Using Next-Generation Genomic Datasets In Disease Association

Luke Jostins

King's College
University of Cambridge

September 2012

This dissertation is submitted for
the degree of Doctor of Philosophy

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 Attributions section or the text. It does not exceed the word limit set out 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 for any degree, diploma or any other qualification.

Luke Jostins
18^{th} September 2012

This is a post-viva dissertation, containing some minor corrections to that submitted on 18^{th} September 2012. The corrections were suggested by Stephen Sawcer and Peter Donnelly.

Luke Jostins
24^{th} January 2012

How are you going to be successful in treatment, if you do not understand the real essence of each disease?

— Galen, On Natural Faculties

Using Next-Generation Genomic Datasets In Disease Association

Luke Jostins
King's College, Cambridge
Ph.D Thesis, September 2012

Abstract

The first generation of genome-wide association studies (GWAS) uncovered thousands of genetic risk factors for hundreds of complex human diseases. However, over the past five years new high-throughput techniques, including next-generation sequencing and low-cost custom genotyping, have allowed us to expand disease association studies into larger sample sizes and across the entire spectrum of human variation. This thesis will explore the potential of these new technologies, and in particular their application to the study of Inflammatory Bowel Disease (IBD) genetics. After reviewing the historical context of complex disease genetics, I introduce the statistical methods and models used in this thesis, and demonstrate how they can be placed into a unified framework of genetic risk models. I then detail three analysis projects that focus on identifying risk variants that the first generation of GWAS was unable to study. The first investigates how genotype imputation, coupled with high-density sequencing reference sets, can aid locus discovery in both European and African populations. The second discusses the use of a custom genotyping chip (the Immunochip) to discover risk variants with low effect sizes, by allowing low-cost genotyping of a very large number of samples. The third investigates the use of next-generation sequencing of multiply affected (or "multiplex") families in order to identify low-frequency, high penetrance risk alleles. Throughout these three projects I describe the discovery of a large number of novel IBD risk loci, and discuss how statistical and biological interrogation of these risk loci can help us to develop and expand biological hypotheses.

Acknowledgements

First and foremost I would like to thank my adviser Jeff Barrett for more things than I could reasonably list here. Needless to say, his dedication to statistical rigour, free sharing of data, clear scientific writing and public communication have provided me with a model scientist that I will always seek to emulate.

I would like express my gratitude to my other advisers, thesis committee members and first year examiners, including Richard Durbin, Julian Parkhill, Chris Tyler-Smith and Ralph McGinnis. I would especially like to thank my external adviser John Todd for the confidence he has shown in me over the last three years.

I would also like to thank the rest of Medical Genomics/ Barrett Group/ Statistical and Computational Genetics/ Team 143 (we'll settle on a name one day): Kate Morley, for teaching me most of what I know, James Morris, for providing the perfect combination of well-planned software development, wry humour and cake, Yang Luo for Chinese sweets and old-school Cambridge amiability (dan shi...), Iris Kolder for her irrepressible sense of mischief and Isabelle Claynen for her infectious passion for biology.

I'd also like to thank all the people that I have got to know during my time at Sanger: Annabel Smith, Christina Hedberg-Delouka and Alex Bateman for running such a tight ship. The Anderson Group, including Carl Anderson, Jimmy Liu and Jamie Floyd, for tea and laughs. Daniel MacArthur for acting as my perennial foil, and Liz Murchison for being a true friend. Sanger's army of pipeline developers past and present, including Thomas Keane, Shane McCarthy, Petr Danecek, Jim Stalker, Josh Randall and Martin Pollard, for keeping my life (and the lives of hundreds of others) running smoothly, Chloe Noble and Anja Kolb-Kokocinski for somehow keeping Human Genetics from constantly falling over. Anne Wombwell, for taking the pain out of conference travel. All the other people who have made Sanger what is it for me, including (but definitely not limited to) Nicole Soranzo, Aaron Day-Williams, Aarno Palotie, Dan Gaffney, Matt Hurles and Helena Kilpinen.

This thesis would not have been possible without my collaborators, and I would like to thank them all: The clinicians and analysts of the UKIBDGC, for making me feel part of something important, including Miles Parks, Charlie Lees, James Lee, Tim Raine, John Mansfield, Christopher Matthew and Jack Satsangi. The many researchers of the IIBDGC, for showing me just how big "Big Science" can get, including Judy Cho, Andre Franke, Rinse Weersma, Sarah West, Yashoda Sharma, Phil Schumm, Kaida Ning, Severine Vermeire, and Dermot McGovern. All the people that make Boston a powerhouse of genetic analysis, including Stephan Ripke, Johan Essers, Hailiang Huang, Mark Daly, Soumya Raychaudhuri, Xinli Hu and Lizzy Rossin. Closer to home, the statistical know-how of the Diabetes and Inflammation Laboratory, including David Clayton and Chris Wallace. The dedication of the MalariaGEN researchers, Chris Spencer, Gavin Band, Katja Kivinen, Kirk Rockett, Quang Li and Dominic Kwiatkowski, for always doing the right thing, instead of the easy thing. I'd especially like to thank Adam Levine and Tony Segal from UCL for giving me a taste of real biology.

I would like to thank my fellow members of Genomes Unzipped, including Joe Pickrell, Don Conrad, Dan Vorhaus, Caroline Wright and Vincent Plagnol, for fighting all the good fights we so love to fight. And the many people I have met through blogging and twitter for drawing me into their community, and in particular Chris Gunter, Zoe McDougall and Razib Khan for all their support.

I'd also like to thank Daniel Stretch and Stephen Eglen from the Cambridge Computational Biology Institute, and Jules Griffin and Sarah Lummis from King's College, for letting me loose on five years worth of Cambridge students. And I'd like to thank everyone at King's, for taking in an overconfident Physics student just out of A-levels and producing an overconfident geneticist eight years later.

I'd like to thank all of my friends in Cambridge who have been so important to forming how I think about the world, and my parents for putting me on it the first place. I'd also like to thank Hannah "Pixie" Price, for everything.

I would like to thank the research bodies who have funded the research in the thesis, including the Wellcome Trust, the Medical Research Council, the National Association for Colitis and Crohns disease, the Charles Wolfson Charitable Trust and the Gates Foundation, as well as the many international funders who supported the IIBDGC. I'd also like to thank Gil McVean, Anne Pratt, Ali Momin, Helen Thompson and the Wellcome Trust for ensuring that I have somewhere to go next.

Last, but certainly not least, I would like to thank the 104,341 (give or take a few) patients and volunteers who donated their DNA to all the studies that make up this thesis.

Attributions

Many of the projects described in this thesis are of a collaborative nature, and many were performed as part of large, international consortia. Below is a summary of the contributions of other scientists to the work described in this thesis.

Chapter 3

The African data used for testing imputation were generated as part of the MalariaGEN Consortial Project 1 (CP1). Samples were collected by partners from 12 centres, and full information on all sample collections is available from `http://www.malariagen.net/projects/cp1`. Genotyping was carried out at the Wellcome Trust Sanger Institute. Pre-imputation data processing, quality control and calling of genotypes was carried out by Kirk Rockett, Katja Kivinen, Gavin Band, Si Quang Le and Chris Spencer.

The list of loss-of-function (LoF) variants was generated by the 1000 Genomes LoF Group (now Functional Integration Group). The list of high quality, polymorphic LoF variants was generated by Daniel MacArthur.

Chapter 4

The samples used in the combined GWAS-Immonchip analysis were collected by the research groups of the International IBD Genetics Consortium (`http://www.ibdgenetics.org/groups.html`). Genotyping was performed across multiple centres, summarised in Table 4.8.

Quality control, genotype imputation and association testing for the

GWAS collections were performed by Stephan Ripke. The optiCall genotyping program used to call Immunochip samples was developed by Tejas Shah and Carl Anderson. The tag SNP meta-analysis technique was developed in collaboration with Stephan Ripke and Mark Daly, and implemented by Stephan Ripke. The phenotype likelihood modelling approach was developed and implemented by Jonah Essers.

The DAPPLE analysis was carried out by Stephan Ripke and Lizzie Rossin. The eQTL and cSNP analyses were carried out by Kaida Ning. The immune cell expression enrichment analysis was carried out by Xinli Hu and Soumya Raychaudhuri, and the gene expression network analysis was carried out by Ken Hui and Eric Shadt.

Immunochip cluster plots were manually inspected by Mitja Mitrovic, Jeff Barrett, Carl Anderson, Emilie Theatre, Tobias Balschun, Sarah West, Kaida Ning, Zhi Wei, Karin Fransen, Kyle Bailey, Isabelle Cleynen, Suzanne van Sommeren, Philippe Goyette, Sok Meng Evelyn Ng and Martin Ladouceur.

Chapter 5

All research described in Chapter 5 was carried in close collaboration with Adam Levine. There is no single element of this project that I carried out entirely without Adam's input, and virtually all of Section 5.4 was carried out jointly.

However, the following were carried out by Adam with little or no input from me:

- The collection of all samples and phenotype data
- The processing and analysis of the gene expression data
- The calling and quality control of genome-wide genotype data
- The calling of SNPs and indels in the exome data

Gavin Sewell selected the SNPs used to test the load of common IBD risk variants in the family. CytoSNP 12 genotyping was carried out at University College, London, and Sequenom genotyping, whole-genome sequencing and whole-exome sequencing was carried out at the Wellcome Trust Sanger Institute. QC of the exomes was carried out by Martin Pollard.

Chapter 6

The whole genome sequencing experiment was designed in collaboration with the UK IBD Genetics Consortium, who provided the samples. Sample selection was performed by James Lee. Sequencing was carried out at the Wellcome Trust Sanger Institute, and sequence data was processed and quality controlled by Martin Pollard and Josh Randall. Variant calling, imputation, stringent sample and variant QC and association analysis were carried out by Yang Luo.

Publications

From this Dissertation

- Band, G., Quang, S. L., **Jostins, L.**, Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., *et al* (2012) Imputation based meta-analysis of severe malaria in three African populations. (Under review).

- **Jostins, L.**, Ripke, S., Weersma, R., Duerr, R. H., McGovern, D. P, Hui, K. Y., Lee, J. C., *et al* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease *Nature* **491**(7422):119-124.

- MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., **Jostins, L.**, *et al* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**(6070):823-828

- **Jostins, L.** and Barrett, J. C. (2011) Genetic risk prediction in complex disease. *Hum Mol Genet.* **20** (R2):R182-8.

- **Jostins, L.**, Morley, K. I. and Barrett, J. C. (2011) Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet.* **9**(6):662-666.

Arising elsewhere

- **1000 Genomes Project Consortium** (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422):56-65.

- Liu, J. Z., Almarri, M. A., Gaffney, D. J., Mells, F. G., **Jostins, L.**, Cordell, H. J., Ducker, S. J., *et al* (2012) Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis *Nat Genet.* **44**(10):1137-1141.

- **Jostins, L.** (2012) Dispatches from the functional phase of genome biology *Genome Biol.* **13**(6):316.

- **Jostins, L.**, Pickrell, J. K., Macarthur, D. G. and Barrett J. C. (2012) Misuse of hierarchical linear models overstates the significance of a reported association between OXTR and prosociality. *Proc Natl Acad Sci USA* **109**(18):E1048.

- Sewell, G. W., Rahman, F. Z., Levine, A. P., **Jostins, L.**, Smith, P. J., Walker, A. P., Bloom, S. L., Segal, A. W. and Smith, A. M. (2012) Defective tumor necrosis factor release from Crohn's disease macrophages in response to toll-like receptor activation: Relationship to phenotype and genome-wide association susceptibility loci. *Inflamm Bowel Dis.* **8**(11):2120-2127.

- Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., ***et al*** (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* **42**(12):1118-1125.

- **1000 Genomes Project Consortium** (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061-1073.

- Krawitz, P., Rödelsperger, C., Jäger, M., **Jostins, L.**, Bauer, S. and Robinson, P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics* **26** (6):722-729.

Contents

Chapter 1

Introduction and Background

1.1 Justifications of phenotype and approach

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

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

1.1.1 Why study inflammatory bowel disease?

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

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

It is a poorly understood disease with a high burden

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

As well as increased mortality, IBD is a life-long disease that is diagnosed early in life (mean age of diagnosis is 27). 40-50% of patients will require surgery within 10 years of diagnosis, and most will require drug therapy throughout their lives (Bernstein, 2011). As well as the costs in human suffering, it is estimated that in Europe each patient with Crohn's disease costs €2898-6960 in direct health costs, and a total of up to €16.7 billion per year in economic costs (Yu et al., 2008).

