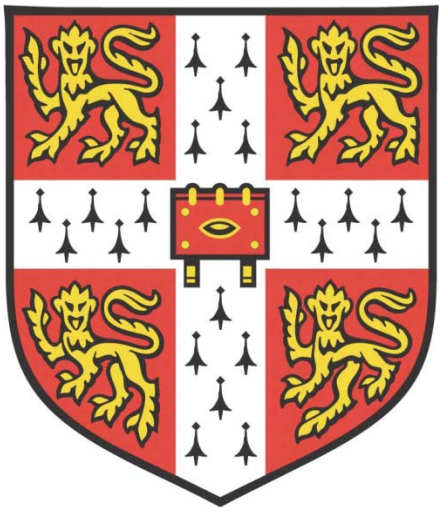
The Genetic Architecture of Immune-Mediated Complex Diseases



Jimmy Zhenli Liu

Darwin College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2014

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the Contributions section within each chapter. It does not exceed the word limit set by the Degree Committee for the Faculty of Biology, and is not substantially the same as any work that has been, or is being, submitted to any other university degree, diploma or any other qualification.

Jimmy Liu

18 September, 2014

That’s the whole problem with science. You’ve got a bunch of empiricists trying to describe things of unimaginable wonder.

— Bill Watterson, Calvin and Hobbes

Abstract

Complex disease risk is characterised by a combination of multiple genetic factors along with the environment. Since 2005, genome-wide association studies have discovered thousands of genetic variants associated with hundreds of such diseases. Following on from these types of studies, custom genotyping arrays with dense SNP content have allowed for greater refinement across risk loci, while their low cost has enabled powerful locus discovery projects and cross-phenotype comparisons in very large sample sizes. Combining risk loci with disease-relevant functional genomic data allows for insights into the biology of disease. In this dissertation, I explore locus discovery, cross-phenotype comparisons and functional data integration across four immune-mediated complex diseases – primary biliary cirrhosis, primary sclerosing cholangitis, and the two forms of inflammatory bowel disease – Crohn's disease and ulcerative colitis.

In Chapter 1, I provide a historical background of our understanding of how genetic variation contributes to phenotypic variation, and the technological and theoretical advances in the last twenty years that have lead to the large-scale high-throughput locus discovery projects of today.

In Chapter 2, I describe a locus discovery project using the Immunochip custom genotyping array for primary biliary cirrhosis. In addition to identifying three new risk loci and refining associated variants within known risk loci, I explore how integrating association results with functional genomic annotations across various cell lines from the ENCODE Project can provide insights into the cell types and genomic features most relevant to disease.

In Chapter 3, I describe a similar locus discovery project using the Immunochip for primary sclerosing cholangitis (PSC), where nine novel risk loci were identified. Over 80% of PSC patients are also diagnosed with inflammatory bowel disease, the majority of which is ulcerative colitis. I explore genetic factors

that may explain this overlap, and show that despite this high comorbidity, around half of PSC risk loci appear unique to PSC.

In Chapter 4, I describe a trans-ethnic genome-wide association meta-analysis for inflammatory bowel disease (IBD) comprising individuals of European, East Asian, Indian and Iranian ancestry genotyped on a combination of genome-wide arrays and the Immunochip. Forty new IBD loci were discovered associated with Crohn's disease, ulcerative colitis or both. I show that there exists pervasive sharing of IBD risk loci between European and non-European populations, while also noting specific loci where effect sizes differ between populations. The study demonstrates the utility of performing large-scale GWAS meta-analyses across different populations to identify novel susceptibility loci.

I then move beyond locus discovery in Chapter 5, where I describe a simple method for integrating differential gene expression datasets with disease risk loci. I applied the method to two gene expression datasets reflecting the genes that are involved in maintaining intestinal T cell homeostasis, and those triggered in the gut in 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 autoimmune disorders, 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 of the preceding chapters on unravelling the genetic architecture of complex diseases. I then look to the types locus discovery projects that will shape the field in the coming years, and the potential for these to be ultimately translated into better treatment outcomes for patients.

Acknowledgements

First and foremost, I thank my supervisor, Carl Anderson, for giving me the opportunity to pursue this PhD. It has been an absolute privilege. This dissertation would not have been possible without your continued guidance, enthusiasm and ceaseless faith in me throughout these years. With your stubborn attention to detail and unrelenting loyalty to rigour, you stand as a role model scientist for myself and no doubt many others to come.

I also thank my secondary supervisor, Jeff Barrett, and my degree committee, Stephen Sawcer and Ines Barroso for their guidance over these years. Thank you also to Christina Hedberg-Delouka, Annabel Smith, Alex Bateman and Julian Rayner for keeping the Sanger PhD program such a well-oiled machine.

To members of the Anderson Group (both past and present) - Tejas Shah, Eva Serra, Sun-Gou Ji and Jamie Floyd - I could not have asked for nicer folks to share an office with. It's been an absolute pleasure working with you all; thank you for putting up with me.

The work presented in this dissertation would not have been possible without the efforts of collaborators both at Sanger and around the world, of which there are far too many to list here. But for their hard work, dedication and willingness to share the spoils of research, I am especially indebted to Mohammed Al Marri, Luke Jostins, Daniel Gaffney, Richard Sandford, Trine Folseraas, Tom Hemming Karlsen, Johannes Roksund Hov, Eva Ellinghaus, Andre Franke, Tim Raine, Adam Reid, Suzanne van Sommeren, Rinse Weersma and Hailiang Huang. Thank you also to legions of doctors, nurses, researchers and administrators of the UK PBC Consortium, the International PSC Genetics Consortium and the International IBD Genetics Consortium for their tireless efforts in bringing together groups around the world towards the common noble goal of advancing disease research. None of this of course would have been possible without the >100,000 donors whose DNA were used in these projects, for which I will be forever grateful.

I would also like to thank the many friends and colleagues I've gotten to know during my time at Sanger. Chris Franklin, Yang Luo, James Morris, Scott Shooter, Isabelle Cleynan, Mari Niemi and everyone in Morgan N333 and N309 - cheers for the lunchtime banter, post-lunchtime strolls, post-stroll tea breaks, mingles and thoughts. To everyone I've enjoyed playing with (and competing against) in the Genome Campus volleyball, football and cricket leagues, I tried my best and can only apologise. Thank you to the fellow PhD students and Sanger and those I've gotten to know in Cambridge for making my time here so enjoyable.

I also thank the Wellcome Trust for generously funding my time at Sanger, as well as the various funding bodies for making these projects possible.

Lastly, I want to thank my family, especially my parents, Hua and Xiaoyu, for the unconditional love, support, understanding and encouragement that they have shown me in everything that I do.

Table of Contents

Publications

From this dissertation

Liu J.Z., van Sommeran S., Huang H., Ng S.C. *et al.*, Association study discovers 38 susceptibility loci for inflammatory bowel disease and shows pervasive sharing of genetic risk across diverse populations. *Under review.*

Raine T., Liu J.Z., Anderson C.A., Parkes M. and Kaser A., Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. Gut. *In press.*

Liu J.Z. and Anderson C.A., Genetic studies of Crohn's disease: past, present and future. Best Practice & Research: Clinical Gastroenterology, 28:373-386, 2014.

Foth B.J., Isheng J.T., Reid A.J., Bancroft A., *et al.*, Whipworm genomes and dual-species transcriptome analysis provide molecular insights into an intimate host-parasite interaction. Nature Genetics, 46:693-700, 2014.

Liu J.Z., Hov J.R., Folseraas T., Ellinghaus E., *et al.*, Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nature Genetics 45:670-675, 2013.

Liu, J.Z., Almarri, M.A., Gaffney, D.J., Mells, G.F., Jostins, L., Cordell, H.J., Ducker, S.J., Day, D.B., Heneghan, M.A., Neuberger, J.M., *et al*. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nature Genetics, 44:1137-1141, 2012.

Arising elsewhere

Curtis J., Luo Y., Zenner H.L., Cuche-Lourenco D., *et al.,* Susceptibility to tuberculosis is associated with the ASAP1 gene that regulates dendritic cell migration. *Under Review.*

Houldcroft C.J., Petrova V., Liu J.Z., Frampton D., *et al.,* Host genetic variants and gene expression patterns associated with Epstein-Barr virus copy number in lymphoblastoid cell lines. PLoS One 9:e108384, 2014*.*

Robles-Espinoza C.D., Harland M., Ramsay A.J., Aoude L.G., *et al.,* POT1 loss-of-function variants predispose to familial melanoma. Nature Genetics, 46:478-481, 2014.

Shah T.S., Liu J.Z., Floyd J.A.B., Morris J.A., *et al.,* optiCall: A robust genotyping-calling algorithm for rare, low frequency and common variants. Bioinformatics 28:1598-1603, 2012.

List of tables

List of figures

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).

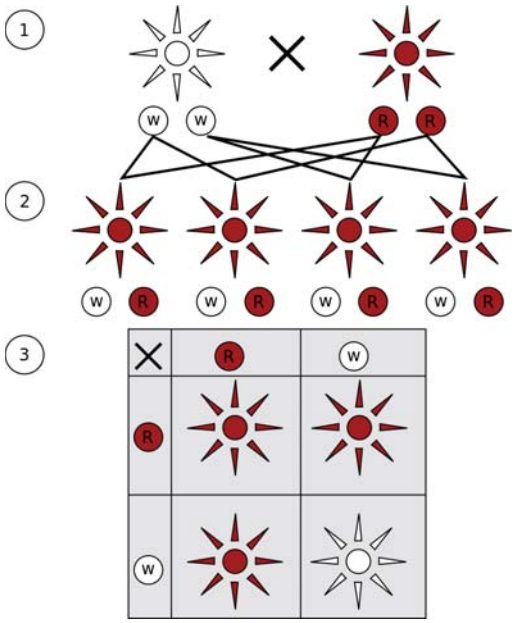


Figure 1.1. Mendel's laws of inheritance. In this example, there are two alleles: W and R which give rise to either a white or red phenotype respectively when both copies are present. Red is dominant and white is recessive. In (1) the parental generation, the parents are homozygotes for each of the alleles. 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).

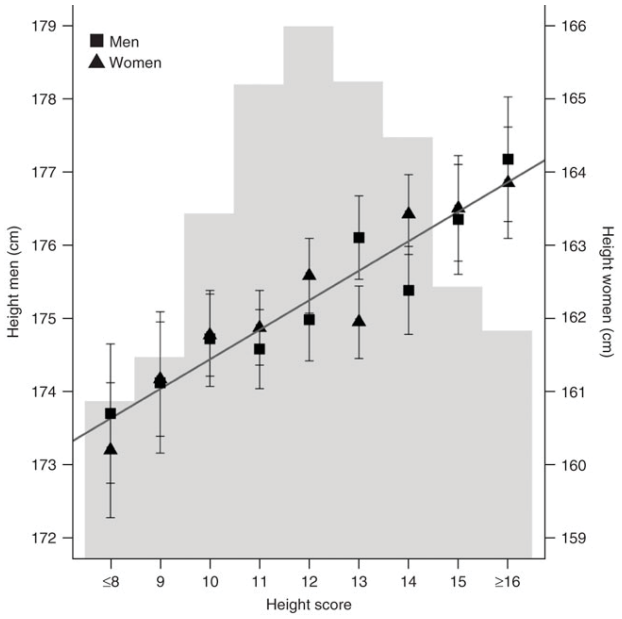


Figure 1.2. Polygenic inheritance in a normally distributed trait: height. Using 12 SNPs associated with height, 7,566 individuals were grouped according to the number of height-increasing alleles they carried (height score on x-axis). The gray bars represent the fraction of individuals in each height 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)

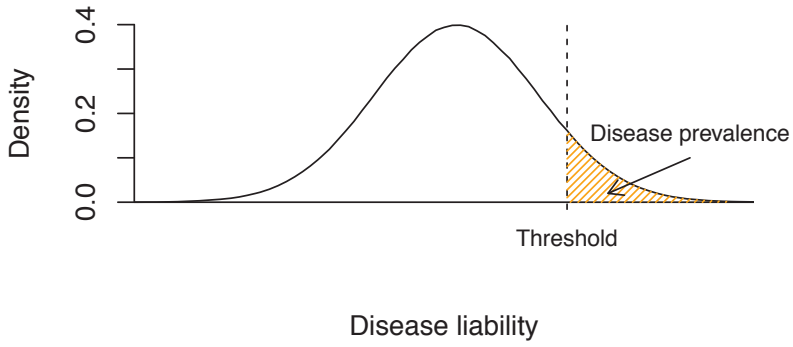


Figure 1.3. The liability threshold model. Disease liability can be thought of as a continuous trait that incorporates all environmental and genetic risk factors of a disease and is normally distributed in the population. Individuals who exceed a given threshold (dashed vertical line) will be affected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

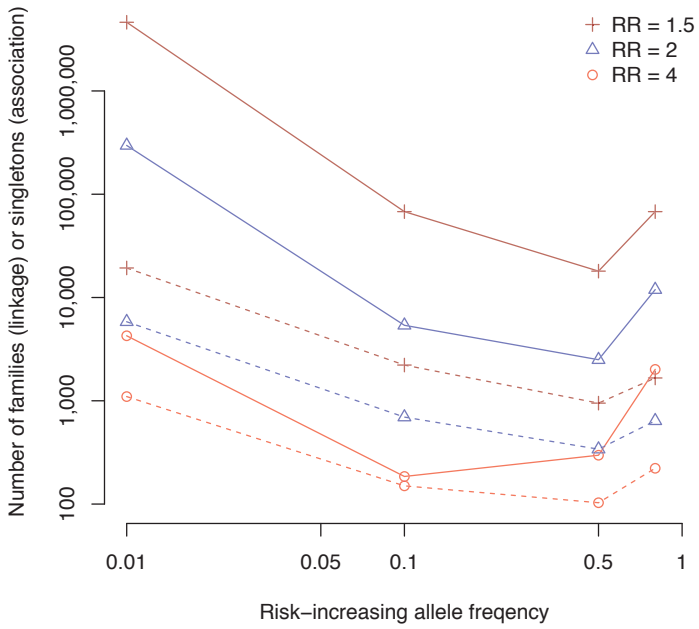


Figure 1.4. Power of linkage vs. association outlined in Risch and Merikengas (1996). The minimum number of samples required to detect a genetic variant with genotypic relative risks of 1.5, 2 and 4 at 80% power (at genome-wide significance) are plotted for linkage studies using related individuals (solid 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.

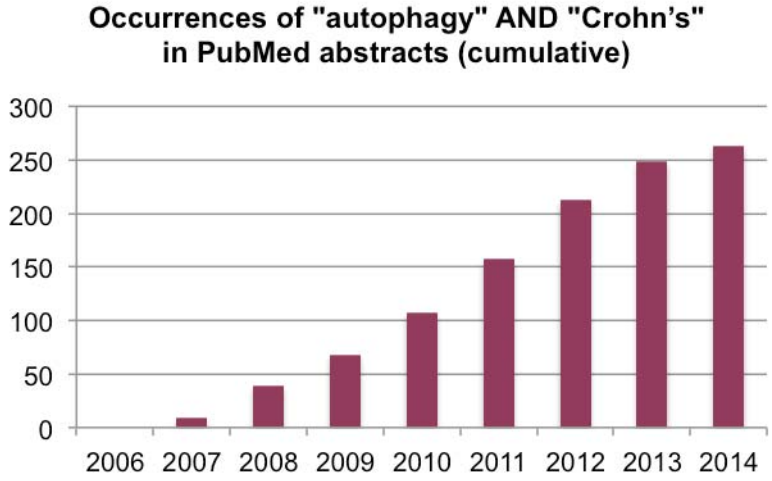


Figure 1.5. Number of publications indexed in PubMed with the terms "autophagy" and "Crohn's" in the abstract since 2006.

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.

Chapter 2. Discovery, refinement and functional genomics integration of primary biliary cirrhosis risk loci using the Immunochip

2.1 Introduction

Primary biliary cirrhosis (PBC) is characterized by the immune-mediated destruction of intra-hepatic bile ducts, resulting in chronic cholangitis, liver fibrosis and ultimately cirrhosis (Kaplan and Gershwin, 2005). With a UK prevalence of 35:100,000, rising to 94:100,000 women over 40 years of age, it is the most common autoimmune liver disorder (James *et al.,* 1999; Kaplan and Gershwin, 2005). Family-based studies indicate a substantial genetic component to PBC susceptibility, with a sibling recurrence risk of ~10.5 in the UK (Jones *et al.,* 1999). Genome-wide association studies (GtwitWAS) have identified 22 PBC risk loci, and highlighted the role of NFkB signaling, T-cell differentiation, Toll-like receptor and tumor necrosis factor signalling in disease pathogenesis (Hirschfield *et al.,* 2009; Liu *et al.,* 2010b; Mells *et al.,* 2011). Sixteen of these loci are also associated with other immune-mediated diseases such as multiple sclerosis, celiac disease and type 1 diabetes (T1D), shedding light on the involvement of common genes and pathways across these diseases (Zhernakova *et al.,* 2009). Despite these advances, the specific causal variant at many of these loci remains unknown.

To better define known risk variants and identify additional susceptibility loci, I performed an association study in 2,861 cases from the UK PBC

Consortium and 8,514 UK population controls from the 1958 British Birth Cohort and National Blood Service. All samples were genotyped using the Immunochip, an Illumina Infinium array containing 196,524 variants (718 small insertions/deletions and 195,806 SNPs). Two thirds of these variants reside in 186 loci with known associations with one or more autoimmune disorders, while most of the remaining variants were included as part of GWAS replication efforts for various autoimmune disorders (Cortes and Brown, 2011; Trynka *et al.,* 2011a). Compared with GWAS arrays, the Immunochip has increased marker density within known autoimmunity-associated loci, increasing the power to detect PBC associations within these selected candidate loci and providing a powerful means of fine mapping known PBC loci, as causal variants are more likely to be directly genotyped.

2.1.1 Chapter overview

In this chapter, I describe the results from an association study for PBC risk loci. In total, 19 loci reach genome-wide significance ($P < 5\times10^{-8}$), three of which are novel. One of these novel loci includes a low-frequency non-synonymous SNP in *TYK2*, further implicating JAK/STAT and cytokine signalling in PBC pathogenesis. Multiple independent common, low frequency and rare variant associations were found at five loci. Further investigation of one of these regions (3q25) showed that the most significantly associated signal in the locus was driven by a shared haplotype with two other SNPs, and that this top signal was no longer genome-wide significant when testing for association using a joint model of all signals in the region. Imputation and association testing of HLA haplotypes also confirmed three known independent genome-wide significant associations. Finally, I observed that 15 of the 26 independent non-HLA association signals overlapped with regions of open chromatin in B-lymphoblastoid cell lines as identified in the ENCODE project, though this was not significantly different compared to other cell lines when taking LD and the SNP composition on the Immunochip into account ($P = 0.06$).

2.1.2 Contributions

The study design was conceived by the Wellcome Trust Case Control Consortium 3 (WTCCC3) and the UK PBC Consortium. Case ascertainment and phenotyping were performed by the UK PBC Consortium. Controls were ascertained from the UK National Blood Service and the 1958 Birth Cohort Controls group. See Supplementary Note in Liu *et al*. (2012) for the full list of contributors. Sample and SNP quality control was performed by Mohamed Almarri. All other analyses, unless stated, were performed by myself.

2.2 Methods

2.2.1 Samples, DNA extraction and genotyping

All subjects were of self-declared British or Irish ancestry. Cases were collected by the UK PBC Consortium, which consists of 142 NHS trusts including all UK liver transplant centers. All individuals were over 18 years of age with probable or certain PBC. Three criteria were applied to diagnose the condition: a) a positive test for the presence of anitmitochondrial antibodies (titer 1:40 or higher), b) liver biopsy histology consistent with PBC, and c) liver biochemistry consistent with PBC (i.e. a higher level of bilirubin, aspartate transaminase, alanine transaminase, alkaline phosphatase or gamma-glutamyl transferase compared to the upper reference level). Diagnosis was documented as probable when two criteria were satisfied and certain if all three criteria were satisfied. A total of 2,981 cases were supplied by the UK PBC Consortium. 8,970 control samples were ascertained from the 1958 British Birth Cohort and the National Blood Service. This study contains 1,838 cases and 2,356 controls that were also included in a recent PBC GWAS (Mells *et al.,* 2011).

DNA was extracted from blood or saliva. Blood samples from PBC patients were extracted by the East Anglian Medical Genetics Service, while saliva samples were collected using an Oragene kit and DNA extracted at Source BioScience Healthcare. DNA samples were plated, normalized and shipped to the Wellcome Trust Sanger Institute for sample quality control.

Samples were genotyped on an Illumina iSelect HD custom genotyping array (Immunochip). All 2,981 cases and 4,537 controls were genotyped at the Wellcome Trust Sanger Institute. A further 4,433 control samples were genotyped at the Center for Public Health Genomics at the University of Virginia. Genotyping of control samples was coordinated by the Immunochip consortium for use in several Immunochip projects. The NCBI build 36 (hg18) map was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm). Normalized probe intensities were extracted for all samples passing standard laboratory QC thresholds and genotypes were called using optiCall (Shah *et al.,* 2012). Genotypes with an individual posterior probability lower than 0.7 were defined as unknown. optiCall was chosen because we found it to be more accurate in calling common and low-frequency variants on Immunochip compared to other established algorithms such as Illuminus (Teo *et al.,* 2007) and GenoSNP (Giannoulatou *et al.,* 2008; Shah *et al.,* 2012)

2.2.2 Quality control

Sample quality control (QC) was performed for each sample set separately. All monomorphic SNPs were removed prior to QC. Samples with a call rate lower than 98% and heterozygosity more than three standard deviations from the mean were excluded. A set of LD-pruned SNPs with minor allele frequency (MAF) > 20% were used to estimate identity by descent (IBD) and ancestry. For each pair of individuals with an estimated IBD > 18.75%, the sample with the lower call rate was removed. Principal component analysis was used to exclude samples of non-European ancestry (Price *et al.,* 2006) (Figure 2.1).

Following sample QC 2,861 cases and 8,514 controls remained (Table 2.1). SNPs with a minor allele frequency less than 0.1%, Hardy-Weinberg equilibrium $P < 10^{-6}$ in controls, call rate lower than 98%, or significantly different ($P < 10^{-5}$) call rate in cases vs. controls (or between the two control sets) were excluded. After marker QC 143,020 polymorphic SNPs were available for analysis (Table 2.2).

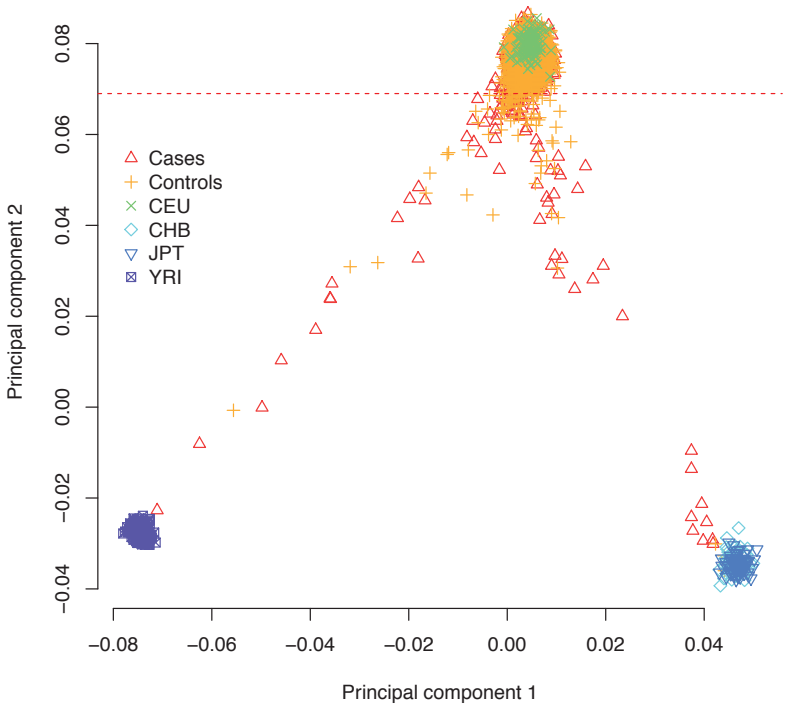


Figure 2.1. Principal component analysis of PBC cases and controls. The first two principal components were calculated for 18,995 SNPs that had MAF > 20% on the Immunochip and overlap those for the CEU, CHB, JPT and YRI HapMap samples. The red horizontal line indicates the exclusion threshold on the second principal component.

Sample	Heterozygosity/ missingness	Relatedness	Ancestry	Total[a]
Cases	29	47	65	140
Controls 1	70	187	32	224
Controls 2	37	169	53	232
Total	136	403	150	596

Table 2.1. Sample quality control. [a]Some samples failed more than one QC metric

Sample	HWE[a]	Call rate	MAF[b]	NRM[c]	Total Remaining[f]
Cases	-	8,301	39,504	9,362[d]	143,020
Controls 1	1,721	6,871	39,954	4,605[e]	
Controls 2	1,771	7,372	40,048		

Table 2.2. SNP quality control. [a]Hardy-Weinberg equilibrium. [b]Minor allele frequency. [c]Non-random missingnes (dbetween cases and controls, [e]between both sets of controls). [f]Some SNPs failed more than one QC metric.

The Immunochip contains 2,258 SNPs that were included as a replication panel for non-immune-mediated disorders. These SNPs were used as null markers to estimate the overall inflation of the distribution of association test statistics (Devlin and Roeder, 1999).

2.2.3 Imputation

Additional genotypes were imputed using 90,977 SNPs from the 186 Immunochip high density regions with the 1000 Genomes Phase I (interim) June 2011 release reference panel and IMPUTE2 (Howie *et al.,* 2009). Imputation was performed separately in three batches of 3,792, 3,792 and 3,791 individuals, with the case/control ratio constant across batches. SNPs with a posterior probability less than 0.9, IMPUTE INFO score < 0.5 and those with differential missingness ($P < 10^{-5}$) between the three batches were removed, as were those SNPs that failed the same exclusion thresholds used for the original Immunochip QC. After imputation, a total of 237,619 SNPs were available for analysis.

2.2.4 Association analysis

Case-control association tests were implemented using a standard one-degree of freedom Cochran-Armitage test for trend in PLINK v1.07 (Purcell *et al.,* 2007). Secondary associations were identified using step-wise logistic regression analysis conditioning on the allelic dosage of the primary signal in each significant locus. The process was repeated, conditioning on all independent genome-wide significant SNPs, until all genome-wide significant signals were accounted for (Cordell and Clayton, 2002). Cluster plots for all SNPs $P < 5\times10^{-6}$ were manually checked using Evoker (Morris *et al.,* 2010), and poorly called SNPs were removed from further study.

2.2.5 HLA Imputation

Imputation of six classic HLA alleles (class I: HLA-A, HLA-B and HLA-C, class II: HLA-DQA1, HLA-DQB1 and HLA-DRB1) was performed using the prediction algorithm proposed by Leslie et al. and implemented in the program HLA*IMP (Dilthey *et al.,* 2011; Leslie *et al.,* 2008). The imputation reference panel includes

~2,500 individuals of European ancestry with both genotype and classical HLA-allele type data. Case-control association was performed on HLA allele posterior probabilities generated from HLA*IMP using logistic regression to account for genotype uncertainty following imputation. Stepwise conditional logistic regression was used to identify independent association signals among the 21 HLA-alleles that reached $P < 0.0001$.

2.2.6 Variance in disease risk explained

The variance in disease risk explained by the 26 independent genome-wide significant SNPs and four HLA-alleles was estimated using a disease liability threshold model (Falconer and Mackay, 1996; So *et al.,* 2011) assuming a disease prevalence of 40/100,000 and log-additive risk. A review of population-based epidemiological studies of PBC found prevalence rates varied from 1.9 to 40.2 per 100,000 depending on the surveyed population, time of survey and phenotype definitions (Boonstra *et al*., 2012). The choice of using 40/100,000 in variance explained calculations was based on three recent large population-based surveys in European populations, where prevalence was estimated to be 38.3-40.2/100,000 (Podda *et al*., 2013).

2.2.7 eQTL analysis

Expression quantitative trait loci (eQTLs) within genome-wide significant loci were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) and a study by Gaffney *et al.,* (2012). The eQTL Browser contains significant eQTLs that were identified in recent studies across multiple cell lines and populations, while Gaffney *et al.,* reanalysed gene expression data from 210 lymphoblastoid cell lines using a total of 13.6M SNPs from the 1000 Genomes Project. For more details, see Gaffney, *et al.* (2012) and references listed in http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/.

2.2.8 Enrichment of open chromatin regions

The Encyclopedia of DNA Elements (ENCODE) project annotated regions of open chromatin using the direct sequencing of DNase-I hypersensitive sites (DNase-seq: sixteen different cell lines) (Myers *et al.,* 2011; Song *et al.,* 2011). The approach involves isolating nucleosome-depleted regions of DNA and mapping reads from next-generation sequencing to determine their location. I estimated the amount of enrichment for open chromatin peaks among significant PBC risk loci across the ENCODE cell lines. SNPs were first grouped into independent loci – beginning with the most strongly associated SNP (the "lead SNP"), I assigned SNPs in moderate LD with the lead SNP ($r^2 > 0.1$) to the associated locus, while those in high LD ($r^2 > 0.8$) were also considered candidate causal SNPs. The process then proceeds to the next most significantly associated SNP (that had not already been assigned to a locus), and assigned to the next locus this SNP along with those in moderate and high LD to this new locus, and so on. After the addition of each new locus, I calculated E,

$$E = \frac{OC_{loci;}/N_{loci}}{OC_{ichip}/N_{ichip}}$$

where, for a given cell line, OC_{loci} and N_{loci} are the number of candidate causal SNPs ($r^2 > 0.8$ with the lead SNP(s)) that lie within open chromatin peaks across the selected loci and the total number of SNPs within the loci ($r^2 > 0.1$ with the lead SNP(s)), respectively. OC_{ichip} and N_{ichip} are the equivalent measures across all SNPs within Immunochip high density regions. I only included the high density regions to increase the likelihood that the causal variant was assayed, and excluded SNPs in the HLA and those with MAF < 0.05 to avoid possible biases due to LD structure. To compare *E* between cell lines, the number of candidate causal SNPs in open chromatin ($OC_{loci:allcells}$) and the total number SNPs in open chromatin ($OC_{ichip:allcells}$) were first calculated for the union of open chromatin peaks across all cell lines other than that being evaluated. I then tested the alternative hypothesis that, for a given cell line, the proportion $OC_{loci}/OC_{ichip} > OC_{loci:allcells}/OC_{ichip:allcells}$ using a one-sided binomial test.

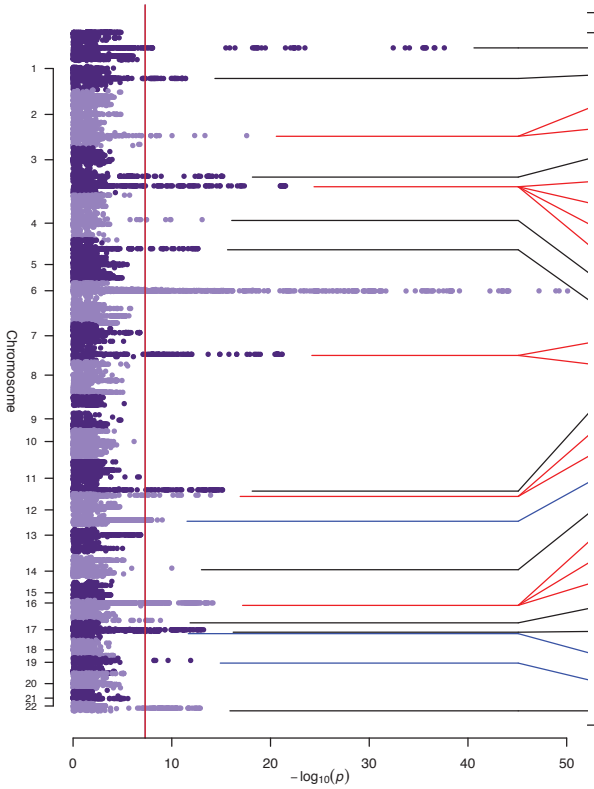
To ensure that the test was well calibrated under the null hypothesis I undertook 1000 permutations of PBC case control labels, repeating the association and enrichment analyses for each permutation. Comparing the observed level of enrichment at the top 21 loci to the equivalent from the permutations I obtained a similar, non-significant empirical P-value of 0.073 indicating that the proposed enrichment analysis is well calibrated under the null. A 95% confidence interval for *E* was estimated using the permutations.

2.3 Results and discussion

Following quality control, 143,020 polymorphic SNPs were available across 2,861 cases and 8,514 controls. (Table 2.1, Table 2.2, Figure 2.1). A further 94,559 SNPs in the Immunochip fine-mapping regions were imputed using genotypes from the 1000 Genomes June 2011 release. The inflation factor inferred from 2,258 SNPs not associated with autoimmune disease showed only a modest inflation (λ=1.096), similar to that reported in a previous GWAS study that included 4,194 overlapping samples (Mells *et al.,* 2011).

2.3.1 Replicating known PBC risk loci

Sixteen of the 22 known PBC risk loci reached genome-wide significance ($P < 5\times10^{-8}$) (Figure 2.2) and four showed nominal evidence of association ($5\times10^{-8} < P < 5\times10^{-4}$). Two PBC risk loci, 14q32 and 19q13, were not included on Immunochip as the array was designed before the publication of the most recent PBC GWAS (Mells *et al.,* 2011). At 12 of the genome-wide significant loci, the most associated SNP was different to that previously reported (Figure 2.3). There was little difference in the effect-size estimates between the GWAS tagging SNP and the most strongly associated Immunochip SNP, although this may partly be due to a large proportion of overlapping samples between the two studies. Nevertheless, the similarities in ORs despite the denser coverage of the Immunochip suggests that the ORs of tag-SNPs in GWAS adequately reflect the true ORs, and that synthetic associations are unlikely to explain the associations at these risk loci (Anderson *et al.,* 2011b).



SNP[a]	RA[b]	Chr:Position[c]	RAF[d]	OR[e]	P–value[f]	Gene(s)	Annotation[g]
rs72678531	G	1:67,571,033	0.17	1.61	2.47×10^{-38}	*IL12RB2*	OC
rs2488393	A	1:195,988,863	0.21	1.28	4.29×10^{-12}	*DENND1B,CRB1*	
rs3024921	A	2:191,651,517	0.06	1.62	2.59×10^{-18}	*STAT1,STAT4*	
rs7574865	A	2:191,672,878	0.22	1.31	1.38×10^{-13}	*STAT1,STAT4*	
rs2293370	G	3:120,702,624	0.80	1.39	6.84×10^{-16}	*TMEM39A,POGLUT1,TIMMDC1,CD80*	NS
rs2366643	A	3:161,219,179	0.57	1.35	3.92×10^{-22}	*IL12A*	OC
rs62270414	G	3:161,122,353	0.15	1.41	5.74×10^{-17}	*IL12A*	OC
rs668988	G	3:161,198,245	0.43	1.26	4.73×10^{-09}	*IL12A*	OC
rs80014155	A	3:161,108,087	0.004	3.44	2.64×10^{-11}	*IL12A*	
rs7665090	G	4:103,770,651	0.52	1.26	8.48×10^{-14}	*MANBA,NFKB1*	NS,eQTL
rs6871748	A	5:35,921,739	0.72	1.30	2.26×10^{-13}	*IL7R,CAPSL,SPEF2,UGT3A1*	NS
rs35188261	A	7:128,470,775	0.17	1.52	6.52×10^{-22}	*IRF5,TNPO3*	OC
rs3807307	G	7:128,366,438	0.47	1.22	4.12×10^{-09}	*IRF5,TNPO3*	OC,eQTL
rs80065107	A	11:118,187,977	0.79	1.39	7.20×10^{16}	*DDX6*	OC
rs1800693	G	12:6,310,270	0.40	1.27	1.18×10^{-14}	*TNFRSF1A,LTBR,SCNN1A*	OC
rs11064157	A	12:6,362,910	0.25	1.23	1.69×10^{-09}	*TNFRSF1A,LTBR,SCNN1A*	
rs11065979	A	12:110,543,940	0.44	1.20	2.87×10^{-09}	*ATXN2,BRAP,SH2B3*	NS
rs911263	A	14:67,823,346	0.71	1.26	9.95×10^{-11}	*RAD51B*	
rs1646019	G	16:11,267,181	0.71	1.31	6.72×10^{-15}	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
rs12708715	C	16:11,085,325	0.68	1.29	2.69×10^{-08}	*SOCS1,CLEC16A,PRM1,PRM2*	OC,eQTL
rs80073729	A	16:11,281,298	0.004	2.96	2.69×10^{-08}	*SOCS1,CLEC16A,PRM1,PRM2*	OC
rs11117433	G	16:84,577,017	0.77	1.26	1.41×10^{-09}	*IRF8*	OC
rs8067378	G	17:35,304,874	0.52	1.26	6.05×10^{-14}	*ORMDL3,ZPBP2,GSDMB,IKZF3*	NS,OC,eQTL
rs17564829	G	17:41,362,429	0.24	1.25	2.15×10^{-09}	*CRHR1,MAPT*	NS,OC,eQTL
rs34536443	G	19:10,324,118	0.95	1.91	1.23×10^{-12}	*TYK2*	NS
rs2267407	A	22:38,076,996	0.23	1.29	1.29×10^{-13}	*SYNGR1,PDGFB,RPL3*	OC,eQTL

Figure 2.2. Manhattan plot and table of genome-wide significant PBC risk loci. Novel loci are coloured in blue. Loci with multiple independent risk loci are coloured in red. [a]Most significantly associated SNP in locus. [b]Risk allele. [c]Base-pair position (NCBI36). [d]Risk allele frequency. [e]Odds ratio. [f]P-value for primary signals calculated from the Cochran-Armitage test for trend. Secondary signals calculated from stepwise logistic regression. [g]Whether SNPs in high linkage disequilibrium (r2 > 0.8) with the lead SNP overlap one of more of the following annotations: eQTL (expression quantitative trait loci), NS (non-synonymous SNP), OC (open chromatin).

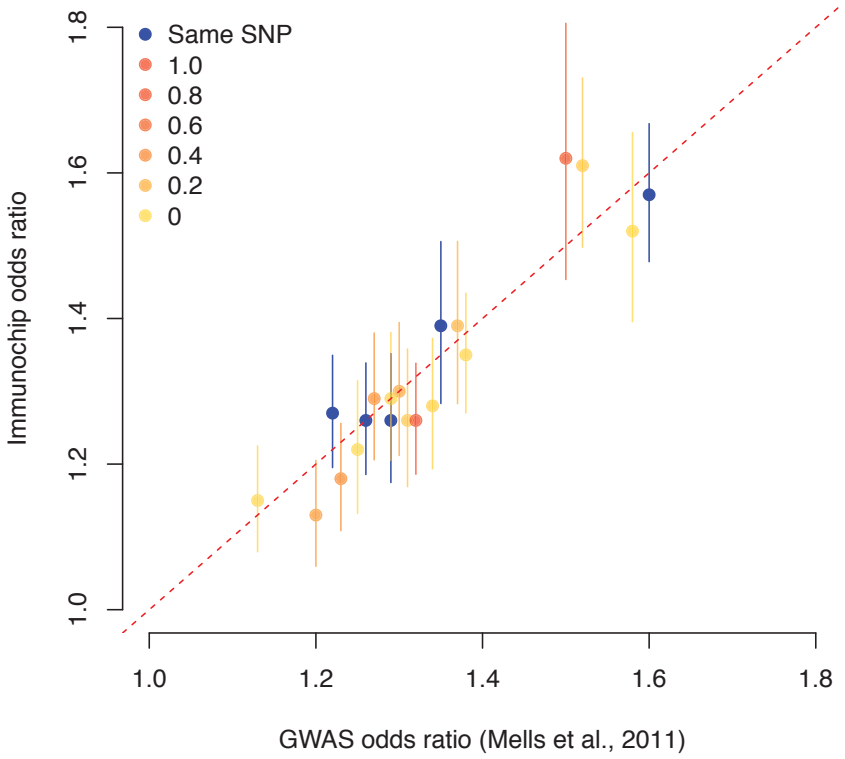


Figure 2.3. PBC risk loci odds ratios from this study vs. those from Mells *et al.* (2011). Colours denote the LD (r^2) between the SNP reported in Mells *et al.* and the most significant SNP in the same locus in this study. Error bars represent OR 95% confidence intervals. The red dashed line is $y = x$.