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

It is highly heritable, and well characterised via twin studies

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

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

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

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

It has been well studied by linkage and GWAS

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

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

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

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

1.1.2 Why study complex disease genetics?

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

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

To directly understand biology

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

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

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

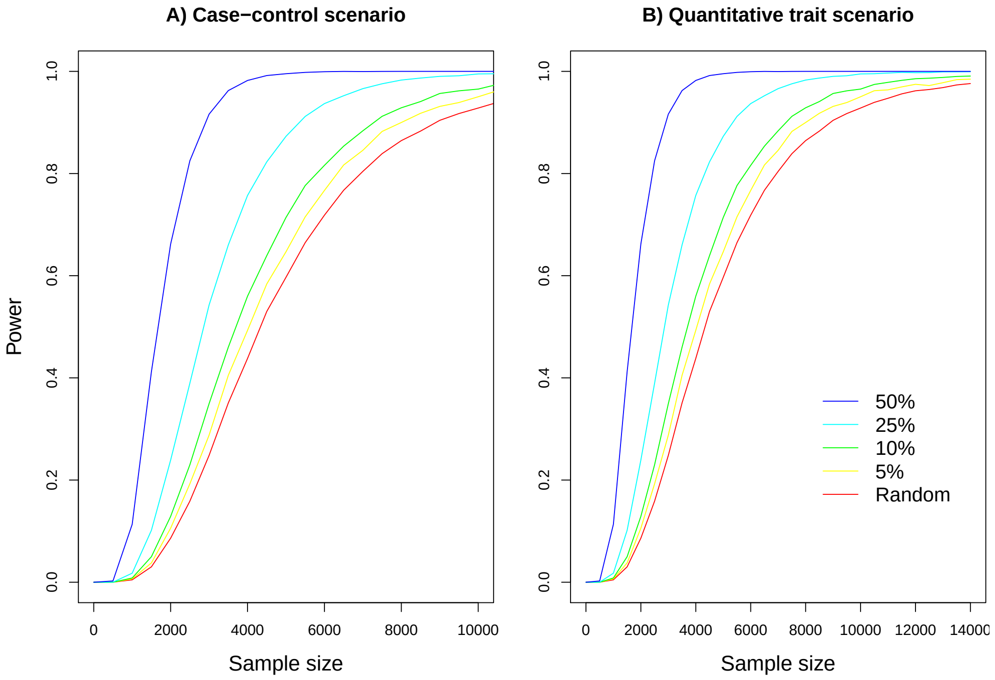


Figure 1.1: Improvement in power curves gained by prioritising samples based on genetic risk scores with different predictive powers. The colour of the line represents the proportion of total variance captured by the risk score, with the red line representing a random (i.e. non-prioritised) selection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

To facilitate further research

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

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

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

To predict disease genetically

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

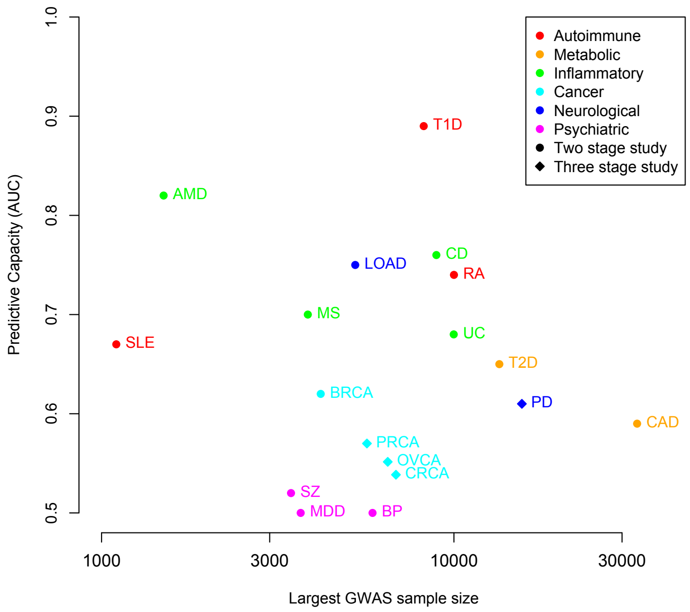


Figure 1.2: The predictive accuracy of variants discovered by genome-wide association studies, as a function of the effective sample size ($= \frac{2}{1/N_{case}+1/N_{control}}$), adjusted for the number of stages in the study (three stage studies have a smaller fraction of samples with GWAS data, and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

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

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

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

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

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

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

1.2 A brief history of human disease genetics

1.2.1 The age of molecular disease: 1940 to 1980

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

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

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

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

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

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

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

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

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

Linkage studies have a sophisticated statistical heritage. In the 1930s both Haldane (1934) and Fisher (1935) described statistical methods for detecting genetic linkage between dominant traits. Linkage has undergone constant statistical refinement for over half a century, with the development of the parametric LOD score (Morton, 1955), pedigree likelihood modelling (Elston and Stewart, 1971), the multipoint Lander-Green algorithm (Lander and Green, 1987), Non-Parametric Linkage (Kruglyak et al., 1996) and the development of sparse gene flow trees (Abecasis et al., 2002). Each of these statistical developments has been in response to the development of linkage from small-scale breeding experiments to massive whole-genome meta-analyses with hundreds of markers and thousands of individuals.

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

Technological revolutions during the 1970s provided a platform for linkage studies of human disease to come of age. This began with the development of amplification in DNA within viral or bacterial vectors (Jackson et al., 1972), and developed rapidly to sequencing of entire genes by dye termination (Sanger) sequencing (Sanger et al., 1977). These developments meant that, if the location of a gene could be identified, it could theoretically lead to the gene being cloned and sequenced, its protein sequence determined and its tissue expression distribution characterised. The development of the Southern blot during the same period (Southern, 1975) allowed easy genotyping of RFLPs (variants in the DNA that interfered with the action of restriction enzymes). This was the first time that genotypes could be efficiently measured from DNA itself, and led to the development of human linkage maps without the need for mutation phenotypes (Botstein et al., 1980). Suddenly, discovering disease loci by linkage became both possible, and potentially highly biologically informative.

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

The number of samples and variants typed in these early studies were counted in double digits, and the researchers only managed to discover mutations with extremely high penetrance in diseases with simple genetic architecture. The methods used to solve these diseases required monumental effort to use, and seem primitive and laborious by modern standards. However, in other ways they contained many of the essential principles of modern genetics. They used direct typing of DNA, without requiring any prior knowledge of the disease biology, to uncover disease loci. They utilised state-of-the-art technology, combined with rigorous statistical analysis, and in many cases shared data, samples and expertise across large, international consortia.

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

1.2.3 The beginning of complex disease genetics: 1994-2005

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

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

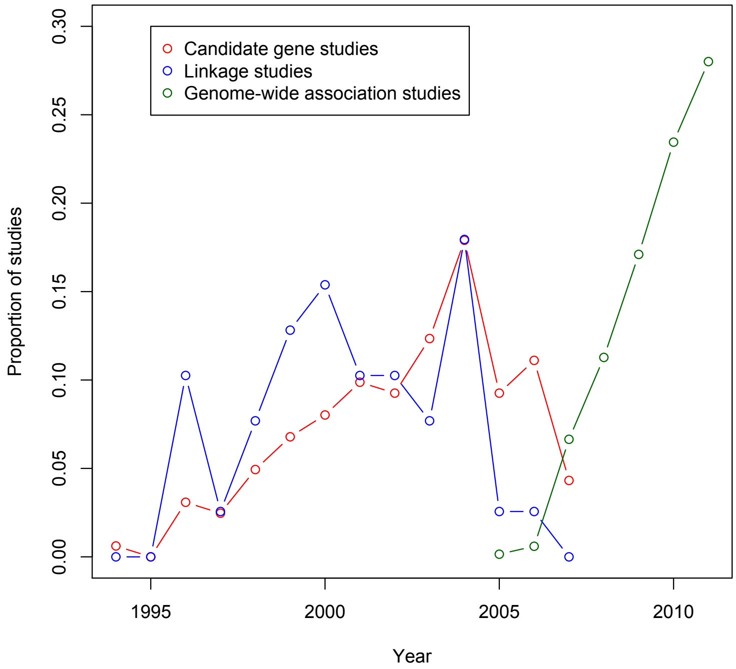


Figure 1.3: A timeline of complex disease genetics. Candidate gene studies (prior to 2007) are taken from reviews by Bosker et al. (2011) and Morgan et al. (2007). Linkage studies (prior to 2007) are taken from reviews by Guan et al. (2008) and Baumgart and Carding (2007). GWAS taken form the NHGRI GWAS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

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

In 1986, Kary Mullis and colleagues published the polymerase chain reaction (PCR), a method for rapidly amplifying specific DNA sequences in vivo (Mullis et al., 1986). This revolutionised the study of DNA. In 1989, Variable Number Tandem Repeats (VNTRs) were described as a class of variant easily genotyped by PCR (Weber and May, 1989), and linkage maps based on VNTRs appeared soon after. Additionally, by 1993 the TaqMan system was being used to genotype SNPs and small indels using PCR (Lee et al., 1993). These new techniques allowed genotyping of denser maps, in many more samples, at much lower cost than the old techniques. As well as making studies into many more Mendelian diseases affordable, this new technology also drove an explosion of studies into the genetics of complex disease, including both genome-wide linkage studies and association studies of candidate genes (see Figure 1.3).

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

Despite these successes, however, many of the linkage peaks discovered were sporadic, and could not be consistently replicated. Even more disappointing was the failure of linkage meta-analysis. The Genome Search Meta-analysis (GSMA) method (Wise et al., 1999) was introduced in 1999 to allow the results of linkage scans to be combined without sharing genotyping data, and theory created the possibility of very highly powered linkage studies. However, when the large linkage meta-analyses arrived, including thousands of affected families and representing millions of dollars of total investment, they produced almost no significant, novel results (van Heel et al., 2004; Guan et al., 2008; Concannon et al., 2009).

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

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

The reasons for this failure have been widely debated. The use of post-hoc adjustment to push p-values into nominal significance has been suggested (as has been demonstrated in other fields (Masicampo and Lalande, 2012)), often with an implication that this is a result of "hypothesis driven" investigators pushing their pet gene. However, I believe that most of the failure of candidate gene studies follows naturally from the sample size, p-value thresholds and the (then unknown) distribution of effect sizes in truly associated loci.

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

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

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

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

In 1986, a description of the first automated DNA sequencing machine was published (Smith et al., 1986). This machine used 4-colour dye termination, separated fragments through gel electrophoresis and imaged them digitally. It was commercialised as the ABI 370-series, and at its peak a single machine could produce 7200 bp (base pairs) of sequence per hour (Dovichi, 1997). In 1996 ABI released its first capillary sequencing machine, the ABI 310, followed two years later by the 96-capillary ABI 3700-series, capable of producing approximately 80kbp of sequence per hour (Dovichi, 1997). This was the technology that drove the sequencing of the human genome, the first full drafts of which were published in 2001 (Lander et al., 2001; Venter et al., 2001).

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

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

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

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

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

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

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

However, while the early days of GWAS were characterised by dramatic successes, they also suffered some teething troubles, driven mostly by a lack of a standardised GWAS protocol. For instance, in 2006 a genome-wide study of 649 individuals reported an association between a variant in the gene *INSIG2* (Herbert et al., 2006) and childhood obesity. This association did not meet the modern definition of "genome-wide significant" (GWS) ($p < 5 \times 10^{-8}$), and reports soon came in that the association did not replicate in independent cohorts (Dina et al., 2007; Loos et al., 2007; Rosskopf et al., 2007). Another early GWAS reported an association between memory performance and a variant in the gene *KIBRA* that did not meet genome-wide significance (Papassotiropoulos et al., 2006), which itself spawned a series of contradictory and inconclusive candidate gene studies (Schaper et al., 2008; Need et al., 2008; Bates et al., 2009) (exactly the situation GWAS was designed to prevent). Other early genome-wide association studies employed statistical techniques that seem somewhat unusual by modern standards (e.g. Liu et al. (2006)).

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

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

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

Over the last five years the number of GWAS per year has increased linearly (Figure 1.3). As the number of association studies increased, the next logical step was to combine studies together into meta-analyses (as was done during the linkage era). Early GWAS meta-analyses often consisted of pairwise collaborations, such as Samani et al. (2007), and often did not produce many more significant hits than the original GWAS. However, meta-analyses soon started producing startling results. The first Crohn's disease meta-analysis, consisting of three studies, discovered 21 new loci, bringing the total to 30 (Barrett et al., 2008) (more than the entire WTCCC), and the type 2 diabetes meta-analysis discovered six new loci for the previously very hard to crack disease (Zeggini et al., 2008). In 2009 the type 1 diabetes meta-analysis broke the record for the disease with the largest number of associations, with 40 loci (Barrett et al., 2009a), topped by the 71 Crohn's disease loci in 2010 (Franke et al., 2010). For almost all diseases studies, the majority of associations now came from large consortium meta-analyses.

1.2.6 Technological advances post-GWAS: 2004-Present

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

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

This technology spawned a new breed of systematic resequencing studies of human reference populations. In 2007 the 1000 Genomes Project was founded, to perform low-coverage (2-4X) sequencing on thousands of human genomes. The project started with a pilot that detected 16M SNPs, indels and structural variants in 180 HapMap samples (Project, 2010). The full project will eventually sequence 2500 individuals from 25 populations, with the first phase producing calls for nearly 40M variants across 1092 individuals (Project, 2012). Unlike the HapMap, this dataset is a near-complete map of genetic variation in these samples, including all common SNPs and indels genotyped in all individuals, as well as an extensive catalogue of low frequency variation.

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

Other technological advances in sequencing followed these developments. In 2007 NimbleGen published their sequence capture technology (Albert et al., 2007), which used microarrays to pull down a specified subset of the genome, allowing low cost sequencing of a subset of the genome. This birthed the field of "whole exome sequencing", in which only the 1% of the genome coding that codes for proteins is sequenced. Interestingly, the benefits of this technology were first seen in the field of Mendelian diseases, where exome sequencing can identify all coding mutations in an individual's genome, and public databases (such as the 1000 Genomes Project) can exclude all polymorphic markers, leaving a small number of candidate causal mutations. The discovery of the causal mutation for Miller syndrome by exome sequencing (Ng et al., 2010) was rapidly followed by other successes, and this method is now the dominant method for solving Mendelian diseases (Bamshad et al., 2011).

1.2.7 Next-generation GWAS and post-GWAS studies

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

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

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

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

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

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

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

1.2.8 Conclusions

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

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

One effect of this technological drive is a tendency for statistical arguments to be raised, settled and often forgotten decades before the technology catches up. This can lead to a certain amount of historical blindness. Discussions of rare variants and genetic heterogeneity, for instance, seem to wax and then wane away every 10 years or so (with early family studies, with RFLP studies, with the failure of complex disease linkage, and in the GWAS era). Another effect is that methods can become ingrained, and used without proper thought to what they mean. This was one of the reasons behind the failure of candidate gene studies, where a rule-of-thumb (a p-value threshold of 0.05) became a blindly applied law even in cases where it was not appropriate.

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

1.3 Outline of this thesis

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

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

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

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

In Chapter 3, I investigate the use of genotype imputation algorithms in genome-wide association studies. As we saw above, genotype imputation can allow disease association to be tested with far more SNPs than have been genotyped in a GWAS, facilitating meta-analysis and increasing power. I begin by investigating the impact of reference set size and diversity on imputation in Europeans, using the HapMap data, with particular focus on the imputation of low frequency variants. I then investigate how effective the same reference sets are at performing imputation in African populations. Next, I expand this analysis to new datasets, looking at how well 1000 Genomes project data can impute low-frequency variation in a diverse African population. Finally, I show how imputation of variants from the 1000 Genomes pilot can be used to draw conclusions about disease biology, by estimating the influence of loss-of-function variants on 7 complex diseases.

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

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

In contrast to the study of common variants of small effect, Chapter 5 describes a set of approaches to discover rare variants of large effect by using large, multiplex families. I begin by producing a joint model of common polygenic and rare dominant penetrant genetic risk in families, and exploring how the probability of observing multiplex families of a certain size varies depending on heritability and penetrance. I then lay out a method of performing genetic risk prediction in families, and show that this method can effectively distinguish between multiplex families that do or do not harbour a penetrant mutation.

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

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

Chapter 2

Statistical methods and models of genetic risk

2.1 Introduction

The field of complex disease genetics is inherently statistical, both in the sense that it studies a phenomenon (complex disease) that is by definition probabilistic, and in the sense that it relies on statistical methods to make inferences from the data under study. Examples of these statistical methods include risk prediction (either using relative risks or odds ratios), regression analyses (usually using logistic regression) and family analyses (generally using liability threshold models). Each of these methods is built around assumptions, and these assumptions themselves form a model (either explicit or implicit) about the distributions of genetic risk in the population. In many cases, these methods imply very different and mutually incompatible assumptions.

In the last few years the interest in statistical models of genetic risk has increased dramatically. Recent papers include general discussions of modelling issues arising from GWAS (e.g. Sawcer and Wason (2012)), and detailed examinations of specific models (e.g. Wray et al. (2010)). Two recent reviews (Wray and Goddard, 2010; Clayton, 2012) have made broad comparisons of different models of genetic risk, noting a number of inconsistencies between models and describing different implications for association studies and risk prediction. However, neither provided a systematic survey of the properties of genetic risk models, and in particular neither gave a detailed investigation into the relationships between different models, and between models and statistical methods. The time is thus ripe for a unified analysis that places different statistical methods and models of risk into a single framework.

In this chapter I will lay out a simple framework for classifying such models, and discuss three major models of genetic risk. Together, these three models underlie most standard models and methods used in the field. I will investigate how these models differ, how suitable each is to the tasks that they have been used for, and how their predictions about the distributions of genetic risk differ from each other.

In the introduction, I will formulate a general description of a model of genetic risk, and discuss a specific family of models that are specified in terms of a normally distributed genetic risk score and a link function. In Section 2.2, I will go on to discuss in more detail the relationship between locus-based models of genetic risk (such as those fitted in GWAS) and continuous risk scores. Sections 2.3-2.5 will discuss and critically assess three specific models of risk that correspond to three link functions (the log, probit and logit models), and in Section 2.6 I will compare how these models differ in their predictions about the distribution of genetic risk. In the final section I will discuss how confusion between these models can generate real problems in statistical genetics, as well as discussion some of the limitations of this approach.

2.1.1 Definition of a genetic risk model

In general, a model of genetic risk has two properties. Firstly, it specifies a distribution of a genetic risk value $p_i \in [0, 1]$ for a randomly selected individual i

$$p_i \sim Distribution(\theta), \tag{2.1}$$

where the probability of an individual developing the disease is equal to p_i, or

$$P(d_i = 1|p_i) = p_i. \tag{2.2}$$

Here, d_i is an indicator variable taking on value $d_i = 1$ if the individual i has the disease (if we are modelling the prevalence) or will develop the disease in their lifetime (if we are modelling the lifetime risk).

Secondly, a model of genetic risk specifies a joint distribution for genetic risk values p_i and p_j for individuals i and j that share a family relationship r_{ij}

$$(p_i, p_j) \sim Distribution(\theta, r_{ij}). \tag{2.3}$$

For a purely genetic model, we make the additional assumption that disease incidence is independent in families conditional on their genetic risk, i.e.

$$\begin{aligned} P(d_i = 1, d_j = 1 | p_i, p_j) &= P(d_i = 1 | p_i) P(d_j = 1 | p_j) \\ &= p_i p_j. \end{aligned} \tag{2.4}$$

In essence, we assume that relatives have no shared environmental risk. In this chapter we will almost exclusively consider purely genetic models. In the case where environmental and genetic risks act independently, these models can be reasonably interpreted as the behaviour of the genetic component, and are easily extended to include environmental risk (as discussed in Section 2.4.1). In the presence of strong gene-environment interaction, however, these purely genetic models will become inaccurate, and the true model will depend on the form of the interaction.

We refer to p_i as the genetic risk or the genetic disease probability. Its distribution can be discrete or continuous, though we will only consider continuous distributions in this chapter.

2.1.2 Observable parameters of a genetic risk model

While each model of genetic risk has its own set of parameters θ, there are a number of common parameters that we can calculate for any model, which in turn are measurable in real populations.

The first parameter I will consider is the population prevalence of the disease, or the probability that a randomly selected individual has the disease in question. This is equal to

$$\begin{aligned} K &= P(d=1) \\ &= \int_p P(d=1|p)f(p)dp \\ &= \int_p pf(p)dp \\ &= E[p], \end{aligned} \tag{2.5}$$

where $f(\cdot)$ is the probability density function of p.

A more complicated measure is how "genetic" a disease is. This concept is relatively ill defined. The heritability of liability h^2 is often used for this purpose, which is equal to the proportion of variance in the total risk that can be attributed to genetics, where risk is measured on the liability scale (discussed in Section 2.4.1). However, this parameter is model specific.

Instead, for comparison across models we will use the relative recurrence risk, equal to the fold enrichment of disease prevalence in relatives of affected individuals. For relatives of type r_{ij}, this is calculated as

$$\begin{aligned} \lambda_r &= \frac{P(d_i=1|d_j=1,r_{ij})}{P(d_i=1)} \\ &= \frac{P(d_i=1,d_j=1|r_{ij})}{P(d_i=1)^2} \\ &= \frac{\int_{p_i}\int_{p_j} P(d_i=1|p_i)P(d_j=1|p_j)f(p_i,p_j|r_{ij})dp_i dp_j}{K^2}. \end{aligned} \tag{2.6}$$

This can in theory be measured directly from population data, if common environment can be controlled for. Regardless of whether or not it can actually be measured, the definition is model independent, and acts as a useful benchmark to compare across models.

Finally, we will be interested in the distribution of p in cases and controls

$$\begin{aligned} f(p|d=1) &= \frac{P(d=1|p)f(p)}{P(d=1)} \\ &= \frac{p}{K}f(p) \qquad (2.7) \\ f(p|d=0) &= \frac{1-p}{1-K}f(p). \qquad (2.8) \end{aligned}$$

2.1.3 Genetic risk scores and link functions

In this chapter, we will consider a specific family of continuous genetic risk models. These models have two components, firstly a normally distributed genetic risk score

$$\eta \sim N(\mu, \sigma^2) \tag{2.9}$$

and secondly a link function g that connects this genetic risk score to the genetic risk probability

$$p = g(\eta). \tag{2.10}$$

We can thus write down the probability density function of p as

$$f(p) = \frac{d\eta}{dg}\frac{1}{\sigma}\phi\left(\frac{\eta-\mu}{\sigma}\right), \tag{2.11}$$

where ϕ is the density of the standard normal distribution.

In the following section, we will describe the relationship between discrete genotypes $\vec{x}$ and risk scores η. We will then consider three link functions: the log link $g(\eta) = exp(\eta)$, the logit link $g(\eta) = (1 + exp(-\eta))^{-1}$ and the probit link $g(\eta) = \Phi^{-1}(\eta)$.

2.2 From discrete genotypes to continuous risk

The conversion from discrete genotypes to a continuous genetic trait was first outlined by Fisher (1918), who showed not only that a large number of discrete genetic factors can give rise to a continuous trait, but also that certain correlation structures in this continuous trait exist between family members as a consequence of Mendelian inheritance. In this section I will outline the relationship between discrete genetic risk factors and a continuous risk score, and outline the distribution and parameters of this score.

Note that in the following section I will use lowercase x and y to refer to random variables that represent genotype dosages (i.e. $x \in (0, 1, 2)$), uppercase X and Y to refer to general random variables, and lowercase z to refer to a standard normal random variable.

The above described η score is constructed from a combination of genotypes across n loci, $\vec{x} = (x_1, ..., x_n)$. The general form is

$$\eta = t(\vec{x}), \tag{2.12}$$

where t is the function that maps from genotype to score. Note that, in this general formulation, there is no requirement that η be normally distributed (as described in Equation 2.9).

We can simplify this by assuming that the loci are all independent, and each contributes independently to η, i.e.

$$\eta = a_0 + \sum_{l=1}^{n} t_l(x_l). \tag{2.13}$$

As the random variables x_l are independent, and providing that the transformed variables $t_l(x_l)$ have finite means and variances that are independent of the indicator variable l, it follows from the central limit theorum that η

tends to a normal distribution as n increases.

We can modify the score to include interaction terms between genotypes, e.g. by including second-degree interaction

$$\eta = a_0 + \sum_{i=1}^{n} \sum_{j=i}^{n} t_{ij}(x_i, x_j). \tag{2.14}$$

In this section we discuss the particulars of going from a combination of genotypes to a continuous risk score. We will discuss the problem in general in terms of the properties of sums of independent variables, and then discuss the specific case where η is a linear function, i.e. $f_l(x_l) = a_l x_l$. Finally, we will discuss issues with non-linear functions.

2.2.1 Properties of a sum of independent variables

Suppose we have two sets of random variables, X_1 and X_2, and Y_1 and Y_2, such that $X_i \perp Y_j \; \forall (i, j)$.