2.3.2 Multiple independent signals

Stepwise conditional regression (Cordell and Clayton, 2002) revealed multiple independent signals at five loci, with 16p13 harbouring three, and 3q25 four such associations (Figure 2.2, Figure 2.4). At the 16p13 locus, the third independent signal, rs80073729, is a rare SNP (MAF < 0.5%) recently associated with celiac disease (Trynka *et al.,* 2011). In the same study, Trynka et al. (2011) also identified multiple independent signals at 3q25, though rs80014155, a rare SNP that best tags the fourth independent PBC association at this locus, was not among them.

Further dissection of the four independent signals in 3q25 region revealed a complex genetic architecture. While stepwise conditional regression revealed

four independently associated SNPs (henceforth referred to as SNPs 1 through 4, as ordered according to P-value), jointly modelling the four SNPs in a multiple logistic regression model revealed large differences in the strength of association for SNPs 1 and 2 when compared with univariate regression (Figure 2.4, Table 2.3). For SNP 1, the strength of association fell from $P = 5.58\times10^{-22}$ to $P = 5.23\times10^{-7}$. Conversely, the strength of association for SNP 2 increased from $P = 6.75\times10^{-12}$ to $P = 2.93\times10^{-25}$. These changes are in part driven by the LD patterns within this region. SNP 1 is in moderate LD with SNPs 3 ($r^2 = 0.322$, D' = 0.75) and 4 (r2 = 0.004, D' = 0.91), such that the risk increasing alleles for all three SNPs reside more often on the same haplotype background. Thus the strong association signal for SNP 1 from a univariate association test is partly driven by its correlation with SNPs 3 and 4. Conversely, SNPs 2 and 3 are also in moderate LD ($r^2 = 0.10$, D' = 0.90), although in this case, the risk increasing allele of SNP 2 more often shares the same haplotype as the risk decreasing allele of SNP 3. Hence, the signal for SNP 2 is diluted by the risk decreasing effects of SNP 3 when performing a univariate test. Indeed, by accounting for the effects of independent SNPs in this region, it appears that SNP 2 is the most strongly associated signal ($P = 2.93 \times 10^{-25}$) while SNP 1 is no longer genome-wide significant ($P = 5.23 \times 10^{-7}$).

	SNP	RAF[a]	RA[b]	uncond OR[c]	uncond P[c]	cond OR[d]	cond P[d]
1	rs2366643	0.57	A	1.36	5.58×10^{-22}	1.22	5.23×10^{-7}
2	rs62270414	0.15	G	1.32	6.75×10^{-12}	1.59	2.93×10^{-25}
3	rs668998	0.44	G	1.26	1.98×10^{-14}	1.31	3.05×10^{-11}
4	rs80014155	0.004	A	3.07	6.66×10^{-10}	3.44	2.64×10^{-11}

Table 2.3. Unconditioned and conditioned association results for the four independent signals at 3q25. aRisk allele frequency. bRisk allele. cOdds ratios and P-values from univariate (unconditioned) association tests. dOdds ratios and P-values from multiple logistic regression model that includes all four SNPs as covariates.

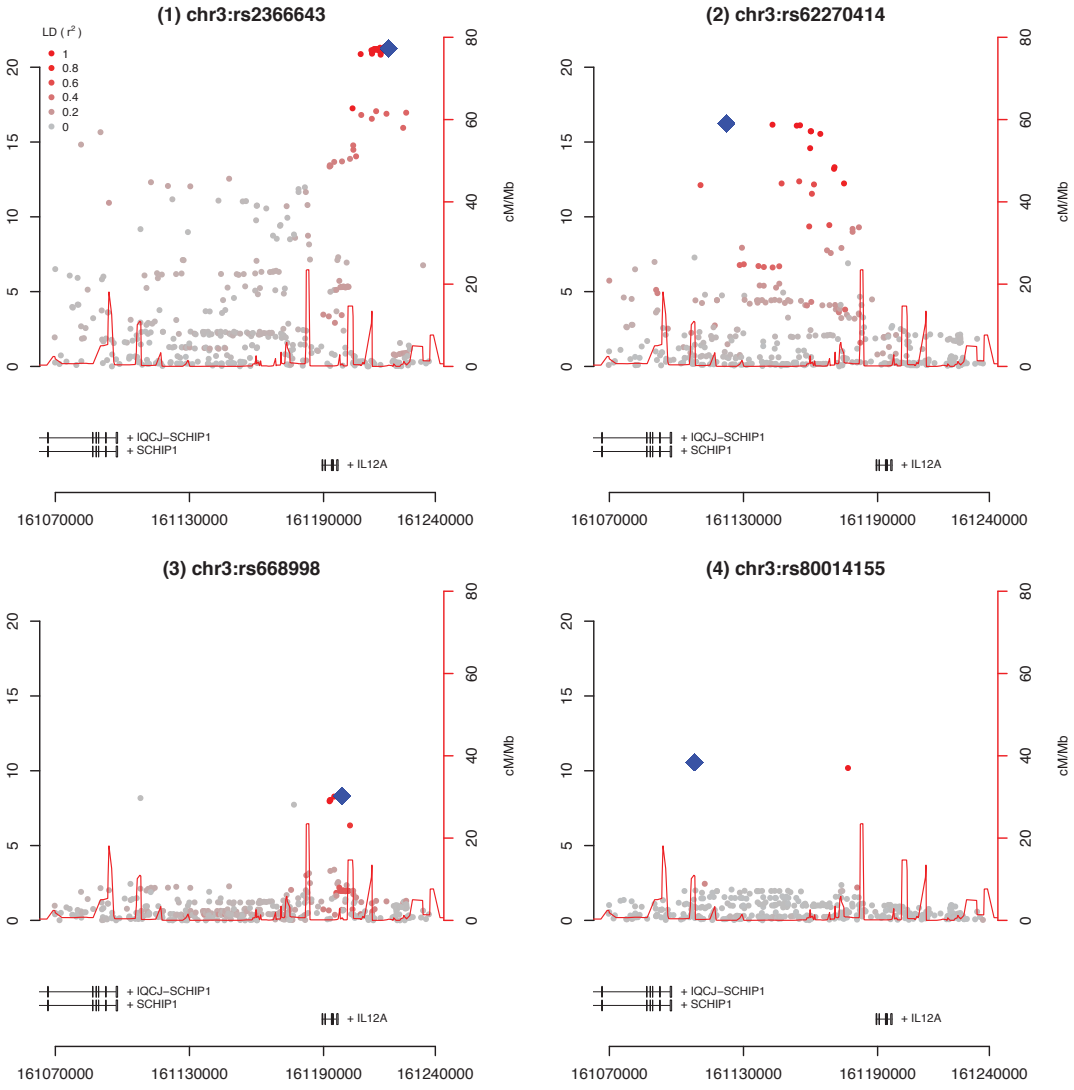


Figure 2.4. Multiple independent signals at 3q25 from stepwise conditional regression. Panel 1 shows the regional association plot of all SNPs with no conditioning. Panel 2 shows association results when conditioned on rs2366632. Panel 3 shows association results when conditioned on rs2366632 and rs62270414. Panel 4 shows association results when conditioned on rs2366632, rs62270414 and rs668998. The colour gradient (from red to grey) represents the strength of LD between the lead SNP (in blue) and others in the region.

The architecture of the 3q25 region demonstrates how the haplotype structure between associated SNPs can both dilute and inflate marginal association signals, such that stepwise regression may not completely reveal the true effect sizes of multiple associated SNPs within a region. In these situations, a two-stage procedure consisting of SNP-selection using conditional regression

and then performing joint multiple regression including the selected SNPs, may be more appropriate.

The identification of multiple independent signals show that resequencing efforts in large number of cases across known GWAS loci will be a powerful means of identifying additional independent signals (Hunt *et al.,* 2013). It is likely that the two rare SNP associations at 3q25 and 16p13 would have been overlooked using standard GWAS arrays due to poor tagging, unless they were directly genotyped. For example, in a case control study of 10,000 cases and 10,000 controls, there is only 0.07% power to detect association with the closest tagging SNP of rs80073739 on the Illumina Human1M chip, rs11649025 (minor allele frequency = 10%, r^2 = 0.04, D' = 1), at $P < 5\times10^{-8}$. These additional independent association signals thus yield a more complete understanding of the genetic architecture of PBC and enable more informative genotype-based recall and fine-mapping studies to be conducted.

2.3.3 Novel PBC risk loci

Three newly-associated PBC risk loci reached genome-wide significance (Figure 2.2). The strongest association on 19p12, rs34536443 (OR = 1.91, $P = 1.24\times10^{-12}$), is a low-frequency (MAF = 0.05) non-synonymous SNP in the tyrosine kinase 2 gene (*TYK2*), and is also associated with multiple sclerosis (Ban *et al.,* 2009). The locus has also been implicated in T1D (Wallace *et al.,* 2010), psoriasis (Strange *et al.,* 2010) and Crohn's disease (Franke *et al.,* 2010), although rs34536443 was not genotyped as part of these studies. For T1D and psoriasis, the strongest associations were to common SNPs that reside on the same haplotype (rs2304256: r^2 = 0.06, D' = 0.9 and rs280519: r^2 = 0.03, D' = 1). The most associated SNP in Crohn's disease and the second psoriasis signal (rs12720356) is independent of rs34536443 (r^2 = 0, D' = 0.003). The 12q24 locus has been associated with celiac disease (Hunt *et al.,* 2008; Trynka *et al.,* 2011a), rheumatoid arthritis (Stahl *et al.,* 2010) and T1D (Barrett *et al.,* 2009), though it was a non-synonymous SNP in *SH2B3*, rs3184504 (OR = 1.19, $P = 1.11\times10^{-8}$), rather than the most significant SNP in this study, rs11065979 (OR =

1.2, P = 2.87×10^{-9}), that was most strongly associated. The two SNPs are in high LD (r^2 = 0.81) and further studies are required to narrow the set of potential causal variants underlying the PBC association signal at this locus. The most associated SNP in the 17q21 region, rs17564829 (OR = 1.25, P = 2.15×10^{-9}), is located in *MAPT*, a gene that has been associated with cognitive symptoms in Parkinson's disease. While cognitive symptoms sometimes associated with PBC, it remains to be seen if the true causal variant at the locus has its functional effect through *MAPT*, and whether this functional effect then results in cognitive changes in PBC patients.

Both *TYK2* and *SH2B3* are involved in the production of cytokines, adding to the evidence that cytokine imbalances play a role in PBC and other autoimmune diseases (Rong *et al.,* 2009; Wang *et al.,* 2010a). *TYK2* is a member of the Janus kinase family, which transduce cytokine signals by phosphorylating STAT transcription factors. Couturier *et al.* (2011) showed that heterozygotes for rs34536443 have significantly reduced *TYK2* activity, which promotes the secretion of Th2 cytokines (Couturier *et al.,* 2011). For *SH2B3*, carriers of the A risk allele of rs3184504 show a moderate increase in production of cytokines and stronger activation of the NOD2 recognition pathway compared to carriers of the G allele (Zhernakova *et al.,* 2010), suggesting a possible role in helping prevent bacterial infection.

2.3.4 Associations with HLA haplotypes

Candidate genes studies have implicated several HLA-DR alleles in PBC susceptibility, particularly the DRB1*08 allele (Donaldson *et al.,* 2006; Invernizzi *et al.,* 2008; Mullarkey *et al.,* 2005; Wassmuth *et al.,* 2002). However, such studies were hindered by small sample sizes resulting in low power. As the Immunochip includes much denser SNP coverage of the MHC, it is expected that more HLA-types will be able to be imputed at greater accuracy than using traditional GWAS SNP chips. Here, the classical HLA alleles (HLA-A, B, C, DQA1, DQB1 and DRB1) were imputed from genotyped SNPs in the MHC (Dilthey *et al.,* 2011; Leslie *et al.,* 2008). Fourteen HLA-alleles reached genome-wide significance and conditional

analysis clustered these associations into four independent signals (Table 2.4). The most significant association was the HLA-DQA1*0401 allele (OR = 3.06, P = 5.9×10^{-45}), which forms a haplotype with two other HLA class II alleles (DQB1*0402 and DRB1*0801) and is an established PBC risk locus (Donaldson *et al.,* 2006; Invernizzi *et al.,* 2008; Mullarkey *et al.,* 2005; Wassmuth *et al.,* 2002). The second and third most significant clusters, DQB1*0602 (OR = 0.64, P = 2.32×10^{-15}) and DQB1*0301 (OR = 0.70, P = 6.48×10^{-14}) both have protective effects, confirming previous studies showing suggestive associations between these loci and PBC susceptibility (Donaldson *et al.,* 2006; Mullarkey *et al.,* 2005). The fourth most associated cluster, DRB1*0404 (OR = 1.57, P = 1.22×10^{-9}) has not been previously associated with PBC. The variance in disease liability explained by the 26 independent SNPs and four HLA-types are 4.9% and 1.4% respectively.

Haplotype	HLA type	Freq Cases	Freq Controls	OR	P-value
1	HLA*DQA1:0401	0.063	0.022	3.07	5.90×10^{-45}
	HLA*DQB1:0402	0.06	0.021	3.04	1.91×10^{-42}
	HLA*DRB1:0801	0.054	0.018	3.18	1.14×10^{-40}
	HLA*B:3905	0.01	0.003	5.48	4.81×10^{-12}
2	HLA*DQB1:0602	0.09	0.132	0.64	2.32×10^{-15}
	HLA*DRB1:1501	0.092	0.135	0.65	2.78×10^{-15}
	HLA*DQA1:0102	0.136	0.184	0.69	4.19×10^{-15}
	HLA*B:0702	0.109	0.144	0.73	4.93×10^{-10}
3	HLA*DQB1:0301	0.134	0.179	0.7	6.48×10^{-14}
	HLA*DRB1:1101	0.015	0.032	0.33	2.14×10^{-13}
	HLA*DQA1:0501	0.193	0.24	0.75	4.76×10^{-12}
	HLA*DRB1:1104	0.008	0.018	0.24	3.72×10^{-9}
4	HLA*DRB1:0404	0.072	0.052	1.57	1.22×10^{-9}
	HLA*DQB1:0302	0.133	0.104	1.34	6.96×10^{-9}

Table 2.4. Genome-wide significant HLA-type associations. Conditional analysis revealed four independent haplotypes.

2.3.5 Functional annotations and enrichment of open chromatin regions among risk loci

To identify candidate causal variants, I searched for non-synonymous variants in high LD ($r^2 > 0.8$) with the most associated variants at each PBC risk locus. I identified 39 such variants (of which 13 were directly genotyped) within seven risk loci (Figure 2.2), including two of the novel PBC associations identified in this study, *TYK2* and *SH2B3*. Functional follow-up studies are needed before

these non-synonymous variants can be confirmed as the causal disease variants at these loci.

As variation in gene expression is also likely to influence PBC risk, I evaluated the extent to which the most associated SNP at each locus tags expression quantitative trait loci (eQTLs) or regions of open chromatin. Known eQTLs were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) and Gaffney *et al*. (2012). Open chromatin regions in a range of cell lines were identified as part of the Encyclopedia of DNA Elements (ENCODE) project (Myers *et al.,* 2011; Song *et al.,* 2011) using DNase I hypersensitive sites sequencing (DNase-seq). Of the 26 independent non-HLA genome-wide significant SNPs identified in this study, 15 have an r^2>0.8 with SNPs that overlap DNase-seq peaks in a B-lymphoblastoid cell line (GM12878), and seven are also significant eQTLs in the same cell line (Figure 2.2).

To test if the enrichment of GM12878 open chromatin in regions was significantly greater than that for all other cell lines, associated SNPs were grouped into independent loci, and an enrichment score calculated for all loci that contained a genome-wide significant SNP (Section 2.2.8). Overall, GM12878 had the highest enrichment score compared with the other cell lines, though the difference in enrichment was non-significant (P = 0.068) (Figure 2.5).

The enrichment analysis protocol described here is predicated on the observation that the majority of complex disease risk loci do not lie within protein coding regions, and are likely to influence disease through their effects on gene expression, perhaps in a cell-specific manner. GWAS loci are indeed enriched for eQTLs (Nicolae *et al.,* 2010), though assigning causality to an associated variant based on eQTLs remains challenging due to LD and uncertainty over the precise regulatory mechanisms. Integrating functional genomic annotations may help bridge this gap between disease risk loci and eQTLs. Here, I used regions of open chromatin (as measured by DNase-I hypersensitivity) as it is a general indicator of potential regulatory activity (Bell *et al.,* 2011). These accessible regions make up 1-2% of the genome of a given

cell type, and are correlated with a range of other regulatory factors such as promoter and enhancer histone marks and transcription factor binding sites. Genetic variation in these regions have been shown to modify chromatin accessibility and transcription factor binding, which in turn lead to changes in gene expression (Degner *et al.,* 2012; Kasowski *et al.,* 2010). As such, variants within these regions that are in high LD with disease risk variants are good causal candidates, and enrichment in certain cell types may point to the relevant cells of interest in a particular disease.

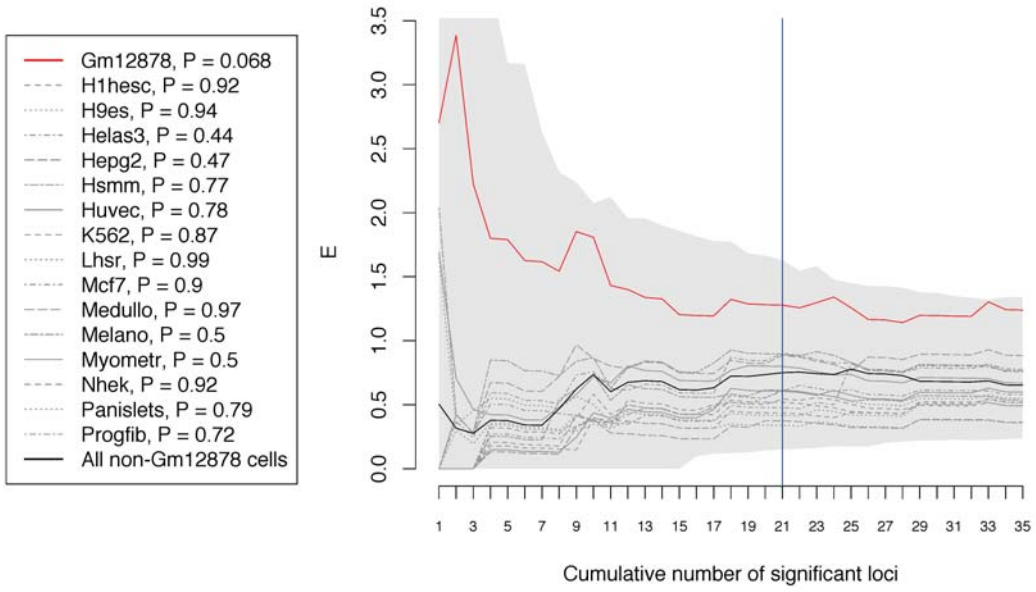


Figure 2.5. Enrichment of DNase-seq peaks among PBC risk loci in Gm12878 compared to other ENCODE cell lines. The relative enrichment (E) of SNPs within DNase-seq peaks was calculated across the 21 most associated loci. There is suggestive, though non-significant, evidence that genome-wide significant loci ($P < 5\times10^{-8}$ - vertical blue line) are more likely to lie within DNase-seq peaks in B-lymphoblastoid cell lines (solid red line) than they are to lie within the union of all other annotated cell lines (solid black line) (P = 0.068). Dotted grey lines denote E for other annotated cell lines. The shaded grey area represents the 95% confidence interval of E for Gm12878 from 1000 permutations. Cell types: Gm12878: B-lymphoblastoid, H1hesc: embryonic stem cells, H9es: embryonic stem cells, Helas3: cervical carcinoma, Hepg2: liver carcinoma, Hsmm: skeletal muscle myoblasts, Huvec: umbilical vein endothelial cells, K562: leukemia, Lhsr: prostate epithelial cells, Mcf7: mammary gland adenocarcinoma, Medullo: medulloblastoma, Melano: epidermal melanocytes, Myometr: Myometrial cells, Nhbe: bronchial epithelial cells, Nhek: epidermal keratinocytes, Panislets: pancreatic islets, Progfib: fibroblasts.

Fifteen of the 25 non-HLA PBC risk loci overlap regions of open chromatin in the GM12878 B-lymphoblastoid cell lines. While this number appears not to be significant when compared with the other ENCODE cell lines (P = 0.068), as a classical autoimmune disorder with a well-defined antibody presence (Jones, 2003), PBC risk is likely to be influenced by B cell activity. Moreover, it is important in these types of analyses not to bias results due to LD. Had I naively performed the enrichment analysis based on association P-value thresholds rather than pre-binning SNPs into independent loci, the evidence for enrichment with GM12878 open chromatin would have been much stronger (P = 0.0012) (Figure 2.6). This P-value enrichment approach, while seen in other studies (Maurano *et al.,* 2012), has the potential to bias results when SNPs that are moderately correlated with each other are counted multiple times if they overlap with functional annotations. In contrast, the approach presented here only considers a single potential causal variant per locus (i.e. SNPs that are in LD with the most strongly associated SNP is removed from further consideration), with all other variants in moderate LD are excluded from further analysis.

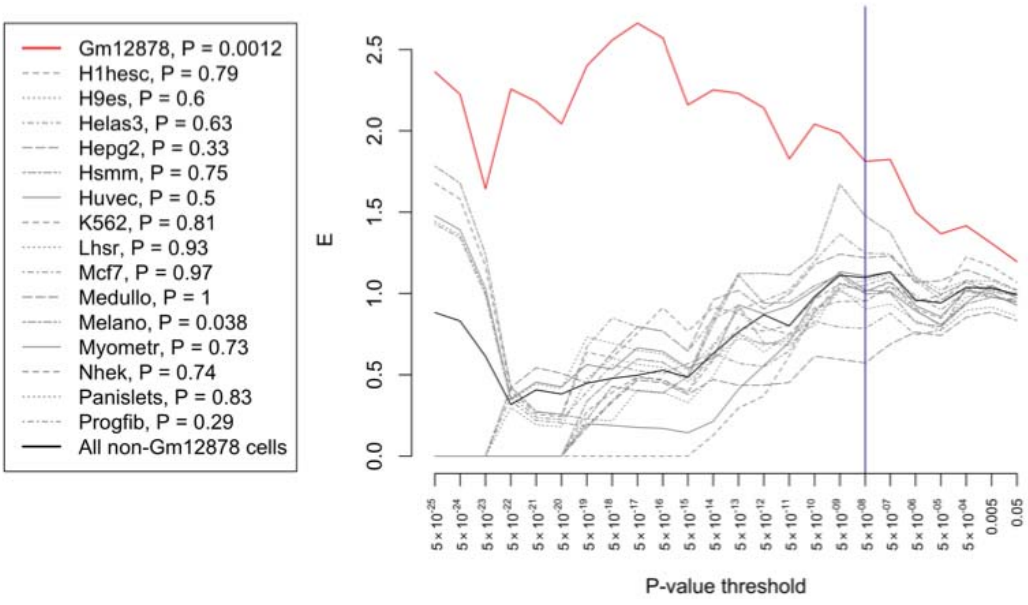


Figure 2.6. Enrichment of DNase-seq peaks among PBC risk loci calculated from P-value bins.

The thresholds used to define causal candidates at associated loci ($r^2 > 0.8$ with most associated SNP) as well as SNPs in LD to exclude ($r^2 > 0.1$) are somewhat subjective choices. The $r^2 > 0.8$ cut-off is based on a ubiquitous definition for which a SNP is "tagged" (Wang *et al.*, 2005), and is used throughout this chapter and the remainder of this thesis when defining causal candidate SNPs. This threshold has previously been shown to be effective at trading off power and the number of SNPs that need to be genotyped (de Bakker *et al.*, 2005), assessing the coverage of genotyping arrays (Barrett and Cardon, 2006) and resolving haplotypes (Carlson *et al.*, 2004). The threshold of $r^2 > 0.1$ to remove SNPs in LD with the most associated SNP in a locus (the lead SNP) was used to ensure that genome-wide significant SNPs whose signals are driven by their moderate LD with a much more strongly associated SNP are not considered causal candidates, while also allowing for additional truly independent variants to be counted. This approach yielded 21 genome-wide significant loci (Figure 2.5). Of the 26 independently associated SNPs identified in this study (Figure 2.2), 22 reside in the Immunochip high density regions considered in this enrichment analysis. The one ostensibly independently associated SNP that was excluded (rs668988) had $r^2 = 0.32$ with another lead SNP, rs2366643. Raising the r^2 threshold of 0.1 would have meant more SNPs denoted as "independently loci" even if their signals were entirely driven by a more strongly associated SNP nearby. For instance, at the rs72678531 ($P = 2.47\times10^{-36}$) locus, a second genome-wide significant SNP (rs17129749; $P = 3.50\times10^{-8}$) would have been declared independent had a minimum r^2 threshold of 0.3 been used. On the other hand, a lower threshold may exclude truly independently associated SNPs. Overall, any choice of LD thresholds involves trade-offs between excluding SNPs in LD and capturing truly independent association signals.

Finally, it should be noted that enrichment in a certain cell type does not automatically implicate that cell in disease. A certain amount of enrichment may be expected given that many gene promoters are active across multiple cell types (e.g. housekeeping genes). Between two given cell types, 30-40% of open chromatin regions may be shared (Song *et al.,* 2011). Moreover, lack of enrichment cannot rule out that cell's involvement in disease. This study was

also limited by the availability of cell types where the same functional genomic annotations were obtained in a consistent manner. It is likely that similar studies in autoimmune disorders will incorporate annotations from a range of immune cells (e.g. various types T cells, monocytes, dendritic cells, macrophages). For instance, recent approaches examining gene expression in murine immune cells found significant enrichment for B cell expressed genes among systemic lupus erythematosus risk loci, CD4 T cell genes among rheumatoid arthritis loci, and dentritic cell genes among Crohn's disease loci (Hu *et al.,* 2011; Jostins *et al.,* 2012). Moreover, the power to detect enrichment will only increase as the list of associated risk loci ever expands.

2.4 Conclusion

Through genotyping of 2,861 PBC cases and 8,514 controls on the Immunochip genotyping array, three novel PBC risk loci were identified, including a low-frequency non-synonymous SNP in *TYK2*, further implicating the JAK-STAT and cytokine signalling in disease pathogenesis. Together, these newly discovered risk loci in conjunction with 16 previously known loci offer further leads into the biological pathways that underlie PBC risk.

Within the 186 high density regions, the Immunochip includes ~90,000 directly genotyped SNPs compared with ~10,000 SNPs on the Illumina Human-660W Quad array used in Mells *et al.*, (2011). This denser coverage suggests that common causal variants are more likely to be genotyped directly or offer better tagging than SNPs from GWAS arrays. Reassuringly, odds ratios at known loci did not significantly differ to those from Mells *et al.*, (2011) despite the lead SNP changing at all but five of these loci. This suggests that the loci discovered in this study were primarily driven by sample size rather than SNP density, and that further GWAS of ever larger sample sizes will continue to discovery new risk loci.

The dense coverage also allows for greater refinement of the genetic architecture at risk loci. Multiple independent association signals were identified at five loci, including low-frequency and rare variants that are poorly tagged on GWAS arrays. At the 3q25 locus, four independent signals were identified,

though the effect sizes at two of the SNPs varied significantly when assessed under a joint model than when considering SNPs one at a time, highlighting that the haplotype structure of these regions with multiple signals should be considered when reporting association results.

Finally, I also explored the potential of integrating association results with large-scale functional genomic annotations to identify the cell types in which PBC associated variants are likely to be influencing disease. Future association studies in larger sample sizes in combination with disease-relevant functional genomic datasets will greatly improve the understanding of PBC and other complex disorders.

Chapter 3. Discovery of primary sclerosing cholangitis risk loci and the genetic relationship with inflammatory bowel disease

3.1 Introduction

Primary sclerosing cholangitis (PSC) is a severe liver disease of unknown etiology that results in the fibrotic destruction of the bile ducts (Aadland *et al.,* 1987; Broome *et al.,* 1996; Farrant *et al.,* 1991). The pathogenesis of PSC is poorly understood, and due to the lack of effective medical therapy, PSC remains a leading indicator for liver transplantation in Northern Europe and the US (Karlsen *et al.,* 2010b), despite the relatively low prevalence (~10/100,000). Affected individuals are diagnosed at a median age of 30-40 years and suffer from an increased frequency of inflammatory bowel disease (IBD) (60-80%) (Karlsen and Kaser, 2011; Karlsen *et al.,* 2010b) and autoimmune diseases (25%) (Saarinen *et al.,* 2000). Conversely, approximately only 5% of patients with IBD develop PSC (Karlsen and Kaser, 2011; Karlsen *et al.,* 2010b). A 9-39-fold sibling recurrence risk indicates a strong genetic component to PSC risk (Bergquist *et al.,* 2008). In addition to multiple strong associations within the human leukocyte antigen (HLA) complex, recent association studies have identified genome-wide significant loci at 1p36 (*MMEL1*/*TNFRSF14*), 2q13 (*BCL2L11*), 2q37 (*GPR35*), 3p21 (*MST1*), 10p15 (*IL2RA*) and 18q21 (*TCF4*) (Ellinghaus *et al.,* 2012; Folseraas *et al.,* 2012; Karlsen *et al.,* 2010a; Melum *et al.,* 2011; Srivastava *et al.,* 2012).

In order to identify additional risk loci associated with PSC risk, 3,789 PSC cases from Europe and North America, along with 25,079 population matched controls, were genotyped on the Immunochip. The IBD status were also available for 3,283 of the PSC cases, and, along with results from a recent GWAS of IBD (Jostins *et al.,* 2012), allowed for powerful cross-phenotype genetic comparisons.

3.1.1 Chapter overview

In this chapter, I discuss the identification of twelve genome-wide significant PSC risk loci outside the HLA region, nine of which are implicated in PSC risk for the first time. Within the HLA region, HLA-allele imputation revealed five independent associations. Due to the high comorbidity with IBD (72% of cases have Crohn's disease (CD), ulcerative colitis (UC) or indeterminate IBD), investigating the shared and unique genetic basis between the two disorders has implications in understanding shared biology and disease classification. I investigated this sharing at PSC risk loci, and considered in aggregate IBD risk and variants genome-wide, showing the presence of both overlapping and distinct genetic architectures for PSC and IBD.

3.1.2 Contributions

The study design was conceived by the International PSC Genetics Study Group (IPSCSG). Cases and controls were ascertained through the IPSCSG and the International IBD Genetics Consortium (IIBDGC). Genotyping was performed at various centres described in section 3.2.1 and the Supplementary Note of Liu et al. (2013). GRAIL analysis was performed by Trine Folseraas. Quality control on unpublished GWAS data was performed by Sun-Gou Ji. All other analyses were performed by myself.

3.2 Methods

3.2.1 Samples, DNA extraction and genotyping

Recruitment of PSC cases was performed in 14 countries in Europe and North America (Table 3.1). Diagnosis of PSC was based on standard clinical,

biochemical, cholangiographic and histological criteria with exclusion of secondary causes of sclerosing cholangitis (Chapman *et al.,* 1980). Controls were recruited from blood donors, population-based studies as part of this study, or via the International Immunochip Consortium. See Supplementary Note of Liu *et al.* (2013) for details.

	Controls	PSC cases	Total
Scandinavia	4,324	917	5,241
North Central Europe	9,438	1,136	10,574
Southern Europe	580	115	695
UK	8,663	1,033	9,696
North America	2,074	588	2,662
Total	25,079	3,789	28,868

Table 3.1. Post-QC patient and control panels. PSC cases and controls in the study sorted by broad geographic panels (based on participating centre information, not genotypes). Scandinavia: Finland, Norway, Sweden; North Central Europe: Belgium, Germany, The Netherlands, Poland; Southern Europe: France, Greece, Italy, Spain; UK: United Kingdom; North America: Canada, USA.

DNA was extracted from whole blood, transformed lymphocytes or liver tissue using commercially available kits or an in-house out-salting method. DNA samples were genotyped using the Immunochip according to Illumina protocols. The NCBI build 36 (hg18) reference was used and normalised probe intensities were extracted for all samples passing standard laboratory quality control thresholds. All genotypes were called specifically for this study using optiCall (Shah *et al.,* 2012), but separately across each genotyping batch. Genotypes with a posterior probability lower than 0.7 were defined as unknown. All PSC cases were genotyped at the Institute of Clinical Molecular Biology in Kiel, Germany, or at the department of Genetics, University of Groningen and University Medical Centre Groningen, The Netherlands.

3.2.2 Quality control

SNPs with a call rate < 80% were removed prior to sample QC (n = 235). Per individual genotype call rate and heterozygosity rate were calculated using PLINK (Purcell *et al.,* 2007) and outlying samples were identified using Aberrant (Figure 3.1) (Bellenguez *et al.,* 2012), which identifies outliers from otherwise Gaussian distributions. A set of 20,837 LD-pruned (r^2 < 0.1) SNPs with minor

allele frequency > 10% present in both the Immunochip and the Illumina Omni2.5-8 array used in the 1000 Genomes Project (Genomes Project *et al.,* 2012) were used to estimate identity by descent and ancestry. For each pair of individuals with estimated identity by descent ≥ 0.9, the sample with the lower call rate was removed (unless case/control status was discordant between the pair, in which case both samples were removed, n = 92). Related individuals (0.1875 < identity by descent < 0.9) remained in the analysis to maximize power because the mixed model association analysis can correctly account for the relatedness between individuals. Principal components analysis was performed using SMARTPCA (Patterson *et al.,* 2006). Principal components were defined using population samples from the 1000 Genomes Project genotyped using the Illumina Omni2.5-8 genotyping array and then projected into PSC cases and controls, with non-European outliers identified using Aberrant and removed (Figure 3.2). Following sample QC, 3,789 PSC cases and 25,079 remained. SNPs with a minor allele frequency less than 0.1%, Hardy-Weinberg equilibrium $P < 10^{-5}$ in controls, call rate lower than 98%, or significant differential missing data rate between cases and controls ($P < 10^{-5}$) were excluded. After completion of marker QC, 131,220 SNPs were available for analysis – further reduced to 130,422 after cluster plot inspection of nominally associated SNPs. The genomic inflation factor (Devlin *et al.,* 1997) was calculated using 2,544 "null" SNPs. These SNPs were included on the Immunochip as part of replication panels for bipolar disease and other non-immune-related studies.

3.2.3 Imputation

Using 85,747 post-QC SNPs located in the Immunochip high density regions, additional genotypes were imputed using IMPUTE2 with the 1000 Genomes Phase 1 (March, 2012) reference panel of 1,092 individuals (Genomes Project *et al.,* 2012) and 744,740 SNPs. Imputation was performed separately across ten batches, with the case:control and country of origin ratios constant across batches. SNPs with a posterior probability less than 0.9 and those with differential missingness ($P < 10^{-5}$) between the 10 batches were removed, as

were SNPs failing the exclusion thresholds used for genotyped SNP QC. After imputation, a total of 208,852 SNPs were available for analysis.

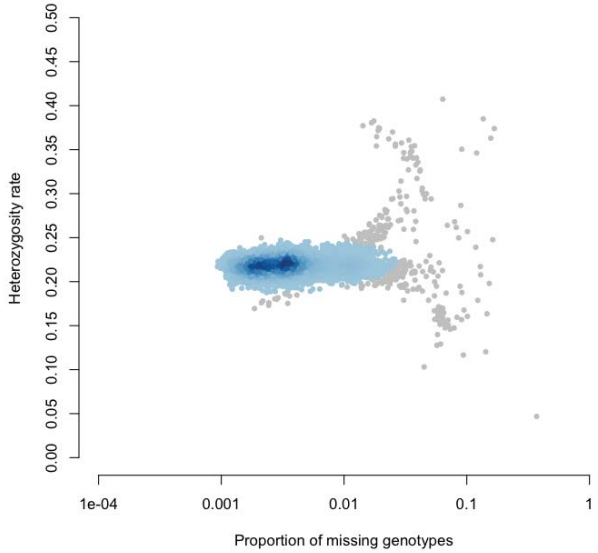


Figure 3.1. Heterozygosity rate and proportion of missing genotypes for PSC cases and controls. The grey points represent outlying individuals. Heterozygosity proportions and missingness were calculated using PLINK (Purcell et al., 2007). Outliers were detected using Aberrant (Bellenguez et al., 2012).

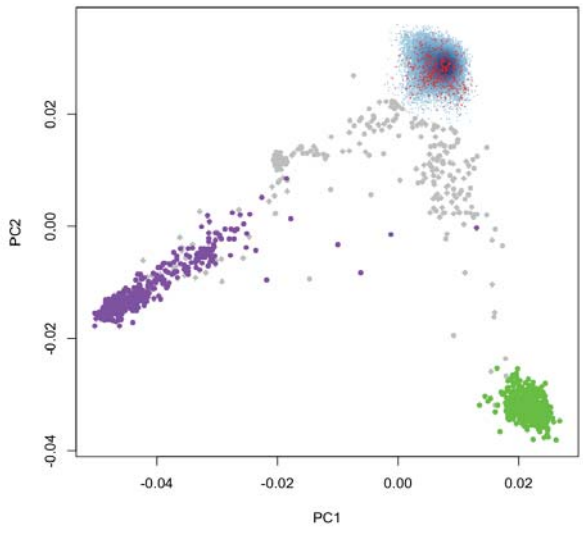


Figure 3.2. Principal components analysis of PSC cases and controls with 1000 Genomes Omni2.5-8 data. The red, purple and green points represent 1000 Genomes CEU (Utah residents with Northern and Western European ancestry), YRI (African) and CHB+JPT (Han Chinese and Japanese) populations respectively. The blue points represent the included PSC cases

and controls, overlapping the CEU population, with the grey points showing those who were identified as ancestry outliers (and therefore excluded). The principal components were generated using 20,837 common (MAF>0.10) SNPs overlapping between the Immunochip (this study) and the Omni2.5-8 array.

3.2.4 Association analysis

Case-control association tests were performed using a linear mixed model as implemented in MMM (Pirinen *et al.,* 2012). A covariance matrix, *R*, of a random effects component was included in the model to explicitly account for confounding due to population stratification and cryptic relatedness between individuals. This method has been shown to better control for population stratification than correction for principal components or meta-analyses of matched subgroups of cases and controls (Korte *et al.,* 2012; Sawcer *et al.,* 2011). *R* is a symmetric n×n matrix with each entry representing the relative sharing of alleles between two individuals compared to the average in the sample, and is typically estimated using genome-wide SNP data. To avoid biases in the estimation of *R* due to the design of the Immunochip, SNPs were first pruned for LD ($r^2 < 0.1$). Of the remaining SNPs, I then removed those that lie in the HLA region or have a minor allele frequency < 10%. Finally, I excluded SNPs that showed modest association ($P < 0.005$) with PSC in a linear regression model fitting the first 10 principal components as covariates. A total of 17,260 SNPs were used to estimate *R*. The following parameters were used in MMM: logOR = 2 (more accurate when genotypes are coded 0,1,2 and no predictors other than genotyes), mean_center = 1 (genotypes are mean-centred), impute_missing = 1 (missing genotypes are set to mean of non-missing genotypes), min_d = 0.1 (lower bound for accepted eigenvalues of *R*).

Due to computational limitations, I estimated the *R* matrix and performed all association analyses separately for UK (n = 9,696) and non-UK (n = 19,172) samples, and then combined the results using a fixed-effects (inverse-variance weighting) meta-analysis. This reduced the λ_{GC} (estimated using the 2,544 "null" SNPS and using the first 10 PCs as covariates) from 1.13 to 1.02 (Figure 3.3), suggesting population stratification was well-controlled for. Stepwise conditional regression was used to identify possible independent associations at

genome-wide significant loci. SNP×SNP interactions between all pairs of genome-wide significant SNPs were tested using the PLINK --epistasis command. Signal intensity plots of all non-HLA loci with association $P < 5\times10^{-6}$ were visually inspected using Evoker (Morris *et al.,* 2010). SNPs that clustered poorly were removed (N = 800).