We construct scores by adding these variables together, i.e. $\eta_i = X_i + Y_i$. The expectation and variance of this score are given by

$$\begin{aligned} E[\eta_i] &= E[X_i + Y_i] \\ &= E[X_i] + E[Y_i] \qquad (2.15) \\ Var[\eta_i] &= Var[X_i + Y_i] \\ &= Var[X_i] + Var[Y_i], \qquad (2.16) \end{aligned}$$

and the covariance are given by

$$\begin{aligned}
cov(\eta_1, \eta_2) &= E[\eta_1\eta_2] - E[\eta_1]E[\eta_2] \\
&= E[(X_1 + Y_1)(X_2 + Y_2)] - E[X_1 + Y_1]E[X_2 + Y_2] \\
&= E[X_1X_2] + E[Y_1]E[X_2] + E[X_1]E[Y_2] + E[Y_1Y_2] - \\
&\quad E[X_1]E[X_2] - E[X_1]E[Y_2] - E[X_2]E[Y_1] - E[Y_1]E[Y_2] \\
&= E[X_1X_2] - E[X_1]E[X_2] + E[Y_1Y_2] - E[Y_1]E[Y_2] \\
&= cov[X_1, X_2] + cov[Y_1, Y_2]. \qquad (2.17)
\end{aligned}$$

We can generalise this to the sum of n variables $\eta_i = \sum_{j=1}^{n} X_{ij}$ such that $X_{ab} \perp X_{cd} \forall a, c, b \neq d$, to give

$$\begin{aligned}
E[\eta_i] &= \sum_{j=1}^{n} E[X_{ij}] \qquad (2.18) \\
Var[\eta_i] &= \sum_{j=1}^{n} Var[X_{ij}] \qquad (2.19) \\
cov(\eta_1, \eta_2) &= \sum_{j=1}^{n} cov(X_{1j}, X_{2j}). \qquad (2.20)
\end{aligned}$$

If the X_{ij}'s have finite mean and variance, then when n is large we can approximate (η_1, η_2) as a multivariate normal with

$$\begin{aligned}
\vec{\mu} &= (E[\eta_1], E[\eta_1]) \qquad (2.21) \\
\Sigma &= \begin{pmatrix} Var[\eta_1] & cov(\eta_1, \eta_2) \\ cov(\eta_1, \eta_2) & Var[\eta_1]. \end{pmatrix} \qquad (2.22)
\end{aligned}$$

If we imagine that the X_{ij}'s are functions of allele count for independently segregating genetic risk loci, we can see that to calculate the covariance of a function that is a sum of such functions only requires the calculation of the covariance of each function individually.

2.2.2 General covariance for linear functions of allele count

A linear, or additive, risk score has the form

$$\eta_i = a_0 + \sum_{l=1}^{n} a_l x_{il}. \tag{2.23}$$

Again we will assume that the variants in the score are in linkage equilibrium, and thus the allele counts at different loci are independent ($x_{ia} \perp x_{ib} \forall a \neq b$). The score η_i has expectation and variance

$$\begin{aligned} E[\eta_i] &= a_0 + \sum_{l=1}^{n} a_l E[x_{il}] \\ &= a_0 + \sum_{l=1}^{n} a_l 2 f_l \qquad (2.24) \\ Var[\eta_i] &= \sum_{l=1}^{n} a_l^2 Var[x_{il}] \\ &= \sum_{l=1}^{n} a_l^2 2 f_l (1 - f_l), \qquad (2.25) \end{aligned}$$

where f_i is the allele frequency of variant l.

To calculate the covariance, suppose two individuals i and j have a coefficient of relatedness ρ_{ij}. This is equal to the probability that any given allele on any given chromosome will be shared IBD (with $\rho_{ij} = 0.5$ for siblings i and j, etc).

For a variant with allele frequency f, we can denote the allele count for individual i as $x_i = x_{i1} + x_{i2}$, where x_{ik} are allele counts on individual chromosomes $k = 1, 2$ for individual i. We will use $S_k^{ij} = 1$ to denote that this allele is shared IBD between individuals i and j on chromosome k, with $P(S_k^{ij} = 1) = \rho_{ij}$. For now I will assume $S_1^{ij} \perp S_2^{ij}$, i.e. that the IBD sharing states for the two chromosomes are independent (as is the case for siblings, for example). The next section will generalize this to arbitrary IBD distributions.

The joint distribution of genotypes on a particular chromosome k for two individuals i and j with coefficient of relatedness ρ_{ij} is given by

$$P(x_{ik}, x_{jk}) = \rho_{ij} P(x_{ik}, x_{jk} | S_k^{ij} = 1) + (1 - \rho_{ij}) P(x_{ik}, x_{jk} | S_k^{ij} = 0), \tag{2.26}$$

where

$$P(x_{ik}, x_{jk} | S_k^{ij} = 1) = \begin{cases} P(x_{ik}) & \text{if } x_{ik} = x_{jk}; \\ 0 & \text{otherwise}, \end{cases} \tag{2.27}$$

$$\tag{2.28}$$

and

$$P(x_{ik}, x_{jk} | S_k^{ij} = 0) = P(x_{ik}) P(x_{jk}). \tag{2.29}$$

We can calculate the covariance in allele counts between two individuals of x_{ik} and x_{jk} by first calculating

$$\begin{aligned}
E[x_{ik}x_{jk}] &= \sum x_{ik}x_{jk}P(x_{ik},x_{jk}) \\
&= P(x_{ik}=1,x_{jk}=1) \\
&= \rho_{ij}P(x_{ik}=1,x_{jk}=1|S_k^{ij}=1)+(1-\rho_{ij})P(x_{ik}=1,x_{jk}=1|S_k^{ij}=0) \\
&= \rho_{ij}f+(1-\rho_{ij})f^2, \qquad (2.30)
\end{aligned}$$

and then by using this to calculate the covariance

$$\begin{aligned}
cov[x_{ik},x_{jk}] &= E[x_{ik}x_{jk}]-E[x_{ik}]E[x_{jk}] \\
&= \rho_{ij}f+(1-\rho_{ij})f^2-f^2 \\
&= \rho_{ij}f(1-f).
\end{aligned}$$

We can use Equation 2.17 to give $cov[x_i,x_j]=2\rho_{ij}f(1-f)$. Note that $var[x_i]=2f(1-f)$, so $cor[x_i,x_j]=\rho_{ij}$. This means that, as well as being the probability of sharing any given allele IBD, the coefficient of relatedness is also equal to the correlation in genotype counts.

We can therefore give the covariance of η_i and η_j as

$$\begin{aligned}
cov[\eta_i,\eta_j] &= cov[\sum_{l=1}^{n}a_lx_{il},\sum_{l=1}^{n}a_lx_{jl}] \\
&= \sum_{l=1}^{n}cov[a_lx_{1l},a_lx_{2l}] \\
&= \rho_{ij}\sum_{l=1}^{n}a_l^2 2f_l(1-f_l) \\
&= \rho_{ij}var[\eta_j]. \qquad (2.31)
\end{aligned}$$

Again, note that $cor[\eta_i, \eta_j] = \rho_{ij}$.

When I refer to "additive" genetic risk throughout this thesis, I refer a to risk score which can be expressed thus on some scale. This assumption of additivity is important because it allows us to assume that the correlation between individuals on this scale is equal to their coefficient of relatedness.

2.2.3 Covariance for non-linear functions of allele count

The coefficient of relatedness is not sufficient to give the full joint genotype distribution for two individuals. For instance, while full siblings and parent-offspring pairs both have the same coefficient of relatedness ($\rho = 0.5$), they have distinct patterns of allele sharing due to the fact that parent-offspring always share exactly one allele IBD, but siblings can share zero, one or two.

We write the proportion of alleles shared IBD 1 and 2 as p_1, p_2 (with $1 - p_1 - p_2$ with IBD 0). We can calculate the coefficient of relatedness from the IBD probabilities as $\rho = \frac{1}{2}p_1 + p_2$. Parent-offspring pairs have $p_1 = 1$ and $p_2 = 0$, siblings have $p_1 = 0.5$ and $p_2 = 0.25$.

The table below shows the joint genotype distributions depending on IBD status.

Genotype (x_i,x_j)	IBD = 0	IBD = 1	IBD = 2
0,0	$(1-f)^4$	$(1-f)^3$	$(1-f)^2$
0,1	$2f(1-f)^3$	$f(1-f)^2$	0
0,2	$f^2(1-f)^2$	0	0
1,1	$4f^2(1-f)^2$	$f(1-f)$	$2f(1-f)$
1,2	$2f^3(1-f)$	$f^2(1-f)$	0
2,2	f^4	f^3	f^2

Note that certain genotype combinations can occor multiple ways. For instance, there are two possible ways of having one individual with two alleles and one with zero, $(x_i,x_j) = (0,2)$ and $(x_i,x_j) = (2,0)$. This means that the probability of being in either of these two states is equal to $2f^2(1-f)^2$.

We can then calculate the expected values of various non-linear functions of genotype count. For instance, the product of genotype values has expectation

$$\begin{aligned}
E[x_ix_j|\text{IBD}=0] &= 4f^2(1-f)^2 + 8f^3(1-f) + 4f^4 \\
&= 4f^2 \qquad (2.32)\\
E[x_ix_j|\text{IBD}=1] &= f(1-f) + 4f^2(1-f) + 4f^3 \\
&= f(1+3f) \qquad (2.33)\\
E[x_ix_j|\text{IBD}=2] &= 2f(1-f) + 4f^2 \\
&= 2f(1+f) \qquad (2.34)\\
E[x_ix_j] &= (1-p_1-p_2)E[x_ix_j|\text{IBD}=0] \\
&\quad +p_1E[x_ix_j|\text{IBD}=1] + p_2E[x_ix_j|\text{IBD}=2] \\
&= (1-p_1-p_2)4f^2 + p_1(f(1+3f)) + p_22f(1+f) \\
&= f(p_1+2p_2) + f^2(4-p_1-2p_2) \\
&= 2f\rho + 2f^2(2-\rho). \qquad (2.35)
\end{aligned}$$

As we saw above, the expectation of the product is dependent only on ρ, and not on the specific IBD distribution.

The expectation of $x_ix_j^2$ is given by

$$\begin{aligned}
E[x_i x_j^2|\text{IBD}=0] &= 4f^2(1-f)^2+12f^3(1-f)+8f^4 \\
&= 2f^2(2(1-f)^2+6f(1-f)+4f^4) \\
&= 4f^2(1+f) && (2.36) \\
E[x_i x_j^2|\text{IBD}=1] &= f(1-f)+6f^2(1-f)+8f^3 \\
&= f(1+5f+2f^2) && (2.37) \\
E[x_i x_j^2|\text{IBD}=2] &= 2f(1-f)+8f^2 \\
&= 2f(1+3f) && (2.38)
\end{aligned}$$

$$\begin{aligned}
E[x_i x_j^2] &= (1-p_1-p_2)E[x_i x_j^2|\text{IBD}=0]+p_1 E[x_i x_j^2|\text{IBD}=1] \\
&\quad +p_2 E[x_i x_j^2|\text{IBD}=2] \\
&= (1-p_1-p_2)4f^2(1+f)+p_1 f(1+5f+2f^2)+p_2 2f(1+3f) \\
&= (p_1+2p_2)f+(4+p_1+2p_2)f^2+2(4-p_1-2p_2)f^3 \\
&= 2\rho f+2(2+\rho)f^2+4(2-\rho)f^3. && (2.39)
\end{aligned}$$

Again, this expression is only dependent on ρ. Finally, the expectation of $x_i^2 x_j^2$ is given by

$$\begin{aligned}
E[x_i^2x_j^2|\text{IBD}=0] &= 4f^2(1-f)^2+16f^3(1-f)+16f^4 \\
&= 4f^2(1+f)^2 \qquad (2.40) \\
E[x_i^2x_j^2|\text{IBD}=1] &= f(1-f)+8f^2(1-f)+16f^3 \\
&= f(1+7f+8f^2) \qquad (2.41) \\
E[x_i^2x_j^2|\text{IBD}=2] &= 2f(1-f)+16f^2 \\
&= 2f(1+7f) \qquad (2.42)
\end{aligned}$$

$$\begin{aligned}
E[x_i^2x_j^2] &= (1-p_1-p_2)E[x_i^2x_j^2|\text{IBD}=0] \\
&\quad +p_1E[x_i^2x_j^2|\text{IBD}=1]+p_2E[x_i^2x_j^2|\text{IBD}=2] \\
&= (1-p_1-p_2)4f^2(1+f)^2+p_1f(1+7f+8f^2)+p_22f(1+7f) \\
&= f(p_1+2p_2)+f^2(4+3p_1+10p_2) \\
&\quad +8f^3(1-p_2)+4f^4(1-p_1-p_2). \qquad (2.43)
\end{aligned}$$

And these in turn allow to us to calculate covariance and correlations of non-linear functions of allele count between relatives. For instance, consider the non-linear function $\eta_i = x_i + bx_i^2$ with dominance term b.

$$\begin{aligned}
E[\eta_i] &= E[x_i] + bE[x_i^2] \\
&= 2f(1+b) + 2bf^2 \qquad (2.44) \\
E[\eta_i]^2 &= E[x_i] + bE[x_i^2] \\
&= 4f^2(1+b)^2 + 8f^3b(1+b) + 4b^2f^4 \qquad (2.45) \\
E[\eta_i^2] &= E[(x_i + bx_i^2)^2] \\
&= E[x_i^2] + 2bE[x_i^3] + b^2E[x_i^4] \\
&= 2f(1+b)^2 + 2f^2(1+6b+7b^2) \qquad (2.46) \\
Var[\eta_i] &= E[\eta_i^2] - E[\eta_i]^2 \\
&= 2f(1+b)^2 - 2f^2(1-2b-5b^2) \\
&\quad -8f^3b(1+b) - 4f^4b^2 \qquad (2.47) \\
E[\eta_i\eta_j] &= E[(x_i + bx_i^2)(x_j + bx_j^2)] \\
&= E[x_ix_j] + 2bE[x_ix_j^2] + b^2E[x_i^2x_j^2] \qquad (2.48) \\
cov[\eta_i, \eta_j] &= E[\eta_i\eta_j] - E[\eta_i]E[\eta_j] \qquad (2.49) \\
cor[\eta_i, \eta_j] &= \frac{cov[\eta_i, \eta_j]}{var[\eta_i, \eta_j]}. \qquad (2.50)
\end{aligned}$$

We can find the maximum and minimum values of the correlation by differentiating $cor[\eta_i, \eta_j]$ with respect to b. We find that the correlation takes on the minimum value of p_2 when $b = \frac{-1}{1+2f}$, and a maximum value of ρ when $b = 0$.

As Figure 2.1 shows, low frequency variants show very little drop off in correlation until very high degrees of dominance, whereas higher frequency variants show a smoother drop off in correlation. Dominance effects thus have a stronger impact on the risk score correlation when the variants have higher frequency.

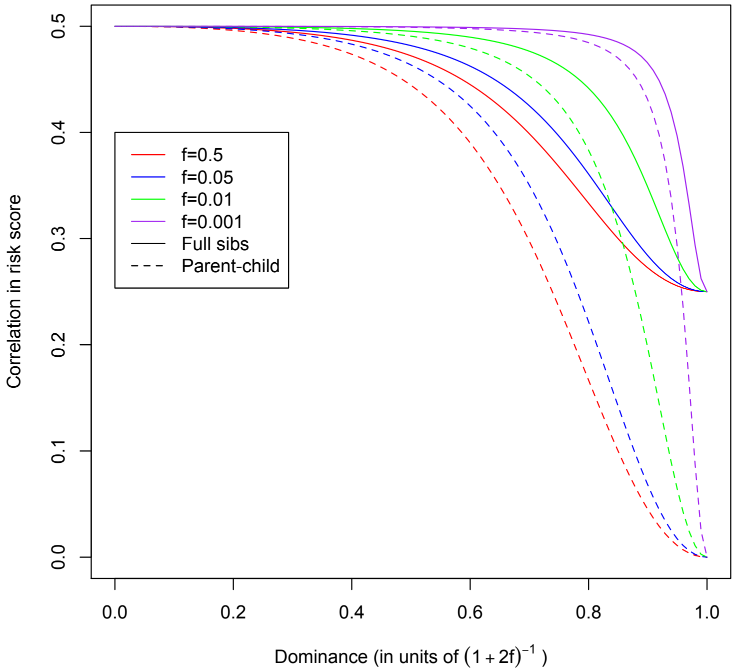


Figure 2.1: The decrease in correlation in risk score for siblings and parent-child pairs with increasing value of the dominance term (normalised such that the maximum value is $\frac{-1}{1+2f}$). Different colour lines represent variants with different allele frequencies.

2.3 The log risk model

The log risk model was defined by Pharoah et al. (2002) and more recently elaborated on by Clayton (2009). It has most commonly been used to make inferences about the utility of genetic risk prediction (Clayton, 2009; Sawcer et al., 2010; Chatterjee et al., 2011), though it has also been used to estimate sibling recurrence ratios in twin studies (Clayton, 2009).

As we will see, the model is asymptotically equivalent to the Risch multi-locus model of genetic risk. It is also equivalent to the assumption of multiplicative combination of relative risk that is often used in genetic risk prediction (e.g. by the genetic testing company deCODEme).

This model is the least realistic of the models that I will consider, due to the fact that the probability is not bounded, though it is also one of the more widely used, probably due to its analytic tractability.

The link function for the log risk model is

$$p = \exp(\eta). \tag{2.51}$$

Substituting this into Equation 2.11, the density function for p is given by

$$f(p) = \frac{1}{p\sigma}\phi\left(\frac{log(p)-\mu}{\sigma}\right). \tag{2.52}$$

2.3.1 Calculating parameters

The prevalence parameter K is given by

$$\begin{aligned} K &= E[p] \\ &= \int \exp(\mu+\sigma x)\phi(x)dx \\ &= \int \frac{1}{\sqrt{2\pi}}\exp(\mu+\sigma x-\frac{1}{2}x^2)dx \\ &= \int \frac{1}{\sqrt{2\pi}}\exp(\mu+\sigma^2/2-\frac{1}{2}(x-\sigma)^2)dx \\ &= \exp(\mu+\frac{\sigma^2}{2})\int \phi(x-\sigma)dx \\ &= \exp(\mu+\frac{\sigma^2}{2}), \end{aligned} \tag{2.53}$$

i.e. the expectation of the log-normal distribution (Johnson et al., 1994).

As we saw in Section 2.2.2, under additivity the correlation in log risk score is equal to their coefficient of relatedness ρ. We can thus express the genetic risk for relatives p_1 and p_2 as

$$\begin{aligned} p_1 &= \exp(\mu + \sigma z_1) && (2.54) \\ p_2 &= \exp(\mu + \rho\sigma z_1 + \sqrt{1-\rho}\sigma z_2), && (2.55) \end{aligned}$$

where z_i are standard normal variables.

The probability of both relatives developing the disease given z_is is

$$\begin{aligned} P(d_1 = 1, d_2 = 1|\eta_1, \eta_2) &= p(d_1|\eta_1)p(d_2|\eta_2) \\ &= p_1 p_2 \\ &= \exp(\mu + \sigma z_1 + \mu + \rho\sigma z_1 + \sqrt{1-\rho^2}\sigma z_2) \\ &= \exp(2\mu + \sigma(1+\rho)z_1 + \sigma\sqrt{1-\rho^2}z_2). \quad (2.56) \end{aligned}$$

The mean rate of co-occurrence is thus

$$\begin{aligned}
E[p_1p_2] &= \int p_1\phi(z_1)p_2\phi(z_2)dz_1dz_2 \\
&= \exp(2\mu)\int \exp(\frac{1}{2}(z_1^2 - 2\sigma(1+\rho)))p_2\phi(z_2)dz_1dz_2 \\
&= \exp(2\mu)\int \exp(\frac{1}{2}((z_1-\sigma(1+\rho))^2 - \sigma^2(1+\rho)^2))p_2\phi(z_2)dz_1dz_2 \\
&= \exp(2\mu+\frac{1}{2}\sigma(1+\rho))^2)\int \phi(z_1-\sigma(1+\rho))p_2\phi(z_2)dz_1dz_2 \\
&= \exp(2\mu+\frac{1}{2}\sigma(1+\rho)^2) \\
&\quad \times\int \phi(z_1-\sigma(1+\rho))\exp(\frac{1}{2}(z_2^2-2\sigma\sqrt{1-\rho^2}))dz_1dz_2 \\
&= \exp(2\mu+\frac{1}{2}\sigma(1+\rho)^2) \\
&\quad \times\int \phi(z_1-\sigma(1+\rho))\exp(\frac{1}{2}((z_2-\sigma\sqrt{1-\rho^2})^2-\sigma^2(1-\rho^2)))dz_1dz_2 \\
&= \exp(2\mu+\frac{1}{2}\sigma(1+\rho)^2+\frac{1}{2}\sigma^2(1-\rho^2))) \\
&\quad \times\int \phi(z_1-\sigma(1+\rho))\phi(z_2-\sigma\sqrt{1-\rho^2})dz_1dz_2 \\
&= \exp(2\mu+\frac{1}{2}\sigma(1+\rho)^2+\frac{1}{2}\sigma^2(1-\rho^2))) \\
&= \exp(2\mu+\sigma(1+\rho)). \qquad (2.57)
\end{aligned}$$

The recurrence ratio in relatives is thus

$$\begin{aligned}
\lambda_r &= \frac{E[p_1p_2]}{K^2} \\
&= \exp(\rho\sigma). \qquad (2.58)
\end{aligned}$$

We can rearrange equations 2.53 and 2.58 to give parameters μ and σ, given a prevalence K and a sibling recurrence ratio λ_s

$$\sigma^2 = 2log(\lambda_s) \tag{2.59}$$

$$\mu = log(K) - \sigma^2. \tag{2.60}$$

2.3.2 Case and control distributions

The distribution of η in cases is given by the probability density function

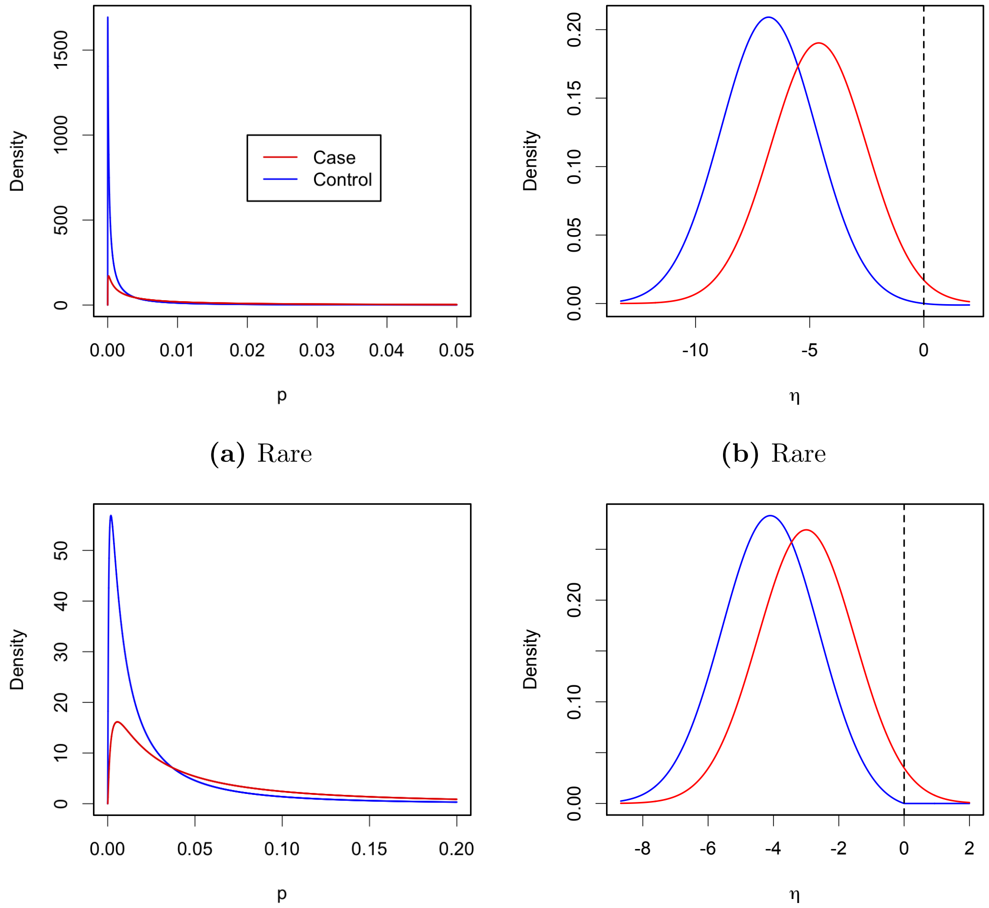
$$\begin{aligned} P(\eta|d=1) &= \frac{P(d=1|\eta)P(\eta)}{P(d=1)} \\ &= \frac{e^{\eta}\phi(\frac{\eta-\mu}{\sigma})}{\sigma K}. \end{aligned} \tag{2.61}$$

This can be simplified to

$$\begin{aligned} P(\eta|d=1) &= \frac{\exp(\eta)\frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-(\eta-\mu)^2}{2\sigma^2})}{\exp(\mu+\sigma^2)} \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp(\eta + \frac{-(\eta-\mu)^2}{2\sigma^2} - \mu - \sigma^2) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-(\eta-(\mu+\sigma^2))^2}{2\sigma^2}) \\ &= \frac{1}{\sigma}\phi(\frac{\eta-(\mu+\sigma^2)}{\sigma}), \end{aligned} \tag{2.62}$$

i.e. normally distributed with a mean $\mu + \sigma^2$ and a variance σ^2. Thus, the distribution of log risk for cases is the same as for the population as a whole, but shifted upwards by σ^2.

The distribution for risk in controls is given by



(a) Rare

(b) Rare

(c) Common

(d) Common

Figure 2.2: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$)

$$P(\eta|d = 0) = \frac{(1 - e^{\eta})\phi(\eta)}{\sigma(1 - K)}. \tag{2.63}$$

The distribution of probability and risk score in cases and controls is shown for example parameters (simulating a common and rare disease) in Figure 2.2. Note that, in both parameter sets, a not insignificant number of cases have a value of $\eta > 0$ and therefore $p > 1$ (see Section 2.3.5 for more on this issue).