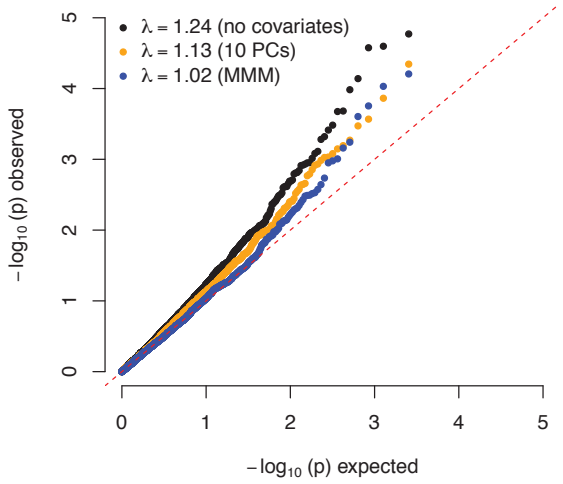


Figure 3.3. Quantile-quantile plots and genomic inflation factors of observed vs. expected P-values. Association tests were compared for logistic regression with no covariates, logistic regression with the first 10 principal components as covariates, and a linear mixed model implemented in MMM (Pirinen et al., 2012). Tests were performed on 2,544 "null" SNPs with no evidence for association with immune-related phenotypes. The dashed red line is y = x.

3.2.5 Functional annotation of risk loci

Gene regulatory elements from the Encyclopedia of DNA elements (ENCODE) and coding SNPs were annotated using HaploReg (Ward and Kellis, 2012). For each risk locus, SNPs in high linkage disequilibrium ($r^2 > 0.8$) with the most significantly associated SNP were assessed as to whether they lie within regions with promoter and enhancer marks, DNAse-I hypersensitivity, protein binding or regulatory motifs in one or more of 147 cell types. Expression quantitative trait loci (eQTLs) were collated from the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/).

3.2.6 GRAIL and DAPPLE analyses

To assess the functional relationship among established genomic PSC risk regions, we performed a GRAIL pathway analysis. GRAIL is a statistical tool that uses text mining of published abstracts in the PubMed database to identify and quantify functional similarity among genes within disease-associated regions (Raychaudhuri *et al.,* 2009). The output GRAIL score is a significance score, Ptext, which is adjusted for multiple hypothesis testing. Sixteen PSC risk loci (7 known and 9 novel) were used as input for this analysis.

Similarly, DAPPLE assesses functional similarity through constructing networks of protein-protein interactions (Rossin *et al.,* 2011). Gene connectivity is assessed based on the number of direct and indirect (via other proteins) connections and a permuted P-value is calculated. The 16 PSC risk loci were used as input into DAPPLE. Genes with $P < 0.05$ were listed as causal candidates.

3.2.7 HLA imputation and association analysis

Imputation of HLA class I and II genes was performed using HLA*IMPv2 (Dilthey *et al.,* 2011; Leslie *et al.,* 2008). The imputation reference panel includes ~2,500 individuals of European ancestry with both genotype and classical HLA-allele type data. Cluster plots for all SNPs contributing to the imputation of HLA types were manually inspected and poorly clustered SNPs were removed. Case-control association was performed on HLA allele posterior probabilities using the mixed model framework described previously. Stepwise conditional regression was used to determine independent HLA association signals.

3.2.8 Heritability explained

The proportion of variance explained by the genome-wide significant SNPs and HLA alleles was calculated using a disease liability threshold model (Falconer and Mackay, 1996; So *et al.,* 2011) assuming a disease prevalence of 10/100,000 and multiplicative disease risk.

3.2.9 Prediction of PSC using IBD risk loci

Odds ratios (ORs) for Crohn's disease and ulcerative colitis in 163 IBD-associated SNPs were obtained from Jostins *et al.* (2012). I used the R package Mangrove (http://cran.r-project.org/web/packages/Mangrove) to generate risk scores and estimate each individual's probability of developing PSC among the 3,789 PSC cases and 25,079 controls assuming additive risk (log-additive OR). The performance of the predictor using either Crohn's disease or ulcerative colitis ORs was assessed by constructing a receiver operating characteristic (ROC) curve, which shows the proportion of true and false positives at each probability threshold. The area under the curve (AUC) was calculated to compare the predictive power of the ulcerative colitis and Crohn's disease ORs.

The DeLong method was used to test if the AUC using ulcerative colitis ORs was significantly different to the AUC using Crohn's disease ORs (DeLong *et al.,* 1988). The method is a non-parametric approach for test the alternative hypothesis that two (or more) AUCs estimated from different sets of predictors in the same samples are significantly different. As the AUC is equivalent to the Mann Whitney U statistic for comparing the distribution values from two samples, variances of correlated U statistics can be estimated using the approach of Sen (1960). The method is equivalent to a jackknife resampling approach for estimating the variance of the AUC (DeLong *et al.*, 1988).

3.2.10 Genetic correlation between PSC and IBD

Genome-wide SNP data were available for 5,322 Crohn's disease cases, 6,307 ulcerative colitis cases and 12,164 population matched controls genotyped as part of previous GWAS meta-analyses (Anderson *et al.,* 2011a; Franke *et al.,* 2010; Jostins *et al.,* 2012). Genome-wide data from an ongoing GWAS for PSC (2,871 cases and 12,019 controls) were also obtained (Sun-Gou Ji, personal communication). All datasets were obtained post-QC. The IBD dataset was imputed using the HapMap phase 2+3 reference panel, while the PSC dataset was imputed using a combined 1000 Genomes Phase I plus UK10K reference panel. Additional QC included removing 35 cases 3803 controls from the IBD dataset

that were duplicated or related with individuals in the PSC dataset using PLINK (pi_hat > 0.1). SNPs with a missingness rate of greater than 2% in the combined data were also removed. In total, 721,733 autosomal SNPs that overlap the two datasets remained. The top 20 Principal components estimated from the 1000 Genomes Phase I individuals were projected onto all IBD and PSC cases and controls.

The proportion of genetic variation (as tagged by common genome-wide SNPs) that is shared between PSC and IBD was estimated using the bivariate linear mixed-effects model implemented in GCTA (Lee *et al.,* 2012). The method uses genome-wide SNPs to estimate genetic similarities between pairs of individuals, and uses bivariate restricted maximum likelihood to estimate covariance components (r_G) of the linear mixed model. In all, each of four PSC subphenotypes (all PSC cases, PSC cases with UC, PSC cases with CD and PSC cases with no IBD) were tested against CD and UC. To test whether r_G is significantly different from 0 (i.e. there is no genetic overlap between the two phenotypes), r_G was fixed at 0 and a likelihood ratio test comparing this constrained model and the unconstrained model was applied.

3.3 Results and discussion

Following quality control and imputation, 208,852 SNPs from 3,789 cases and 25,079 population controls were available for analysis, of which 80,183 SNPs located in the Immunochip high density regions were imputed using the 1000 Genomes reference panel. Case-control association testing was performed using a linear mixed model as implemented in MMM to minimise the effect of population stratification (λ_{GC} = 1.02, estimated using 2,544 "null" SNPs).

3.3.1 Locus discovery

Twelve non-HLA genome-wide significant ($P < 5\times10^{-8}$) PSC susceptibility loci were identified, nine of which were implicated in PSC for the first time (Table 3.2, Figure 3.4, Figure 3.5). The most associated SNP within each locus was a common variant (all risk allele frequencies >0.18) of moderate effect (ORs

between 1.15 and 1.4) (Table 3.2). Genotype imputation and stepwise conditional regressions within each locus did not identify additional independent genome-wide significant signals, nor did genotype-genotype or sex-genotype interaction analyses.

For seven of the nine novel loci, the most significantly associated SNP in the locus was the same SNP or was in strong linkage disequilibrium (LD; $r^2 > 0.8$) with the original association reports for another disease (Table 3.3). The two exceptions were 11q23, where only independent disease associations ($r^2 < 0.01$) have so far been reported for colorectal cancer (Peters *et al.,* 2012), and 6q15, where the most significantly associated PSC variant, rs56258221 (OR = 1.23, P = 8.36×10^{-12}), is in low-to-moderate LD with the previously reported *BACH2* variants in Crohn's disease ($r^2 = 0.23$) and type 1 diabetes ($r^2 = 0.12$).

Chr	SNP[a]	RA[b]	RAF cases[c]	RAF controls[c]	*P*-value	OR (95%CI)	LD region[d] (Kb)	RefSeq genes in LD region	Notable nearby gene(s)[e]	Functional annotation[f]
1p36	rs3748816	A	0.698	0.656	7.41×10^{-12}	1.21 (1.14-1.27)	2,398-2,775	9	*MMEL1, TNFRSF14*	eQTL,MS, OC, PB, HM
2q33	**rs7426056**	A	0.277	0.229	1.89×10^{-20}	1.3 (1.23-1.37)	204,155-204,397	1	*CD28*	HM, OC
3p21	rs3197999	A	0.352	0.285	2.45×10^{-26}	1.33 (1.26-1.4)	48,388-51,358	90	*MST1*	eQTL,MS, OC, PB HM
4q27	**rs13140464**	C	0.871	0.836	8.87×10^{-13}	1.3 (1.21-1.4)	123,204-123,784	4	*IL2, IL21*	OC, PB
6q15	**rs56258221**	G	0.213	0.183	8.36×10^{-12}	1.23 (1.16-1.31)	90,967-91,150	1	*BACH2*	OC, PB
10p15	rs4147359	A	0.401	0.349	8.19×10^{-17}	1.24 (1.18-1.3)	6,070-6,206	2	*IL2RA*	PB
11q23	**rs7937682**	G	0.298	0.265	3.17×10^{-09}	1.17 (1.11-1.24)	110,824-111,492	19	*SIK2*	OC, PB, HM
12q13	**rs11168249**	G	0.506	0.466	5.49×10^{-09}	1.15 (1.1-1.21)	46,442-46,534	3	*HDAC7*	OC, PB, HM
12q24	**rs3184504**	A	0.527	0.488	5.91×10^{-11}	1.18 (1.12-1.24)	110,186-111,512	16	*SH2B3, ATXN2*	MS, OC, HM
18q22	**rs1788097**	A	0.518	0.483	3.06×10^{-08}	1.15 (1.1-1.21)	65,633-65,721	2	*CD226*	MS, OC, PB, HM
19q13	**rs60652743**	A	0.864	0.836	6.51×10^{-10}	1.25 (1.16-1.34)	51,850-51,998	6	*PRKD2, STRN4*	OC, PB, HM
21q22	**rs2836883**	G	0.777	0.728	3.19×10^{-17}	1.28 (1.21-1.36)	39,374-39,404	-	*PSMG1*	OC, PB, HM

Table 3.2. Association results of twelve non-HLA genome-wide significant risk loci for PSC. [a]SNPs from novel PSC-associated loci are shown in bold. [b]Risk increasing allele. [c]Risk allele frequency. [d]LD regions around lead SNPs were calculated by extending in both directions a distance of 0.1 centimorgans as defined by the HapMap recombination map. [e]Select candidate gene(s) within same LD region as the associated SNPs. [f]Denotes if there are SNPs with r2>0.8 with the hit SNP that have functional annotations: eQTL: expression quantitative trait locus, HM: overlaps a region of histone modification MS: missense mutation; OC: open chromatin; PB: protein binding.

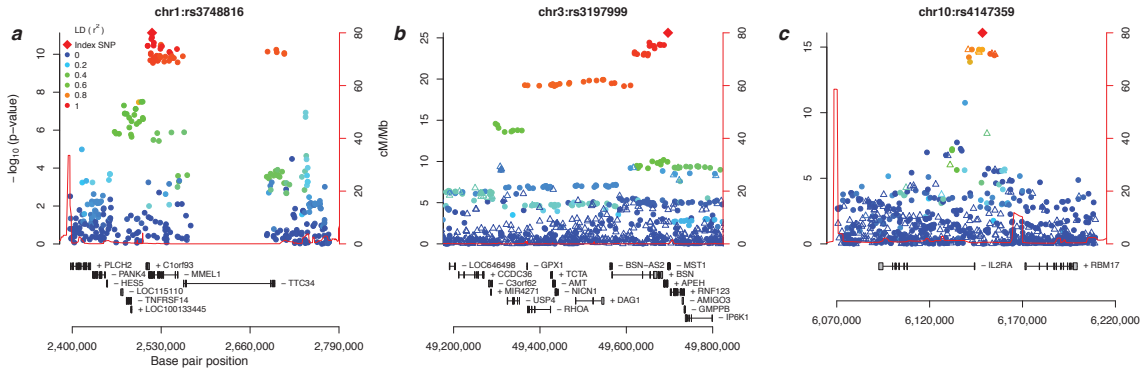


Figure 3.4. Regional association plots for genome-wide significant associations at previously established PSC risk loci. Filled-in circles are directly genotyped and hollow-triangles are imputed SNPs. The colour of the marker (see legend in panel a) illustrates the linkage disequilibrium between the most associated SNP and others in the locus.

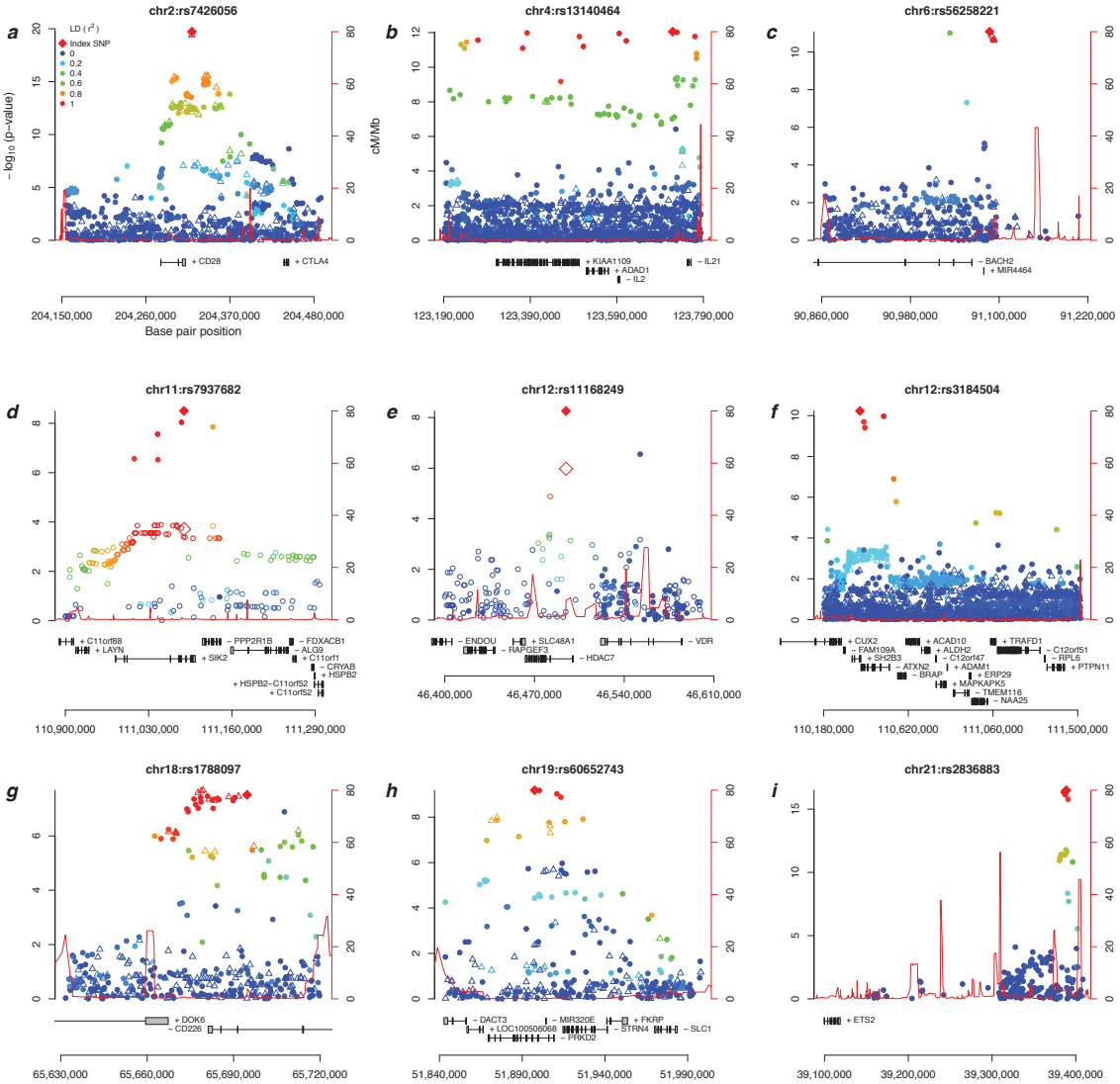


Figure 3.5. Regional association plots of nine newly associated PSC risk loci. In panels d and e, the most associated SNPs are located outside Immunochip fine-mapping regions. Association signals from the discovery panel of the largest PSC GWAS to date are shown as hollow circles and the most associated SNP as a hollow diamond (genotyped and imputed to HapMap release 22 SNPs, cases overlap with the current study).

Locus	SNP	Same signal	Different signal
1p36	rs3748816	CeD,RA,UC	
2q33	rs7426056	CeD	AA,CeD,CHD,GD,Ht,MI,RA,T1D
3p21	rs3197999	**CD**,**UC**	
4q27	rs13140464	CD,CeD,RA,UC	AA,T1D
6q15	rs56258221		CD,CeD,MS,T1D,Vi
10p15	rs4147359	AA,MS,RA	T1D,Vi
11q23	rs7937682		Colorectal cancer
12q13	rs11168249	**CD**,**UC**	
12q24	rs3184504	**BP**,**CeD**,Ch,CKD,**EC**,He,Hg,**Ht**,PBC,RVC,**T1D**	
18q22	rs1788097	T1D	
19q13	rs60652743	T1D	CLL
21q22	rs2836883	AS,**CD**,**UC**	

Table 3.3. Association of genome-wide significant PSC risk loci with other diseases. A SNP association for another disease is defined to be the same signal if this SNP is in high LD (r2 > 0.8) with one or more genome-wide significant PSC associated SNPs in the locus. Diseases highlighted in bold denote associations where the lead SNP is the same in both diseases. Previously associated SNPs were obtained from the Catalog of Published Genome-wide Association Studies (http://www.genome.gov/gwastudies). SNPs reported in other Immunochip experiments were available for CeD, CD, UC and PBC (Jostins et al., 2012; Liu et al., 2012; Trynka et al., 2011b). AA: Alopecia areata, AS: Ankylosing spondylitis, BP: Blood pressure, CD: Crohn's disease, CeD: Celiac disease, Ch: Cholesterol, CHD: Coronary heart disease, CKD: Chronic kidney disease, CLL: Chronic lymphocytic leukaemia, EC: Eosinophil counts, GD: Grave's disease, He: Haematocrit, Hg: Haemoglobin, HT: Hypothyroidism, MI: Myocardial infarction, MS: Multiple sclerosis, PBC: Primary biliary cirrhosis, RA: Rheumatoid arthritis, RVC: Retinal vascular calibre, T1D: Type 1 diabetes, UC: Ulcerative colitis, Vi: Vitiligo

3.3.2 Associations at previously reported non-HLA PSC risk loci

In the main association analysis, three out of six previously reported genome-wide significant ($P < 5\times10^{-8}$) non-HLA risk loci (rs3748816 at 1p36, rs3197999 at 3p21 and rs4147359 at 10p15) (Folseraas *et al.,* 2012; Melum *et al.,* 2011; Srivastava *et al.,* 2012) were genome-wide significant (Table 3.2, Figure 3.4). In a fourth locus, the genome-wide significant SNPs from the previous study (rs3749171 and rs4676410 at 2q37) (Ellinghaus *et al.,* 2012) failed genotyping in one of the genotyping batches and was excluded. However, the peak SNP in this dataset (rs2011743) was in moderate linkage disequilibrium (r2 = 0.29) with the lead SNP from the previous study (rs3749171) and showed nominal association, ($P = 5.0\times10^{-5}$, OR = 1.17, 95% CI 1.08-1.26). The previously reported PSC associations at 2q13 and 18q21 were not covered on the Immunochip (Ellinghaus *et al.,* 2012; Melum *et al.,* 2011).

3.3.3 Candidate gene prioritisation

To prioritize candidate genes within the non-HLA genome-wide significant loci, I searched for nonsynonymous coding and known eQTLs among the SNPs in high LD ($r^2 > 0.8$) with the most associated SNPs. Risk loci were als functionally annotated using data from the ENCODE project (Ward and Kellis, 2012). Networks were constructed based on known protein-protein interactions (DAPPLE) (Rossin *et al.,* 2011) and the text mining published literature (GRAIL) (Raychaudhuri *et al.,* 2009) to identify potentially important disease-relevant genes. For four of the 12 genome-wide significant loci, the same gene (*MMEL1*, *MST1*, *SH2B3*, and *CD226*) was annotated by more than one method (Table 3.4), suggesting these as candidates for further investigation at these loci.

Two newly associated loci are located outside of the Immunochip fine mapping regions (Figure 3.5). At 11q23, the most strongly associated SNP, rs7937682 (OR = 1.17, $P = 3.18\times10^{-9}$), is located in an intron of salt-inducible kinase 2 (*SIK2*), which both influences the expression of interleukin-10 in macrophages and Nur77, an important transcription factor in leukocytes (Hanna *et al.,* 2011). The association at 12q13 is with an intronic SNP (rs11168249, OR = 1.15, $P = 5.49\times10^{-9}$) within the histone deacetylase 7 (*HDAC7*) gene, which has also been associated with IBD (Jostins *et al.,* 2012). *HDAC7* has been implicated in negative selection of T cells in the thymus (Kasler *et al.,* 2011), a key factor in the development of immune tolerance. A role for *HDAC7* in PSC etiology is supported by the novel association at 19q13, where the most associated SNP, rs60652743 (OR = 1.25, $P = 6.51\times10^{-10}$) is located within an intron of serine-threonine protein kinase D2 (*PRKD2*). When T cell receptors of thymocytes are engaged, *PRKD2* phosphorylates *HDAC7*, leading to nuclear exclusion of *HDAC7* and loss of its gene regulatory functions, ultimately resulting in apoptosis and negative selection of immature T cells (Dequiedt *et al.,* 2003; Dequiedt *et al.,* 2005). Interestingly, this negative selection takes place due to a loss of HDAC7-mediated repression of Nur77 (regulated by SIK2) (Clark *et al.,* 2012), linking three novel PSC loci to this pathway.

Locus	SNP	ENCODE	eQTL[b]	Missense[b]	GRAIL[c]	DAPPLE[c]	No. of genes
1p36	rs3748816	P,E,D,PB,RM	*MMEL1*	*MMEL1*			1
2q33	rs7426056	E,D,PB,RM			*CD28*		1
3p21	rs3197999	P,E,D,PB,RM	*USP4*	*BSN,MST1*	*GPX1,MST1*		5
4q27	rs13140464	D,PB,RM			*IL2*		1
6q15	rs56258221	D,PB,RM			*BACH2*		1
10p15	rs4147359	PB,RM			*IL2RA*		1
11q23	rs7937682	P,E,D,PB,RM			*CRYAB,HSPB2*	*SIK2*	3
12q13	rs11168249	E,D,PB,RM			*VDR*		1
12q24	rs3184504	P,E,D,RM		*SH2B3*	*SH2B3,TRAFD1*	*C12orf51*	3
18q22	rs1788097	E,D,PB,RM		*CD226*	*CD226*		1
19q13	rs60652743	P,E,D,PB,RM					0
21q22	rs2836883	E,D,PB,RM			*ETS2*		1

Table 3.4. Candidate functional annotations and genes among genome-wide significant PSC risk loci. [a]SNPs in high LD ($r^2 > 0.8$)with the lead SNP that overlap one or more of the following ENCODE annotations in at least one of 147 cell types identified using HaploReg (Ward and Kellis, 2012). P: promotor histone markers; E: enhancer histone markers; D: DNAse-I hypersensitivity; PB: protein binding; RM: regulatory motifs. [b]SNPs in high LD with the most significantly associated SNP in the locus that are either known eQTLs or missense mutations. [c]Genes implicated by GRAIL, DAPPLE or functional similarity networks that show nominally significant ($P < 0.05$) number of connections.

3.3.4 HLA association

The associations at the HLA complex at 6p21 were refined by imputing HLA haplotypes at *HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1*, *HLA-DQB1*, *HLA-DQA* and *HLA-DPB1* (Dilthey *et al.,* 2011). Imputation was highly accurate at 2 digit level for *HLA-B* and *HLA-DRB1*, with >96% and 98% concordance respectively when compared with previous in-house sequencing-based HLA typing data (Karlsen *et al.,* 2007; Melum *et al.,* 2011). The lead SNP in the HLA complex (rs4143332; $P = 6.39\times10^{-249}$) was in perfect linkage disequilibrium with the lead SNP in the previous genome-wide association study (rs3134792, $r^2 = 1.0$) (Melum *et al.,* 2011), and in almost perfect linkage disequilibrium with HLA-B*08:01 ($r^2 = 0.996$ with imputed HLA- B*08:01 in this dataset). HLA-B*08:01 is encoded on the ancestral HLA-B*08:01-DRB1*03:01 haplotype (AH8.1) which is associated with multiple autoimmune diseases (Candore *et al.,* 2002).

Stepwise conditional analysis was performed including both SNP and HLA haplotypes. The SNP rs4143332 (tagging HLA-B*08:01) and a complex HLA class II association signal determined by HLA-DQA1*01:03 and SNPs rs532098, rs1794282 and rs9263964 explain all of the genome-wide significant HLA

association signals in the data (Figure 3.6). Stepwise conditional regression with only HLA alleles showed significant associations with the established PSC haplotypes HLA-B*08:01, HLA-DQA*01:03, HLA-DQA*05:01, DRB1*15:01 and DQA*01:01, confirming previously reported associations with HLA haplotypes in PSC (Table 3.5) (Chapman *et al.,* 1983; Donaldson *et al.,* 1991; Donaldson and Norris, 2002; Wiencke *et al.,* 2007).

The HLA-DRB1*15:01 association overlaps with that of ulcerative colitis (risk increasing) and Crohn's disease (risk decreasing) (Okada *et al.,* 2011; Stokkers *et al.,* 1999). Since imputed genotypes at the class II region were only available for four (HLA-DRB1, HLA-DQB1, HLA-DQA1 and HLA-DPB1) out of 20 loci (Horton *et al.,* 2004), further studies involving direct sequencing of all HLA class II loci along with assessments of their protein structure and peptide binding are required to causally resolve the link between this HLA subregion and PSC development (Hov *et al.,* 2011; Hovhannisyan *et al.,* 2008).

		Per-allele model		Full model	
HLA allele	MAF	OR	P-value	OR	P-value
B*08:01	0.12	2.82	3.70×10^{-246}	2.53	3.79×10^{-80}
DQA*01:03	0.07	2.23	1.20×10^{-100}	3.66	7.43×10^{-167}
DQA*05:01	0.16	2.39	6.00×10^{-175}	1.87	5.41×10^{-36}
DRB1:15:01	0.14	1.04	0.28	1.57	7.41×10^{-35}
DQA*01:01	0.09	0.83	1.20×10^{-6}	1.31	6.60×10^{-15}

Table 3.5. Odds ratio and P-value of independent HLA allele associations with PSC. Five independent HLA allele associations were identified via stepwise conditional analysis. The per-allele model denotes ORs and P-values of each HLA-allele from a univariate model (no covariates), while the full model includes all five HLA alleles as covariates in a multivariate model. Association testing was performed using the linear mixed model implemented in MMM (Pirinen *et al.*, 2012).

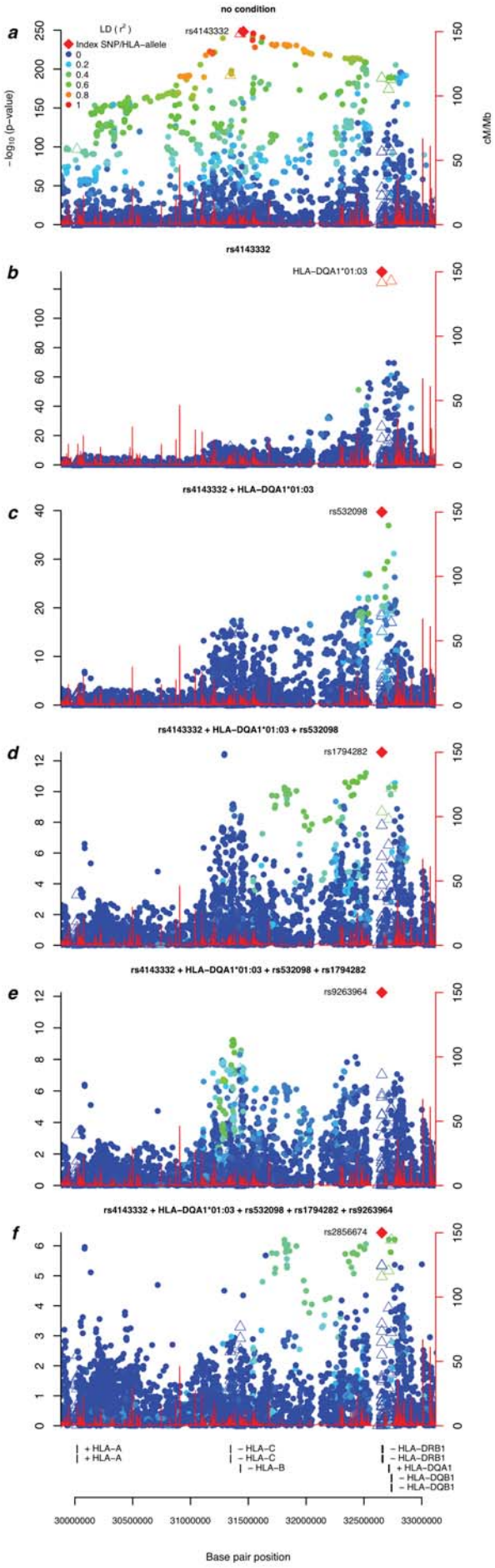


Figure 3.6. Regional association plots from stepwise conditional regression in the HLA complex in PSC. Regional association plots show both SNPs (filled circles and imputed HLA alleles (hollow triangles). Panel a displays associations with no conditioning, showing the peak association with rs4143332, which is in strong linkage disequilibrium with HLA-B*08:01. Panel b shows the association results when conditioned on rs4143332, panel c conditioned on rs4143332 and HLA-DQA1*01:03 and so on. Points are coloured according to linkage disequilibrium with the most strongly associated variant (see panel a for colour legend), which is sown as a filled in red diamond. Recombination rates in the region are shown by the red lines (in cM/MB).

3.3.5 Genetic overlap with IBD

IBD subphenotypes were available for 3,285 of the 3,789 cases (Table 3.6). Although 72% of the PSC patients in this study have a diagnosis of concomitant IBD, only half of the genome-wide significant loci were associated with IBD in the recent International IBD Genetics Consortium (IIBDGC) GWAS meta-analysis (Jostins *et al.,* 2012), despite the greater sample size of that study (~75,000 cases and controls) (Figure 3.7). Three of the six PSC risk alleles with no evidence of association in IBD (*BACH2*, *IL2RA* and *PRKD2*) contain other variants nearby that are associated with IBD. Across the 12 PSC loci, there was greater similarity between the OR estimates for PSC and ulcerative colitis than for PSC and Crohn's disease. Indeed, all but one of the CD/UC ORs for PSC-only risk alleles are > 1, suggesting that some of these may also be IBD risk loci, or that these ORs are partly driven by the small number of IBD cases in Jostins *et al.* who also have PSC.

Subphenotype	N
Crohn's disease	355
Ulcerative colitis	1898
Indeterminate IBD	108
No IBD	922
Unknown	506
	3789

Table 3.6. IBD Subphenotypes among PSC cases.

Significant genetic overlap between PSC and IBD was also observed at the 163 known IBD risk loci. While only six of these loci exceeded genome-wide significance in the PSC association analysis, 123 of the 163 IBD risk loci showed the same direction of effect ($P = 5.07\times10^{-11}$) (Figure 3.8). If the two phenotypes were unrelated, this fraction would be closer to 50%. This positive correlation in the direction of effects was stronger for loci associated with just UC (74% concordance, P = 0.0053) than those only associated with CD (60% concordance, P = 0.1). The greatest concordance was seen for loci that were associated with both CD and UC (80% concordance, $P = 1.1\times10^{-11}$).

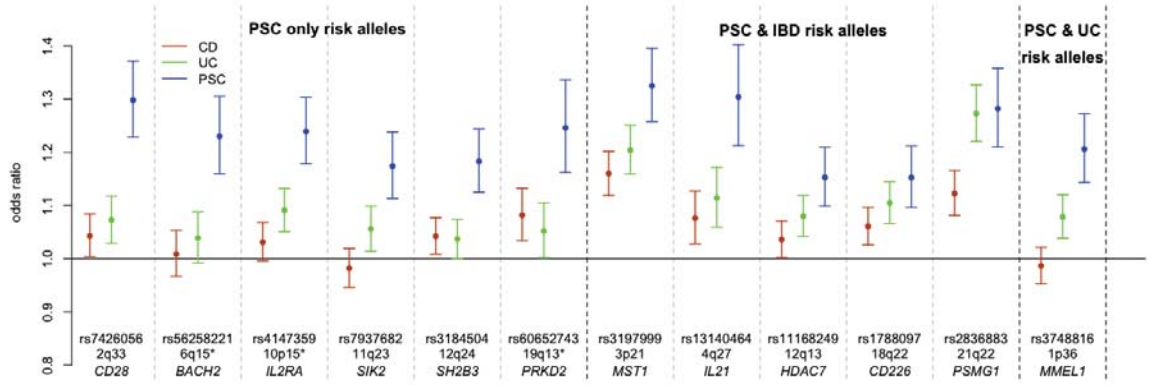


Figure 3.7. Odds ratio comparisons for PSC risk loci in IBD. IBD ORs and designation of loci as UC, CD or both (IBD) were obtained from Jostins et al. (2012). Error bars represent 95% confidence intervals. *The PSC associated alleles at 6q15 (*BACH2*), 10p15 (*IL2RA*) and 19q13 (*PRKD2*) are independent of the reported IBD associations (r2 < 0.3) but are located in the same broad genetic regions as the IBD-associated SNPs.

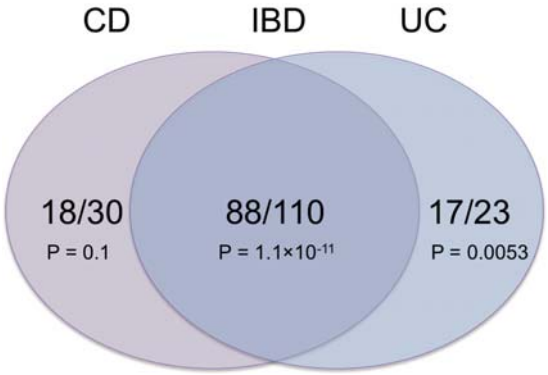


Figure 3.8. Venn diagram of directions of effect in PSC of SNPs associated with either CD, UC or both (IBD). The numbers within each segment denote the number of variants that have the same direction of effect in PSC as CD/UC/IBD over the total number of CD/UC/IBD variants. P-values were obtained from a binomial test (H1: proportion ≠ 0.5).

This trend for a greater genetic similarity between PSC and UC than CD also extends to the aggregate effect sizes at these loci. I used the Crohn's disease and ulcerative colitis OR estimates for the 163 IBD-associated loci to generate risk scores and predict case/control status in the PSC sample. There was a significantly greater area under the receiver operating characteristic curve (AUC) when prediction was performed using UC ORs compared to CD ORs (UC AUC = 0.62, CD AUC = 0.56, $P = 1.2 \times 10^{-57}$) (Figure 3.9).

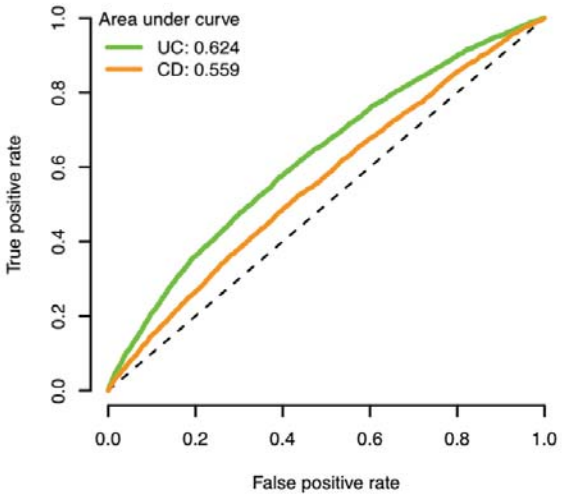


Figure 3.9. Predicting PSC using OR estimates from CD and UC risk loci. The green and orange lines represent the ROC curves for discriminating PSC cases from population controls using UC and CD ORs estimated in Jostins et al. (2012) respectively. The dashed diagonal line us y = x, and specifies the ROC curve of a random predictor.

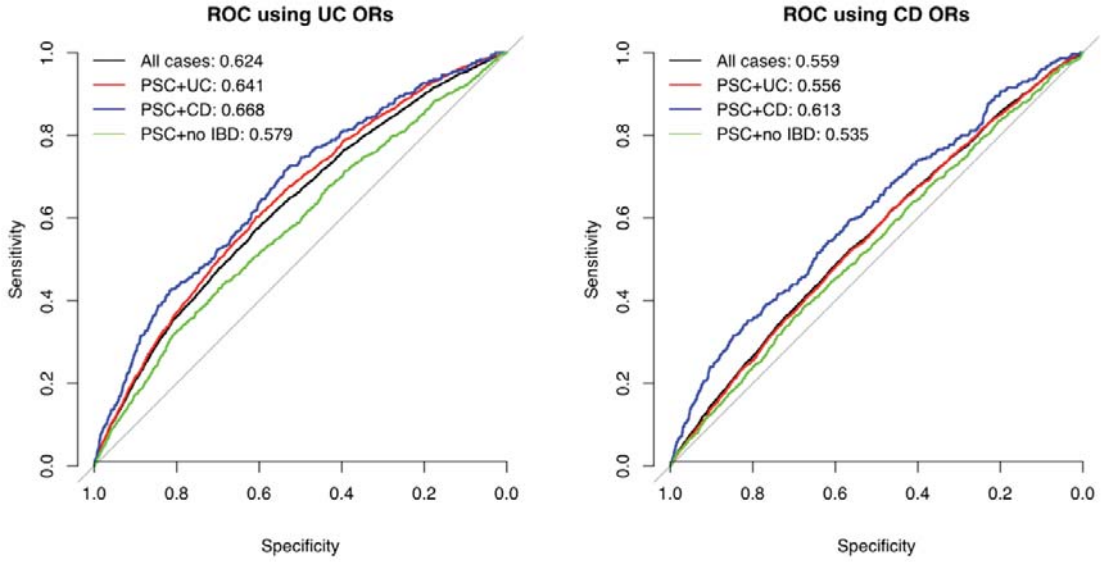


Figure 3.10. Predicting the IBD subphenotypes of PSC patients using OR estimates from CD and UC risk loci. PSC cases were divided into whether they have also been diagnosed with CD, UC, or no IBD, and the performance of the UC and CD ORs predictors assessed for each subphenotype.

That prediction accuracy is greater when performed using UC ORs suggests that PSC is genetically more similar to UC than CD, consistent with clinical observations of greater comorbidity between PSC and UC than CD. However, this

conclusion creates a slight circular argument. It may well be that the higher number of PSC patients with UC than CD is driving this improved prediction. To test this, I repeated the analysis on predicting subsets of PSC cases. PSC cases were divided into whether they have UC, CD, or no IBD (hence referred to as PSC+UC, PSC+CD or PSC+no IBD repsectively) (Table 3.6), and an AUC was estimated using UC and CD ORs on their ability to distinguish each PSC-IBD subset with controls. Notably, the results show that the better predictive performance using UC ORs extends to both PSC+UC and PSC+CD, with little difference in AUCs between the two subsets (PSC+UC AUC = 0.64, PSC+CD AUC = 0.67) (Figure 3.10). This suggests that the previous predictive performance on all PSC cases using UC ORs was not driven by the greater comorbidity between PSC and UC than with CD.