2.3.3 Relationship to Risch model

The log risk model can be seen as an approximation to the Risch multilocus model, introduced by Risch (1990), that has been used to make inferences about genetic risk prediction (Wray et al., 2007). The Risch model assumes that n loci exist, each with the same relative risk r and a risk allele frequency f. An individual's disease probability is based on the number of risk alleles they carry x, and is given by

$$\begin{aligned} p &= p_0 r^x \\ &= \exp\left[log(p_0) + xlog(r)\right], \end{aligned} \tag{2.64}$$

where p_0 is the disease probability in individuals with zero risk alleles.

x is binomially distributed, with $x \sim Binom(2n, f)$. As we saw above, as n grows larger, x tends in distribution to $N(2nf, 2nf(1-f))$, and thus

$$p \rightarrow \exp\left(\eta\right) \;\; \text{where} \; \eta \sim N\left(log(p_0) + 2nf, 2nf(1-f)log(r)^2\right), \tag{2.65}$$

i.e. the Risch model is asymptotically equivalent to the log risk model with $\mu = log(p_0) + 2nf$ and $\sigma^2 = 2nf(1-f)log(r)^2$.

2.3.4 Relationship to multiplicative relative risk model and log-linked regression

A commonly used risk prediction method is the multiplicative relative risk model (also known as the log-linear relative risk model). This is the most widely used of the relative risk models in epidemiology (Breslow and Storer, 1985), and has been used in genetics, as a model for genetic risk prediction (Lu and Elston, 2008). Notably, it is the model used by the genetic testing company deCODEme to produce individual disease probabilities given a customer's genotypes (deCODEme, 2012).

Under the multiplicative relative risk model, we have n loci, with each having a frequency f_i and a genotypic relative risk r_i. The probability for an individual who has allele counts x_i is given by

$$\begin{aligned} p &= f_0 \prod r_i^{x_i} \\ &= \exp\left[log(f_0) + \sum_{i=1}^{n} x_i log(r_i) \right]. \end{aligned} \tag{2.66}$$

Note that this can be seen as a generalisation of the Risch model, with identical $f = f_i$ and $r = r_i$ for all i, and $x = \sum_i x_i$.

As long as the values r_i are finite, the terms $x_i log(r_i)$ will have finite mean and variance, and thus the central limit theorum states that the summation above will tend towards a normal distribution as n grows, giving

$$p \to \exp\left(\eta\right) \ \text{ where } \eta \sim N\left(\mu, \sigma\right), \tag{2.67}$$

where

$$\mu = log(f_0) + \sum_{i=1}^{n} 2 f_i log(r_i) \tag{2.68}$$

$$\sigma^2 = \sum_{i=1}^{n} 2 f_i (1 - f_i) log(r_i)^2, \tag{2.69}$$

i.e. equivalent to the log risk model with μ and σ.

2.3.5 Problem with probabilities greater than 1

Wray and Goddard (2010) noted a problem with the log risk and Risch models, in that they can predict probabilities greater than one. The authors suggest a modified version of the model, where probabilities are capped at 1. In practice, capping at 1 may not be conservative enough: the genetic testing company deCODEme cap their genetic risk probabilities at 90% (deCODEme, 2012). In contrast, Clayton (2012) argued that this is not a major problem with the model, as for relatively uncommon diseases probabilities greater than 1 are relatively rare in the general population.

However, I will show that this is a real problem with the model in many circumstances. It is true that unless the disease is very common, the total number of individuals with $p > 1$ is small. For a disease with $K = 0.01$ and $\lambda_s = 9$, less than 0.1% of individuals have $p > 1$, and even for a disease with $K = 0.05$ and $\lambda_s = 3$ only 0.3% of individuals have this property (Figure 2.2). However, these values rise dramatically if we only consider cases, to 0.5% and 2.2% respectively, and if we consider identical twins where both are affected, 7% and 23% of twin pairs have a probability greater than 1.

So, while probabilities for randomly selected individuals are unlikely to suffer from this problem, the individuals in those groups we are often most concerned with (i.e. those with a family history and those who will go on to develop the disease) are far more likely to. In particular, the very high proportion of doubly-affected twin pairs with probabilities greater than 1 is concerning given that the expectation of the product of these probabilities is used to calculate the sibling recurrence ratio in Equation 2.58. Because this expectation is likely to be overestimated due to the greater-than-one

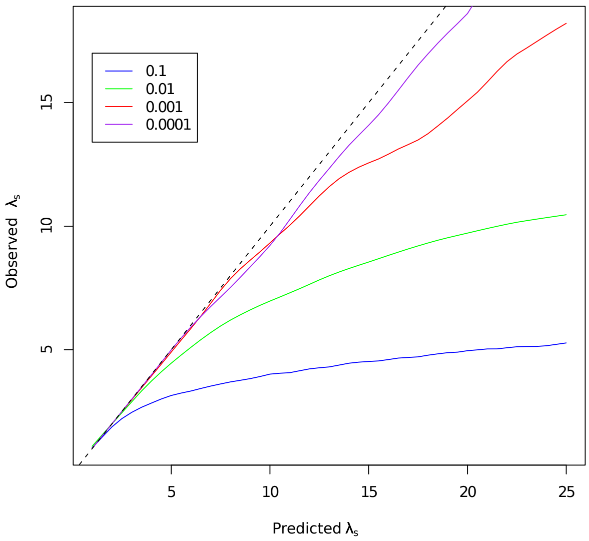


Figure 2.3: The λ_s predicted under the log risk model compared to the observed value under the truncated log model with all probabilities greater than 1 set to 1, for varying prevalence.

probabilities, it will follow that the value of λ_s could be greatly overestimated, and likewise the size of the genetic variance and parameter σ could be underestimated given a value λ_s.

To investigate the degree to which this will lead to errors, I simulated families under a truncated model (i.e. setting all $p > 1$ to $p = 1$), and compared the observed λ_s values to those predicted by Equation 2.58. Figure 2.3 shows that the log risk model significantly overestimates virtually all values of λ_s when $K = 0.1$, all values of $\lambda_s > 5$ for $K = 0.01$, and values of $\lambda_s > 10$ for $K = 0.001$. Only for very rare diseases ($K < 0.0001$) does the log risk model perform well regardless of the value of λ_s.

2.4 The probit risk model

The probit model of risk, also called the liability threshold model, was introduced by Falconer (1965), and further refined by Reich et al. (1972) and Falconer and Mackay (1996). Due to its compatibility with structural equation modelling and the popularity of the Mx program (Neale and Cardon, 1992), it has come to be used as the dominant model for twin studies of binary traits (Rijsdijk and Sham, 2002). Outside of family studies, it has also been used to study the potential limits of genetic risk prediction (Wray et al., 2010), and has even been important in influencing how many non-statisticians develop their theories of disease (see for instance Haegert (2004)).

The link function for the probit risk model is

$$p = \Phi(\eta), \tag{2.70}$$

where Φ is the cumulative distribution function of the standard normal distribution. Substituting this into Equation 2.11 gives a probability density of

$$f(p) = \frac{1}{\sigma\phi\left(\Phi^{-1}(p)\right)}\phi(\frac{\Phi^{-1}(p) - \mu}{\sigma}), \tag{2.71}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution (or quantile) function of the standard normal distribution.

2.4.1 Relationship to the liability threshold model

The probit risk distribution in Equation 2.70 is derived from the liability threshold model. The liability threshold model assumes that individuals have a liability score $L \sim N(0, 1)$, and an individual is assumed to have the disease if L is larger than some threshold T. A simple form of the liability model assumes that L can be expressed in terms of an additive genetic component A and an environmental component E as

$$L = A + E, \tag{2.72}$$

where $A \sim N(0, h^2)$, $E \sim N(0, 1 - h^2)$ and $A \perp E$.

We can express $A = hz$ where $z \sim N(0, 1)$, and thus the distribution of genetic disease probabilities is

$$\begin{aligned} p &= P(A + E > T) \\ &= P(E > T - A) \\ &= \Phi\left(-\frac{T - hz}{\sqrt{1 - h^2}}\right) \\ &= \Phi(\eta). \end{aligned} \tag{2.73}$$

We thus see that the liability threshold model is equivalent to the probit model with

$$\mu = -\frac{T}{\sqrt{1 - h^2}} \tag{2.74}$$

$$\sigma = \sqrt{\frac{h^2}{1 - h^2}}, \tag{2.75}$$

and likewise

$$T = -\frac{\mu}{\sqrt{1+\sigma^2}} \tag{2.76}$$

$$h^2 = \frac{\sigma^2}{1+\sigma^2} \tag{2.77}$$

A note on the ACDE liability model

Liability threshold modelling is often extended to partition the liability in more detail. A general formulation is the “ACDE” model, where

$$L = A + C + D + E, \tag{2.78}$$

and where A is an additive genetic risk score, D is a dominant genetic risk score, C is an environmental risk shared between family members and E is non-shared environmental risk. All these terms have their own individual variances σ_X^2, and $\sum_X \sigma_X^2 = 1$.

As we have already seen, the correlation in additive risk score A is ρ_{ij}, and as we saw in Section 2.2.3 the correlation in a fully dominant risk score is $p_2 = p(IBD = 2)$. The correlation in common environment is by definition 1. It is this formulation that is generally used in twin studies, where the model is fitted (ideally by maximum likelihood, though often by approximate methods) to a set of identical twins (i.e. $\rho_{ij} = 1$ and $p_2 = 1$) and non-identical twins (i.e. ρ_{ij} =0.5 and p_2 =0.25). In practice, having only two distinct levels of relatedness means that only two parameters can be fitted, so in general we either set $D = 0$ (the “ACE” model), or $C = 0$ (the “ADE” model, generally used for twins reared apart). Note that this formulation is not specific to the liability threshold model, similar covariance relationships can be defined for any model that is expressed in terms of a normally distributed risk score.

2.4.2 Calculating parameters

By definition, the threshold T is selected such that a proportion K of individuals have a value greater than T, i.e. $T = \Phi^{-1}(1-K)$. We can thus write K in terms as μ and σ as

$$K = 1 - \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right). \tag{2.79}$$

The heritability, provided by Wray et al. (2010) using equations derived by Reich et al. (1972), is given by

$$h^2 = 2\frac{T - T_s\sqrt{1-(T^2-T_s^2)(1-T/z)}}{z + T_s^2(i-T)}, \tag{2.80}$$

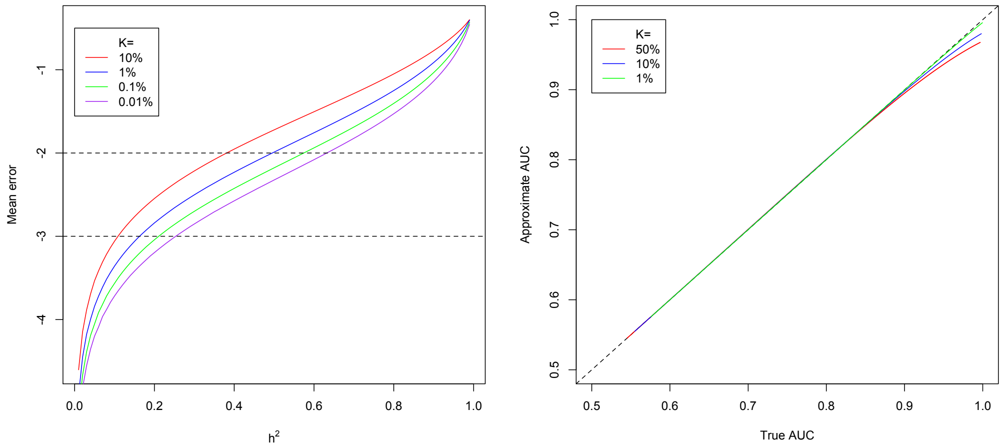
where $T_s = \Phi^{-1}(1-\lambda_s K)$, and $z = \frac{\phi(T)}{K}$.

2.4.3 Case and control distributions

Wray et al. (2010) calculated an approximate normal density for the genetic liability A in cases as

$$P(A|d=1) \approx \frac{1}{\sqrt{h^2(1-h^2z(z-T))}}\phi(\frac{zh^2-A}{\sqrt{h^2(1-h^2z(z-T))}}). \tag{2.81}$$

This is an approximation to the exact density



(a) **(b)**

Figure 2.4: a) The $\log_{10}$ mean error (average squared distrance from the true value) of the normal density approximation to the genetic liability in cases $P(A|d = 1)$, as a function of prevalence K and heritability h^2. b) The Area Under the ROC Curve calculated using the exact and approximate equations, as a function of K and h^2

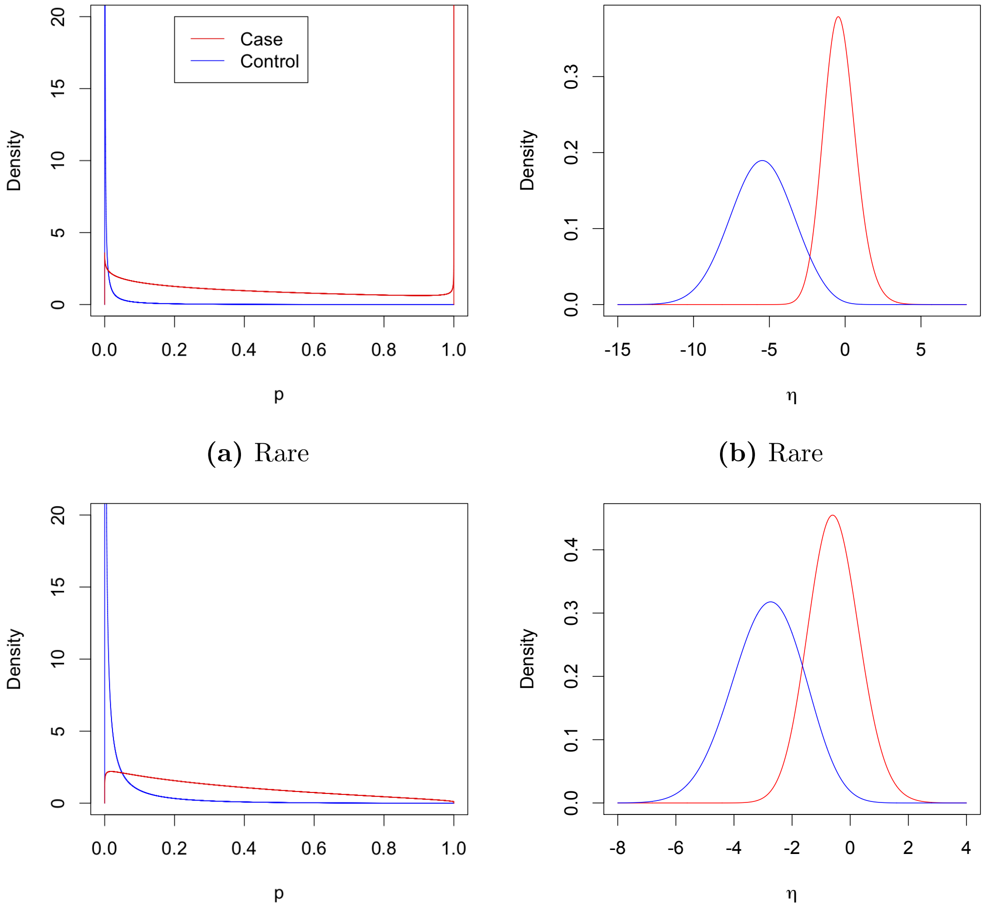
$$P(A|d = 1) = \frac{P(A + E > T|A)P(A)}{P(A + E > T)} \tag{2.82}$$

$$= \frac{1}{K}\Phi\left(\frac{A - T}{\sqrt{1 - h^2}}\right)\phi\left(\frac{A}{\sqrt{h^2}}\right). \tag{2.83}$$

Similar expressions exist for the genetic liability in controls.

Figure 2.4A shows the mean accuracy of this normal approximation as a function of the heritability and prevalence. Note that there is significant error in this approximation at high heritabilities, particularly if the prevalence is also high.

This approximation is used by Wray et al. (2010) to calculate the maximum possible predictive capacity of genetic risk prediction for various dis-



(a) Rare

(b) Rare

(c) Common

(d) Common

Figure 2.5: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$), under the probit model.

eases. The error in this function for highly heritable common diseases suggests that these values could be in error. However, Figure 2.4B shows that, in practice, this error only serves to slightly underestimate the very largest AUCs for very common $K > 0.1$ diseases, which does not substantially change the conclusions drawn from these results.

Examples of the distributions of η and p in cases and controls are shown in Figure 2.5.

2.4.4 Relationship to probit regression and latent variable modelling

Probit regression is a form of latent variable regression introduced by Bliss (1935) in 1935 as a model for bio-assay analysis. It was the dominant method of analysis for dichotomous traits until the 1960s, when the logistic regression model began to overtake it (see discussion of the logistic model below).

The probit model is a latent variable model, based on a continuous score

$$y = \beta_0 + \sum_i \beta_i x_i + e, \tag{2.84}$$

where $\vec{\beta}$ are parameters of the model, $\vec{x}$ are observed variables, and $e \sim N(0,1)$ is an unobserved (or latent) variable. The observed outcome is a binary indicator variable

$$d(y) = \begin{cases} 1 & \text{if } y > 0; \\ 0 & \text{otherwise.} \end{cases} \tag{2.85}$$

The probit regression model is fitted to determine the values of $\vec{\beta}$.

We write $X = \beta_0 + \sum_i \beta_i x_i$, which, given a large number of predictors, can be approximated as $X \sim N(\mu_x, \sigma_x^2)$, where

$$\mu_x = \beta_0 + \sum_i 2 f_i \beta_i \tag{2.86}$$

$$\sigma_x^2 = \sum_i 2 f_i (1 - f_i) \beta_i^2. \tag{2.87}$$

We can write the probability of $d = 1$ as

$$\begin{aligned} P(d=1) &= p(y>0) \\ &= P(X+e>0) \\ &= P\left(\frac{X-\mu_x}{\sqrt{1+\sigma_x^2}} + \frac{e}{\sqrt{1+\sigma_x^2}} > -\frac{\mu_x}{\sqrt{1+\sigma_x^2}}\right) \\ &= P(A+E>T), \end{aligned} \tag{2.88}$$

i.e. equivalent to the liability threshold model where

$$\begin{aligned} h^2 &= \frac{\sigma_x}{\sqrt{1+\sigma_x^2}} \\ &= \frac{\sum_i 2f_i(1-f_i)\beta_i^2}{\sqrt{1+\sum_i 2f_i(1-f_i)\beta_i^2}} \end{aligned} \tag{2.89}$$

$$\begin{aligned} T &= -\frac{\mu_x}{\sigma_x^2} \\ &= -\frac{\beta_0 + \sum_i 2f_i\beta_i}{\sqrt{1+\sum_i 2f_i(1-f_i)\beta_i^2}}. \end{aligned} \tag{2.90}$$

We can use this to fit the liability threshold or probit model directly from the results of probit regression, and thus calculate the variance explained by a set of genetic markers. While this is generally not used as a method for calculating heritability, if the liability threshold model is, in fact, the true model of genetic risk, this method should give the best approximation to the true variance explained by a set of genetic predictors.

2.5 The logit risk model

The general logit-normal (or logistic-normal) distribution was first defined by Mead (1965) in 1965, who noted that its moments have no analytic closed form, and its parameters can only be estimated iteratively (and even then only with some difficulty). However, the logit link itself is much older, having been used in bio-assay since the 1930s (see discussion of logistic regression below).

The logistic-normal distribution has been used previously to model serial observations under a random effects model (Stiratelli et al., 1984), but I believe has only been used directly in quantitative genetics once. Commenges et al. (1995) used a logistic-normal model to test hypotheses about familial aggregation in Alzheimer's disease conditional on known risk factors, much like the standard use of the probit model described above.

The implicit importance of the logit model is much larger than its lack of direct application may suggest. The most common methods used in modern statistical genetics, multiplicative odds ratio analysis and logistic regression, both implicitly assume the existence of logit-normally distributed risk. In essence, a model of genetic risk in the population is implicitly assumed by the methodology of almost all human complex disease genetics, but almost never directly investigated. This disconnect between the common usage of the regression technique and the infrequent use of the limiting normal has been noted in other fields (Frederic and Lad, 2003).

The link function for the logit risk model is

$$p = (1 + \exp(-\eta))^{-1} : \eta \sim N(\mu, \sigma), \tag{2.91}$$

and the density is

$$f(p) = \frac{1}{\sigma p(1-p)}\phi\left(\frac{1}{\sigma}log\left(\frac{p}{1-p}\right) - \frac{\mu}{\sigma}\right). \tag{2.92}$$

Note that η is equal to the log-odds of disease

$$\begin{aligned} log(O) &= log\left(\frac{p}{1-p}\right) \\ &= \eta. \end{aligned} \tag{2.93}$$

2.5.1 Calculating parameters

As with the moments of the logit-normal, none of the parameters of the logit normal have closed-form analytic solutions. Instead, they must be calculated by numeric integration.

The prevalence is given by

$$\begin{aligned} K &= E[p] \\ &= \int_0^1 pf(p)dp. \end{aligned} \tag{2.94}$$

To calculate the relative recurrence ratio, we need to look at the bivariate distribution. Suppose we have two individuals with a relatedness coefficient ρ. We model their genotypic risks as

$$p_i = \frac{1}{1+\exp(-\eta_i)}, \tag{2.95}$$

where $\vec{\eta} = (\eta_1, \eta_2)$ are jointly normally distributed with a mean of μ and a covariance

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{2.96}$$

We can transform $\vec{\eta}$ into independent standard normals $\vec{x}$ by noting that

$$\vec{\eta} = \mu + B\vec{x}, \tag{2.97}$$

where B is the Cholesky decomposition of Σ, such that $BB' = \Sigma$ and thus

$$B = \sigma \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}. \tag{2.98}$$

From this, we can transform p_i, giving

$$x_1 = \frac{1}{\sigma}\left[log\left(\frac{p_1}{1-p_1}\right) - \mu\right] \tag{2.99}$$

$$x_2 = \frac{1}{\sigma\sqrt{1-\rho^2}}\left[log\left(\frac{p_2}{1-p_2}\right) - \mu - \sigma\rho x_1\right]. \tag{2.100}$$

The determinant of the Jacobian of this transformation is

$$\left|\frac{d\vec{x}}{d\vec{y}}\right| = \frac{1}{\sigma^2\sqrt{1-\rho^2}\prod_{i=1}^{2} p_i(1-pi)}, \tag{2.101}$$

thus the joint density of risk is given by

$$g(p_1, p_2) = \frac{\phi(\vec{x})}{\sigma^2\sqrt{1-\rho^2}p_1(1-p_1)P_2(1-p_2)} \tag{2.102}$$

From this we can calculate λ_R

$$\lambda_R = \frac{\int\int p_1 p_2 f(p_1, p_2) dp_1 dp_2}{K^2}. \tag{2.103}$$

2.5.2 Fitting the logit risk model numerically

To find parameters μ and σ given parameters K and λ_s, we find values that minimize the error function

$$Error(\mu,\sigma) = \left(\sqrt{E[p_1p_2|\mu,\sigma]} - \sqrt{\lambda_s K^2}\right)^2 + (E[p|\mu,\sigma] - K)^2 . \quad (2.104)$$

I use the Nelder-Mead algorithm (Nelder and Mead, 1965) implemented in the statistical language R. Note that the convergence speed and reliability of this procedure can be very dependent on the initial values of μ and σ. We can get a good initial guess by expressing the logit risk in terms of the probit model

We can express the probit model on the logit scale

$$\eta_{probit} = \Phi^{-1}((1+e^{-\eta_{logit}})^{-1}) \quad (2.105)$$

$$\frac{d\eta_{probit}}{d\eta_{logit}} = \left[\phi(\Phi^{-1}((1+e^{-\eta_{logit}})^{-1})(1+e^{-\eta_{logit}})(1+e^{\eta_{logit}})\right]^{-1} (2.106)$$

We can then get the density of the logit risk score given the probit model

$$f(\eta_{logit}|\mu_{probit},\sigma_{probit}) = f(\eta_{probit}|\mu_{probit},\sigma_{probit})\frac{d\eta_{probit}}{d\eta_{logit}}, \quad (2.107)$$

which can in turn give us the expectation and variance of the logit risk variable under the probit model, which we use as an initial guess for the parameters μ and σ under the logit risk model

$$\mu_{init} = \int_{\eta} \eta f(\eta | \mu_{probit}, \sigma_{probit}) d\eta \tag{2.108}$$

$$\sigma_{init} = \int_{\eta} (\eta - \mu_{init})^2 f(\eta | \mu_{probit}, \sigma_{probit}) d\eta. \tag{2.109}$$

2.5.3 Case and Control Distributions

There is no particularly elegant way of describing the distribution of the probability p and the risk score η in cases and controls. Instead we can only use the general equations given in Section 2.1.2.