Thus far, I have used the PSC risk loci identified in this study and the IBD risk loci identified in Jostins *et al.* (2012) to illustrate genetic risk factors that are shared and those that are unique to the two diseases. I next considered the degree of sharing that exists genome-wide. Using a linear mixed model that simultaneously considers the effects of all genome-wide SNPs on a phenotype, it is possible to estimate the size of additive genetic variance component, or the total proportion of variance explained, of these SNPs (Yang *et al.,* 2010). In a bivariate extension of the method, it is also possible to estimate additive covariance components due to the SNPs, and provide an estimate of the genetic correlation (r_G) between two phenotypes (Lee *et al.,* 2012). I estimated the degree of genetic correlation between PSC and IBD using individual-level genotype data from an on-going PSC GWAS (2,871 cases and 12,091 controls) (Sun-Gou Ji, personal communication) and from previous IBD GWAS meta-analyses (5,322 CD caess, 6,307 UC cases and 12,164 controls) (Franke *et al.,* 2010; Anderson *et al.,* 2011a; Jostins *et al.,* 2012).

When considering all PSC cases, the genetic correlation was higher between PSC and UC (r_G = 0.47) than PSC and CD (r_G = 0.21), in line with the previous results showing that overlap at specific risk loci (Figure 3.11). Repeating the analysis on subsets of PSC according to their IBD diagnoses (Table 3.6) also showed similar levels of genetic correlation, with the exception of between

PSC+no IBD patients and CD (r_g = 0.038), which was not significantly different from 0 (P = 0.23). Removing the HLA region from the analysis increased estimates of r_G for both between PSC and CD (r_G = 0.26) and PSC and UC (r_G = 0.55), suggesting that variants in the HLA complex confer different effects on PSC and IBD. This is not surprising given differences in effect sizes between HLA variants in PSC and UC. For instance, rs4143332, which tags the HLA-B*08:01 haplotype, shows no evidence of association in UC (P = 0.12; UC data from Chapter 4), while it is by far the strongest associated variant in PSC in this study ($P = 6.39\times10^{-249}$) (Table 3.5 and Figure 3.6).

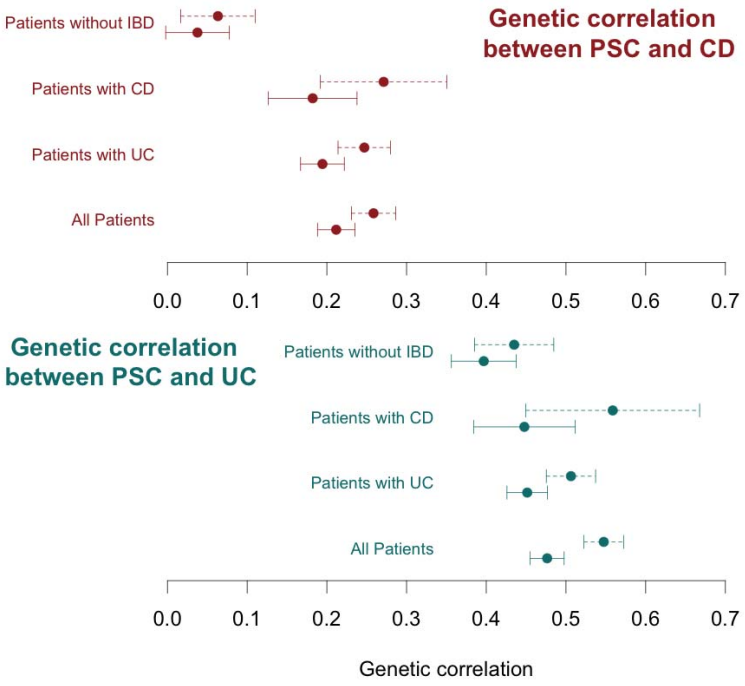


Figure 3.11. Genetic correlation (r_G) estimates using genome-wide SNP data between CD/UC and PSC subphenotypes. Error bars represent standard errors. The dashed error bars and points represent r_G estimates when the HLA region is excluded.

The previous sets of analyses looked at three levels of the genetic overlap between PSC and IBD: within 12 PSC risk loci, within 163 IBD risk loci, and genome-wide. Taken together, the results demonstrate that there is indeed a high degree of genetic overlap between PSC and IBD, that this overlap is stronger between PSC and UC, and does not appear dependent on the IBD-status of the

PSC patient. Given the unclear aetiology of PSC, this raises questions about how pleiotropy can arise. Is PSC a direct result of IBD (and in particular, UC), in which a number of genetic and environmental modifiers affecting existing IBD patients give rise to PSC? Or is PSC a distinct disorder in its own right that shares phenotypic features and genetic risk factors with IBD, much in the same way that CD and UC are considered distinct?

In order to answer help answer these nosological questions, it is important to distinguish between the various situations in which pleiotropy can arise. If it is assumed that a single causal variant underlies a locus that is associated with two correlated phenotypes, the observed pleiotropy can either be mediated by shared biology (biological pleiotropy) or via only one of the phenotypes (mediated pleiotropy). In the former case, the causal variant may reflect molecular processes that result in distinct pathological features (e.g. in different cell types), leading to increased risk for both diseases. In the later case, apparent pleiotropy will be observed if the first phenotype directly causes the second, such that associations with the second phenotype are due entirely to this phenotypic correlation. These two models are illustrated in Figure 3.12, where PSC and UC are modelled as two distinct phenotypes that share a causal genetic variant, or where PSC is a direct consequence of UC. Mediated pleiotropy can be tested by looking for an association in the second phenotype in individuals where the first phenotype is not present. If the association signal persists, then the observed pleiotropy is more likely due to shared biology rather than being mediated by one of the phenotypes (Solovieff *et al.,* 2013). More generally, Mendelian randomisation can also be used to tease out the causal relationships, however, neither approach can distinguish between the models of pleiotropy when there exists one or more confounding factors that affect both the phenotypes and is also influenced by genotype (Lawlor *et al.,* 2008).

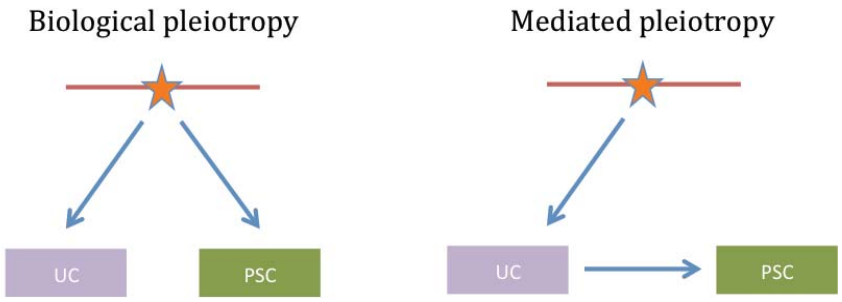


Figure 3.12. Two models of pleiotropy. The star represents a causal genetic variant tagged by a SNP associated to both UC and PSC. The arrows indicate the directions of causality between the SNP and phenotypes. Figure adapted from Solovieff et al. (2013).

Of the 12 genome-wide significant PSC risk variants identified here, six were also reported to be associated with UC (Figure 3.7). If PSC is partly mediated via IBD (and UC in particular), then it may be that these observed PSC associations at UC SNPs are due to mediated rather than biological pleiotropy. To test this, I again stratified PSC cases into subsets of whether they were also diagnosed with UC (n = 1,898) or had no IBD (n = 922) (Table 3.6). I then repeated the association analysis for each subset against controls. There was no evidence for any differences in odds ratios at any of the PSC and UC-associated variants, nor for that matter, any of the other six genome-wide significant PSC risk variants (Figure 3.13). It would have also been possible to stratify PSC cases into those with CD or indeterminate IBD, though there were much fewer samples of these and hence little power to detect any differences.

While these results suggest that common biology rather than phenotypic correlation explains the pleiotropy between PSC and IBD at these loci, caution must be applied when extrapolating these to all PSC and IBD associated loci. Firstly, this analysis was only performed on risk loci that were detected using all PSC cases. Hence variants that affect all PSC individuals are much more likely to be discovered than those associated with only a subphenotypes of PSC. Secondly, it remains to be seen how many of the non-IBD PSC cases will go on to develop IBD. The two diseases share some gastrointestinal symptoms, and while IBD precedes PSC in the majority of cases, the onset of both conditions may be separated be several years (Saich and Chapman, 2008). Finally, larger sample sizes will be required to obtain an accurate assessment of whether any

associations at additional risk loci, especially those with known IBD associations, are driven by the correlated phenotypes or shared biology.

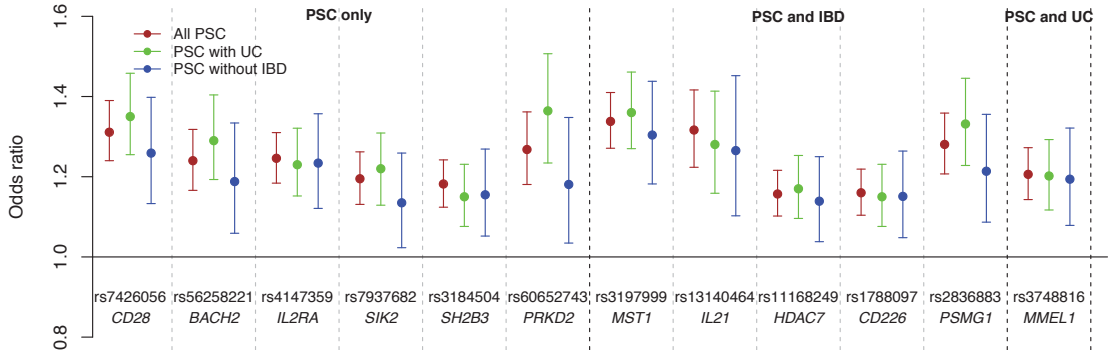


Figure 3.13. Odds ratios of PSC risk loci calculated using all PSC cases compared with odds ratios calculated using PSC+UC and PSC+no IBD subphenotypes. The error bars represent 95% confidence intervals. Designation of whether loci are associated with just PSC or both PSC and IBD follows that of Figure 3.7.

Overall, I showed that the genetic overlap between PSC and IBD is pervasive, and that this overlap is stronger between PSC and UC than between PSC and CD, mirroring the phenotypic comorbidity between the diseases. Within PSC risk loci, the genetic effects appear independent of whether UC was diagnosed along with PSC, suggesting that these loci reflect shared biology between the two diseases rather than a UC to PSC causal relationship. Incorporating association results with disease relevant functional genomic datasets may provide leads in uncovering the mechanisms behind this pleiotropy: does the causal variant result in distinct pathological features in different cells types, and do these differences reflect different disease states? I will explore approaches of integrating functional genomic datasets with disease risk loci to help answer these types of questions in Chapter 5.

3.4 Conclusion

Through genotyping of 3,789 PSC cases and 25,079 controls using the Immunochip, this study identified 12 non-HLA genome-wide significant loci, of

which nine are implicated in PSC for the first time. Network analysis using GRAIL and DAPPLE, along with searching for known eQTLs and coding variants revealed at least one candidate gene in at 11 of these loci, three of which are linked by genes that interact with each other to mediate T cell apoptosis (*SIK2*, *HDAC7* and *PRKD2*), offering new leads into the pathogensis of PSC.

The data also convincingly show pervasive overlap between genetic variants that affect PSC and IBD, and that this overlap is greater between PSC and UC than between PSC and CD, reflecting the observed comorbidity between the disorders. As many as half the variants are shared between PSC and UC when considering PSC and UC risk loci, as well as the genetic covariance between the two disorders tagged by SNPs genome-wide. Stratifying PSC cases into those with and without UC strongly suggests that this overlap is due to biological rather than mediated pleiotropy. This study demonstrates the utility of cheap high-density genotyping arrays in discovering novel loci and enabling powerful cross-phenotype comparisons.

Chapter 4. Trans-ethnic meta-analysis for inflammatory bowel disease risk loci and population comparisons

4.1 Introduction

Inflammatory bowel disease (IBD) describes chronic inflammatory conditions that affect the gastrointestinal tract. Crohn's disease (CD) and ulcerative colitis (UC) are the two main forms of IBD. In CD, inflammation can occur in patches anywhere along the gastrointestinal tract, while in UC, inflammation occurs continuously and is restricted to the colon. The exact causes of IBD are unknown, though it is likely to involve a disrupted immunological response to gut microbiota in genetically susceptible individuals (Khor *et al.,* 2011). There is currently no known cure, and disease is managed by a combination of immune-suppressing medications, dietary changes or surgery.

The prevalence of IBD in European populations ranges from 26-322 cases per 100,000 for CD and 24-505 per 100,000 for UC (Loftus, 2004; Molodecky *et al.,* 2012). The prevalence of IBD in Asian populations is lower (1-18 per 100,000 for CD; 5-57 per 100,000 for UC) though has been rapidly increasing in recent decades (Molodecky *et al.,* 2012; Prideaux *et al.,* 2012). This increase is hypothesised to be a result of lifestyle changes such as westernisation of diet, improved hygiene, vaccinations and antibiotics use, as well as genetic differences between Europeans and Asians (Prideaux *et al.,* 2012).

In 2012, a GWAS meta-analysis of IBD in ~75,000 European individuals identified 163 loci (representing 193 independent signals) associated with CD, UC or IBD (both CD and UC) at genome-wide significance ($P < 5\times10^{-8}$) (Jostins *et al.,* 2012). Smaller GWAS in populations from Korea, Japan and India (Asano *et al.,* 2009; Juyal *et al.,* 2014; Yamazaki *et al.,* 2013; Yang *et al.,* 2014b) have revealed six associated risk loci at genome-wide significance. Three of these loci overlap with those identified in Europeans (13q12, *FCGR2A* and *SLC26A3*), while the remaining three are nominally associated in Europeans ($P < 5\times10^{-4}$) and also show consistent directions of effect (Jostins *et al.,* 2012). This sharing of risk loci suggest that combining samples from different populations will give greater power to identify risk loci. Nevertheless, despite the much smaller sample sizes (typically a discovery cohort of a few hundred cases), these studies also hinted at genes that differ in their effect on European and Asian IBD. These differences include variants that confer significantly different effect sizes (e.g. *TNFSF15*, HLA), established susceptibility genes with no evidence of associations in East Asians (e.g. *NOD2*, *ATG16L1*), and vice versa (e.g. *ATG16L2*).

Here, I describe a trans-ethnic genetic association study of 10,216 individuals (2,043 CD, 2,801 UC and 5,372 controls) of East Asian, Indian and Indo-European descent and 65,642 European individuals (17,897 CD, 13,768 UC and 33,977 controls – an extension of Jostins *et al.* (2012)) genotyped on the Immunochip. I combined Immunochip data with the Jostins *et al.* GWAS data (5,956 CD, 6968 UC and 21,770 controls) in a transethnic meta-analysis with a total of 96,620 individuals (13,654 European samples were genotyped on both Immunochip and GWAS arrays and removed from the Immunochip cohort). In addition to locus discovery, I also used Immunochip data to compare the effects of IBD risk loci between European and non-European populations in an effort to identify both commonalities and differences in the genetic risk of IBD between the populations.

4.1.1 Contributions

The study design was conceived by the International IBD Genetics Consortium (IIBDGC). Cases and controls were ascertained through the IIBDGC and the

International Multiple Sclerosis Genetics Consortium. Genotyping was performed at various centres described in Jostins *et al.* (2012). Immunochip SNP and sample quality control were performed by Suzanne van Sommeren and Hailiang Huang. Association studies in individual non-European populations on the Immunochip were performed by Suzanne van Sommeren. GWAS QC, meta-analysis and imputation in Europeans were performed by Stephan Ripke and described in Jostins et al. 2012. GRAIL and DAPPLE analyses was performed by Hailiang Huang. Coding variant analyses were performed by Atshushi Takahashi. All other analyses were performed by myself.

4.2 Methods

4.2.1 Sample collection and genotyping

Non-European IBD patients and matched controls were recruited from centres in Japan, China, Hong Kong, South Korea, India, Iran and the UK. Recruitment of European patients and matched controls genotyped on the Immunochip was performed in 15 countries in Europe, North America, Australia and New Zealand. GWAS samples were originally obtained from seven CD and eight UC collections. See Jostins et al. (2012), Anderson et al. (2011) and Franke et al. (2010) for details. Controls consisted of blood donors or population-based studies. IBD diagnosis was based on accepted radiologic, endoscopic and histopathologic evaluations. All included cases fulfil clinical criteria for IBD.

4.2.2 Immunochip quality control

Quality control on Immunochip samples was performed separately for each cohort (European, East Asian, Indian and Iranian). SNP QC consisted of removing SNPs with a low call rate (< 98% across all genotyping batches in the ethnic population, or < 90% in one batch), SNPs that fail Hardy Weinberg equilibrium in controls ($P < 10^{-5}$), SNPs that have heterogeneous allele frequencies among the different genotyping batches within one ethnic population ($P < 10^{-5}$), SNPs that are not present in 1000 genomes phase 1, SNPs with a different missingness rate between cases and controls ($P < 10^{-5}$) and monomorphic SNPs. Following SNP

QC, 108,803 SNPs remained in the East Asian dataset, 146,785 SNPs in the Indian dataset, 153,982 in the Iranian dataset and 143,098 in the European dataset. The fewer number of SNPs in the East Asian cohort is primarily driven by the greater number of monomorphic SNPs. For the sample QC, samples with a low call rate (<98%) and outlying heterozygosity rate ($P < 0.01$) were removed. To identify duplicated and related samples, a subset of SNPs that 1) did not contain SNPs in high-LD regions, 2) have a minor allele frequency (MAF) of <0.05 and 3) pruned for LD ($r^2 < 0.1$), was used to estimate identity by descent. Sample pairs with an identity by descent of >0.8 were considered duplicates, pairs with an identity by descent of >0.4 where considered related. For these pairs, the sample with the lowest genotype call rate was removed.

Principal component analysis (PCA) was performed with the first two PCs estimated from 1000 Genomes Phase I samples and projected onto each of the non-European samples (Price *et al.,* 2006). A clear separation of the populations can be seen, with the samples clustering as expected (Figure 4.1).

After sample QC, 65,642 European (17,897 CD, 13,768 UC and 33,977 controls), 6,543 East Asian (1,690 CD, 1,134 UC and 3,719 controls), 2,413 Indian (184 CD, 1,239 UC and 990 controls) and 1,260 Iranian (169 CD, 428 UC, 663 controls) individuals remained (Table 4.1). Compared with the samples used in Jostins *et al.* (2012), this transethnic study includes an additional 3,548 cases and 16,406 controls (Figure 4.2).

	Immunochip samples			
Population	**CD**	**UC**	**Controls**	**Total**
European	17,897	13,768	33,977	65,642
East Asian	1,690	1,134	3,719	6,543
Indian	184	1,239	990	2,413
Iranian	169	428	663	1,260

Table 4.1. Post-QC patient and control panels genotyped on the Immunochip.

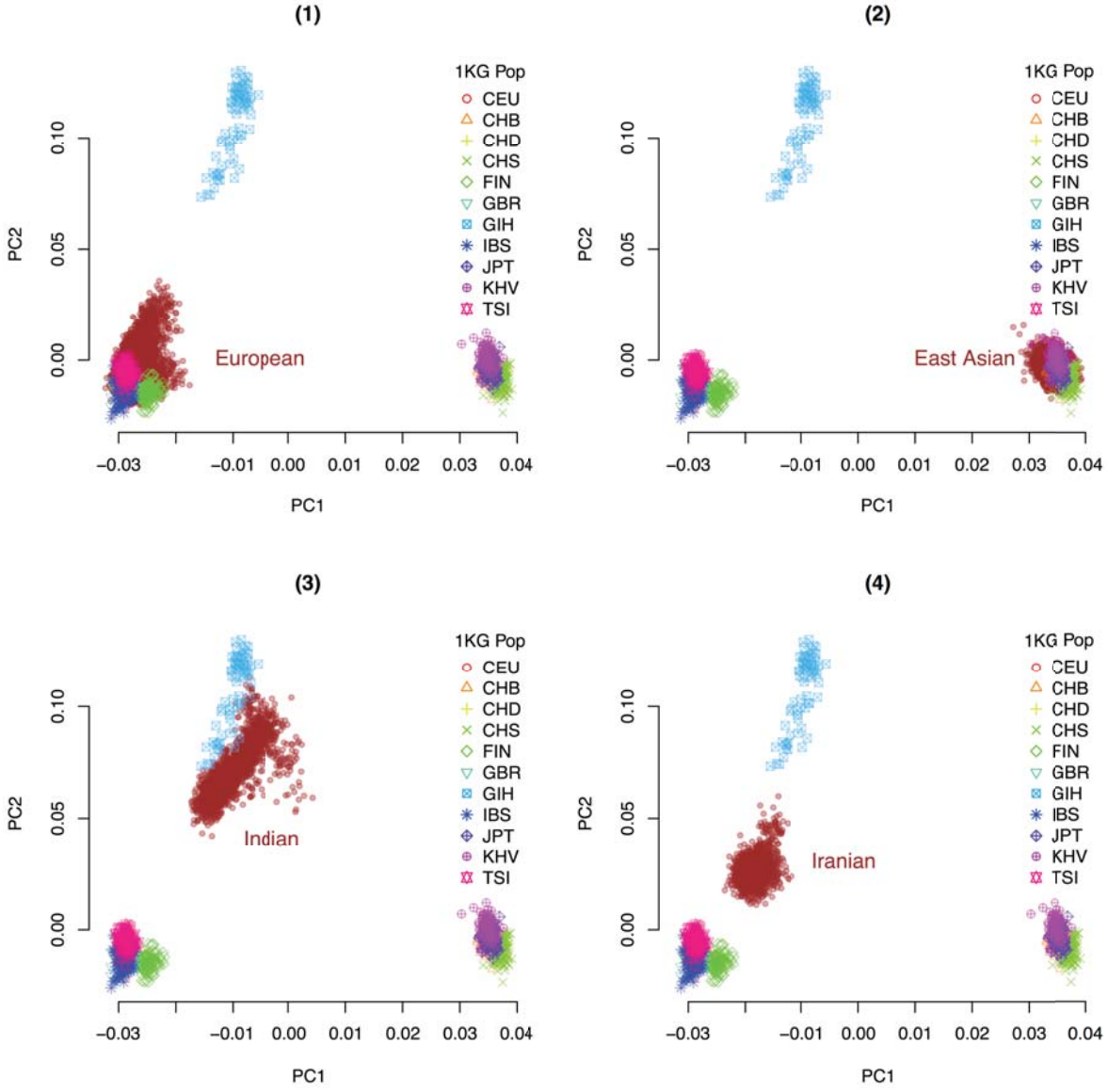


Figure 4.1. Principal components analysis of (1) European, (2) East Asian, (3) Indian and (4) Iranian IBD patients and controls. PCs 1 and 2 are plotted for each cohort as brown circles along with those from the 1000 Genomes Phase I samples.

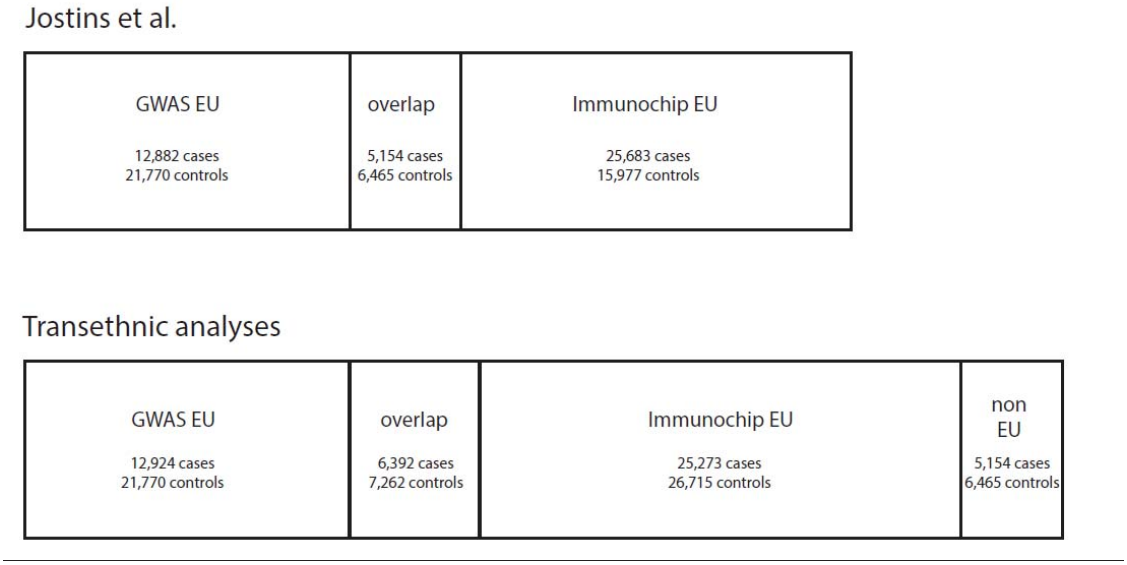


Figure 4.2. Comparison of samples used in this study with those from Jostins et al. (2012).

4.2.3 Per-population association analysis

Case-control association tests per population (European, East Asian, Indian and Iranian) per phenotype (CD, UC and IBD combined) were performed using a linear mixed model implemented in MMM (Pirinen *et al.,* 2012). The random effects component covariance matrix, *R*, was calculated using a set of SNPs with MAF > 0.1, pruned for LD ($r^2 < 0.2$) and showed no evidence of association using logistic regression with 10 PCs as covariates (P > 0.005). A total of ~14,000 SNPs were used for calculating *R* (varies between populations). For European samples, two separate association analyses were performed – one including all European Immunochip individuals (used for population comparisons), and one where 13,654 samples that overlap or are related to GWAS individuals were removed (used in the GWAS Immunochip meta analysis).

4.2.4 Transethnic meta-analysis

For European samples, association results for 1000 Genomes-imputed GWAS and Immunochip individuals (with overlaps removed) were combined using an inverse variance weighted fixed-effects meta-analysis for each of the three phenotypes. These European meta-analysis results were combined with the East Asian, Indian and Iranian association results using MANTRA (Morris, 2011), a

transethnic GWAS meta-analysis method that allows for heterogeneity of effect sizes between distantly related populations. In total, this transethnic meta-analysis was performed on 96,856 individuals and 126,990 SNPs that overlap the Immunochip and GWAS (Table 4.2). Signal intensity plots for all non-HLA loci with P-value < 10^{-7} (in the per-population association tests) or $\log_{10}$ Bayes factor (BF) > 6 in the meta-analysis were visually inspected using Evoker, and SNPs that clustered poorly were removed (Morris *et al.,* 2010).

Significantly associated loci were defined by an LD window of $r^2 > 0.6$ from the most associated SNP in the region with a per-population association P < 5×10^{-8} or $\log_{10}$ BF > 6. Regions less than 250 kb apart from each other were merged into a single associated locus.

Population	CD	CD controls	UC	UC controls	IBD	IBD controls
European GWAS	5,956	14,927	6,968	20,464	12,882	21,770
European Immunochip	14,594	26,715	10,679	26,715	25,273	26,715
Non-European Immunochip	2,043	5,372	2,801	5,372	4,844	5,372
Total	22,593	47,014	20,448	52,551	42,999	53,857

Table 4.2. Post-QC case and control panels used in the transethnic meta-analysis.

Associated loci were classified according to their strength of association with CD, UC or both using a multinomial logistic regression likelihood modelling approach within the Europeans only (Jostins *et al.,* 2012). Four multinomial logistic regression models with parameters β_{CD} and β_{UC} were fitted with the following constraints:

1. CD-specific model: $\beta_{UC} = 0$ (1 d.f.)

2. UC-specific model: $\beta_{CD} = 0$ (1 d.f.)

3. IBD unsaturated model: $\beta_{CD} = \beta_{UC} = \beta_{IBD}$ (1 d.f.)

A fourth unconstrained model with 2 d.f. was also estimated with β_{CD} and β_{UC} both fitted by maximum likelihood. Log-likelihoods were calculated for each model, and three likelihood-ratio tests were performed comparing models 1-3 against the unconstrained model. If the P-values of all three tests were less than

0.05, the SNP was classified as associated with both CD and UC but with evidence of different effect sizes. Otherwise, of the three constrained models, the SNP was classified according to the model with the largest likelihood. If 'IBD unsaturated' is the best fitting model the locus can be interpreted as associated with both CD and UC but with no evidence for different effect sizes.

4.2.5 Gene prioritisation

Two functional annotations: coding variants and expression quantitative trait loci (eQTLs), and two network approaches: GRAIL (Raychaudhuri *et al.,* 2009) and DAPPLE (Rossin *et al.,* 2011), were used to prioritise candidate genes within novel associated loci. Coding SNPs were identified if a missense or nonsense SNP was in high LD ($r^2 > 0.8$) with a lead SNP in either the 1000 Genomes Phase 1 European (CEU, FIN, GBR and IBS samples) or East Asian (CHB, CHS and JPT samples) populations (Genomes Project *et al.,* 2012). Expression quantitative trait loci were collated from the University of Chicago eQTL browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl). New IBD associated SNPs with $r^2 > 0.8$ (1000 Genomes European or East Asian cohort) with a known eQTL were reported.

4.2.6 Variance explained

The proportion of variance explained by each associated locus per population was calculated using a liability threshold model (So *et al.,* 2011) assuming a disease prevalence of 500 per 100,000 and log-additive disease risk.

4.2.7 Heterogeneity of effect sizes and allele frequencies between populations

For an associated SNP, differences in the effect size between two populations were tested using a t-test for a significant difference in log odds ratios (ORs). Overall heterogeneity between all four populations was assessed using Cochran's Q test, and the percentage of differences in ORs due to heterogeneity rather than chance was evaluated using the fixed effects I^2 statistic (Higgins and Thompson, 2002). Fixation index (F_{st}) values for a SNP between two populations were

calculated using the Weir and Cockerham method on allele frequencies in control samples only (Weir and Cockerham, 1984).

4.2.8 Genetic correlation

The proportion of genetic variation tagged by Immunochip SNPs that is shared between pairs of each of four populations was estimated using the bivariate linear mixed-effects model implemented in GCTA (Lee *et al.,* 2012). The method uses high-density SNP data to estimate genetic similarities between pairs of individuals to estimate covariance components (r_G) of the mixed model. I applied the method across Immunochip individuals for all pairwise combinations of population comparisons for CD and UC with 20 PCs as covariates, assuming a disease prevalence of 0.005. To test whether r_G is significantly different from 0 (or 1), r_G was fixed at 0 (or 1) and a likelihood ratio test comparing this constrained model with the unconstrained model was applied. An r_G of 0 means that no genetic variants are shared between the two populations, while a value of 1 means that all the genetic variance tagged in one population is shared with the other. In Europeans, only 10,000 cases and 10,000 controls (selected at random) were included due to computation limitations, while all non-Europeans samples were included.

4.2.9 Gene-based likelihood ratio test

Due to the much larger sample sizes, there is greater power to detect loci with multiple independent signals in Europeans than the non-European populations. However, if these independent SNPs within a locus are also associated in a non-European population, there may be greater power to detect these signals by jointly modelling them in the non-European population rather than single-SNP tests. To investigate this, I describe an approach that 1) identifies independently associated SNPs among SNPs within the Immunochip high-density regions in the European cohort, 2) assign the independently associated SNPs to genes, and 3) for genes with multiple associated SNPs, tests these SNPs jointly in a per-gene manner for association in a non-European cohort.

1) Independently associated SNPs were identified using the conditional and joint multi-SNP approach implemented in GCTA (GCTA-COJO) (Yang *et al.,* 2012). GCTA-COJO uses summary association statistics and LD information from a reference panel to approximate independently associated signals. GCTA-COJO was applied to CD, UC and IBD summary statistics from the Immunochip European analysis using the same European individuals as the reference panel. A joint association $P < 5\times10^{-6}$ and $r^2 < 0.9$ were used as cut-offs for assigning independent signals. It has been shown that the LD-based approximation approach of GCTA-COJO generates almost identical results to conditional logistic regression when the individual genotypes used in the association study and the reference panel are identical, as was the case in this study (Yang *et al.,* 2012). Significant independently associated SNPs that were identified via this approach were taken forward.

2) The independently associated SNPs identified in 1) were grouped according to their proximity to genes. A SNP was assigned to a gene if it lies within ±50kb of that gene's transcript start/stop positions (GENCODE 17 definitions) (Harrow *et al.,* 2012). Due to some genes overlapping each other, some SNPs may be assigned to multiple genes. Genes with more than one assigned SNP were taken forward for joint modelling in the non-European cohorts.

3) For a gene where more than one independently associated SNP was identified, the *K* independent SNPs were modelled jointly in a multiple logistic regression model (for the phenotype in which it was originally identified in) in each of the non-European populations and the total log-likelihood for the model calculated. I then performed a likelihood ratio test (with $K - 1$ degrees of freedom) comparing the log-likelihoods of this joint model with *K* SNPs and one from a null model without SNP effects. Genes with P-values less than 5×10^{-5} (equivalent to a 5% Bonferroni correction for ~1000 genes – roughly the number tagged by SNPs on the Immunochip) were considered statistically significant.

4.3 Results and discussion

4.3.1 Per-population association and transethnic meta-analysis

Per-population association analysis and the meta-analysis across all populations identified 40 novel risk loci at genome-wide significance (MANTRA $\log_{10}$ BF > 6 or per-population association $P < 5\times10^{-8}$ in at least one of the phenotypes) (Table 4.3). Likelihood modelling classified eight of these to be only associated with CD, four with UC, and 28 with IBD (both UC and CD). Of the 28 IBD loci, eight showed significant evidence of different CD/UC effect sizes (Table 4.3). Owing to the much larger sample sizes, 25 of the 40 novel loci were genome-wide significant in Europeans alone. Indeed, only three loci showed stronger evidence of association in a non-European population than European (rs10774482: IBD European P = 0.30, Iranian $P = 2.17\times10^{-7}$, Indian $P = 1.12\times10^{-3}$; rs2072711: CD European P = 7.51×10-3, East Asian $P = 2.17\times10^{-7}$; rs6856616: IBD European P = 9.72×10^{-7}, East Asian $P = 1.33\times10^{-7}$). Of these, rs6856616 was previously reported as a novel CD risk locus in a GWAS in Korean individuals (Yang *et al.,* 2014b).

The strongest signal in the European-only analysis was rs395157 (IBD $P = 2.22\times10^{-20}$). The magnitude of this association was unexpectedly high, given that the number of Europeans in this study was only modestly greater than that of Jostins et al. (2012) (86,640 vs. 76,312), such that this SNP should have exceeded genome-wide significance and reported in the previous study. The reason why this was not originally reported in Jostins et al. was a result of an error in the GWAS and Immunochip meta-analysis, where discordant alleles were merged (and effects cancelled out). This was due to the SNP having an allele frequency very close to 0.5, such that the minor allele of the GWAS and Immunochip were different. No other associated signals appeared to be affected by this issue.

Chr.	SNP	Base pair position	[a]Best trait	[b]LR trait	[c]Log_{10} BF	[d]Het I^2	Eur. OR	Eur. P	Eas. OR	Eas. P	Ind. OR	Ind. P	Ira. OR	Ira. P
1	rs1748195	62822181	CD	CD	6.08	0	1.07	7.13×10^{-8}	1.04	0.41	1.11	0.36	1.05	0.73
1	rs34856868	92326871	IBD	IBD_U	6.16	0	0.82	9.80×10^{-9}	0.11	0.43	1.47	0.34	1.36	0.69
1	rs11583043	101238642	UC	IBD_U	8.34	66.51	1.08	6.05×10^{-8}	1.18	0.032	1.27	3.80×10^{-3}	1.46	9.80×10^{-3}
1	rs6025	167785673	IBD	IBD_U	6.43	0	0.84	2.51×10^{-8}	-	-	0.81	0.41	0.7	0.31
1	rs10798069	185142082	CD	IBD_S	7.24	0	0.93	4.25×10^{-9}	0.94	0.12	1.06	0.59	1.01	0.92
1	rs7555082	196865286	CD	IBD_U	7.97	0	1.13	1.47×10^{-10}	0.6	0.67	1.02	0.92	0.85	0.44
2	rs11681525	145208852	CD	CD	8.8	59.3	0.86	4.08×10^{-11}	-	-	1.5	0.12	0.69	0.22
2	rs4664304	160502254	IBD	IBD_U	6.34	0	1.06	2.61×10^{-8}	1.01	0.77	1.04	0.51	1.18	0.12
2	rs3116494	204300266	UC	IBD_S	7.03	0	1.08	1.30×10^{-7}	1.17	0.1	1.21	0.043	1.19	0.15
2	rs111781203	228368356	IBD	IBD_U	10.04	0	0.94	2.16×10^{-10}	0.91	0.031	0.88	0.033	0.98	0.84
2	rs35320439	242386014	CD	IBD_S	7.71	0	1.09	9.89×10^{-10}	1.04	0.37	1.07	0.54	1.03	0.81
3	rs113010081	46432416	UC	IBD_U	7.45	0	1.14	9.02×10^{-10}	0.02	0.5	0.84	0.38	1.12	0.71
3	rs616597	103052416	UC	UC	6.68	54.68	0.93	9.34×10^{-6}	0.85	1.04×10^{-3}	0.84	0.029	0.79	0.044
3	rs724016	142588260	CD	CD	7.41	70.87	1.06	3.36×10^{-6}	1.21	5.56×10^{-6}	1.13	0.3	0.97	0.86
4	rs2073505	3414301	IBD	IBD_U	6.87	0	1.1	1.46×10^{-7}	1.14	6.83×10^{-3}	1.04	0.62	0.95	0.76
4	rs4692386	25741459	IBD	IBD_U	6.47	0	0.94	1.21×10^{-8}	0.97	0.49	0.98	0.7	0.9	0.27
4	rs6856616	38001431	IBD	IBD_U	9.78	61.59	1.1	9.72×10^{-7}	1.24	1.33×10^{-7}	1.07	0.35	1.18	0.31
4	rs2189234	106294947	UC	UC	8.85	0	1.08	1.95×10^{-10}	1.11	0.033	0.98	0.76	1.06	0.61
5	rs395157	38903489	IBD	IBD_U	19.5	0	1.1	2.22×10^{-20}	1.09	0.027	1.12	0.065	0.99	0.93
5	rs4703855	71729655	IBD	IBD_U	6.83	70.26	0.93	7.16×10^{-11}	1	0.97	1.04	0.52	1.15	0.18
5	rs564349	172257584	IBD	IBD_U	8.12	37.54	1.06	1.54×10^{-7}	1.15	1.54×10^{-4}	1.09	0.22	1.07	0.51
6	rs7773324	327559	CD	IBD_U	7.67	0	0.92	1.06×10^{-9}	0.97	0.53	0.88	0.27	1	0.98
6	rs13204048	3365405	CD	IBD_S	7.23	53.54	0.93	2.89×10^{-8}	0.94	0.13	0.6	3.23×10^{-3}	0.97	0.85
6	rs7758080	149618772	CD	IBD_S	7.88	0	1.08	7.27×10^{-9}	1.11	0.017	1.06	0.62	0.93	0.63
7	rs1077773	17409204	UC	UC	5.86	76.72	0.93	5.96×10^{-9}	1.11	0.053	1.01	0.85	1.05	0.66
7	rs2538470	147851381	IBD	IBD_U	10.93	54.64	1.07	3.00×10^{-11}	1.15	9.78×10^{-4}	0.97	0.63	1.22	0.059
8	rs17057051	27283471	IBD	IBD_U	6.74	15.92	0.94	5.50×10^{-8}	0.9	0.022	1.02	0.7	0.87	0.16
8	rs7011507	49291795	UC	IBD_U	7.49	39.32	0.9	6.40×10^{-8}	0.82	7.42×10^{-4}	0.94	0.47	1.13	0.43
10	rs3740415	104222706	IBD	IBD_U	6.26	0	0.95	1.03×10^{-7}	0.93	0.073	0.98	0.75	1	0.99
12	rs10774482*	971525	IBD	CD	6.02	91.3	1.01	0.3	1	0.97	1.21	1.12×10^{-3}	1.63	2.17×10^{-7}
12	rs7954567	6361386	CD	CD	8.25	0	1.09	1.30×10^{-9}	1.17	0.076	1.12	0.35	1.12	0.47
12	rs653178	110492139	IBD	IBD_U	6.57	49.67	1.06	1.11×10^{-8}	0.02	0.042	1.15	0.13	0.97	0.72
12	rs11064881	118631308	IBD	IBD_U	7.02	31.65	1.1	5.95×10^{-8}	0.01	0.29	1.22	0.053	1.4	0.03
13	rs9525625	41916030	CD	CD	8.55	37.25	1.08	1.41×10^{-9}	1.07	0.22	1.11	0.34	1.46	7.08×10^{-3}
17	rs3853824	52235992	CD	IBD_S	8.46	50.42	0.92	1.17×10^{-10}	0.95	0.32	0.88	0.29	1.31	0.066
17	rs17736589	74248713	UC	UC	6.53	53.41	1.09	4.34×10^{-8}	1.05	7.30×10^{-5}	1.03	0.73	1.34	0.026
18	rs9319943	55030807	CD	CD	6.33	33.39	1.08	9.05×10^{-7}	1.19	2.03×10^{-3}	0.95	0.69	1.21	0.22
18	rs7236492	75321604	CD	IBD_S	6.6	0	0.91	9.09×10^{-9}	1.44	0.68	1.14	0.62	0.84	0.64
22	rs2072711	35598501	CD	IBD_S	6.12	91.56	0.96	7.51×10^{-3}	1.26	2.17×10^{-7}	1	0.98	1.28	0.17
22	rs727563	40197323	CD	CD	7.1	76.01	1.1	1.88×10^{-10}	0.95	0.23	0.93	0.52	0.93	0.61

Table 4.3. Table of novel IBD risk loci from MANTRA transethnic meta-analysis or individual per-population analyses. [a]Phenotype with the largest MANTRA Bayes factor. [b]Likelihood modelling classification. IBD_S and IBD_U refer to IBD saturated and unsaturated respectively. [c]MANTRA log10 Bayes factor. [d]Heterogeneity I2 percentage. Per-population ORs and P-values refer the Best trait column.

4.3.2 Candidate genes

Candidate genes for each of the novel loci were identified using two SNP annotations: coding SNPs, known eQTLs, and two network approaches: GRAIL and DAPPLE. These methods identified at least one candidate gene in 28 of 40 novel risk loci, four of which harbour genes identified by multiple methods (Table 4.4 A-B). Including the new 40 loci in GRAIL and DAPPLE analyses with known IBD risk loci revealed additional candidate genes with significant connectivity scores ($P < 0.05$ in either GRAIL and DAPPLE) at 34 of the 163 known loci that weren't reported in Jostins et al. (Table 4.5). A visual inspection of the GRAIL network plot reveals the interconnectedness between the novel and known IBD risk loci (Figure 4.3).

Many of the genes associated with IBD highlight the importance of T cells in IBD pathogenesis. T cells are an integral component in the adaptive immune response, and become activated in response to MHC-bound antigens via signalling through the T cell receptor. This process depends on PRKCQ signalling, which results in increased expression of CD44. Co-stimulation via other ligands such as CD28, CD81 and CD27 are also required for T cells to generate memory. Impaired immune responses may occur from inappropriate co-stimulation, and is characterised by increased expression of PDCD1. Other processes that can impair immune responses also include apoptosis (implicating UBASH3A) and recruitment of immunosuppressive regulatory T cells, driven partly by the chemokine CCL20. The genes mentioned are all within loci associated with IBD risk from this study and others, highlighting the importance of genetic risk factors in T cell responses in IBD pathogenesis, and may provide targets for development of future therapies.

Chr.	SNP	Cis-eQTL	Nonsynonymous coding	GRAIL	DAPPLE	Genes implicated by multiple methods
1	rs1748195	*DOCK7,AF086387, ANGPTL3*				
1	rs34856868		*BTBD8*			
1	rs11583043			*EDG1*		
1	rs6025			*SELP, SELE, SELL*		
1	rs10798069			*PTGS2,PLA2G4A*		
1	rs7555082			*PTPRC*		
2	rs4664304	*LY75*	*PLA2R1*	*LY75*		*LY75*
2	rs3116494			*ICOS ,CD28, CTLA4*		
2	rs111781203			*CCL20*		
2	rs35320439			*PDCD1, ATG4B*		
3	rs113010081			*FLJ78302,LTF,CCR1, CCR3,CCR5*	*CCR2*	
3	rs616597			*NFKBIZ*		
4	rs2073505		*HGFAC*			
5	rs395157			*OSMR,FYB*	*LIFR, OSMR*	*OSMR*
5	rs564349			*DUSP1*		
6	rs7773324			*IRF4,DUSP22*		
6	rs7758080			*MAP3K7IP2*		
7	rs1077773			*AHR*		
7	rs2538470	*CNTNAP2*				
8	rs17057051	*PTK2B*		*PTK2B*		*PTK2B*
10	rs3740415	*PDCD11,TMEM180, ACTR1A*		*NFKB2*		
12	rs7954567			*CD27,TNFRSF1A,LTBR*		
12	rs653178			*SH2B3*		
12	rs11064881	*PRKAB1*				
13	rs9525625			*TNFSF11*		
18	rs7236492			*NFATC1*		
22	rs2072711	*CSF2RB*	*NCF4*	*CSF2RB*	*IL2RB,CSF2RB*	*CSF2RB*
22	rs727563	*MEI1,PHF5A, NFP2L1,TOB2*				

Table 4.4A. Candidate genes implicated by coding variants, eQTLs, GRAIL and DAPPLE in 28 of the 40 novel IBD risk loci.

Chr.	SNP	eQTL SNP	LD (r^2)	Gene	Type	Tissue
1	rs1748195	rs1748195	1	*DOCK7*	Cis	Monocytes
		rs10889353	0.99	*AF086387*	Cis	Liver
		rs1168089	1	*ANGPTL3*	Cis	Liver
2	rs4664304	rs7601374	0.97	*LY75*	Cis	Liver
8	rs17057051	rs17057051	1	*PTK2B*	Cis	Monocytes
10	rs3740415	rs3740415	1	*PDCD11*	Cis	LCLs
		rs7342070	0.98	*TMEM180*	Cis	Liver
		rs5870	0.93	*ACTR1A*	Cis	LCLs
12	rs11064881	rs11064881	1	*PRKAB1*	Cis	LCLs
		rs11064881	1	*PRKAB1*	Cis	Monocytes
22	rs2072711	rs2072711	1	*CSF2RB*	Cis	LCLs
22	rs727563	rs12165508	1	*MEI1*	Cis	LCLs
		rs203319	0.99	*PHF5A*	Cis	Monocytes
		rs202628	0.96	*NHP2L1*	Cis	Liver
		rs202614	0.94	*TOB2*	Cis	Liver

Table 4.4B. Known eQTLs tagged by novel IBD associated SNPs. eQTL SNPs, gene, eQTL type (cis or trans) and tissue studied were extracted from publications collated in the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/). LD (r^2) values were extracted from the European and East Asian cohorts of the 1000 Genomes Project Phase I (the larger r^2 of the two cohorts are reported).

Chr.	SNP	New GRAIL	New DAPPLE	[a]Uniquely new
1	rs35675666	*PARK7, ERRFI1*		*PARK7, ERRFI1*
1	rs6426833	*PLA2G2A*		*PLA2G2A*
1	rs4845604		*S100A11*	*S100A11*
1	rs3024505	*IL10*		
2	rs7608910	*REL*		
2	rs10865331		*COMMD1*	*COMMD1*
2	rs2382817	*IL8RA, IL8RB, IL8RBP*		*IL8RA, IL8RB, IL8RBP*
5	rs7702331		*BTF3*	*BTF3*
5	rs2188962	*RAD50*	*RAD50, IL5*	*RAD50, RAD50*
6	rs3851228		*FYN*	
6	rs212388	*TAGAP*	*EZR*	*TAGAP ,EZR*
7	rs10486483	*SKAP2*		*SKAP2*
7	rs1456896	*IKZF1*		*IKZF1*
7	rs9297145	*SMURF1*		*SMURF1*
8	rs7015630	*NBN*		*NBN*
8	rs1991866		*FAM49B*	*FAM49B*
9	rs4743820		*SYK*	*SYK*
10	rs2227564	*PLAU*	*VCL*	*PLAU, VCL*
11	rs10896794		*ZFP91*	*ZFP91*
11	rs11230563	*GPR44*		*GPR44*
11	rs2231884	*SIPA1*		*SIPA1*
12	rs11612508	*DUSP16*		*DUSP16*
12	rs11168249	*RAPGEF3, SENP1*		*RAPGEF3, SENP1*
12	rs7134599		*IL22, IL26*	
13	rs9557195	*EBI2*		*EBI2*
15	rs17293632		*SMAD3*	
17	rs2945412	*NOS2A*		*NOS2A*
17	rs3091316	*CCL1,CCL7*		*CCL1, CCL7*
18	rs1893217	*PTPN2*		*PTPN2*
18	rs727088	*DOK6*		*DOK6*
19	rs11879191	*ICAM3*		*ICAM3*
19	rs17694108		*CEBPG*	
19	rs4802307		*CALM3*	*CALM3*
19	rs1126510	*PTGIR*		*PTGIR*
20	rs6142618	*HCK*		*HCK*
20	rs4911259		*COMMD7*	*COMMD7*
20	rs6088765	*PROCR*		*PROCR*
20	rs913678	*PTPN1, TMEM189-UBE2V1*		*PTPN1, TMEM189-UBE2V1*
21	rs2284553		*IL10RB, IFNAR2*	
21	rs7282490	*AIRE*		*AIRE*
22	rs2266959		*MAPK1*	
22	rs2413583	*MAP3K7IP1*		*MAP3K7IP1*

Table 4.5. New genes in known IBD risk loci implicated from GRAIL and DAPPLE network analyses. [a]New genes that weren't previously implicated by either GRAIL or DAPPLE

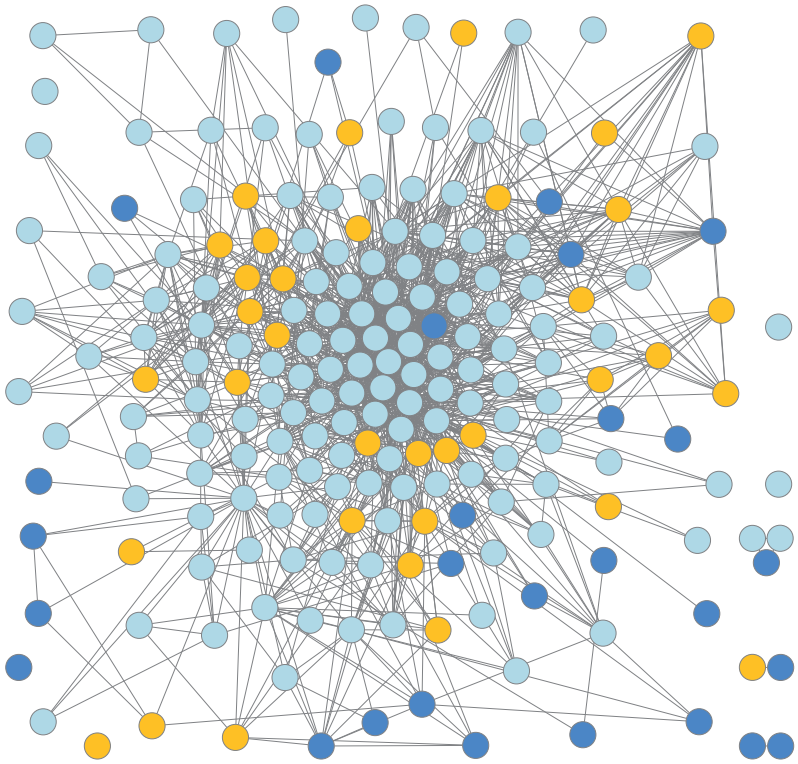


Figure 4.3. GRAIL network for all genes with GRAIL P < 0.05. Yellow nodes represent newly associated genes, light blue nodes represent known genes, dark blue genes represents new genes in known loci that now reach GRAIL P < 0.05 after including the novel loci.

4.3.3 Validation of known loci

Of the 163 IBD risk loci identified in Jostins et al. (2012), all but 16 exceeded genome-wide significance ($P < 5\times10^{-8}$) in the European only analysis here. Fifteen of these loci continue to show suggestive levels of significance ($P < 1.44\times10^{-6}$). This is equivalent to a false discovery rate of < 0.001, and not beyond what's expected given the initially reported P-values for these SNPs in Jostins et al. ($3.60\times10^{-9} < P < 3.71\times10^{-8}$) and the sampling variability in replication vs. discovery P-values (Lazzeroni *et al.,* 2014). However, one SNP, rs2226628, fell to P = 0.0023 in this analysis, suggesting that this may have been an initial false positive report, and larger samples will be required to unequivocally implicate this locus. Nevertheless, as expected, the majority of signals (107/163) become more significant with the additional European samples (Figure 4.4).

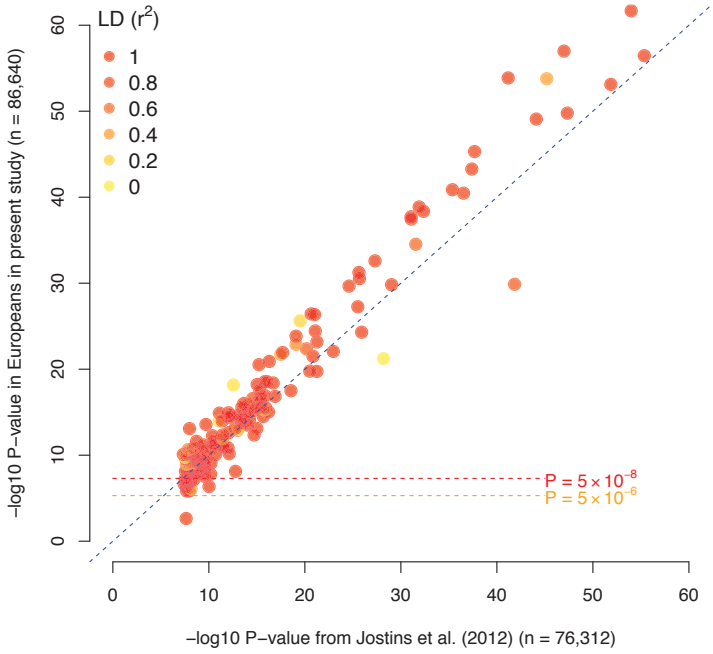


Figure 4.4. Comparison of association P-values reported in Jostins *et al.* (2012) and Europeans in this present study. LD r^2 values are between the SNP reported here and that from Jostins et al. Different SNPs may have been reported if there was stronger signal was found in this study or the previously reported SNP was removed during QC. The blue dashed line represents y = x.

The discrepancy in the rs2226628 GWAS + Immunochip meta-analysis between our study and Jostins et al. is driven almost entirely by the Immunochip samples (Immunochip IBD Jostins et al. $P = 7.52 \times 10^{-7}$ vs. P = 0.012 in this study). Several factors may be driving this discrepancy. Firstly, in the Jostins et al. study, it was later found that ~1,200 samples were mistakenly included in both the initial GWAS and the subsequent Immunochip replication effort. This may have led to an inflation of the P-values for rs2226628 and other SNPs, for which we have now corrected in this latest analysis. Another factor may be the different association methods used on Immunochip samples. In Jostins *et al.,* association was performed using logistic regression with 4 PCs as covariates, while in this study, we applied a linear (logistic) mixed model. If the SNP shows within-European population stratification that was not adequately captured by the first 4 PCs, then this may have also lead to an inflated P-value. Indeed, this SNP does appear to show varying frequencies across the European populations in the 1000

Genomes data (MAF = 0.2 in GBR to 0.47 in FIN) (Genomes Project *et al.,* 2012). In our Immunochip samples, using logistic regression with 4 PCs as covariates did result in this SNP being more significant than the mixed model (P = 0.012 vs. P = 1.85×10^{-4}), though it still did not reach the same level of significance as that of Jostins et al. Finally, the Jostins et al. meta-analysis was performed using two different SNPs – the final reported P-value was a meta-analysis of rs6592362 from the GWAS cohort and rs2226628 from the Immunochip samples. This was done since the original GWAS hit SNP, rs6592362, was not present on Immunochip and rs2226628 was selected as it was the best tag (r^2 = 0.50). In this study, I only combined GWAS and Immunochip at rs2226628, though would have achieved a more significant signal had I combined the two different SNPs (P = 7.38×10^{-6}). Notably, rs2226628 is non-significant in the GWAS (P = 0.08), and it may be the case that combining two different SNPs that are only in moderate LD with each other did not reflect the true signal in this region (if there is one).

4.3.4 Population comparisons

Recent large-scale transethnic genetic studies of complex diseases have shown that the majority of risk loci originally identified in Europeans are shared across other populations (Dastani *et al.,* 2012; Okada *et al.,* 2014; Replication *et al.,* 2014; Teslovich *et al.,* 2010). The true extent of sharing is difficult to characterise as the GWAS sample sizes in non-European populations are often much smaller than their European counterparts, limiting power to detect associated loci. Despite this study including over 10,000 non-European samples and being the largest non-European study of its type, this still pales in comparison with the European sample size of over 85,000. As such, we expect that the majority of known risk loci will not replicate in the non-European populations at genome-wide significance. Nevertheless, there were significant trends both in terms of directions of effect and strength of the correlation across all three phenotypes when comparing the 233 independently associated SNPs in Europeans and the individual non-European populations (Figure 4.5).

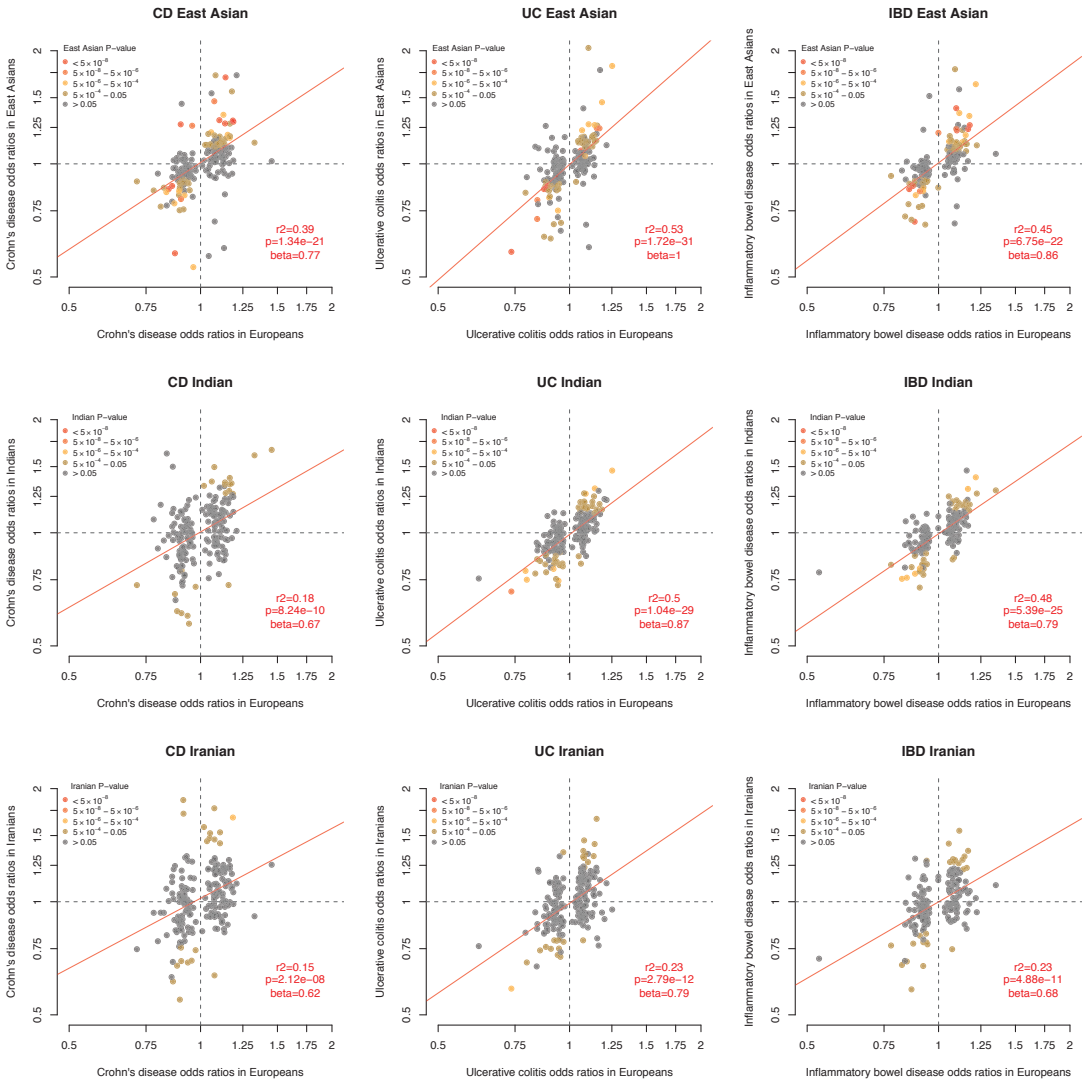


Figure 4.5. Odds ratio comparison between European and non-European populations at 233 SNPs associated with CD, UC other both. For each SNP, ORs (on log-scale) were taken for the corresponding phenotype in the European and non-European population if it was classified as associated with that phenotype in the likelihood modelling (section 4.2.4). Points are coloured according to the strength of association for the respective phenotype in the non-European population. The red line indicates the best-fitting linear regression line, weighted by the inverse variance of the log ORs in the non-European population. Regression coefficients, significance and goodness of fit are listed in the bottom right corner of each plot.

Consistent with the concordant effect sizes at associated SNPs, there were high genetic correlations (r_G) when considering all SNPs on the Immunochip for all pairwise population comparisons (Table 4.6). Estimates of r_G ranged from 0.42 (between East Asian and Indian CD) to 0.92 (between Indian and Iranian CD). Given that rare SNPs are more likely to be population-specific, high r_G values

also support the notion that the majority of causal variants are common. It is also unsurprising that r_G is significantly smaller than 1 for all pairwise comparisons (apart for those involving Iranian CD, though with only 169 cases, this most likely reflects lack of power) as there are examples of IBD risk loci that are not present in some populations, or where there are differences in effect size between populations (discussed below). Nevertheless, r_G is significantly greater than 0 ($P < 0.021$) for all pairwise population comparisons across both CD and UC. Together, these results indicate that a large proportion of IBD risk loci are shared across different populations, though accurate assessments of the actual number of shared loci and their effect sizes will require much larger sample sizes.

Phenotype	Population 1	Population 2	r_G	Standard Error	P-value (H_1: $r_G > 0$)	P-value (H_1: $r_G < 1$)
Crohn's disease	East Asian	Indian	0.42	0.13	8.02×10^{-4}	3.45×10^{-4}
	East Asian	Iranian	0.73	0.26	8.56×10^{-4}	0.223
	European	East Asian	0.76	0.04	0	4.47×10^{-14}
	European	Indian	0.56	0.09	6.58×10^{-10}	3.43×10^{-4}
	European	Iranian	0.82	0.34	5.06×10^{-7}	0.357
	Indian	Iranian	0.92	0.63	0.0209	0.456
Ulcerative colitis	East Asian	Indian	0.83	0.08	0	0.011
	East Asian	Iranian	0.56	0.12	1.37×10^{-5}	4.59×10^{-4}
	European	East Asian	0.79	0.04	0	6.61×10^{-9}
	European	Indian	0.84	0.05	0	8.23×10^{-4}
	European	Iranian	0.67	0.08	2.61×10^{-15}	6.75×10^{-4}
	Indian	Iranian	0.53	0.14	1.11×10^{-4}	2.64×10^{-3}

Table 4.6. Pairwise genetic correlation (rG) tagged by Immunochip SNPs.

While there was significant correlation in the effect sizes of IBD loci between different populations, identifying loci that differ in their effects between populations may reveal differences in disease pathogenesis. As discussed, a comprehensive comparison of effect sizes will require much larger sample sizes in non-Europeans than the one in this study. However, there was sufficient power to detect genetic heterogeneity between our East Asian and European cohorts at several alleles with reported large effect size in Europeans. For instance, consistent with previous genetic studies of Crohn's disease in East Asians (Ng *et al.,* 2012), the three coding variants in *NOD2* (nucleotide-binding oligomerisation domain-containing protein 2) with the largest effect sizes in

Europeans are all monomorphic in East Asians. Furthermore, across all NOD2 variants, no association signals were observed in the East Asian cohort beyond what is expected under a null distribution given the number of SNPs (83) assayed in this region on the Immunochip (minimum P = 7.18×10^{-4}). Similarly, at the *IL23R* (interleukin 23 receptor) gene, previous studies have shown that the most associated variants in Europeans are either monomorphic or do not appear to be associated in East Asians, though there is evidence of additional variants in *IL23R* that are associated in East Asians (Ng *et al.,* 2012). In line with these observations, the *IL23R* SNP with the largest effect in European CD and UC (rs11209026) is monomorphic in East Asians, while two secondary *IL23R* variants observed in Europeans were also non-significant (rs6588248, P = 0.65; rs7517847, P = 0.04) in East Asian IBD. Nevertheless, there was strong evidence for an association at rs76418789 with both CD and UC in East Asians (IBD P = 1.83×10^{-13}). The same variant was previously implicated in a GWAS of CD in Koreans (Yang *et al.,* 2014). This variant demonstrates suggestive evidence of association in European IBD (P = 3.99×10^{-6}, OR = 0.66), though has a much lower allele frequency than in East Asian populations (MAF = 0.004 vs. 0.07).

The identification of CD risk variants in *ATG16L1* (autophagy-related protein 16-1), first implicated autophagy as an important process in CD pathogenesis (Hampe *et al.,* 2007; Parkes *et al.,* 2007; Rioux *et al.,* 2007). At *ATG16L1*, the variant most strongly associated with Crohn's disease in Europeans (rs12994997) has a risk allele frequency (RAF) of 0.53 and OR of 1.27. The variant shows no evidence of association in East Asians, (P = 0.21), driven at least in part by a significant difference in allele frequency (RAF = 0.24, F_{st} = 0.15). However, assuming the effect size at this SNP in the East Asian cohort was equal to that seen in the European cohort, there would have more than 80% power to detect association of suggestive significance (P < 5×10^{-5}) in this study. Indeed, there was also evidence for heterogeneity of odds at this SNP (East Asian OR = 1.06; P = 8.45×10^{-4}). Association in European individuals to a locus containing *IRGM* further implicated authophagy in IBD risk, and the most associated SNP at this locus in Europeans shows only nominally significant evidence of association in East Asian CD (rs11741861, European P = 5.89×10^{-44}, East Asian P = $2.62\times10^{-}$

^{3}) as well as evidence of heterogeneity of effect (European OR = 1.33 vs. East Asian OR = 1.13; heterogeneity P = 1.2×10^{-3}). Given these results it is tempting to speculate that autophagy plays a lesser role in East Asian IBD compared to European IBD. However, a previous GWAS in a Japanese population identified suggestive evidence of association near another autophagy-related gene, *ATG16L2* (Yamazaki *et al.,* 2013), though this finding was unable to be confirmed because the reported variant (rs11235667) is monomorphic in Europeans and the locus is not covered on the Immunochip.

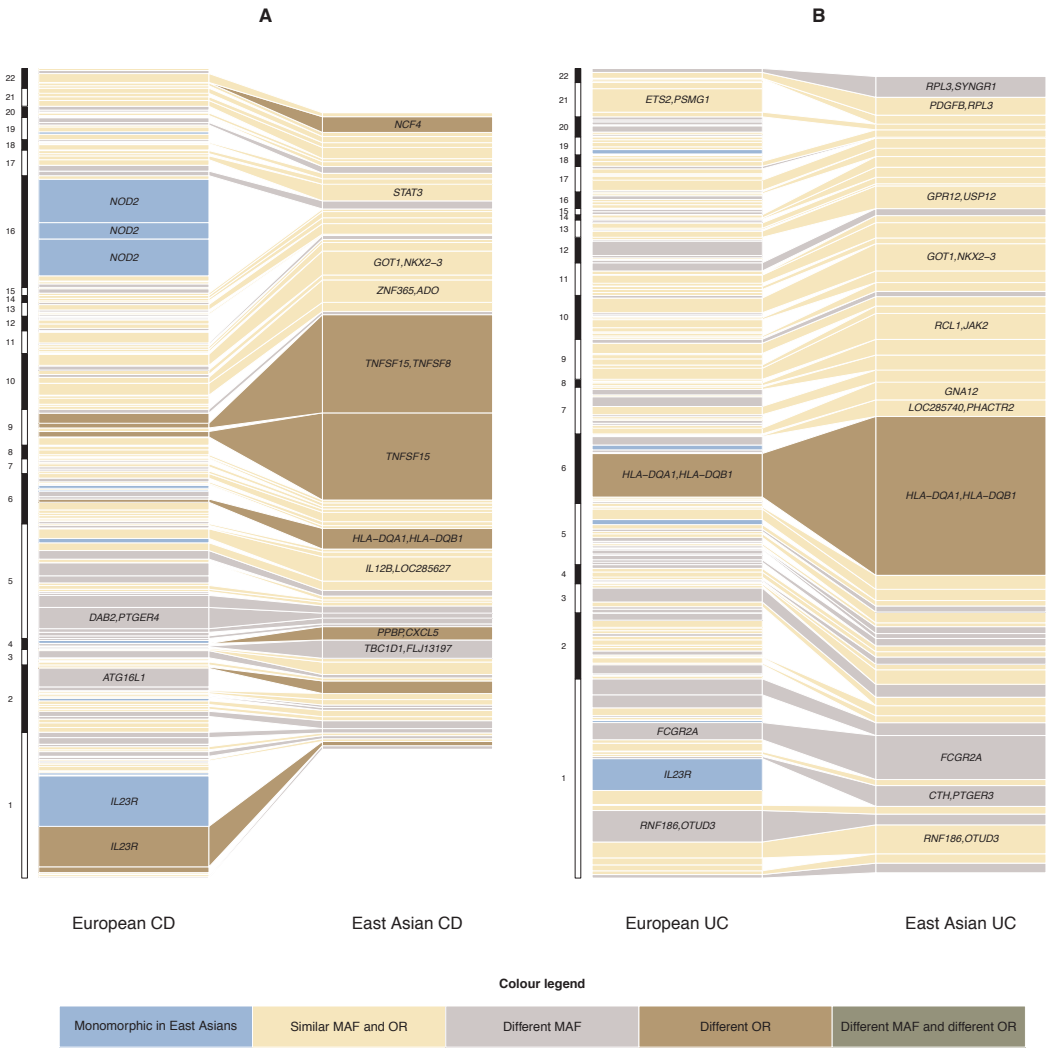


Figure 4.6. Belgravia plot of (A) CD and (B) UC risk variants in Europeans and East Asians. Each box represents an independent association for each disease. The East Asian panel only contains SNPs with association P < 0.01. The size of the box is proportional to the amount of variance explained in disease risk (liability scale) for that variant. The colours of the boxes

represent whether any difference in variance is due to differences in allele frequencies (Fst > 0.1), odds ratios ($P < 2.5\times10^{-4}$) or both.

Inflammatory cytokines may play a more important role in East Asian CD, with the greatest variance in disease risk explained for any IBD risk variant observed at the *TNFSF15*/*TNFSF8* (tumour necrosis factor superfamily 15/8) locus. Compared with the modest effect sizes in Europeans, two of the three independent signals at *TNFSF15*/*TNFSF8* showed much larger effects in East Asians: rs4246905 (European OR = 1.14 [95%CI: 1.11-1.18], East Asian OR = 1.73 [1.57-1.91], $P_{het} = 5.91\times10^{-15}$) and rs13300483 (European OR = 1.14 [1.11-1.17], East Asian OR = 1.70 [1.57-1.84], $P_{het} = 1.98\times10^{-19}$) despite similar allele frequencies. The third variant was non-significant in East Asians (rs11554257, P = 0.21).

An experiment testing the effect size of these variants in East Asian CD cases and controls who are >2nd generation immigrants in Western countries will help disentangle the role of environment. If differences still persist, this raises the intriguing possibility that genetic factors are the cause of this heterogeneity. Alternative explanations include gene-gene interactions with other population-specific variants, or that these differences are explained by as-yet undetermined causal variant(s) that may reflect different patterns of LD with the reported SNPs. It is not possible to rule out this hypothesis using the data in this study. Although the Immunochip provides dense coverage at 186 loci with known associations to at least one immune-mediated disease, the selection of SNPs was based on low-coverage sequence data from the pilot release of the 1000 Genomes Project and only incorporates variants identified in the CEU (European ancestry) cohort. Approximately 240,000 SNPs were selected for inclusion with and array design success rate of 80%. A further ~30% of SNPs were also excluded during QC. Therefore, it remains possible that the causal variants remained untyped, and the chances of this occurring are greater in the populations of non-European ancestry. Until the causal variants that underlie these associated loci have been identified (or all SNPs within these loci are included in association tests) the possibility that differential tagging of untyped causal variants are driving this heterogeneity of effect cannot be ruled out.

4.3.5 Gene-based likelihood ratio test

In the previous section, I discussed how the small sample size of the non-European cohorts limits our ability to estimate the effect of known IBD risk loci in these populations. In loci where there a multiple known independent signals in Europeans, it may be possible to use this prior information and test whether the aggregate of these signals show significant associations in non-European populations. Gene-based aggregate approaches for common variants are potentially more powerful than single-SNP approaches for situations where multiple SNPs within a gene are independently associated, and also due to a less stringent gene-wide P-value threshold (Neale and Sham, 2004). By only aggregating SNPs with prior evidence of association in the European cohort, this approach may also have greater power than traditional gene-based tests for common variants that consider all SNPs within a gene (Liu *et al.,* 2010a; Huang *et al.,* 2011). To do this, I first identified loci with multiple independent associations in Europeans, and then modelled these SNPs jointly within each gene in each of the non-European populations. Significance of the model was tested using a likelihood ratio test.

Genes were first selected if they have transcript start/stop boundaries (±50 kb) that overlap the most associated SNP in each locus and were located within the Immunochip high-density regions. Within each gene, independent associations were identified using the conditional and joint multi-SNP model selection approach implemented in GCTA (Yang *et al.,* 2012). I applied this to European Immunochip chip samples within each of the three phenotypes: CD UC and IBD, and identified 111 genes with more than one independent signal. When considering the overlap between genes (a SNP may be assigned to multiple genes), this corresponds to 41 non-overlapping loci. Performing the likelihood ratio tests on SNPs in these loci in the non-European samples revealed nine loci with significant evidence of association ($P < 5\times10^{-5}$). At six of these loci, the P-value form the likelihood ratio test was smaller than the smallest univariate SNP P-value in the non-European cohort. Nevertheless, this power improvement is only marginal, as with the exception of the *TNFSF15*/*TNFSF8* locus, significance

of the likelihood ratio test never exceeded the univariate SNP P-value by more than one order of magnitude.

Gene	Chr.	Gene start	Gene stop	Pop	Pheno	SNPs	Gene P-value	Best SNP P-value	Locus number
TMCO4	1	19.83	20.05	IND	IBD	3	3.37×10^{-5}	8.33×10^{-5}	1
TMCO4	1	19.83	20.05	EAS	UC	3	3.18×10^{-7}	2.36×10^{-6}	1
TMCO4	1	19.83	20.05	IND	UC	3	2.97×10^{-5}	2.95×10^{-4}	1
RNF186	1	19.96	20.06	IND	IBD	3	3.37×10^{-5}	8.33×10^{-5}	1
RNF186	1	19.96	20.06	EAS	UC	3	3.18×10^{-7}	2.36×10^{-6}	1
RNF186	1	19.96	20.06	IND	UC	3	2.97×10^{-5}	2.95×10^{-4}	1
FCGR2A	1	159.69	159.81	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
HSPA6	1	159.71	159.81	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
FCGR3A	1	159.73	159.84	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
IL10	1	204.96	205.06	EAS	UC	2	1.22×10^{-8}	5.72×10^{-7}	3
IL19	1	204.99	205.13	EAS	UC	2	1.22×10^{-8}	5.72×10^{-7}	3
IL18RAP	2	102.35	102.49	EAS	IBD	3	4.81×10^{-6}	9.09×10^{-7}	4
MIR4772	2	102.37	102.47	EAS	IBD	2	1.12×10^{-6}	9.09×10^{-7}	4
SLC9A4	2	102.41	102.57	EAS	IBD	2	1.12×10^{-6}	9.09×10^{-7}	4
LOC285626	5	158.64	158.77	EAS	CD	3	8.46×10^{-10}	3.46×10^{-10}	5
LOC285626	5	158.64	158.77	EAS	IBD	3	1.03×10^{-9}	3.70×10^{-9}	5
LOC285627	5	158.76	158.88	EAS	CD	2	6.01×10^{-10}	3.46×10^{-10}	5
LOC285627	5	158.76	158.88	EAS	IBD	2	1.35×10^{-9}	3.70×10^{-9}	5
TNFSF15	9	116.54	116.66	EAS	CD	2	2.80×10^{-49}	2.83×10^{-45}	6
TNFSF15	9	116.54	116.66	EAS	IBD	3	1.40×10^{-30}	1.65×10^{-30}	6
TNFSF8	9	116.65	116.78	EAS	IBD	2	3.08×10^{-19}	1.13×10^{-17}	6
DKFZP434A062	9	138.29	138.39	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
DKFZP434A062	9	138.29	138.39	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
GPSM1	9	138.29	138.42	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
GPSM1	9	138.29	138.42	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
DNLZ	9	138.33	138.43	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
DNLZ	9	138.33	138.43	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
CARD9	9	138.33	138.44	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
CARD9	9	138.33	138.44	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
SNAPC4	9	138.34	138.46	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
SNAPC4	9	138.34	138.46	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
SDCCAG3	9	138.37	138.47	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
SDCCAG3	9	138.37	138.47	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
PMPCA	9	138.37	138.49	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
PMPCA	9	138.37	138.49	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
C9orf163	9	138.45	138.55	IND	IBD	3	2.52×10^{-5}	1.04×10^{-4}	7
ADO	10	64.18	64.29	EAS	CD	2	9.13×10^{-8}	4.24×10^{-9}	8
EGR2	10	64.19	64.30	EAS	CD	2	6.56×10^{-8}	3.05×10^{-9}	8
NKX2-3	10	101.23	101.34	EAS	CD	2	4.67×10^{-8}	2.00×10^{-9}	9
NKX2-3	10	101.23	101.34	EAS	IBD	2	2.45×10^{-10}	2.79×10^{-11}	9
NKX2-3	10	101.23	101.34	EAS	UC	2	3.89×10^{-5}	3.07×10^{-6}	9

Table 4.7. Genes that exceeded P < 5×10-5 in at least one non-European cohort in the likelihood ratio locus-based test.

The likelihood ratio approach described here is similar to polygenic risk modelling, a commonly used method for identifying pleiotropy between a pair of phenotypes in genotyped individuals (International Schizophrenia Consortium *et al.,* 2009). Here, rather than comparing two phenotypes, I compared the same

phenotype in two populations. In polygenic risk modelling, the effect sizes for a set of SNPs (for example, those with association $P < 5\times10^{-8}$) are first estimated for one phenotype, and then used to construct risk scores based on genotypes for each individual in a second trait from a non-overlapping population. The degree to which these risk scores are correlated with phenotype in this second population are then assessed via linear regression (or logistic regression for dichotomous traits), where the size of the pleiotropic effect and its significance can be estimated.

It is possible to apply the polygenic risk score method to this study, where for a given gene, effect sizes estimated in Europeans are used to generate risk scores in a non-European cohort. However, this type of analysis assumes that LD patterns between the two cohorts tested are identical (or the SNPs being tested are in linkage equilibrium in both populations), which is often not the case when comparing divergent populations. Significant independent SNPs estimated in one population may be correlated with each other in another population, making the true pleiotropic effect difficult to interpret. The likelihood ratio testing approach overcomes this potential bias due to LD by only considering independent signals in the European cohort, and then re-estimating their effects jointly in the non-European cohort. These joint effect sizes will reflect the patterns of LD. Indeed, in situations where LD patterns and allele frequencies are identical between the two cohorts, the likelihood ratio method and the polygenic risk score should provide almost identical results. Of course, neither method is suitable in situations where there are heterogeneous effects exist between the two populations.

4.3.6 Conclusions

In this, the largest trans-ethnic study of IBD in 96,856 individuals of European, East Asian, Indian and Iranian populations, 40 newly associated risk loci were identified, bringing the total number of IBD risk loci to 203. The large number of risk loci shared between populations and high genetic correlations also suggests that the underlying causal variants are common (allele frequencies > 5%), thus adding further weight to the growing number of arguments against the synthetic

association model for explaining common variant associations (Dickson *et al.,* 2010; Anderson *et al.,* 2011b; Wray *et al.,* 2011).

The population comparisons at known IBD risk loci also identified several associated loci that are population specific. For instance, variants in *NOD2* and *IL23R* with major effects in Europeans are monomorphic in East Asians. Given the smaller sample size of the non European cohorts, and that Immunochip SNP selection was based on resequencing data from individuals of European ancestry, there was little power in this study to identify variants that are monomorphic in Europeans but are associated in non-Europeans. Other loci polymorphic across populations also showed evidence for differences in effect size (for instance, *TNFSF15* in Europeans and East Asians; $P_{het} = 1.98\times10^{-19}$). Loci with large differences in effect size raises the intriguing possibility of gene-environment interactions, though the presence of untyped causal alleles cannot be ruled out.

The newly identified loci along with the concordance in directions of effect between populations demonstrates that trans-ethnic association studies are a powerful means of identifying novel risk loci in complex diseases such as IBD. By leveraging imputation based on tens of thousand of reference haplotypes, or directly sequencing large numbers of cases and controls, these studies will more thoroughly survey causal variants and thus have increased ability to model the genetic architecture of IBD across diverse ancestral populations.

Chapter 5. Immune-mediated disease risk loci are enriched for differentially expressed genes from tissue-relevant functional genomic datasets

5.1 Introduction

Identifying the causal variants that are tagged by complex disease risk loci remains challenging. Blocks of linkage disequilibrium often contain multiple correlated association signals that are statistically indistinguishable from each other, and can span dozens of genes with multiple functional candidates. It is clear that the majority of common risk variants do not reside in protein coding regions (Hindorff et al.), suggesting that important aspects of disease etiology are driven by gene expression. While identifying specific causal variants is difficult, approaches that integrate GWAS association results with disease relevant functional genomic datasets may help in narrowing down potential candidate genes and the cell types in which they act.

Expression quantitative trait loci (eQTLs) provide a direct bridge between GWAS and gene expression. These studies measure gene expression across many individuals (typically in a genome-wide approach using microarrays or RNAseq), and then treat the expression level of each gene as a separate quantitative trait to test for association with SNPs – either at the same locus (cis-eQTLs) or genome-wide (trans-eQTLs). Loci that are associated with both gene expression and disease risk implicate particular genes as potential biologically relevant candidates. A limitation of eQTL studies is difficulty in obtaining large sample

sizes in relevant tissues. The largest eQTL studies in over 1000 individuals have generally focused on easy-to-obtain tissue such as heterogeneous cell types within peripheral blood (Hemani *et al.,* 2014; Westra *et al.,* 2013), while smaller studies (typically with sample sizes in the hundreds) have been performed in cell types such as lymphoblastoid cell lines (LCLs), monocytes (Fairfax *et al.,* 2014), dendritic cells (Lee *et al.,* 2014) and heterogeneous tissues such as liver, adipose tissue, skin and brain (Gibbs *et al.,* 2010; Grundberg *et al.,* 2012; Schadt *et al.,* 2008). Despite having identified hundreds of eQTLs, the majority of the heritability of gene expression remains to be uncovered, much like the case with complex disease risk loci. For instance, in a large eQTL study of LCLs, adipose tissue and skin in 856 twins, the reported cis-eQTLs explain on average only 9-12% of the total genetic variance at each gene (Grundberg *et al.,* 2012). Nevertheless, these studies are an invaluable tool for interpreting the findings from GWAS. Indeed, in Chapters 2-4, eQTL datasets were used to prioritise candidate genes at PBC, PSC and IBD risk loci.

Enrichment analysis provides a complementary approach to linking GWAS risk loci with gene expression. These types of analyses ask whether disease risk loci are found disproportionately more often overlapping certain genomic annotations (for example, coding variants, UTRs, or epigenetic marks) than by chance. For instance, GWAS loci across a range of phenotypes appear to be enriched for known eQTLs (Nicolae *et al.,* 2010). Under the further assumption that disease loci act in only a small number of cell types and under certain cell states, questions about the relative importance of specific cells and disease states in disease pathogenesis can also be studied using the enrichment approach. These studies have an advantage over eQTL studies in that genomic annotations can be generated from only a small number of individuals. Such enrichment studies of gene regulatory annotations or genes that are expressed in specific cell types are now common place in the literature (Cowper-Sallari *et al.,* 2012; Ernst *et al.,* 2011; Hu et al.; Liu *et al.,* 2012; Maurano *et al.,* 2012; Trynka *et al.,* 2013).

An important consideration in these types of approaches is the estimation of the null distribution – what amount of overlap, given the number risk loci and

frequency of genomic annotations, is expected just by chance? It is incorrect to assume that functional annotations and risk loci are both randomly distributed across the genome – both are more likely to be found nearer to genes than away from them (Hindorff et al.). Hence it is possible that sets of risk loci associated with any number of traits will be enriched for functional elements purely because of their colocalisation around genes rather than their functional relevance. For this reason, parametric approaches assuming independence or permutation approaches that randomly resample SNPs (while not accounting for LD) or switch case/control labels to construct "null" GWAS datasets may be upwardly biased in their enrichment estimation.

In this study, I combined GWAS results for four immune-mediated and two non-immune related quantitative traits with two differential expression datasets that are relevant to intestinal inflammatory diseases (e.g. Crohn's disease, ulcerative colitis and coeliac disease). The first dataset consists of a gene expression experiment of four intestinal T cell populations and their blood counterparts in healthy individuals (Raine *et al.,* 2014). T cells are the dominant population of immunocytes in the gastrointestinal tract, and display distinct characteristics in their cell surface marker expression, activation pathways and function compared with the blood counterparts. The expression of genes that drive these differences and maintain intestinal homeostasis may be prime candidates to also modulate risk immune-mediated diseases of the gastrointestinal tract.

The second dataset consists of differentially expressed transcripts in mice following infection with the whipworm *Trichuris muris*. Gene expression levels were measured in infected and uninfected populations of heterogeneous cells in cecum tissue (Foth *et al.,* 2014). High dose infections of *T. muris* in mice typically generates a T_H2 response characterised by eosinophil activation, macrophage inhibition and the production of antibodies, such that immunity is acquired. Low dose infection generates a T_H1 response, characterised by macrophage activation other cellular immunity response, ultimately leading to chronic infection. These low dose infections have been used to model the response in humans to infection

by *Trichuris trichuira*, which exhibit striking phenotypic similarities to IBD (Levison *et al.,* 2013; Levison *et al.,* 2010). Early exposure to whipworms in humans is also thought to be protective against IBD, and the hygiene hypothesis suggests that a lack of exposure to pathogens has contributed to the increasing incidences of immune-mediated disorders in developed countries (Elliot *et al.,* 2000; Okada *et al.,* 2010). Furthermore, there is some evidence that by triggering an immune response, whipworms are an effective treatment for IBD (Croese *et al.,* 2006; Summers *et al.,* 2005a; Summers *et al.,* 2005b). For these reasons, if genes that are differentially expressed upon infection are enriched in risk-loci for IBD and other immune-mediated diseases, they may be excellent candidates through which disease is mediated.

5.1.1 Contributions

Generation of gene expression datasets and identification of differential expressed genes were performed by Tim Raine, Adam Reid and others, and are described in Raine *et al.* (2014) and Foth *et al.* (2014). All other analyses were performed by myself.

5.2 Methods

5.2.1 Human T cell transcripts

Differential gene expression data were obtained from Raine *et al.* (2014). Briefly, six healthy subjects underwent biopsy collection at the terminal ileum. These samples were sorted using fluorescence activated cell sorting (FACS), and total RNA from four major T effector memory cell populations isolated: $CD4^+$ and $CD8^+$ expressing intraepithelial lymphocytes (IELs), and $CD4^+$ and $CD8^+$ expressing lamina propria lymphocytes (LPLs). Paired reference $CD4^+$ and $CD8^+$ T cells from the peripheral blood were also isolated. Gene expression was measured using the Affymetrix Gene ST 1.0 microarrays. After QC filtering, expression of 9,468 transcripts that passed in all six cell populations were obtained. Differential expression was analysed pairwise with each gut T cell population paired with its corresponding peripheral blood population taken from the same individual

($CD4^+$ IEL vs. $CD4^+$ blood, $CD4^+$ LPL vs. $CD4^+$ blood, $CD8^+$ IEL vs. $CD8^+$ blood, and $CD8^+$ LPL vs. $CD8^+$ blood). Transcripts that were significantly up-or-down-regulated in either IEL or LPLs vs. blood were taken forward for enrichment analysis.

5.2.2 Mouse cecum transcripts

Differential expression data were obtained from Foth et al. (2014). Briefly, 14 male C57BL/6 were infected with a low dose of *T. muris* (25 eggs by oral gavage) at 6-8 weeks of age. The section of the cecum where the worms reside and those without infection were extracted. Transcriptome libraries for RNA-seq were created following standard Illumina protocols and sequencing was performed on Illumina HiSeq 2000 machines. The number of reads per gene was calculated by summing over all transcripts that map to the gene. Genes that showed differential expression between the infected cases and uninfected controls were estimated at a false discovery rate of 5% using DESeq (Anders and Huber, 2010). Only protein coding genes and those with a unique human orthologue were included for downstream analysis. After filtering, 15,278 genes remained.

5.2.3 GWAS enrichment

The SNP with the strongest association signal (the lead SNP) in each of the associated loci (reported at $P < 5\times10^{-8}$) from the largest published genome-wide association studies (GWAS) were extracted for four immune-mediated complex diseases: Crohn's disease (CD), ulcerative colitis (UC), celiac disease (CeD) and type 1 diabetes (T1D) (Barrett *et al.,* 2009; Jostins *et al.,* 2012; Trynka *et al.,* 2011b), as well as two complex traits: height and body mass index (BMI) (Lango Allen *et al.,* 2010; Speliotes *et al.,* 2010). The two complex traits are unlikely to be strongly influenced by immune-related genes and were included as effective negative controls for the method. For each lead SNP, an associated locus was defined as the genomic region spanning a 0.2cM window either side of the lead SNP, estimated from HapMap Phase II genotypes (The International HapMap Consortium 2007). Where SNPs showed overlapping windows, only the window assigned to the SNP with the most significant p-value was considered.

For each differentially expressed gene, I defined its gene-region spanning ±50kb window from the gene's transcription start/stop site. To account for potential non-random clustering of genes with similar expression patterns and function (Hurst *et al.,* 2004), groups of differentially expressed genes that have overlapping windows were combined into a single window.

For each GWAS phenotype, the number of times a risk locus overlaps with at least one differentially expressed gene-window was counted. To assess the statistical significance of this overlap, I randomly sampled the same number of differentially express genes from the full list of expressed genes. If a sampled gene has a ±50kb window overlapping that of another previously sampled gene, then the windows are merged and these genes are only counted once. I then calculated the number of associated loci that overlap at least one of these randomly sampled lists of genes. The sampling process was repeated 100,000 times for each disease/trait, and the empirical p-value was the number times the overlap with the randomly sampled genes exceeds the overlap with the observed differentially expressed genes, divided by 100,000.

5.3 Results

5.3.1 Human T cell transcripts

Using a 1.4-fold change (adjusted $P < 0.05$), 246, 275, 115 and 142 genes were identified to be upregulated in LPL $CD4^+$, LPL $CD8^+$, IEL $CD4^+$ and IEL $CD8^+$ T cells respectively compared with their counterparts in the blood. Using a P-value cut-off of $P < 2\times10^{-3}$ (equivalent to a 5% Bonferroni correction for 24 tests), a significant enrichment among T1D risk loci were identified for genes upregulated in LPL $CD4^+$ ($P = 10^{-5}$) and LPL $CD8^+$ cells ($P = 10^{-5}$), with 17 and 18 respectively of the 54 associated risk loci overlapping at least one upregulated gene. Strong suggestive evidence for enrichment was also identified for upregulated genes in LPL $CD4^+$ cells in CD ($P = 0.0053$) and CeD ($P = 0.0045$), LPL $CD8^+$ cells in CD ($P = 0.0038$), IEL $CD4^+$ cells in T1D ($P = 0.0053$) and IEL CD8 cells in T1D ($P = 0.001$) (Table 5.1). Only modest levels of enrichment were identified in for LPL T cells in UC ($P = 0.037, 0.029$), almost all of which is driven

by UC risk loci that are also associated with CD (Table 5.3). The lack of enrichment in UC may reflect that fact that inflammation occurs in the colon, while the experiments described here were on cells extracted from small bowel biopsies.

		LPL upregulated vs. blood				IEL upregulated vs. blood			
		CD4+ (246)		CD8+ (275)		CD4+ (115)		CD8+ (142)	
Phenotype	Risk loci	Overlap	P	Overlap	P	Overlap	P	Overlap	P
Crohn's disease	140	23	0.0053	25	0.0038	7	0.56	10	0.33
Ulcerative colitis	133	21	0.0368	23	0.0291	5	0.88	8	0.69
Celiac disease	38	10	0.0045	10	0.0104	5	0.0161	5	0.0841
Type 1 diabetes	54	17	10^{-5}	18	10^{-5}	7	0.0053	10	0.0010
Body mass index	73	2	0.98	3	0.94	5	0.55	6	0.044
Height	192	21	0.87	18	0.99	5	0.98	7	0.99

Table 5.1. Enrichment of genes that are upregulated in gut T cells compared with blood T cells in loci associated with six phenotypes. The numbers in parentheses next to each cell type is the number of upregulated genes in that gut cell type vs. its equivalent in blood.

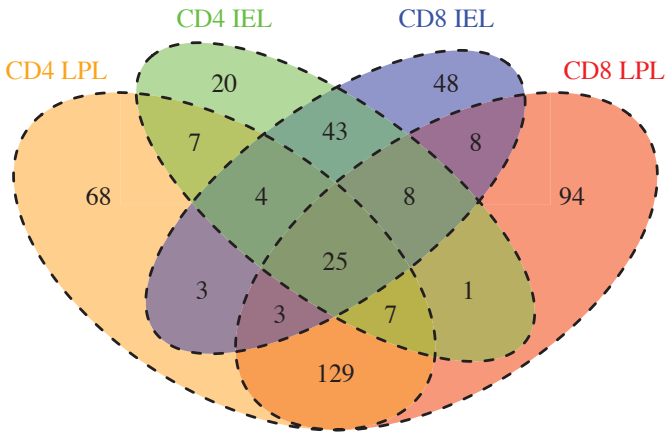


Figure 5.1. Number of upregulated genes that overlap among CD4+ LPL, CD8+ LPL, CD4+ IEL and CD8+ IEL T cells vs. counterparts in blood.

Genes that were downregulated in LPL or IEL T cells compared with their blood counterparts were also tested for enrichment, though no evidence was found for any of the phenotypes ($P > 0.01$) (data not shown). As expected, height and BMI also showed no evidence for enrichment for any of the gene sets tested. These two traits were selected as they include a similar number of associated loci as the immune-mediated diseases tested and, given that immune-related processes are unlikely to play a strong role in these traits, any enrichment

observed in these traits may have been the result of biases in the method that were unaccounted for.

5.3.2 Mouse cecum transcripts

After filtering, 824 genes showed evidence for differential expression (FDR = 5%) between infected and uninfected cecum tissue in C57BL/6 mice. A unique human ortholog was taken forward for 454 of these genes. Significant evidence for enrichment of differentially expressed genes and GWAS risk loci were found for all four immune-related diseases ($P < 0.0024$), the strongest of which were seen in Crohn's disease ($P = 2.0\times10^{-4}$) and ulcerative colitis ($P = 5.7\times10^{-4}$). As with the case for the IEL and LPL T cells, no evidence for enrichment was identified across height or BMI associated loci.

Phenotype	Risk loci	Overlap	P
Crohn's disease	140	34	2.0×10^{-4}
Ulcerative colitis	133	33	6.7×10^{-4}
Celiac disease	38	11	0.0012
Type 1 diabetes	54	15	0.0024
Body mass index	73	6	0.33
Height	192	23	0.52

Table 5.2. Enrichment of genes that are differentially expressed between infected and uninfected cecum tissue among loci associated with six phenotypes.

		CD4+ LPL				CD8+ LPL				CD4+ IEL		CD8+ IEL	Cecum			
Chr.	BP window (Mb)	CD	UC	CeD	T1D	CD	UC	CeD	T1D	CeD	T1D	T1D	CD	UC	CeD	T1D
1	7.78 - 8.27					*ERRFI1*	*ERRFI1*									
1	25.18 - 25.32			*RUNX3*				*RUNX3*								
1	67.90 - 67.9	*IL12RB2, IL23R*	*IL12RB2, IL23R*										*IL12RB2*	*IL12RB2*		
1	77.91 - 78.99	*DNAJB4*				*DNAJB4*										
1	113.82 - 114.62	*PTPN22*			*PTPN22*	*PTPN22*			*PTPN22*		*BCL2L15*	*PTPN22*	*BCL2L15, PTPN22*			*BCL2L15, PTPN22*
1	154.97 - 156.21	*LMNA*	*LMNA*			*LMNA*	*LMNA*						*FDPS*	*FDPS*		
1	160.68 - 161.15												*LY9, SLAMF7*	*LY9, SLAMF7*		
1	172.46 - 172.94									*FASLG*						
1	192.46 - 192.55			*RGS1*	*RGS1*			*RGS1*	*RGS1*	*RGS1*	*RGS1*	*RGS1*				
1	197.19 - 197.94					*DENND1B*	*DENND1B*									
1	206.79 - 207.04					*IL10*	*IL10*		*IL10*							
2	28.58 - 28.67	*FOSL2*	*FOSL2*			*FOSL2*	*FOSL2*									
2	43.45 - 44.05					*ZFP36L2*	*ZFP36L2*									
2	60.78 - 62.12	*REL*	*REL*	*REL*		*REL*	*REL*	*REL*								
2	68.51 - 68.89														*PLEK*	
2	102.69 - 103.27	*IL1R1*	*IL1R1*	*IL1R1*									*IL18RAP*	*IL18RAP*	*IL18RAP*	
2	191.81 - 192.01												*STAT1, STAT4*	*STAT1, STAT4*	*STAT1, STAT4*	*STAT1, STAT4*
2	198.14 - 199.11						*COQ10B*									
2	204.17 - 204.82			*CTLA4*	*CTLA4*			*ICOS*	*ICOS*						*ICOS, CTLA4*	*ICOS, CTLA4*
2	219.30 - 219.3												*CXCR2, SLC11A1*	*CXCR2, SLC11A1*		
2	241.70 - 241.7	*GPR35*	*GPR35*													
3	45.89 - 46.71			*LZTFL1*	*LZTFL1*			*LZTFL1*	*LZTFL1*	*LZTFL1*	*LZTFL1*	*CCR9*			*CCRL2*	*CCRL2*
3	48.17 - 51.83												*BSN, GMPPB, SHISA5, AMT*	*BSN, GMPPB, SHISA5, AMT*		
4	103.37 - 104.25		*NFKB1*													
4	122.90 - 123.77	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*				
5	0.40 - 0.79		*TPPP*													
5	40.18 - 40.98	*PTGER4*	*PTGER4*			*PTGER4*	*PTGER4*									
5	72.38 - 72.59												*TMEM171*			
5	131.28 - 132.14					*KIF3A*	*KIF3A*						*SLC22A4*	*SLC22A4*		
5	150.40 - 150.4	*C5orf62*	*C5orf62*			*DCTN4*	*DCTN4*									
6	0.36 - 0.47			*IRF4*				*IRF4*								
6	127.79 - 128.34												*THEMIS*		*THEMIS*	
6	137.81 - 138.29	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*					
6	167.34 - 167.55	*CCR6*	*CCR6*			*CCR6*	*CCR6*									
7	26.63 - 27.22	*SKAP2*			*SKAP2*											
7	50.36 - 50.75															*DDC*
7	98.71 - 99.37					*ZNF394*	*ZNF394*									
7	107.60 - 107.6													*SLC26A3*		
7	116.78 - 117.45												*CFTR, CTTNBP2*	*CFTR, CTTNBP2*		
7	128.55 - 128.82													*IRF5*		
8	126.44 - 126.63	*TRIB1*	*TRIB1*													
9	93.86 - 94.17												*NFIL3*	*NFIL3*		
10	6.00 - 6.18	*IL2RA*	*IL2RA*		*IL2RA, PFKFB3*	*IL2RA*	*IL2RA*		*IL2RA, PFKFB3*							
10	30.67 - 30.83	*MAP3K8*	*MAP3K8*			*MAP3K8*	*MAP3K8*									
10	35.10 - 35.97	*CREM*	*CREM*			*CREM*	*CREM*									
10	64.30 - 64.76	*EGR2*	*EGR2*			*EGR2*	*EGR2*									
10	80.94 - 81.15	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*							
11	60.55 - 60.97												*SLC15A3, CD6*	*SLC15A3, CD6*		
11	61.37 - 61.76	*FTH1*	*FTH1*			*FTH1*	*FTH1*						*FEN1*	*FEN1*		
11	63.81 - 64.58												*STIP1, FERMT3*	*STIP1, FERMT3*		
11	65.13 - 66.08												*LTBP3, CST6, CTSW*	*LTBP3, CST6, CTSW*		
11	95.97 - 96.47						*MAML2*									
11	118.90 - 118.9												*HYOU1*	*HYOU1*	*PHLDB1*	
12	9.47 - 10.02				*KLRB1, CLEC2B*				*CD69*		*CD69*	*CD69*				
12	12.53 - 12.73	*DUSP16*	*DUSP16*			*DUSP16*	*DUSP16*									
12	56.23 - 56.84				*RPL41*				*IL23A*							
12	57.75 - 58.53				*ARHGAP9*				*ARHGAP9*		*ARHGAP9*	*ARHGAP9*				
12	68.32 - 68.63	*IFNG*	*IFNG*			*IFNG*	*IFNG*									
12	111.54 - 113.13											*C12orf51*				
13	99.61 - 100.11	*GPR183*	*GPR183*		*GPR183*	*GPR18*	*GPR18*		*GPR18*			*GPR18*	*GPR18*	*GPR18*		
14	69.14 - 69.36					*ZFP36L1*	*ZFP36L1*	*ZFP36L1*	*ZFP36L1*							
14	75.42 - 75.75	*FOS*	*FOS*			*FOS*	*FOS*									
14	88.19 - 88.73					*GPR65*	*GPR65*						*GPR65*	*GPR65*		
15	74.55 - 75.94														*SEMA7A*	
15	78.91 - 79.26				*MORF4L1*				*MORF4L1*							*CTSH*
16	11.00 - 11.0												*SOCS1*	*SOCS1*	*CIITA, SOCS1*	*CIITA, SOCS1*
16	28.28 - 29.03											*APOBR*				
16	29.89 - 31.36													*CORO1A*		
16	75.03 - 75.53											*ZNRF1*				
16	85.96 - 86.04												*IRF8*	*IRF8*		
17	32.70 - 32.7												*CCL2, CCL7*	*CCL2, CCL7*		
17	37.35 - 38.25												*NR1D1*	*NR1D1*		*NR1D1*
17	40.29 - 41.05	*STAT3*	*STAT3*										*WNK4*	*WNK4*		
17	58.20 - 58.2					*MIR21*	*MIR21*									
18	46.34 - 46.51					*SMAD7*	*SMAD7*									
19	10.37 - 10.63	*ICAM1*	*ICAM1*		*ICAM1*	*ICAM1*	*ICAM1*		*ICAM1*				*ICAM1*	*ICAM1*		*ICAM1*
19	47.14 - 47.33				*SLC1A5*											
20	44.34 - 44.82												*CD40*	*CD40*		
20	62.18 - 62.49												*PTK6*	*PTK6*		
22	21.71 - 22.27												*CCDC116, SDF2L1*	*CCDC116, SDF2L1*	*CCDC116, SDF2L1*	
22	29.81 - 30.87					*SF3A1*	*SF3A1*		*SF3A1*				*UQCR10*	*UQCR10*		*UQCR10*
22	37.50 - 37.57				*IL2RB*				*IL2RB*							*CYTH4, IL2RB, RAC2*

Table 5.3. Annotation of disease-associated loci that are show nominal levels of enrichment ($P < 0.05$) for genes that show differential expression in healthy gut vs. blood T cells and in infected vs. uninfected mouse cecum tissue. The BP (base pair) window denotes a ±0.2cM around an associated SNP. Windows that overlap were combined into a single window.

5.4 Discussion

The broad patterns of enrichment among disease risk loci and genes expressed in both healthy and in inflamed tissues points to the importance of multiple biological pathways involved in disease risk. The lack of overlap between the expression of genes upregulated in healthy human T cell populations and infected/uninfected mouse cecum samples (Table 5.3) reflects both the different cell composition of the samples and biological processes involved in maintaining homeostasis and responses to infection. That differentially expressed genes in T cells from the gut compared with those from peripheral blood appear to play a role in disease risk serves as an important reminder of the limitations of inferring biology from easily accessible blood cell types. Ideally, further understanding of how gene expression modulates disease risk will involve efforts that combine expression patterns multiple immune cell types under both healthy conditions and disease states.

A major utility of gene expression experiments in relevant tissue types is to identify potential candidate genes among GWAS risk loci. Many of the candidate genes listed here (Table 5.3) were also implicated in other *in silico* approaches reported in the original locus discovery projects. For instance the IBD associated SNP rs1819333 lies 160kb upstream of *CCR6*, a gene that is upregulated $CD4^+$ and $CD8^+$ LPL T cells. *CCR6* is an important regulator of lymphocyte homeostasis in the mucosa (Cook *et al.,* 2000), and was implicated as a candidate gene through the text-mining-based GRAIL network analysis in the original IBD GWAS (Jostins *et al.,* 2012; Raychaudhuri *et al.,* 2009). Similarly, at the IBD associated SNP rs11209026, *IL12RB* was differentially expressed in both $CD4^+$ LPL T cells and cecum tissue. This gene was also implicated in the original IBD GWAS via DAPPLE, a method identifies candidate genes based on reported protein interaction networks (Rossin *et al.,* 2011).

At other loci, the approach also offers new leads at loci with no obvious candidate gene, or alternative candidate genes to those previously proposed. For instance, at the IBD-associated SNP rs35675666, GRAIL analysis originally

suggested *TNFRSF9* as the sole candidate gene at this locus. Here, another nearby gene, *ERRFI1*, was highly expressed in $CD8^+$ LPL T cells. *ERRFI1* belongs to a family of epidermal growth factor receptors that share a common signal transduction pathway through ERK-MAPK with the T cell receptor. This growth factor-mediated signalling has been suggested to modulate intestinal T cell regulation in a murine colitis model (Zaiss *et al.*, 2013), highlighting *ERRFI1* as an alternative candidate gene at this locus. Similarly, at rs17391694, the nearby gene *DNAJB4* was highly expressed in LPL $CD4^+$ and $CD8^+$ T cells. No candidate genes were reported in the original IBD GWAS at this locus, partly reflecting the fact that *DNAJB4* has only recently been described.

Notably, T1D loci also appeared to be enriched for genes differentially expressed among the intestinal tissue described. Even though T1D does not manifest itself in the intestines, part of this enrichment may be a reflection of risk loci that are shared between T1D and the other intestinal diseases tested here. However, several genes residing near T1D-specific risk loci were also observed to be differentially expressed across all the experiments (Table 5.3). There is evidence to suggest that intestinal microbiota not only modulates local inflammation, but also systemic immune-mediated pathologies (Kamada *et al.,* 2013). Moreover, interactions between gut microbiota and the innate immune system have been suggested to partly modulate risk for T1D in mice (Wen *et al.,* 2008). The genes here that appear differentially expressed in populations of intestinal cell types may offer insights in the host-environment interactions across systemic immune-mediated disorders.

The method I described for estimating the degree of enrichment is in line with similar approaches that look to test whether a set of genes is overrepresented by genes from another pre-defined and biologically relevant gene set. Perhaps the most popular of these, Gene Set Enrichment Analysis (GSEA), was developed to estimate whether a set of genes identified from microarray experiments were enriched for genes involved in various biological pathways (Subramanian *et al.,* 2005). The advent of GWAS has spawned a

number of GSEA-type methods for analysing biological pathways that are enriched among GWAS risk loci (reviewed in Wang *et al.* (2010b)).

In the original GSEA approach, a set of genes is first identified and ranked (e.g. according to differential expression P-value between a set of cases and controls), and then tested to see if this rank correlates with a set of genes from another set (e.g. a particular biological pathway) via Kolmogorov-Smirnov-like statistics (Subramanian *et al.,* 2005). Significance is then assessed via permutation of the case-control status and repeating the original analysis in order to obtain a null distribution of correlations. In the context of GWAS, this approach is analogous to permuting case-control status and repeating the GWAS many times – which is both time-consuming and not possible without individual-level genotype data. GWAS adaptations to GSEA have sought to overcome this by only permuting SNP labels on summary GWAS statistics (Zhang *et al.,* 2010), however, this does not account for the correlated structure of SNPs due to LD. Furthermore, neither the phenotype-label nor SNP-label permutation approach takes into account the fact that SNPs that are associated with a complex trait are not randomly distributed throughout the genome, but are rather more likely to be found near functional elements such as genes or regulatory regions.

The approach described here tries to overcome these biases by permuting the set of differentially expressed genes rather than risk loci. While this accounts for both LD and the non-random distribution of risk loci, our method may also be biased by gene size and correlation of expression patterns of certain genes. Larger genes are more likely to overlap with an associated risk locus, such that permuting sets of genes will not be a true reflection of the null distribution. In the T cell datasets, there was modest evidence that differentially expressed genes were longer than the total set of genes tested, potentially inflating enrichment estimates (Figure 5.2 A). The opposite appeared to be the case for the cecum tissue, where the length of differentially expressed genes were shorter than expected, potentially making the test more conservative (Figure 5.2 B).

Similarly, the permutation approach will not truly estimate a null distribution in situations of gene-gene expression correlations. Genes with

coordinated expression are often clustered in areas of low recombination (Hurst *et al.,* 2004), and cis eQTLs may affect the expression of multiple nearby genes. I try to overcome this by combining genes that have overlapping windows (±50kb from the transcript start/stop sites) into a single window. Moreover, the empirical P-value is calculated on the number of risk loci that overlap at least one gene region, not the number of gene regions that overlap at least one risk locus. This distinction is subtle, but in situations where a risk locus overlaps more than one differentially expressed gene region, the test is conservative since these genes only count towards a single overlap, yet multiple genes are sampled during the permutations. Had the empirical P-value been calculated instead on the number of genes that overlap a risk locus, the empirical P-value may have been inflated as now multiple genes can potentially overlap with a single risk locus (Dixson *et al.,* 2014). Nevertheless, the approach will not account for situations where coexpressed genes lie far away from each other.

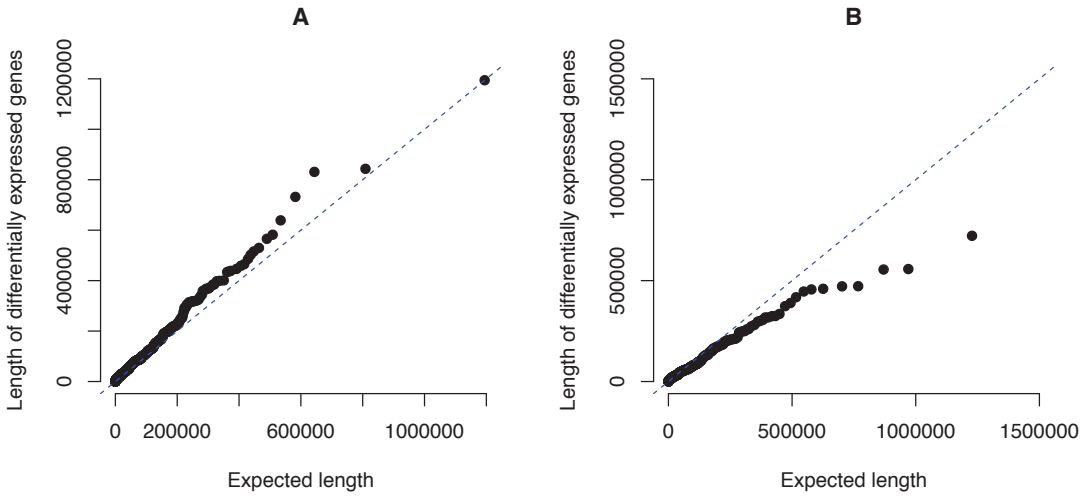


Figure 5.2. Quantile-quantile plots of gene length of differentially expressed genes in (A) gut T cells vs. blood and (B) infected vs. uninfected cecum tissue. The distribution of the expected length was the empirical distribution of all genes tested for differential expression in the respective experiments.

The choice of thresholds when defining locus and gene boundaries is often subjective. In this study, a ±0.2cM window around an associated SNP and a ±50kb window around a gene's transcript start/stop positions were used to

define whether an associated locus overlaps with a gene. The 0.2cM window describes the boundaries in which a causal variant that is tagged by an associated SNP may lie. The same window size was also used in the design of Immunochip high density regions (Tsoi *et* al., 2012 and Jostins, 2012). Similarly, the 50kb gene boundary region was chosen to adequately encompass regions where variants that affect that gene's expression may reside. This window size captures the majority (>93%) identified cis-eQTLs (Veyrieras *et al.*, 2008), though there are examples of some genes with cis-eQTLs greater than 100kb away from a transcription start site (Stranger *et al*., 2012 and Veyrieras *et al*., 2008). Larger windows may lead to more SNPs incorrectly assigned to genes, as well as a greater chance that independent loci overlap. In this study, if SNPs are incorrectly assigned to genes, power will decrease as more noise is introduced. A larger gene-boundary window will also mean that more differentially expressed genes will overlap each other and merged together. Since the resampling process cannot explicitly take this overlap into account, the results may be upwardly biased. On the other hand, using more stringent boundaries may also reduce power if truly regulatory SNPs are not assigned to its corresponding gene.

In Hu *et al.,* (2011) a similar approach looking at the overlap between gene expression in a set of immune cells and GWAS risk loci is described. Promisingly, they try to overcome the potential biases described by estimating the null distribution of enrichment by randomly selecting SNPs from a predefined, LD-pruned set of SNPs that have similar properties to disease-associated SNPs in terms of the number of genes that are located nearby. The accuracy of this approach of course depends on how this set of null SNPs is estimated, and will be more accurate for diseases where there are a large number of associated loci, such that a more representative set of null SNPs can be generated.

5.4.1 Conclusions

In summary, this study describes an approach testing whether disease risk loci are enriched for a set of functionally relevant genes. Evidence for enrichment provides additional candidate genes at associated loci, as well as generating hypotheses as to how these genes mediate disease. There was evidence for

enrichment among risk loci in four immune-mediated disorders with two differential expression datasets – the first comparing T cell subsets in healthy gut tissue with blood counterparts, and the second from samples in the cecum of mice in the presence or absence of *T. muris* infection, implicating processes in both maintaining intestinal homeostasis and response to infection in disease risk. There is a great deal of potential in these integrative approaches as a greater number of functional genomic datasets are generated for a range human tissue across multiple disease states, though care must be taken to ensure that methods employed are unbiased and statistically robust.

Chapter 6. Conclusions and future prospects

This dissertation described four distinct projects that share the common theme of unravelling the genetic basis of complex diseases. In Chapter 1, I gave a historical perspective of our understanding of the genetics of complex traits, from early 20th century efforts at reconciling Mendel's laws with the inheritance of quantitative phenotypes, to attempts throughout the 1980s to early 2000s at identifying complex disease risk loci via linkage scans, and finally to the success of GWAS from the mid-2000s up to the present day. In Chapters 2, 3 and 4, I described such locus discovery projects in PBC, PSC and IBD respectively, much of which was undertaken using the Immunochip custom genotyping array. The dense SNP content of the array has allowed for greater refinement across risk loci, while its low cost has enabled powerful locus discovery projects and cross-phenotype comparisons in very large sample sizes. Once a set of risk loci for a particular disease is found, there is also the question of what to do next. In Chapter 5, I described a simple method of combing disease risk loci with tissue-relevant functional genomic datasets in order to identify candidate genes at these risk loci, as well as potential mechanisms through which they mediate disease.

Pick up any issue of a reputable genetics journal from the past seven years and it may seem that locus discovery in complex traits is routine, if a little tedious for some. Visiting the NHGRI GWAS Catalog (Hindorff *et al.*, 2014) leaves one in no doubt, with 1,961 publications listed and 14,012 reported associated variants as of September, 2014. In the following pages, I will discuss the general lessons learnt from these types of studies, and then will look to future prospects

and challenges for locus discovery, understanding biology, and ultimately translating these findings into better treatment outcomes.

6.1 Effect sizes, power and the genetic architecture of complex traits

For genetic studies of complex traits, sample size is key. With few exceptions (e.g. HLA region in immune-mediated disorders and *NOD2* in CD), the effect of individual common genetic variants on disease risk is modest – allelic odds ratios are typically less than 1.2, and almost always less than 1.5. Robustly identifying these loci requires a combination of large sample sizes, genome-wide coverage and strict statistical criteria for determining significance – three aspects that were overlooked in early linkage and candidate genes studies. Figure 6.1 illustrates the appreciation of the need for large samples in order to robustly identify susceptibility loci – there is clear positive correlation between the number of loci discovered and the sample size of the study.

Chapters 2, 3 and 4 described the largest genetic studies to date for PBC, PSC and IBD respectively in terms of the number of samples recruited. However, the total proportion of variation in disease liability explained by these loci is still modest. Figure 6.2 illustrates the relationship between the cumulative proportion of variance explained and the strength of association. Several conclusions may be drawn from this graph. Firstly, extrapolating these curves clearly suggests that many more risk loci will be discovered as sample sizes get larger, with the total number increasing at an exponential rate (with respect to sample size) while the effect size of each individual locus will get ever smaller (Park *et al.,* 2010). These effect size distributions suggest that there are potentially thousands of susceptibility loci underlying these complex disorders.

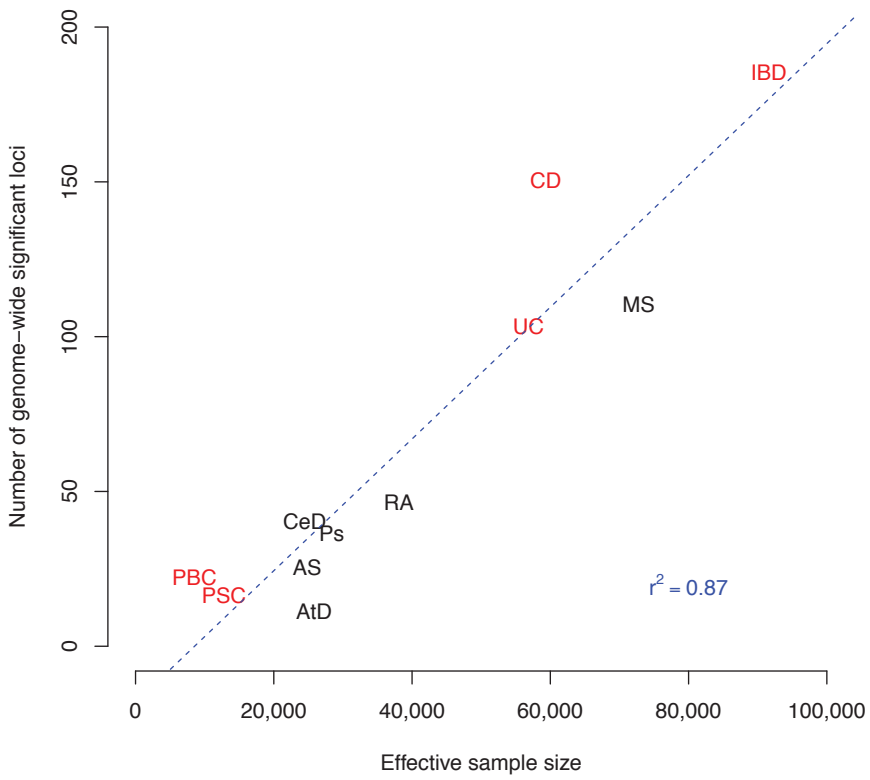


Figure 6.1. Effective sample size vs. number of genome-wide significant risk loci across GWAS and Immunochip studies of nine immune-mediated disorders. Diseases denoted in red formed Chapters 2-4 of this dissertation. The dashed blue line indicates the best fitting line estimated from least squares regression. The effective sample size denotes the cohort with an equal number of cases and controls that have an equivalent power as the sample sizes reported in the original study. This was estimated by iterating sample sizes with a 1:1 case:control ratio until it arrives at the same non-centrality parameter in power calculations as the reported sample size (Purcell et al., 2003). The studies listed are – AS: ankylosing spondylitis (International Genetics of Ankylosing Spondylitis Consortium, 2013), AtD: atopic dermatitis (Ellinghaus et al., 2013a), CeD: coeliac disease (Trynka et al., 2011b), CD: Crohn's disease, UC: ulcerative colitis, IBD: inflammatory bowel disease (Chapter 4), MS: multiple sclerosis (International Multiple Sclerosis Genetics, 2013), PBC: primary biliary cirrhosis (Liu et al., 2012), Ps: psoriasis (Tsoi et al., 2012), PSC: primary sclerosing cholangitis (Liu et al., 2013), RA: rheumatoid arthritis (Eyre et al., 2012).

Secondly, the decreasing effect sizes also raises questions about how much of total heritability can be explained by common variants. Assuming that narrow-sense heritability in Crohn's disease is 50% (Ahmad *et al.,* 2001), extrapolating the risk loci to 20,000 independent common variant associations will still explain

less than half of this heritability (Franke *et al.,* 2010). A similar estimate was arrived at in Lee *et al.* (2011), where the variance explained in Crohn's disease risk tagged by all genotyped variants was only 22%. This suggests that untyped variants (especially rare variants poorly tagged on genotyping arrays) will contribute to the remaining heritability, though that heritability estimates were overestimated in the first place cannot be ruled out. Due to the disease being rare, accurate disease prevalence is hard to estimate. Similarly, familial recurrent risk estimates are often based on ascertained families will multiple affected individuals, potentially overestimating the true familial risk in the population.

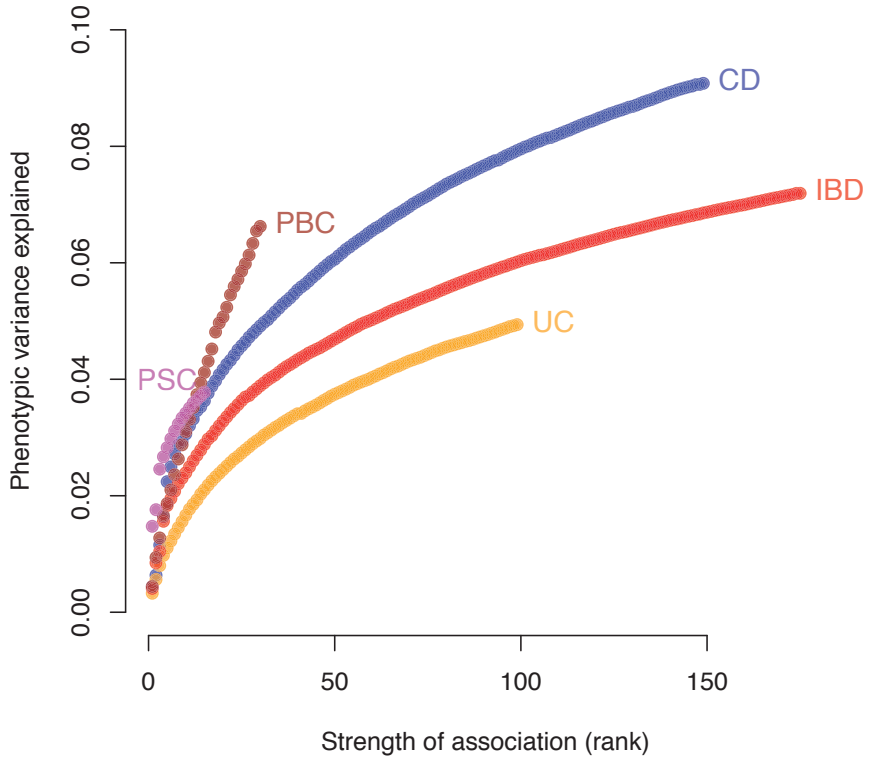


Figure 6.2. Cumulative proportion of variance in disease liability explained by the genome-wide significant loci identified in Chapters 2-4. For PBC and PSC, SNPs on the x-axis are ranked by and plotted by association P-value. For CD, UC and the two combined (IBD), SNPs are ranked by the decreasing MANTRA log10 Bayes factor association signal.

The different trajectories for each of the diseases in Figure 6.2 reflects the different underlying genetic architectures for these disorders and their

tractability to the GWAS approach. For instance, the variance explained by the 26 loci associated with PBC is more than double the equivalent number in UC. While some of these differences may be due to winner's curse (the sample size for PBC was much smaller and there has yet to be any follow-up studies), this also raises interesting questions about factors that shape these differences.

Before discussing these factors, it is interesting to compare the distribution of effect sizes of variants associated with immune-mediated disorders with those from other complex traits. In general, genetic studies for immune-mediated disorders have offered much greater bang-for-genotyping-buck in terms of the number of risk loci discovered and variance explained than other classes of disorders. For instance, the PBC study described in Chapter 2 identified 22 genome-wide significant loci with a sample size of ~2,800 cases and 8,500 controls. In contrast, it required over 5,500 cases and 9,000 controls to identify a single variant associated with endometriosis (Painter *et al.,* 2011). For psychiatric disorders, the story is just as sobering. Despite heritability estimates of ~30-40% in major depressive disorder, only a single borderline genome-wide significant signal was identified in a meta-analysis that included over 16,000 cases and 60,000 controls (Major Depressive Disorder Working Group of the Psychiatric, 2013). Nevertheless, even for these classes of disorders, risk loci will eventually be identified given large enough sample sizes. In schizophrenia, it required 8,000 cases and 19,000 controls to implicate a single locus in disease risk (Shi *et al.,* 2009). Five years later, a GWAS meta-analysis that included 36,000 cases and 113,000 controls increased the number of risk loci to 108 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Why then, is locus discovery in immune-mediated diseases such as PBC more tractable to the GWAS approach than diseases such as schizophrenia and endometriosis? One explanation is related to the hygiene hypothesis – that evolutionary adaptations have not caught up to the rapidly changing environment. Natural selection leaves its indelible footprints on the frequencies of functionally relevant alleles. For much of human history, it's likely that exposure to a range of pathogens was the norm, and alleles that best defend the

host from infection were selected for and rose in frequency throughout the population. In the modern world, vaccinations, better nutrition and awareness of hygiene have greatly reduced our exposure to antigens, leading to an imbalance in the immune system that favours chronic inflammatory conditions (Sironi and Clerici, 2010). It is hypothesised that those same alleles that once protected us from infection are now also those that make us most susceptible to autoimmune disorders.

A second explanation of differences in GWAS tractability is the amount of phenotypic and genetic heterogeneity that underlies complex traits. The presence of heterogeneity in genetic association studies reduces power to detect association and underestimates the effect sizes of risk variants. At the biological level, what is classified as a single disorder may be a result of combinations of different molecular processes, each with its own set of genetic and environmental levers, yet all with similar phenotypic presentations. This may especially be true for psychiatric disorders, where a yes/no diagnosis is still often based on whether a patient shows any *x* number of descriptive symptoms out of a list of *y* (Angst, 2007), resulting in potential for misclassification of cases and controls. This is largely because the most useful biological categories or dimensional categories are still unknown, and a better understanding of the genetic basis of these disorders will help give a clearer picture of disease pathogenesis and diagnoses. Contrast this to an autoimmune disorder such as PBC, where diagnosis is largely based on blood tests and the presence of a specific set of antibodies.

These two hypotheses are not mutually exclusive, and indeed are both likely to play a role in shaping the genetic architecture of complex traits. Future efforts at unravelling this genetic architecture will involve a combination of array-based and sequencing approaches in ever-larger sample sizes. For the remainder of this chapter, I will discuss these approaches, their potential challenges, and ultimately, prospects for translating what we've learnt from locus discovery into more effective treatment outcomes.

6.2 Future prospects for complex disease genetics

6.2.1 Array-based approaches

Genome-wide association studies predominantly focused on identifying common variant associations (variants with minor allele frequencies greater than 5%). There are good economic reasons for why this was the case. There are only so many SNPs that can fit on a genotyping chip, and given the patterns of linkage disequilibrium in the population, the majority of the ~5 million common variants in the genome can be tagged by a selection of ~500,000 SNPs (Barrett and Cardon, 2006; International HapMap *et al.,* 2007). Array-based studies with ever-larger sample sizes will continue to play a role in locus discovery. This is perhaps best exemplified by the UK Biobank's ongoing efforts to genotype their ~450,000 samples on a custom genome-wide genotyping microarray with ~800,000 variants. Individuals recruited to the UK Biobank underwent a range of diagnostic measures and will have their health tracked throughout their lifetime, providing an invaluable resource in the study of complex disease.

The design of the Immunochip, along with similar arrays such as the Metabochip (for metabolic and cardiovascular risk loci) (Voight *et al.,* 2012) and iCOGS (for various cancers) (Sakoda *et al.,* 2013), was also primarily motivated by economics. The ability to include thousands of SNPs for deep replication, high-density regions for fine-mapping, and the genotyping of over 150,000 individuals across multiple disease cohorts (and the sharing of population controls) meant that the Immunochip, at ~$40/sample, was a much more cost-effective platform for locus discovery and fine-mapping than alternative technologies at the time (e.g. Sequenom plexes, whole-genome arrays, pull-down sequencing) (Jostins, 2012).

There are, of course, several limitations to custom high-density arrays such as the Immunochip. Obvious pitfalls include the lack of coverage genome-wide and the ascertainment of variants only present in European populations. Additionally, while ~240,000 variants were initially selected for inclusion on the Immunochip, 196,524 made it onto the final array. Running a typical quality

control protocol will reduce this even further to 130,000-140,000 variants (Liu *et al.,* 2012; Liu *et al.,* 2013), resulting in a total array design success of ~60%. Technical failures explain the majority of these exclusions. SNPs in high-density regions were selected from those identified in the 1000 Genomes Pilot dataset using low-coverage sequencing, such that many of these variants (in particular rare variants) are poorly characterised, either due to being falsely called in the first place and/or poor probe design. Moreover, many variants were missed all together. As demonstrated in Chapters 2 and 3, imputation using the subsequently much larger 1000 Genomes Phase I reference panel almost doubled the number of variants in the high-density regions. Nevertheless, chip design continues to improve, and there are now several custom chips currently being developed or in the analysis phase – e.g. the Exome Chip for coding variants, the "African Power Chip" for African-specific variants, and the "Psych Chip" for risk variants identified in psychiatric disorders. In addition, current genome-wide arrays such as the Illumina Omni2.5 and Omni5 are supplemented with 200,000 and 500,000 custom variants respectively to fit with each researcher's requirements.

6.2.2 Sequencing approaches for rare variant studies

How then, given the state of technology and what we understand about the genetic architecture of complex traits, should one design a locus discovery experiment today? Array-based technologies (whole-genome and targeted arrays) are likely to remain the most cost-effective and efficient methods for identifying common variant associations, though a complete survey of genetic variation in an individual will require high coverage (greater than 30X) whole-genome sequencing – currently costing 1-2 orders of magnitude more per sample than genotyping arrays. These sequencing approaches will be able to capture rare variants (those with minor allele frequencies less than 1%), which are poorly captured on arrays. While most genetic variation in an individual is at common sites, the total number rare variants in the population far outnumber common variants (Keinan and Clark, 2012). In chapter 1 section 1.4.3, I outlined theoretical reasons why rare variants are likely to play a role in complex disease,

and highlighted recent sequencing studies in known risk loci to identify rare variant associations (Hunt *et al.,* 2013; Rivas *et al.,* 2011).

These targeted sequencing studies identified very few novel independent rare variant signals (and that the common variant associations are not driven by nearby rare variants), highlighting the need for larger sample sizes for these types of studies. Under certain assumptions about the effect size distribution of rare variants and selection pressures, well-powered studies may require cohorts of more than 25,000 cases and an equal number of controls, along with equally large numbers for replication (Zuk *et al.,* 2014). Moreover, given the importance of non-coding variation in complex disease risk, there is also a need for whole-genome approaches.

While high-coverage sequencing is still prohibitively expensive, there is currently great potential for low-coverage whole-genome sequencing approaches (less than 6X) as a powerful and cost-effective alternative. In low-coverage sequencing, rare variants are discovered and jointly called across many thousands of individuals, and LD-based imputation methods are used to refine genotype calls. For instance, for a SNP with frequency 0.2% to be discovered, over 2000 individuals need to be sequenced at 30X coverage (60,000 genomes). In contrast, the same SNP can be identified in ~3000 individuals sequenced at 4X (12,000 genomes) – a five fold reduction in sequencing cost (Li *et al.,* 2011). With more sequenced individuals comes greater power to detect associations. Large cohorts of low-coverage sequenced individuals can also be used as reference panels to impute rare variants into new and existing GWAS datasets at much greater accuracy than existing panels. Over the course of 2014-15, it is expected that over 30,000 individuals will be sequenced at low-coverage (www.haplotype-reference-consortium.org). Imputing the millions of new variants discovered from this set into ~25,000 IBD cases (of which ~15,000 have already been genotyped as part of GWAS) will, for the first time, enable dection of association to SNPs with frequencies in the order of 0.1-1% and ORs of 2-3 (Figure 6.3). This sequencing plus imputation approach was demonstrated in a recent study in type 2 diabetes, were variants discovered by sequencing 2,630

samples were imputed in 11,114 cases and 267,140 controls (Steinthorsdottir *et al.,* 2014). The study identified risk variants at several variants with frequencies between 0.65% and 1.5% and ORs of 1.5-2.

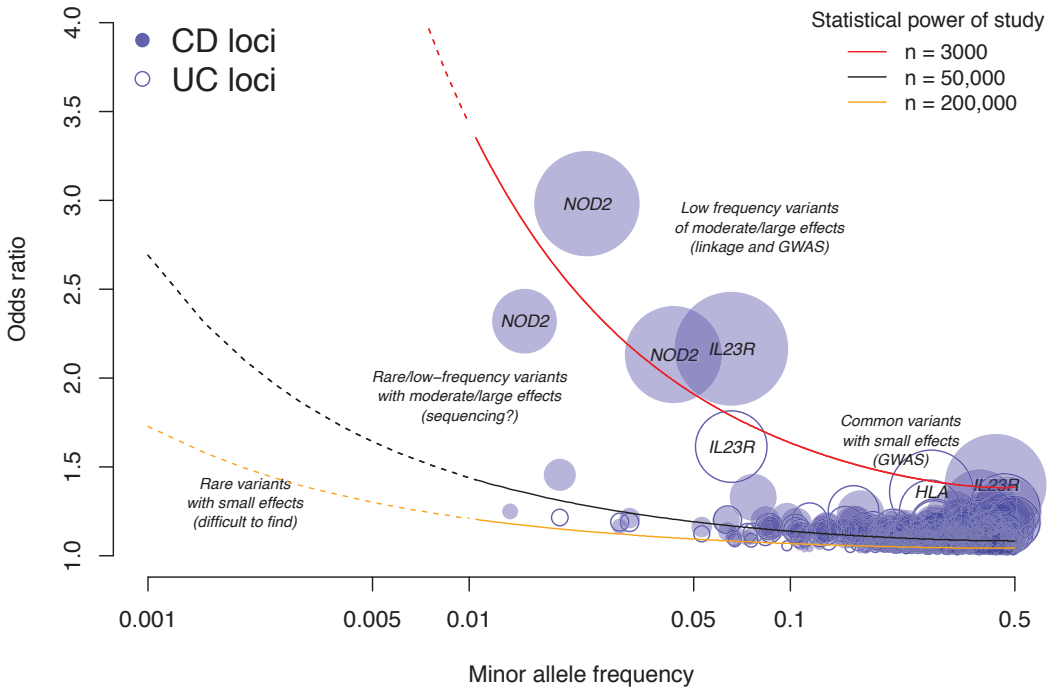


Figure 6.3. The genetic architecture of inflammatory bowel disease. Known CD and UC variants are plotted according to their minor allele frequencies and risk increasing ORs estimated from results in Chapter 4. ORs of risk-decreasing minor alleles were flipped for illustrative purposes. The size of the circles represents the amount variance in disease liability explained by that variant. The red, black and orange lines represent the minimum OR and allele frequency combination for a locus for which a GWAS with 3000, 50,000 or 200,000 individuals (with an equal number of cases and controls) respectively will have greater than 80% statistical power to detect association at $P < 5\times10^{-8}$. The dashed lines represent the allele frequency spectrum of variants that are typically poorly captured on GWAS microarrays (minor allele frequencies less than 1%).

The testing of rare variant associations will also throw up new statistical challenges. Firstly, the established genome-wide significance threshold of $P < 5\times10^{-8}$ is based on a 5% Bonferroni correction on the approximate number of independent regions tagged by common variants genome-wide (in European populations) (International HapMap *et al.,* 2007). Rare variants, on the other

hand, are more numerous and less likely to be in high LD with other variants, such that a genome-wide survey of rare variants will involve many more independent tests than with common variants. It may be the case that a more stringent P-value threshold will be required to avoid too many false-positive reports. Secondly, while methods such as logistic regression are frequently used for common-variant associations, they may not be well-calibrated in rare-variant tests were minor allele counts are low, leading to false type-1 error rates. Ma *et al.* (2013) suggest a minor allele count cut-off of 400 (corresponding to a minor allele frequency of 1% in 20,000 individuals) for when standard logistic regression tests may need to be recalibrated. Thirdly, rare variants are more likely to be population specific, requiring more careful consideration of sample recruitment and study design. There is evidence that the effects of population stratification for rare variants are stronger than for common variants, and that existing methods such as PCA and linear mixed models may not be able to fully account for rare such stratification (Mathieson and McVean, 2012). Family-based association methods, which are robust to population stratification, may once again play an important role. Fourthly, different sequencing studies are likely to involve a range of sequencing technologies and methods, such that differences in coverage, read lengths, variant calling and genotype refinement methods will likely have direct effects on the properties of the variants reported. Methods that account for these differences, especially when cases and controls are sequenced separately, need to be developed. Finally, while I have discussed these challenges in the context of single-variant association tests, they also equally apply to the suite of rare-variant region-based tests. Additional challenges to these region-based methods include the choice of test, defining regions and which SNPs to test and difficulties in assigning causal variants (some of which I discussed in Chapter 1 section 1.4.3).

Despite these hurdles, there is a growing recognition among health policy makers about the importance of sequencing in medical research. In December 2012, the UK Government announced an initiative to sequence 100,000 whole genomes by the end of 2017. Patients will be recruited from NHS centres, and will consist of those with rare diseases or various cancers. The project is

currently in its pilot phase, and it remains to be determined exactly how the samples will be sequenced, but will likely involve high-coverage sequencing of 40,000 patients. For rare diseases, the parents of patients will also be sequenced, and for each cancer patient, two genomes will be sequenced – one from the tumour and one from healthy tissue (Connor, 2014).

The immediate benefits of the UK 100K Project will be felt by patients and their families. For example, the sequencing of rare disease families will help in identifying highly penetrant de novo mutations, perhaps easing parents' concerns about having additional children. Identifying the somatic driver mutations in cancer can also inform the best course of treatment. For complex disease researchers, the project provides an invaluable resource to use with existing datasets (for instance, as population controls or imputation reference panels), as well as a testing ground for methods development. In the long term, the infrastructure, knowhow and experience gained through the project will provide a blueprint for future sequencing efforts. With the announcement of the Illumina HiSeq X Ten in January 2014, the raw cost of sequencing a genome at high-coverage (30X) today is around $1000. It is no stretch to imagine that this price will fall to a few hundred dollars within the decade, well below the price of many routine medical diagnostic tests, such that getting your genome sequenced (if you haven't already) will become part-and-parcel of a trip to the doctor.

What will it mean then, for complex disease research, when every patient with Crohn's disease, type 2 diabetes or schizophrenia in the country will have their genomes available? For locus discovery, the list of disease-associated variants will continue to grow. In a study of say, 300,000 Crohn's disease cases and a million controls, there will be greater than 80% power to detect variants with odds ratios greater than ~1.1 and a minor allele frequency of 1%. For variants with frequencies around 0.1%, odds ratios greater than ~1.4 will be detected. The total number of risk loci will likely be in the thousands, and, by this stage, insights into disease biology will primarily come from the molecular pathways and biological mechanisms that these risk loci cluster into and interact with rather than investigating the genes in isolation. By having entire families

sequenced, it will also allow the identification of any *de novo* and rare highly penetrant variants in complex disease risk. In addition, such sample sizes along with medical records will also enable well-powered studies on disease subphenotypes such as clinical progression and drug response. Variants associated with disease progression may not necessarily also be associated with disease susceptibility (Lee *et al.,* 2013a), and such studies, along with integration with real-time monitoring of gut microbiota and cellular markers such as gene expression and epigenetic marks in disease-relevant cells, will pave the way for personalised treatments based on an individual's genetic makeup.

6.2.3 Genetic studies in non-European populations

As many as 96% of published GWAS up to 2011 were conducted in populations of European descent, yet these populations make up less than 15% of the world (Bustamante *et al.,* 2011). This disparity is primarily driven by resources – Western countries overwhelmingly spend more on scientific research, both in absolute terms and as a proportion of GDP, than do non-Western countries. It is then no surprise that the types of studies that rely on cutting edge technology (while costs are still at a premium) are first undertaken in these countries. Reassuringly, efforts such as the African Genomes Project, targeted funding efforts from research charities such as the Wellcome Trust, as well as the ever growing stream of home grown genetic studies emerging from researchers in Asian and Latin American countries are leading the charge in addressing this imbalance.

There is great scientific value in expanding complex trait genetics to the rest of the world. Firstly, as demonstrated for IBD in Chapter 4, much of the risk loci for complex disorders are likely to be shared across populations. This means that ascertaining samples from non-European populations is an effective way of boosting power to detect association. Of course, researchers will need to be aware of the potential for population stratification, though statistical methods that account for population stratification and potential heterogeneity between populations are now quite mature for common variant associations (Morris, 2011; Yang *et al.,* 2014a). Secondly, genetic differences between populations can

inform biology. SNPs that are monomorphic in Europeans will go undetected in GWAS, yet finding associations at these variants in non-European populations will create new leads in understanding disease pathogenesis. When one considers the genetic diversity of African populations, this will almost certainly be the case. Moreover, variants that show large differences in effect sizes between populations point to potential gene-environment interactions, allowing for insights into the environmental factors that modify disease risk. Different population histories also create different patterns of LD. These patterns will be instrumental in fine-mapping efforts to localise causal variants at associated loci common across populations. Finally, aside from the scientific reasons listed above, there are clear humanitarian arguments for expanding genetic studies to non-Western countries and to study the diseases that most burden them. Those most in need must not be the last to benefit from genetic research (Bustamante *et al.,* 2011).

6.2.4 Genetic prediction

In addition to gaining a better understanding of disease biology, genetic information can also potentially be used for disease risk prediction. Prediction methods for complex diseases typically involve assigning a risk score to an individual based on their genotypes and previously estimated effect sizes (for instance, ORs from GWAS) across risk alleles. Risk alleles can be assigned not only based on known associations, but also include nominally associated variants. Prediction accuracy can be evaluated by methods such as the receiver operating characteristic curve (ROC), which estimates the true and false positive rates of the predictor at various risk score cut-offs (Lasko *et al.,* 2005). The area under the ROC (AUC) is the probability that for a randomly selected pair of diseased and healthy individuals, the diseased individual will have a higher risk score. An AUC of 0.5 means that the prediction method is no better than chance, while a value of 1 means that the method perfectly discriminates between diseased and healthy individuals.

For complex autoimmune diseases, genetic risk prediction is still in its infancy and does not currently offer much in terms of clinical utility. Estimates of

AUC using just family history of disease, genetic risk loci or the two together in Crohn's disease range from 0.56 to 0.74 (Kang *et al.,* 2011; Ruderfer *et al.,* 2010). Including risk factors such as smoking and age into the risk model may improve the AUC. Nevertheless, given its high heritability, the theoretical maximum possible AUC assuming that all Crohn's disease risk loci have been identified and effect sizes are accurately measured is estimated to lie between 0.96-0.98 (Jostins and Barrett, 2011; Wray *et al.,* 2010). However, while this figure seems high, the utility of genetic prediction is limited given the low prevalence of Crohn's and other immune-mediated diseases. Even assuming a generous disease prevalence estimate of 1% and AUC of 0.98, less than 12% of individuals who test positive (using a sensitivity cut-off of 0.93) will develop disease (Jostins and Barrett, 2011). Increasing the threshold will increase the proportion of positively identified individuals but also exclude a higher number of cases from being identified. While never providing any guarantees, the use of genetic prediction in complex diseases may ultimately at best aid in disease diagnosis, and at worst create greater awareness among those most highly at risk for disease.

6.3 From causal variants to treatment outcomes

In Chapter 1, I discussed potential approaches and challenges involved in narrowing down a risk locus into a single causal variant. Assuming now that a set of causal candidates has been identified, what is required to confirm causality? The direct modelling of these variants in cell lines and model organisms are likely to play an important role in answering this question, and emerging technologies such as DNA editing through CRISPR/Cas and engineering induced pluripotent stem cells (iPSCs) are growing in popularity (Cong *et al.,* 2013; Mali *et al.,* 2013; Robinton and Daley, 2012). The CRISPR/Cas system involves guiding a Cas-cleavage enzyme to a specific site of the genome, which is then imprecisely cleaved and repaired, allowing for specific mutations to be introduced. *In vitro* modelling of these mutations in disease relevant cells types (e.g. those generated from iPSCs) allows for the direct investigation of how these mutations affect cellular phenotypes such as gene expression and responses to infection;

generating hypotheses about how these genetic variants lead to disease susceptibility. Knocking down the relevant genes identified in model organisms will further enable understanding of how these genes affect the organism as a whole.

At this point, it is worth discussing about the level of proof required before causality can be confidently assigned to a genetic variant. From a genetic association standpoint, defining causality is straightforward, though identifying it is difficult. A causal variant is one that can explain a statistical association signal on its own, irrespective of its correlation with other variants. Hence an associated tag SNP cannot be called causal. From a disease risk standpoint, however, causality is more nuanced. Given the typical small effect sizes of associated variants, a causal variant is neither necessary nor sufficient to cause disease (Visscher *et al.,* 2012). CRISPR/Cas, iPSCs and gene knockouts might reveal a disease relevant phenotype, but this also does not prove that the phenotype affects disease risk in the population. It may be the case that we will never have the ability to definitively prove that the observed biological effect of a statistically causal variant is also causal in the disease risk sense.

Defining causality may end up being a moot point if the relevant genes that are identified and biological knowledge gained lead to better treatment outcomes. Identifying a gene target and the creation of a therapeutic molecule is difficult. Over 90% of compounds that enter clinical trials fail to gain approval, reflecting the limited predictive value of preclinical disease models and a lack of understanding of the long-term consequences of perturbing specific molecules (Plenge *et al.,* 2013). While GWAS have provided valuable insights into disease biology, little of this has yet translated into more effective therapeutics. Part of this is of course due to time – moving from a gene target through clinical trials to a final approved drug can take well over a decade. Nevertheless, it is hoped that knowledge of the genes that underlie disease risk will lead to more effective treatment outcomes.

There is a strong historical precedence for the use of human genetics in drug development. Before the large-scale identification of susceptibility genes,

epidemiological observations were often the catalyst for identifying potential therapies. Genetic variation in the human population meant that many individuals carried alleles that mimic the effects of potential therapies. For instance, the development of statins to lower LDL cholesterol levels and treat heart disease was based on observations in families with rare hypercholesterolemia who carried mutations in the *LDLR* gene. Members of these families both had higher levels of cholesterol and higher prevalence of heart disease. Importantly, the number of mutations appeared to affect cholesterol levels and risk of heart disease in a dose-dependant manner. It was also known that the HMG-CoA reductase plays an important role in the production of cholesterol in the liver, and natural products that inhibit this enzyme (e.g. compactin and lovastatin) lowered LDL cholesterol levels in animal models (Plenge *et al.,* 2013). Later clinical trials in humans demonstrated the efficacy and safety of statins, and ultimately showed their effectiveness at reducing heart disease risk in individuals with high cholesterol.

The role of human genetics in drug development is also supported retrospectively by drugs that were developed without the use of human genetics, but whose molecular targets have since been supported by their associations with disease. In the statins example, variants in the *HMGCR* gene (which encodes the HMG-CoA enzyme) were found to be associated with LDL cholesterol by GWAS (Kathiresan *et al.,* 2008). Notably, the effect size of the association bears little relationship to its clinical relevance. The HMGCR signal has an effect on LDL cholesterol levels of ~2.5 mg/dl per allele (Teslovich *et al.,* 2010), or, to put another way, approximately one-tenth of a unit of standard deviation – a tiny effect. Yet statin drugs can reduce LDL levels by around 40 mg/dl (Cholestrol Treatment Triallists' Collaborators, 2005). Other retrospective examples include the targeting of *CTLA-4* by abatacept for rheumatoid arthritis (Genovese *et al.,* 2005; Gregersen *et al.,* 2009), *IL12B* by ustekinumab for Crohn's disease (Mannon *et al.,* 2004; Parkes *et al.,* 2007) and *PPARG* by thiazolidinediones for type 2 diabetes (Spiegelman, 1998; Zeggini *et al.,* 2007).

The identification of shared risk loci across different diseases also enables repurposing of existing drugs. By looking at the overlap between GWAS risk loci and current drugs in development, Sanseau *et al.* (2012) identified over 100 targets that were associated with a disease other than the one the drug was being developed for. For instance, *TNFSF11* is currently inhibited by denosumab for treatment of osteoporosis in postmenopausal women. Variants in *TNFSF11* are also associated with Crohn's disease (Franke *et al.,* 2010), and it is tempting to suggest that this drug may be repurposed (Sanseau *et al.,* 2012). The use of existing approved drugs also avoids the need for lengthy safety trials, meaning that treatments can be marketed in a much shorter time frame.

Using GWAS risk loci to guide the development of novel drugs will be a much bigger challenge. Plenge *et al.* (2013) list nine criteria for prioritising risk loci before drug discovery should be considered:

1. The gene harbours a causal variant that is unequivocally associated with a medical trait of interest
2. The biological function of the causal gene and causal variant are known
3. The gene harbours multiple causal variants of known biological function, thereby enabling the generation of genotype–phenotype dose–response curves
4. The gene harbours a loss-of-function allele that protects against disease, or a gain-of-function allele that increases the risk of disease
5. The genetic trait is related to the clinical indication targeted for treatment
6. The causal variant is associated with an intermediate phenotype that can be used as a biomarker
7. The gene target is druggable
8. The causal variant is not associated with other adverse event phenotypes
9. Corroborating biological data support genetic findings

Going through this list, it's clear that, with the exception of point 1, the majority of GWAS risk loci do not yet satisfy any of these criteria. The rationale for many of these points (e.g. the need for loss-of-function or gain-of-function alleles) is that it is simply easier to develop drugs that inhibit certain type of

targets with the current knowledge of assays. This will hopefully change in the future as technology and assays improve. For instance, kinases were once thought to be undruggable, though this is now changing with the development of kinase-inhibitors (Gashaw *et al.,* 2011). Moreover, some GWAS risk loci themselves may not necessarily be the most actionable drug target, but rather can inform molecular pathways that are relevant to disease. For instance, if a risk locus is a ligand, knowing the corresponding receptor (for which drugs are well-suited to exert their effects on) will offer additional potential targets. Ultimately however, understanding the disease-relevant biological functions of risk loci will always remain the first step on the road to drug discovery.

6.4 Concluding remarks

It is almost certain that within the coming decades, low-cost whole genome-sequencing will become routine, and it is not too much of a stretch to imagine locus discovery projects involving hundreds of thousands, perhaps even millions, of whole-genome sequenced cases and controls. While the theoretical framework of association studies as outlined in Risch and Merikengas (1996) are unlikely to change, these types of studies will also throw up new methodological challenges that will need to be overcome. Along with genome-sequencing, large scale functional genomic studies will ever expand to include greater coverage of cell types and disease states, and methods to integrate these data sources will play an important role in understanding biology.

It needs to be emphasised that locus discovery is not an end in itself. Challenges remain in taking what we've learned from genetic studies to build more complete models of disease pathogenesis and ultimately translating these into better patient outcomes.

Bibliography

1000 Genomes Project Consortium., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Aadland, E., Schrumpf, E., Fausa, O., Elgjo, K., Heilo, A., Aakhus, T., and Gjone, E. (1987). Primary sclerosing cholangitis: a long-term follow-up study. Scandinavian journal of gastroenterology *22*, 655-664.

Ahmad, T., Satsangi, J., McGovern, D., Bunce, M., and Jewell, D.P. (2001). The genetics of inflammatory bowel disease. Alimentary Pharmacology & Therapeutics *15*, 731-748.

Anaya, J.M., Corena, R., Castiblanco, J., Rojas-Villarraga, A., and Shoenfeld, Y. (2007). The kaleidoscope of autoimmunity: multiple autoimmune syndromes and familial autoimmunity. Expert review of clinical immunology *3*, 623-635.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A.*, et al.* (2011a). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nature Genetics *43*, 246-252.

Anderson, C.A., Soranzo, N., Zeggini, E., and Barrett, J.C. (2011b). Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biology *9*, e1000580.

Angst, J. (2007). Psychiatric diagnoses: the weak component of modern research. World psychiatry : official journal of the World Psychiatric Association *6*, 94-95.

Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T., Matsumoto, T., Matsui, T., Kakuta, Y., Kinouchi, Y.*, et al.* (2009). A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nature Genetics *41*, 1325-1329.

Asimit, J., and Zeggini, E. (2010). Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics *44*, 293-308.

Bain, S.C., Prins, J.B., Hearne, C.M., Rodrigues, N.R., Rowe, B.R., Pritchard, L.E., Ritchie, R.J., Hall, J.R.S., Undlien, D.E., Ronningen, K.S.*, et al.* (1992). Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. Nature Genetics *2*, 212-215.

Ban, M., Goris, A., Lorentzen, A.R., Baker, A., Mihalova, T., Ingram, G., Booth, D.R., Heard, R.N., Stewart, G.J., Bogaert, E.*, et al.* (2009). Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. European Journal of Human Genetics *17*, 1309-1313.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics *11*, 773-785.

Barrett, J.C., and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. Nature Genetics *38*, 659-662.

Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C.*, et al.* (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics *41*, 703-707.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M.*, et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genetics *40*, 955-962.

Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genetic Epidemiology *35*, 606-619.

Bateson, W., and Mendel, G. (1902). Mendel's principles of heredity; a defence (Cambridge,: University press).

Begovich, A.B., Carlton, V.E., Honigberg, L.A., Schrodi, S.J., Chokkalingam, A.P., Alexander, H.C., Ardlie, K.G., Huang, Q., Smith, A.M., Spoerke, J.M.*, et al.* (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. American Journal of Human Genetics *75*, 330-337.

Bell, O., Tiwari, V.K., Thoma, N.H., and Schubeler, D. (2011). Determinants and dynamics of genome accessibility. Nature Reviews Genetics *12*, 554-564.

Bellenguez, C., Strange, A., Freeman, C., Donnelly, P., and Spencer, C.C. (2012). A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics *28*, 134-135.

Bennett, S.T., Wilson, A.J., Esposito, L., Bouzekri, N., Undlien, D.E., Cucca, F., Nistico, L., Buzzetti, R., Bosi, E., Pociot, F.*, et al.* (1997). Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. Nature Genetics *17*, 350-352.

Bergquist, A., Montgomery, S.M., Bahmanyar, S., Olsson, R., Danielsson, A., Lindgren, S., Prytz, H., Hultcrantz, R., Loof, L.A., Sandberg-Gertzen, H.*, et al.* (2008). Increased risk of primary sclerosing cholangitis and ulcerative colitis in first-degree relatives of patients with primary sclerosing cholangitis. Clinical Gastroenterology and Hepatology *6*, 939-943.

Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E.V., Brunak, S.*, et al.* (2013). A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell *155*, 70-80.

Blumberg, R.S., Dittel, B., Hafler, D., von Herrath, M., and Nestle, F.O. (2012). Unraveling the autoimmune translational research process layer by layer. Nature Medicine *18*, 35-41.

Boonstra, K., Beuers, U., and Ponsioen, C.Y. (2012) Epidemiology of primary sclerosing cholangitis and primary biliary cirrhosis: A systematic review. Journal of Hepatology *56*, 1181-1188.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics *32*, 314-331.

Bouchard, T., Lykken, D., McGue, M., Segal, N., and Tellegen, A. (1990). Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. Science *250*, 223-228.

Brant, S.R. (2011). Update on the heritability of inflammatory bowel disease: The importance of twin studies. Inflammatory Bowel Diseases *17*, 1-5.

Brant, S.R., Fu, Y., Fields, C.T., Baltazar, R., Ravenhill, G., Pickles, M.R., Rohal, P.M., Mann, J., Kirschner, B.S., Jabs, E.W.*, et al.* (1998). American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. Gastroenterology *115*, 1056-1061.

Broome, U., Olsson, R., Loof, L., Bodemar, G., Hultcrantz, R., Danielsson, A., Prytz, H., Sandberg-Gertzen, H., Wallerstedt, S., and Lindberg, G. (1996). Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis. Gut *38*, 610-615.

Bustamante, C.D., De La Vega, F.M., and Burchard, E.G. (2011). Genomics for the world. Nature *475*, 163-165.

Candore, G., Lio, D., Colonna Romano, G., and Caruso, C. (2002). Pathogenesis of autoimmune diseases associated with 8.1 ancestral haplotype: effect of multiple gene interactions. Autoimmunity Reviews *1*, 29-35.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., Nickerson, D.A. (2004) Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analysis Using Linkage Disequilibrium. American Journal of Human Genetics *74*, 106-120.

Castro-Santos, P., Suarez, A., Lopez-Rivas, L., Mozo, L., and Gutierrez, C. (2006). TNF[alpha] and IL-10 Gene Polymorphisms in Inflammatory Bowel Disease. Association of -1082 AA Low Producer IL-10 Genotype with Steroid Dependency. The American Journal of Gastroenterology *101*, 1039-1047.

Cavanaugh, J. (2001). International Collaboration Provides Convincing Linkage Replication in Complex Disease through Analysis of a Large Pooled Data Set: Crohn Disease and Chromosome 16. The American Journal of Human Genetics *68*, 1165-1171.

Cavanaugh, J.A., Callen, D.F., Wilson, S.R., Stanford, P.M., Sraml, M.E., Gorska, M., Crawford, J., Whitmore, S.A., Shlegel, C., Foote, S.*, et al.* (1998). Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16. Annals of Human Genetics *62*, 291-298.

Chapman, R.W., Arborgh, B.A., Rhodes, J.M., Summerfield, J.A., Dick, R., Scheuer, P.J., and Sherlock, S. (1980). Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. Gut *21*, 870-877.

Chapman, R.W., Varghese, Z., Gaul, R., Patel, G., Kokinon, N., and Sherlock, S. (1983). Association of primary sclerosing cholangitis with HLA-B8. Gut *24*, 38-41.

Chen, C.T., Wang, J.C., and Cohen, B.A. (2007). The strength of selection on ultraconserved elements in the human genome. American Journal of Human Genetics *80*, 692-704.

Cho, J.H., Nicolae, D.L., Gold, L.H., Fields, C.T., LaBuda, M.C., Rohal, P.M., Pickles, M.R., Qin, L., Fu, Y., Mann, J.S.*, et al.* (1998). Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: Evidence for epistasis between 1p and IBD1. Proceedings of the National Academy of Sciences *95*, 7502-7507.

Cholestrol Treatment Triallists' (CTT) Collaborators (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. The Lancet 366, 1267-1278.

Clark, K., Mackenzie, K.F., Petkevicius, K., Kristariyanto, Y., Zhang, J., Choi, H.G., Peggie, M., Plater, L., Pedrioli, P.G., McIver, E.*, et al.* (2012). Phosphorylation of CRTC3 by the salt-inducible kinases controls the interconversion of classically activated and regulatory macrophages. Proceedings of the National Academy of Sciences *109,* 16986–16991.

Colhoun, H.M., McKeigue, P.M., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865-872.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A.*, et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Connor, S. (2014). Government backs massive new £300m gene sequencing project. The Independent, August 1, 2014.

ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Internation Genetics of Anklyosing Spondylitis Consortium. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nature Genetics *45*, 730-738.

Cook, D.N., Prosser, D.M., Forster, R., Zhang, J., Kuklin, N.A., Abbondanzo, S.J., Niu, X.-D., Chen, S.-C., Manfra, D.J., Wiekowski, M.T.*, et al.* (2000). CCR6 Mediates Dendritic Cell Localization, Lymphocyte Homeostasis, and Immune Responses in Mucosal Tissue. Immunity *12*, 495-503.

Cooper, G.S., Bynum, M.L., and Somers, E.C. (2009). Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. Journal of Autoimmunity *33*, 197-207.

Cordell, H.J., and Clayton, D.G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. American Journal of Human Genetics *70*, 124-141.

Cortes, A., and Brown, M. (2011). Promise and pitfalls of the Immunochip. Arthritis Research and Therapy *13*, 101.

Couturier, N., Bucciarelli, F., Nurtdinov, R.N., Debouverie, M., Lebrun-Frenay, C., Defer, G., Moreau, T., Confavreux, C., Vukusic, S., Cournu-Rebeix, I.*, et al.* (2011). Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. Brain *134*, 693-703.

Cowper-Sallari, R., Zhang, X., Wright, J.B., Bailey, S.D., Cole, M.D., Eeckhoute, J., Moore, J.H., and Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature Genetics *44*, 1191-1198.

Croese, J., O'Neil, J., Masson, J., Cooke, S., Melrose, W., Pritchard, D., and Speare, R. (2006). A proof of concept study establishing Necator americanus in Crohn's patients and reservoir donors. Gut *55*, 136-137.

Curran, M.E., Lau, K.F., Hampe, J., Schreiber, S., Bridger, S., Macpherson §, A.J.S., Cardon, L.R., Sakul, H., Harris, T.J.R., Stokkers ∥, P.*, et al.* (1998). Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. Gastroenterology *115*, 1066-1071.

Cuthbert, A.P., Fisher, S.A., Mirza, M.M., King, K., Hampe, J., Croucher, P.J.P., Mascheretti, S., Sanderson, J., Forbes, A., Mansfield, J.*, et al.* (2002). The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology *122*, 867-874.

Dastani, Z., Hivert, M.-F., Timpson, N., Perry, J.R.B., Yuan, X., Scott, R.A., Henneman, P., Heid, I.M., Kizer, J.R., Lyytikäinen, L.-P.*, et al.* (2012). Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals. PLoS Genetics *8*, e1002607.

de Bakker, P.I.W., Yelensky, R., Peer, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. Nature Genetics *37*, 1217-1223.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E.*, et al.* (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390-394.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837-845.

Dequiedt, F., Kasler, H., Fischle, W., Kiermer, V., Weinstein, M., Herndier, B.G., and Verdin, E. (2003). HDAC7, a thymus-specific class II histone deacetylase, regulates Nur77 transcription and TCR-mediated apoptosis. Immunity *18*, 687-698.

Dequiedt, F., Van Lint, J., Lecomte, E., Van Duppen, V., Seufferlein, T., Vandenheede, J.R., Wattiez, R., and Kettmann, R. (2005). Phosphorylation of histone deacetylase 7 by protein kinase D mediates T cell receptor-induced Nur77 expression and apoptosis. Journal of Experimental Medicine *201*, 793-804.

Devlin, B., Daniels, M., and Roeder, K. (1997). The heritability of IQ. Nature *388*, 468-471.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997-1004.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biology *8*, e1000294.

Dilthey, A.T., Moutsianas, L., Leslie, S., and McVean, G. (2011). HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics *27*, 968-972.

Dixson, L., Walter, H., Schneider, M., Erk, S., Schäfer, A., Haddad, L., Grimm, O., Mattheisen, M., Nöthen, M.M., Cichon, S.*, et al.* (2014). Retraction for Dixson *et al.,* Identification of gene ontologies linked to prefrontal–hippocampal functional coupling in the human brain. Proceedings of the National Academy of Sciences *Published online before print September 2*

Donaldson, P.T., Baragiotta, A., Heneghan, M.A., Floreani, A., Venturi, C., Underhill, J.A., Jones, D.E., James, O.F., and Bassendine, M.F. (2006). HLA class II alleles, genotypes, haplotypes, and amino acids in primary biliary cirrhosis: a large-scale study. Hepatology *44*, 667-674.

Donaldson, P.T., Farrant, J.M., Wilkinson, M.L., Hayllar, K., Portmann, B.C., and Williams, R. (1991). Dual association of HLA DR2 and DR3 with primary sclerosing cholangitis. Hepatology *13*, 129-133.

Donaldson, P.T., and Norris, S. (2002). Evaluation of the role of MHC class II alleles, haplotypes and selected amino acid sequences in primary sclerosing cholangitis. Autoimmunity *35*, 555-564.

Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A.*, et al.* (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461-1463.

Ellinghaus, D., Baurecht, H., Esparza-Gordillo, J., Rodriguez, E., Matanovic, A., Marenholz, I., Hubner, N., Schaarschmidt, H., Novak, N., Michel, S.*, et al.* (2013a). High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. Nature Genetics *45*, 808-812.

Ellinghaus, D., Folseraas, T., Holm, K., Ellinghaus, E., Melum, E., Balschun, T., Laerdahl, J.K., Shiryaev, A., Gotthardt, D.N., Weismuller, T.J.*, et al.* (2012). Genome-wide association analysis in sclerosing cholangitis and ulcerative colitis identifies risk loci at GPR35 and TCF4. Hepatology *58*, 1074-1083.

Ellinghaus, D., Zhang, H., Zeissig, S., Lipinski, S., Till, A., Jiang, T., Stade, B., Bromberg, Y., Ellinghaus, E., Keller, A.*, et al.* (2013b). Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. Gastroenterology *145*, 339-347.

Elliot, D.E., Urban, J.F., Argo, C.K., and Weinstock, J.V. (2000). Does the failure to acquire helminthic parasites predispose to Crohn's disease? The FASEB Journal *14*, 1848-1855.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M.*, et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Evans, D.M., and Cardon, L.R. (2004). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. American Journal of Human Genetics *75*, 687-692.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K.*, et al.* (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nature Genetics *44*, 1336-1340.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C.*, et al.* (2014). Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science *343*.

Falconer, D., and Mackay, T. (1996). Introduction to Quantitative Genetics (Longman).

Farrant, J.M., Hayllar, K.M., Wilkinson, M.L., Karani, J., Portmann, B.C., Westaby, D., and Williams, R. (1991). Natural history and prognostic variables in primary sclerosing cholangitis. Gastroenterology *100*, 1710-1717.

Farrokhyar, F., Swarbrick, E.T., and Irvine, E.J. (2001). A critical review of epidemiological studies in inflammatory bowel disease. Scandinavian Journal of Gastroenterology *36*, 2-15.

Fisher, R.A. (1930). The genetical theory of natural selection (Oxford,: The Clarendon press).

Folseraas, T., Melum, E., Rausch, P., Juran, B.D., Ellinghaus, E., Shiryaev, A., Laerdahl, J.K., Ellinghaus, D., Schramm, C., Weismuller, T.J.*, et al.* (2012). Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. Journal of Hepatology *57*, 366-375.

Foth, B.J., Tsai, I.J., Reid, A.J., Bancroft, A.J., Nichol, S., Tracey, A., Holroyd, N., Cotton, J.A., Stanley, E.J., Zarowiecki, M.*, et al.* (2014). Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nature Genetics *46*, 693-700.

Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R.*, et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics *42*, 1118-1125.

Gaffney, D.J., Veyrieras, J.B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biology *13*, R7.

Gashaw, I., Ellinghaus, P., Sommer, A., and Asadullah, K. (2011). What makes a good drug target? Drug Discovery Today *16*, 1037-1043.

Genovese, M.C., Becker, J.-C., Schiff, M., Luggen, M., Sherrer, Y., Kremer, J., Birbara, C., Box, J., Natarajan, K., Nuamah, I.*, et al.* (2005). Abatacept for Rheumatoid Arthritis Refractory to Tumor Necrosis Factor α Inhibition. New England Journal of Medicine *353*, 1114-1123.

Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., and Holmes, C.C. (2008). GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics *24*, 2209-2214.

Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J.*, et al.* (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genetics *6*, e1000952.

Greco, L., Romino, R., Coto, I., Di Cosmo, N., Percopo, S., Maglio, M., Paparo, F., Gasperi, V., Limongelli, M.G., Cotichini, R.*, et al.* (2002). The first large population based twin study of coeliac disease. Gut *50*, 624-628.

Gregersen, P.K., Amos, C.I., Lee, A.T., Lu, Y., Remmers, E.F., Kastner, D.L., Seldin, M.F., Criswell, L.A., Plenge, R.M., Holers, V.M.*, et al.* (2009). REL, encoding a member of the NF-[kappa]B family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nature Genetics *41*, 820-823.

Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A.*, et al.* (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature Genetics *44*, 1084-1089.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y.*, et al.* (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.

Halme, L., Paavola-Sakki, P., Turunen, U., Lappalainen, M., Farkkila, M., and Kontula, K. (2006). Family and twin studies in inflammatory bowel disease. World Journal of Gastroenterology *12*, 3668-3672.

Hampe, J., Cuthbert, A., Croucher, P.J.P., Mirza, M.M., Mascheretti, S., Fisher, S., Frenzel, H., King, K., Hasselmeyer, A., MacPherson, A.J.S.*, et al.* (2001). Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet *357*, 1925-1928.

Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J.*, et al.* (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nature Genetics *39*, 207-211.

Hanna, R.N., Carlin, L.M., Hubbeling, H.G., Nackiewicz, D., Green, A.M., Punt, J.A., Geissmann, F., and Hedrick, C.C. (2011). The transcription factor NR4A1 (Nur77) controls bone marrow differentiation and the survival of Ly6C- monocytes. Nature Immunology *12*, 778-785.

Hanson, B., McGue, M., Roitman-Johnson, B., Segal, N.L., Bouchard, T.J., Jr., and Blumenthal, M.N. (1991). Atopic disease and immunoglobulin E in twins reared apart and together. American Journal of Human Genetics *48*, 873-879.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research *22*, 1760-1774.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., *et al.* (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249-253.

Hiatt, R.A., and Kaufman, L. (1988). Epidemiology of inflammatory bowel disease in a defined northern California population. The Western Journal of Medicine *149*, 541-546.

Higgins, J.P., and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine *21*, 1539-1558.

Hindorff, L., Junkins, H., Hall, P., Mehta, J., and Manolio, T. (2014) A Catalog of Published Genome-Wide Association Studies (www.genome.gov/gwastudies)

Hirschfield, G.M., Liu, X., Xu, C., Lu, Y., Xie, G., Gu, X., Walker, E.J., Jing, K., Juran, B.D., Mason, A.L., *et al.* (2009). Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. New England Journal of Medicine *360*, 2544-2555.

Hoggart, C.J., Clark, T.G., De Iorio, M., Whittaker, J.C., and Balding, D.J. (2008). Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology *32*, 179-185.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Jr., Wright, M.W., *et al.* (2004). Gene map of the extended human MHC. Nature Reviews Genetics *5*, 889-899.

Hov, J.R., Kosmoliaptsis, V., Traherne, J.A., Olsson, M., Boberg, K.M., Bergquist, A., Schrumpf, E., Bradley, J.A., Taylor, C.J., Lie, B.A., *et al.* (2011). Electrostatic modifications of the human leukocyte antigen-DR P9 peptide-binding pocket and susceptibility to primary sclerosing cholangitis. Hepatology *53*, 1967-1976.

Hovhannisyan, Z., Weiss, A., Martin, A., Wiesner, M., Tollefsen, S., Yoshida, K., Ciszewski, C., Curran, S.A., Murray, J.A., David, C.S., *et al.* (2008). The role of HLA-DQ8 beta57 polymorphism in the anti-gluten T-cell response in coeliac disease. Nature *456*, 534-538.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics *5*, e1000529.

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. American Journal of Human Genetics *89*, 496-506.

Hu, Y.-J., Berndt, Sonja I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K.E., Ingelsson, E., and Lin, D.-Y. (2013). Meta-analysis of Gene-Level Associations

for Rare Variants Based on Single-Variant Statistics. American Journal of Human Genetics *93*, 236-248.

Huang, H., Chanda, P., Alonso, A., Bader, J.S., and Arking, D.E. (2011). Gene-based tests of association. PLoS Genetics *7*, e1002177.

Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M.*, et al.* (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599-603.

Hugot, J.P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J.M., Lee, J.C., Beaugerie, L., Naom, I., Dupas, J.L., Van Gossum, A., Orholm, M.*, et al.* (1996). Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature *379*, 821-823.

Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O.*, et al.* (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature *498*, 232-235.

Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D.*, et al.* (2008). Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics *40*, 395-402.

Hurst, L.D., Pal, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nature Reviews Genetics *5*, 299-310.

Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M., and Tuomilehto, J. (2003). Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs: A Nationwide Follow-Up Study. Diabetes *52*, 1052-1055.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

International HapMap Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F.*, et al.* (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52-58.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

International Multiple Sclerosis Genetics Consortium. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature Genetics *45*, 1353-1360.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752.

Invernizzi, P., Selmi, C., Poli, F., Frison, S., Floreani, A., Alvaro, D., Almasio, P., Rosina, F., Marzioni, M., Fabris, L.*, et al.* (2008). Human leukocyte antigen polymorphisms in Italian primary biliary cirrhosis: a multicenter study of 664 patients and 1992 healthy controls. Hepatology *48*, 1906-1912.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nature Genetics *29*, 306-309.

James, O.F., Bhopal, R., Howel, D., Gray, J., Burt, A.D., and Metcalf, J.V. (1999). Primary biliary cirrhosis once rare, now common in the United Kingdom? Hepatology *30*, 390-394.

Jawaheer, D., Seldin, M.F., Amos, C.I., Chen, W.V., Shigeta, R., Etzel, C., Damle, A., Xiao, X., Chen, D., Lum, R.F.*, et al.* (2003). Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. Arthritis and Rhematism *48*, 906-916.

Jones, D.E. (2003). Addison's other disease: primary biliary cirrhosis as a model autoimmune disease. Clinical medicine *3*, 351-356.

Jones, D.E., Watt, F.E., Metcalf, J.V., Bassendine, M.F., and James, O.F. (1999). Familial primary biliary cirrhosis reassessed: a geographically-based population study. Journal of Hepatology *30*, 402-407.

Jostins, L. (2012). Using Next-Generation Genomic Datasets In Disease Association (The University of Cambridge).

Jostins, L., and Barrett, J.C. (2011). Genetic risk prediction in complex disease. Human Molecular Genetics *20*, R182-188.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A.*, et al.* (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119-124.

Juyal, G., Negi, S., Sood, A., Gupta, A., Prasad, P., Senapati, S., Zaneveld, J., Singh, S., Midha, V., van Sommeren, S.*, et al.* (2014). Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. Gut *Published Online First 16 May 2014*.

Kamada, N., Seo, S.-U., Chen, G.Y., and Nunez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. Nature Reviews Immunology *13*, 321-335.

Kang, J., Kugathasan, S., Georges, M., Zhao, H., Cho, J.H., and NIDDK IBD Genetics Consotrium (2011). Improved risk prediction for Crohn's disease with a multi-locus approach. Human Molecular Genetics *20*, 2435-2442.

Kaplan, M.M., and Gershwin, M.E. (2005). Primary biliary cirrhosis. New England Journal of Medicine *353*, 1261-1273.

Karlsen, T.H., Boberg, K.M., Vatn, M., Bergquist, A., Hampe, J., Schrumpf, E., Thorsby, E., Schreiber, S., Lie, B.A., and Group, I.S. (2007). Different HLA class II associations in ulcerative colitis patients with and without primary sclerosing cholangitis. Genes and Immunity *8*, 275-278.

Karlsen, T.H., Franke, A., Melum, E., Kaser, A., Hov, J.R., Balschun, T., Lie, B.A., Bergquist, A., Schramm, C., Weismuller, T.J., *et al.* (2010a). Genome-wide association analysis in primary sclerosing cholangitis. Gastroenterology *138*, 1102-1111.

Karlsen, T.H., and Kaser, A. (2011). Deciphering the genetic predisposition to primary sclerosing cholangitis. Seminars in Liver Disease *31*, 188-207.

Karlsen, T.H., Schrumpf, E., and Boberg, K.M. (2010b). Update on primary sclerosing cholangitis. Digestive and Liver Disease *42*, 390-400.

Kasler, H.G., Young, B.D., Mottet, D., Lim, H.W., Collins, A.M., Olson, E.N., and Verdin, E. (2011). Histone deacetylase 7 regulates cell survival and TCR signaling in CD4/CD8 double-positive thymocytes. Journal of Immunology *186*, 4782-4793.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., *et al.* (2010). Variation in Transcription Factor Binding Among Humans. Science *328*, 232-235.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., *et al.* (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature Genetics *40*, 189-197.

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740-743.

Khor, B., Gardet, A., and Xavier, R.J. (2011). Genetics and pathogenesis of inflammatory bowel disease. Nature *474*, 307-317.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., *et al.* (2012). Exome sequencing and the genetic basis of complex traits. Nature Genetics *44*, 623-630.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., *et al.* (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science *308*, 385-389.

Korte, A., Vilhjalmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nature Genetics *44*, 1066-1071.

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics *11*, 241-247.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., *et al.* (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

Lasko, T.A., Bhagwat, J.G., Zou, K.H., and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics *38*, 404-415.

Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine *27*, 1133-1163.

Lazzeroni, L.C., Lu, Y., and Belitskaya-Levy, I. (2014). P-values in genomics: Apparent precision masks high uncertainty. Molecular Psychiatry *19*, 1336-1340.

Lee, J.C., Espeli, M., Anderson, C.A., Linterman, M.A., Pocock, J.M., Williams, N.J., Roberts, R., Viatte, S., Fu, B., Peshu, N., *et al.* (2013a). Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. Cell *155*, 57-69.

Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., *et al.* (2014). Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. Science *343*, 1246949.

Lee, S., Teslovich, Tanya M., Boehnke, M., and Lin, X. (2013b). General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. The American Journal of Human Genetics *93*, 42-53.

Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. American Journal of Human Genetics *88*, 294-305.

Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540-2542.

Leslie, S., Donnelly, P., and McVean, G. (2008). A statistical method for predicting classical HLA alleles from SNP data. American Journal of Human Genetics *82*, 48-56.

Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C., *et al.* (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. Nature Genetics *40*, 584-591.

Levison, S., Fisher, P., Hankinson, J., Zeef, L., Eyre, S., Ollier, W., McLaughlin, J., Brass, A., Grencis, R., and Pennock, J. (2013). Genetic analysis of the Trichuris muris-induced model of colitis reveals QTL overlap and a novel gene cluster for establishing colonic inflammation. BMC Genomics *14*, 127.

Levison, S.E., McLaughlin, J.T., Zeef, L.A., Fisher, P., Grencis, R.K., and Pennock, J.L. (2010). Colonic transcriptional profiling in resistance and susceptibility to trichuriasis: phenotyping a chronic colitis and lessons for iatrogenic helminthosis. Inflammatory Bowel Disease *16*, 2065-2079.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome Research *21*, 940-951.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annual Review of Genomics and Human Genetics *10*, 387-406.

Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., *et al.* (2014). Meta-analysis of gene-level tests for rare variant association. Nature Genetics *46*, 200-204.

Liu, J.Z., Almarri, M.A., Gaffney, D.J., Mells, G.F., Jostins, L., Cordell, H.J., Ducker, S.J., Day, D.B., Heneghan, M.A., Neuberger, J.M., *et al.* (2012). Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nature Genetics *44*, 1137-1141.

Liu, J.Z., Hov, J.R., Folseraas, T., Ellinghaus, E., Rushbrook, S.M., Doncheva, N.T., Andreassen, O.A., Weersma, R.K., Weismuller, T.J., Eksteen, B., *et al.* (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nature Genetics *45*, 670-675.

Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., *et al.* (2010a). A versatile gene-based test for genome-wide association studies. American Journal of Human Genetics *87*, 139-145.

Liu, X., Invernizzi, P., Lu, Y., Kosoy, R., Bianchi, I., Podda, M., Xu, C., Xie, G., Macciardi, F., Selmi, C., *et al.* (2010b). Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nature Genetics *42*, 658-660.

Loftus, E.V., Jr. (2004). Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology *126*, 1504-1517.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580-585.

Ma, C., Blackwell, T., Boehnke, M., Scott, L.J., and the Go, T.D.i. (2013). Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. Genetic Epidemiology *37*, 539-550.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18-21.

Major Depressive Disorder Working Group of the Psychiatric Genetics Consortium. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry *18*, 497-511.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823-826.

Mannon, P.J., Fuss, I.J., Mayer, L., Elson, C.O., Sandborn, W.J., Present, D., Dolin, B., Goodman, N., Groden, C., Hornung, R.L., *et al.* (2004). Anti–Interleukin-12 Antibody for Active Crohn's Disease. New England Journal of Medicine *351*, 2069-2079.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics *11*, 499-511.

Mathew, C.G., and Lewis, C.M. (2004). Genetics of inflammatory bowel disease: progress and prospects. Human Molecular Genetics *13 Spec No 1*, R161-168.

Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nature Genetics *44*, 243-246.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190-1195.

McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H., *et al.* (2008). Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nature Genetics *40*, 1107-1112.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science *304*, 581-584.

Mells, G.F., Floyd, J.A., Morley, K.I., Cordell, H.J., Franklin, C.S., Shin, S.Y., Heneghan, M.A., Neuberger, J.M., Donaldson, P.T., Day, D.B.*, et al.* (2011). Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nature Genetics *43*, 329-332.

Melum, E., Franke, A., Schramm, C., Weismuller, T.J., Gotthardt, D.N., Offner, F.A., Juran, B.D., Laerdahl, J.K., Labi, V., Bjornsson, E.*, et al.* (2011). Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. Nature Genetics *43*, 17-19.

Mirza, M.M., Lee, J., Teare, D., Hugot, J.P., Laurent-Puig, P., Colombel, J.F., Hodgson, S.V., Thomas, G., Easton, D.F., Lennard-Jones, J.E.*, et al.* (1998). Evidence of linkage of the inflammatory bowel disease susceptibility locus on chromosome 16 (IBD1) to ulcerative colitis. Journal of Medical Genetics *35*, 218-221.

Molodecky, N.A., Soon, I.S., Rabi, D.M., Ghali, W.A., Ferris, M., Chernoff, G., Benchimol, E.I., Panaccione, R., Ghosh, S., Barkema, H.W.*, et al.* (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology *142*, 46-54 e42; quiz e30.

Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.F., de Rijk, P., Dewit, O.*, et al.* (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nature Genetics *43*, 43-47.

Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genetic Epidemiology *35*, 809-822.

Morris, J.A., Randall, J.C., Maller, J.B., and Barrett, J.C. (2010). Evoker: a visualization tool for genotype intensity data. Bioinformatics *26*, 1786-1787.

Mullarkey, M.E., Stevens, A.M., McDonnell, W.M., Loubiere, L.S., Brackensick, J.A., Pang, J.M., Porter, A.J., Galloway, D.A., and Nelson, J.L. (2005). Human leukocyte antigen class II alleles in Caucasian women with primary biliary cirrhosis. Tissue Antigens *65*, 199-205.

Murthy, A., Li, Y., Peng, I., Reichelt, M., Katakam, A.K., Noubade, R., Roose-Girma, M., DeVoss, J., Diehl, L., Graham, R.R.*, et al.* (2014). A Crohn/'s disease variant in Atg16l1 enhances its degradation by caspase 3. Nature *506*, 456-462.

Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., Birney, E., Wold, B.*, et al.* (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biology *9*, e1001046.

Neale, B.M., and Sham, P.C. (2004). The future of association studies: gene-based analysis and replication. American Journal of Human Genetics *75*, 353-362.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387-389.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genetics *6*, e1000888.

Nisticò, L., Buzzetti, R., Pritchard, L.E., Van der Auwera, B., Giovannini, C., Bosi, E., Martinez Larrad, M.T., Serrano Rios, M., Chow, C.C., Cockram, C.S.*, et al.* (1996). The CTLA-4 Gene Region of Chromosome 2q33 Is Linked to, and Associated with, Type 1 Diabetes. Human Molecular Genetics *5*, 1075-1080.

Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H.*, et al.* (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature *411*, 603-606.

Ohmen, J.D., Yang, H.-Y., Yamamoto, K.K., Zhao, H.-Y., Ma, Y., Bentley, L.G., Huang, Z., Gerwehr, S., Pressman, S., McElree, C.*, et al.* (1996). Susceptibility Locus for Inflammatory Bowel Disease on Chromosome 16 has a Role in Crohn's disease, but Not in Ulcerative Colitis. Human Molecular Genetics *5*, 1679-1683.

Okada, H., Kuhn, C., Feillet, H., and Bach, J.F. (2010). The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clinical and Experimental Immunology *160*, 1-9.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S.*, et al.* (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376-381.

Okada, Y., Yamazaki, K., Umeno, J., Takahashi, A., Kumasaka, N., Ashikawa, K., Aoi, T., Takazoe, M., Matsui, T., Hirano, A.*, et al.* (2011). HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. Gastroenterology *141*, 864-871 e861-865.

Painter, J.N., Anderson, C.A., Nyholt, D.R., Macgregor, S., Lin, J., Lee, S.H., Lambert, A., Zhao, Z.Z., Roseman, F., Guo, Q.*, et al.* (2011). Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. Nature Genetics *43*, 51-54.

Park, J.-H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics *42*, 570-575.

Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G., Nimmo, E.R., Cummings, F.R., Soars, D.*, et al.* (2007). Sequence

variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nature Genetics *39*, 830-832.

Parkes, M., Cortes, A., van Heel, D.A., and Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature Reviews Genetics *14*, 661-673.

Parkes, M., Satsangi, J., and Jewell, D. (1998). Contribution of the IL-2 and IL-10 genes to inflammatory bowel disease (IBD) susceptibility. Clinical & Experimental Immunology *113*, 28-32.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genetics *2*, e190.

Peters, U., Hutter, C.M., Hsu, L., Schumacher, F.R., Conti, D.V., Carlson, C.S., Edlund, C.K., Haile, R.W., Gallinger, S., Zanke, B.W.*, et al.* (2012). Meta-analysis of new genome-wide association studies of colorectal cancer risk. Human Genetics *131*, 217-234.

Pickrell, Joseph K. (2014). Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. American Journal of Human Genetics *94*, 559-573.

Pirinen, M., Donnelly, P., and Spencer, C. (2012). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Annals of Applied Statistics *7*, 369-390.

Plenge, R.M., Scolnick, E.M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. Nature Reviews Drug Discovery *12*, 581-594.

Podda, M., Selmi C., Lleo, A., Moroni, L., and Invernizzi, P. (2013) The limitations and hidden gems of the epidemiology of primary biliary cirrhosis. Journal of Autoimmunity *46*, 81-87.

Prescott, N.J., Dominy, K.M., Kubo, M., Lewis, C.M., Fisher, S.A., Redon, R., Huang, N., Stranger, B.E., Blaszczyk, K., Hudspith, B.*, et al.* (2010). Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. Human Molecular Genetics *19*, 1828-1839.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics *38*, 904-909.

Prideaux, L., Kamm, M.A., De Cruz, P.P., Chan, F.K., and Ng, S.C. (2012). Inflammatory bowel disease in Asia: a systematic review. Journal of Hastroenterology and Hepatology *27*, 1266-1280.

Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149-150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J*., et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics *81*, 559-575.

Raine, T., Liu, J.Z., Anderson, C.A., Parkes, M., and Kaser, A. (2014). Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. Gut *Published Online First 5 May*.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., and Daly, M.J. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genetics *5*, e1000534.

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M.J., Zhang, W., Below, J.E., *et al.* (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature Genetics *46*, 234-244.

Ricard Cervera, J.C.R., and Munther, K. (2009). Handbook of Systemic Autoimmune Diseases. In Handbook of Systemic Autoimmune Diseases, J.C.R. Ricard Cervera, and K. Munther, eds. (Elsevier), p. ii.

Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W*., et al.* (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature Genetics *39*, 596-604.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N*., et al.* (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature Genetics *43*, 1066-1073.

Robinton, D.A., and Daley, G.Q. (2012). The promise of induced pluripotent stem cells in research and therapy. Nature *481*, 295-305.

Roman, A.L., and Munoz, F. (2011). Comorbidity in inflammatory bowel disease. World Journal of Gastroenterology *17*, 2723-2733.

Rong, G., Zhou, Y., Xiong, Y., Zhou, L., Geng, H., Jiang, T., Zhu, Y., Lu, H., Zhang, S., Wang, P*., et al.* (2009). Imbalance between T helper type 17 and T regulatory cells in patients with primary biliary cirrhosis: the serum cytokine profile

and peripheral cell population. Clinical and Experimental Immunology *156*, 217-225.

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., and Daly, M.J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genetics *7*, e1001273.

Ruderfer, D.M., Korn, J., and Purcell, S.M. (2010). Family-based genetic risk prediction of multifactorial disease. Genome Medicine *2*, 2.

Saarinen, S., Olerup, O., and Broome, U. (2000). Increased frequency of autoimmune diseases in patients with primary sclerosing cholangitis. The American journal of gastroenterology *95*, 3195-3199.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L.*, et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-933.

Saich, R., and Chapman, R. (2008). Primary sclerosing cholangitis, autoimmune hepatitis and overlap syndromes in inflammatory bowel disease. World Journal of Gastroenterology *21*, 331-337.

Sakoda, L.C., Jorgenson, E., and Witte, J.S. (2013). Turning of COGS moves forward findings for hormonally mediated cancers. Nature Genetics *45*, 345-348.

Sanna, S., Jackson, A.U., Nagaraja, R., Willer, C.J., Chen, W.M., Bonnycastle, L.L., Shen, H., Timpson, N., Lettre, G., Usala, G.*, et al.* (2008). Common variants in the GDF5-UQCC region are associated with variation in human height. Nature Genetics *40*, 198-203.

Sanseau, P., Agarwal, P., Barnes, M.R., Pastinen, T., Richards, J.B., Cardon, L.R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. Nature Biotech *30*, 317-320.

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E.*, et al.* (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214-219.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C.*, et al.* (2008). Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biology *6*, e107.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Research *22*, 1748-1759.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421-427.

Shah, T.S., Liu, J.Z., Floyd, J.A., Morris, J.A., Wirth, N., Barrett, J.C., and Anderson, C.A. (2012). optiCall: A robust genotype-calling algorithm for rare, low frequency and common variants. Bioinformatics *28*, 1598-1603.

Sheldon, S., and Poulton, K. (2006). HLA Typing and Its Influence on Organ Transplantation. In Transplantation Immunology, P. Hornick, and M. Rose, eds. (Humana Press), pp. 157-174.

Sen, P.K., (1960). On some convergence properties of U-statistics. Calcutta Statistical Association Bulletin *10*, 1-18.

Sherry, S.T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Research *9*, 677-679.

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J.*, et al.* (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature *460*, 753-757.

Shivananda, S., Lennard-Jones, J., Logan, R., Fear, N., Price, A., Carpenter, L., and van Blankenstein, M. (1996). Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). Gut *39*, 690-697.

Sironi, M., and Clerici, M. (2010). The hygiene hypothesis: an evolutionary perspective. Microbes and Infection *12*, 421-427.

So, H.C., Gui, A.H., Cherny, S.S., and Sham, P.C. (2011). Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genetic Epidemiology *35*, 310-317.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics *14*, 483-495.

Somers, E.C., Thomas, S.L., Smeeth, L., and Hall, A.J. (2009). Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder? American Journal of Epidemiology *169*, 749-755.

Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.K., Sheffield, N.C., Graf, S., Huss, M., Keefe, D.*, et al.* (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Research *21*, 1757-1767.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J.a., Magi, R.*, et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics *42*, 937-948.

Spiegelman, B.M. (1998). PPAR-gamma: adipogenic regulator and thiazolidinedione receptor. Diabetes *47*, 507-514.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics *52*, 506-516.

Srivastava, B., Mells, G.F., Cordell, H.J., Muriithi, A., Brown, M., Ellinghaus, E., Franke, A., Consortium, U.P., Karlsen, T.H., Sandford, R.N.*, et al.* (2012). Fine mapping and replication of genetic risk loci in primary sclerosing cholangitis. Scandinavian Journal of Gastroenterology *47*, 820-826.

Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A.*, et al.* (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nature Genetics *42*, 508-514.

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A.*, et al.* (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nature Genetics *46*, 294-298.

Stokkers, P.C., Reitsma, P.H., Tytgat, G.N., and van Deventer, S.J. (1999). HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. Gut *45*, 395-401.

Strange, A., Capon, F., Spencer, C.C., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., Bellenguez, C., Bergboer, J.G.*, et al.* (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nature Genetics *42*, 985-990.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Guitierrez-Arcelus, M. (2012). Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS Genetics *8*, e1002639.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545-15550.

Summers, R.W., Elliott, D.E., Urban, J.F., Jr., Thompson, R.A., and Weinstock, J.V. (2005a). Trichuris suis therapy for active ulcerative colitis: a randomized controlled trial. Gastroenterology *128*, 825-832.

Summers, R.W., Elliott, D.E., Urban, J.F., Thompson, R., and Weinstock, J.V. (2005b). Trichuris suis therapy in Crohn's disease. Gut *54*, 87-90.

Syvanen, A.-C. (2005). Toward genome-wide SNP genotyping. Nature Genetics *37*.

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P., and Clark, T.G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics *23*, 2741-2746.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J.*, et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707-713.

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G.*, et al.* (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics *43*, 1193-1201.

Trynka, G., and Raychaudhuri, S. (2013). Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. Current Opinion in Genetics & Development *23*, 635-641.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics *45*, 124-130.

Tsoi, L.C., Spain, S.L., Knight, J., Ellinghaus, E., Stuart, P.E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J.E.*, et al.* (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nature Genetics *44*, 1341-1348.

Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N.*, et al.* (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science *230*, 1054-1057.

Vermeire, S., Wild, G., Kocher, K., Cousineau, J., Dufresne, L., Bitton, A., Langelier, D., Pare, P., Lapointe, G., Cohen, A.*, et al.* (2002). CARD15 Genetic Variation in a Quebec Population: Prevalence, Genotype-Phenotype Relationship, and Haplotype Structure. The American Journal of Human Genetics *71*, 74-83.

Veyrieras, J., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genetics *4*, e1000214.

Visscher, P.M., Brown, Matthew A., McCarthy, Mark I., and Yang, J. (2012). Five Years of GWAS Discovery. The American Journal of Human Genetics *90*, 7-24.

Visscher, P.M., Medland, S.E., Ferreira, M.A.R., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W., and Martin, N.G. (2006). Assumption-Free Estimation of

Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genetics *2*, e41.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J.*, et al.* (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genetics *8*, e1002793.

Wallace, C., Smyth, D.J., Maisuria-Armer, M., Walker, N.M., Todd, J.A., and Clayton, D.G. (2010). The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nature Genetics *42*, 68-71.

Wang, D., Zhang, H., Liang, J., Gu, Z., Zhou, Q., Fan, X., Hou, Y., and Sun, L. (2010a). CD4+CD25+ but not CD4+Foxp3+ T cells as a regulatory subset in primary biliary cirrhosis. Cellular and Molecular Immunology *7*, 485-490.

Wang, K., Li, M., and Hakonarson, H. (2010b). Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics *11*, 843-854.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G., Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics *6*, 109-118.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research *40*, D930-934.

Wassmuth, R., Depner, F., Danielsson, A., Hultcrantz, R., Loof, L., Olson, R., Prytz, H., Sandberg-Gertzen, H., Wallerstedt, S., and Lindgren, S. (2002). HLA class II markers and clinical heterogeneity in Swedish patients with primary biliary cirrhosis. Tissue Antigens *59*, 381-387.

Weber, J.L., and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. American Journal of Human Genetics *44*, 388-396.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population-Structure. Evolution *38*, 1358-1370.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

Wen, L., Ley, R.E., Volchkov, P.Y., Stranges, P.B., Avanesyan, L., Stonebraker, A.C., Hu, C., Wong, F.S., Szot, G.L., Bluestone, J.A.*, et al.* (2008). Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature *455*, 1109-1113.

Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E.*, et al.* (2013).

Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics *45*, 1238-1243.

Wiencke, K., Karlsen, T.H., Boberg, K.M., Thorsby, E., Schrumpf, E., Lie, B.A., and Spurkland, A. (2007). Primary sclerosing cholangitis is associated with extended HLA-DR3 and HLA-DR6 haplotypes. Tissue Antigens *69*, 161-169.

Willer, C.J., Dyment, D.A., Risch, N.J., Sadovnick, A.D., Ebers, G.C., and Group, T.C.C.S. (2003). Twin concordance and sibling recurrence rates in multiple sclerosis. Proceedings of the National Academy of Sciences *100*, 12877-12882.

Wray, N.R., Purcell, S.M., and Visscher, P.M. (2011). Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results. PLoS Biology *9*, e1000579.

Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genetics *6*, e1000864.

Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T.*, et al.* (2005). Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Human Molecular Genetics *14*, 3499-3506.

Yamazaki, K., Umeno, J., Takahashi, A., Hirano, A., Johnson, T.A., Kumasaka, N., Morizono, T., Hosono, N., Kawaguchi, T., Takazoe, M.*, et al.* (2013). A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population. Gastroenterology *144*, 781-788.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W.*, et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics *42*, 565-569.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of, A.T.C., Replication, D.I.G., Meta-analysis, C., Madden, P.A., Heath, A.C., Martin, N.G.*, et al.* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics *44*, 369-375, S361-363.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014a). Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics *46*, 100-106.

Yang, S.K., Hong, M., Zhao, W., Jung, Y., Baek, J., Tayebi, N., Kim, K.M., Ye, B.D., Kim, K.J., Park, S.H.*, et al.* (2014b). Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut *63*, 80-87.

Zaiss, Dietmar M.W., van Loosdregt, J., Gorlani, A., Bekker, Cornelis P.J., Gröne, A., Sibilia, M., van Bergen en Henegouwen, Paul M.P., Roovers, Rob C., Coffer, Paul J., and Sijts, Alice J.A.M. (2013). Amphiregulin Enhances Regulatory T Cell-Suppressive Function via the Epidermal Growth Factor Receptor. Immunity *38*, 275-284.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., Freathy, R.M*., et al.* (2007). Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. Science *316*, 1336-1341.

Zhang, K., Cui, S., Chang, S., Zhang, L., and Wang, J. (2010). i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res *38*, W90-W95.

Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G.F., Franke, L., Oosting, M., Barisani, D*., et al.* (2010). Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. American Journal of Human Genetics *86*, 970-977.

Zhernakova, A., van Diemen, C.C., and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature Reviews Genetics *10*, 43-55.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. Proceedings of the National Academy of Sciences *111*, E455-E464.