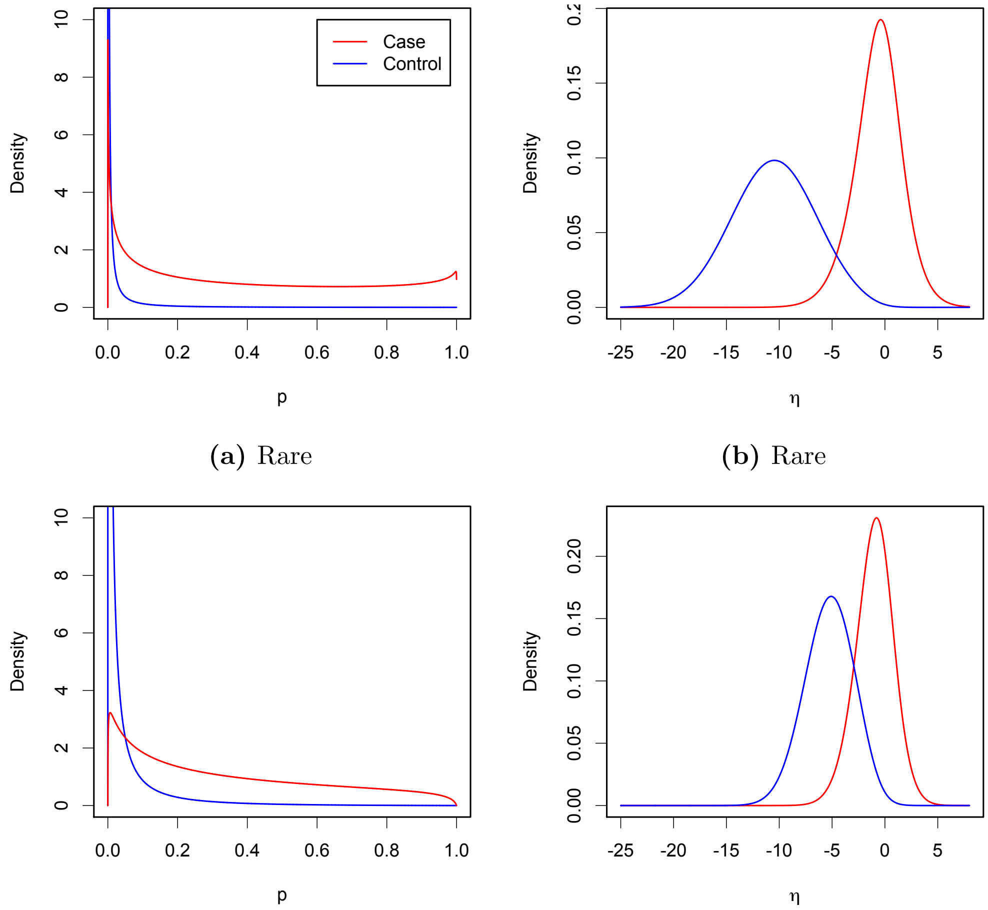
Examples of the distributions of η and p in cases and controls are shown in Figure 2.6.

2.5.4 Relationship to the multiplicative odds ratio model

Odds ratios are widely used to quantify differences between groups, and to make probabilistic predictions for individuals given group membership (see discussion in Morgan and Teachman (1988) for example). Odds ratios are the most widely used summary statistic in medical studies (Bland and Altman, 2000), mostly due to their utility in meta-analyses, though they are not without their detractors (Sackett et al., 1996). In genetics, the odds ratio has become the dominant method for summarising disease associations, largely due to its connection with logistic regression.

Given an exposure $a \in \{0, 1\}$, and an outcome $d \in \{0, 1\}$, we can define the probability conditional on exposure status $a = i$ as $p_i = P(d = 1 | a = i)$. The odds ratio for exposure a is then defined as

$$r_a = \frac{p_1}{1 - p_1} \frac{1 - p_0}{p_0}. \tag{2.110}$$



(a) Rare

(b) Rare

(c) Common

(d) Common

Figure 2.6: The case and control distributions of probability p and risk score η for a rare disease ($K = 0.01, \lambda_s = 9$) and a common disease ($K = 0.05$, $\lambda_s = 3$)

Odds ratios in genetics

Throughout this thesis, I will refer to the effect size of a genetic association in terms of the odds ratios r_{het} and r_{hom}, where

$$r_{het} = \frac{p_1(1-p_0)}{p_0(1-p_1} \tag{2.111}$$

$$r_{hom} = \frac{p_2(1-p_0)}{p_0(1-p_2}, \tag{2.112}$$

where $p_x = P(d = 1|g = x)$ are the disease probabilities conditional on risk allele count x. We will sometimes refer to the genotypic odds ratio $r = r_{het} = \sqrt{r_{hom}}$ (also called the additive odds ratio).

We can rearrange the odds ratio definitions to give expressions for the disease probabilities for non-wild type genotypes in terms of the wild-type disease probability

$$p_1 = \frac{p_0 r_{het}}{1 - p_0 + p_0 r_{het}} \tag{2.113}$$

$$p_2 = \frac{p_0 r_{hom}}{1 - p_0 + p_0 r_{hom}}. \tag{2.114}$$

Given a prevalence K we can get the value of p_0 by solving the equation

$$p_0(1-f)^2 + p_1 2f(1-f) + p_2 f^2 = K, \tag{2.115}$$

which can be solved analytically (but messily), or numerically (counterintuitively, the numeric method is likely to be more accurate (Nievergelt, 2003)).

A common analytic approximation to calculate odds ratios is to normalize the odds ratios such that their population mean is equal to 1

$$\frac{(1-f)^2}{\bar{r}} + \frac{2f(1-f)r_{het}}{\bar{r}} + \frac{f^2 r_{hom}}{\bar{r}} = 1, \tag{2.116}$$

i.e.

$$\bar{r} = (1-f)^2 + 2f(1-f)r_{het} + f^2 r_{hom}. \tag{2.117}$$

We can then set

$$\hat{r}_0 = \frac{1}{\bar{r}} \tag{2.118}$$
$$\hat{r}_1 = \frac{r_{het}}{\bar{r}} \tag{2.119}$$
$$\hat{r}_2 = \frac{r_{hom}}{\bar{r}}, \tag{2.120}$$
$$\tag{2.121}$$

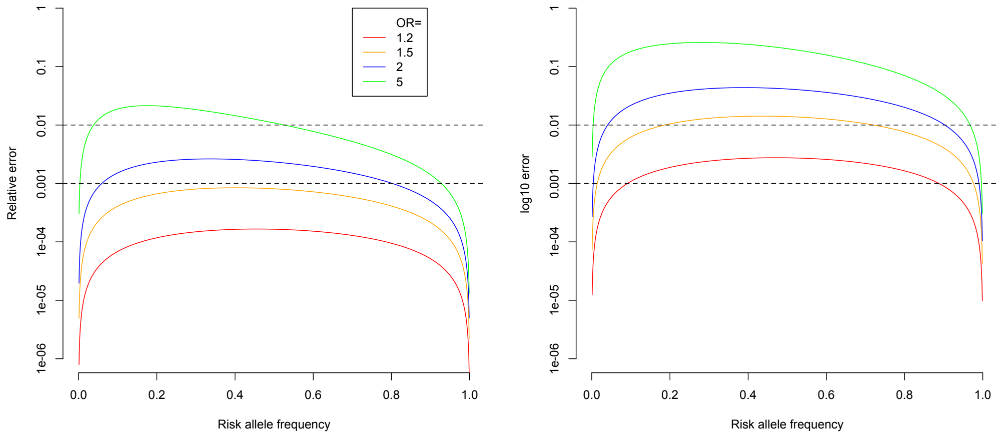
or, given a genotypic odds ratio, $\hat{r_x} = \frac{r^x}{\bar{r}}$.

We can then set the disease probabilities using these normalised odds ratios as

$$p_x = \frac{1}{1 + \frac{1-K}{\hat{r}_x K}}. \tag{2.122}$$
$$\tag{2.123}$$

This is the method used by, for instance, the genetic testing company 23andMe (Macpherson et al., 2007).

The accuracy of this approximation varies depending on the prevalence of the disease in question, and the size of the odds ratio (Figure 2.7). For a rare disease ($K = 0.01$) the approximation is accurate to within 1% for all realistic odds ratios and frequencies (and accurate to within 0.1% or less for $OR < 1.5$). For a more common disease ($K = 0.2$) the approximation is only accurate to within 1% for lower odds ratios ($OR < 1.5$). However, for odds ratios typically found within GWAS (generally $OR < 1.3$) the approximation



(a) K = 0.01 **(b)** K = 0.2

Figure 2.7: The accuracy of the odds ratio normalisation approach to genetic risk prediction, for rare and common diseases, as a function of odds ratio and risk allele frequency.

holds across prevalence and allele frequencies.

Combining independent odds ratios

Suppose we have two exposures a and b, with $p_{ij} = P(d = 1|a = i, b = j)$. A reasonable definition for these two exposures having an independent effect is if the odds ratio r_a does not depend on the value of b, and vice versa, i.e.

$$\begin{aligned} r_a &= \frac{p_{10}}{1 - p_{10}} \frac{1 - p_{00}}{p_{00}} \\ &= \frac{p_{11}}{1 - p_{11}} \frac{1 - p_{01}}{p_{01}}, \end{aligned} \tag{2.124}$$

and

$$\begin{aligned} r_b &= \frac{p_{01}}{1-p_{01}}\frac{1-p_{00}}{p_{00}} \\ &= \frac{p_{11}}{1-p_{11}}\frac{1-p_{10}}{p_{10}}. \end{aligned} \tag{2.125}$$

We can then calculate the joint odds ratio for both exposures, r_{ab}, as

$$\begin{aligned} r_{ab} &= \frac{p_{11}}{1-p_{11}}\frac{1-p_{00}}{p_{00}} \\ &= \left(\frac{p_{11}}{1-p_{11}}\frac{1-p_{01}}{p_{01}}\right)\left(\frac{p_{01}}{1-p_{01}}\frac{1-p_{00}}{p_{00}}\right) \\ &= r_a r_b, \end{aligned} \tag{2.126}$$

i.e. to combine independent odds ratios, multiply them together. Note that this justifies the genotypic odds ratio $r^2 = r_{hom} = r_{het}^2$, as it represents both alleles acting independently at a single locus.

We can generalise this to make a combined odds ratio given genotypes $\vec{x} = \{x_l\}$ across n loci with odds ratios $\vec{r} = \{r_l\}$

$$r_{\vec{x}} = \prod_{l=1}^{n} r_l^{x_l}. \tag{2.127}$$

The disease probability is thus given as

$$
\begin{aligned}
p_{\vec{x}} &= \frac{r_{\vec{x}} p_0}{1 - p_0 + r_{\vec{x}} p_0} \\
&= \frac{1}{1 + \frac{1}{r_{\vec{x}}} \frac{1-p_0}{p_0}} \\
&= \frac{1}{1 + \exp(-\eta)}, \qquad (2.128)
\end{aligned}
$$

where

$$
\begin{aligned}
\eta &= log\left(\frac{p_0}{1-p_0} r_{\vec{x}}\right) \\
&= log\left(\frac{p_0}{1-p_0}\right) + log\left(r_{\vec{x}}\right) \text{ Fixed brackets} \\
&= log\left(\frac{p_0}{1-p_0}\right) + \sum_{l=1}^{n} x_l log(r_l). \qquad (2.129)
\end{aligned}
$$

Again, by the central limit theorem η tends towards a normal distribution with parameters

$$
\begin{aligned}
\mu &= log\left(\frac{p_0}{1-p_0}\right) + \sum_{l=1}^{n} 2 f_l log(r_l) \qquad (2.130) \\
\sigma^2 &= \sum_{l=1}^{n} 2 f_l (1 - f_l) log(r_l)^2 \qquad (2.131)
\end{aligned}
$$

Thus the logit risk model is asymptotically equivalent to the assumption that odds ratios act independently.

2.5.5 Relationship to logistic regression

The logistic function, $g = (1+\exp(-\eta))^{-1}$, has been used since the 19th century as a description of population growth given limited resources (Verhulst, 1838), and in the early 20th century was found to accurately model many physiochemical responses (Reed and Berkson, 1929). It was first used as a regression model by Berkson (1944), who introduced it as an alternative to probit regression (and also introduced the name "logit"). Berkson later laid down in some detail the theoretical and empirical arguments underlying logit and probit link functions (Berkson, 1951).

In the last few decades the logit link has succeeded the probit link as the dominant form of regression model for binary outcomes(Cramer, 2003). It is very widely used in medical literature (though often imperfectly (Bagley et al., 2001)), and is the dominant method for performing genome-wide association studies under the presence of confounding factors, particularly with the rise of principal component methods to control population stratification (Price et al., 2006).

The logistic regression model has the form

$$p = (1+\exp(-\eta))^{-1} \text{ where} \tag{2.132}$$

$$\eta = \beta_0 + \sum_i x_i \beta_i. \tag{2.133}$$

This is equivalent to equations 2.128 and 2.129 with parameters $\beta_0 = \frac{p_0}{1-p_0}$ and $\beta_i = log(r_i)$. We can thus see that, given an arbitrarily large number of predictors, the logistic regression model is approximated by the logit-normal risk model. This also provides us with a method of fitting the logit risk model from genetic data, using the results of logistic regression.

2.6 Comparing models of risk

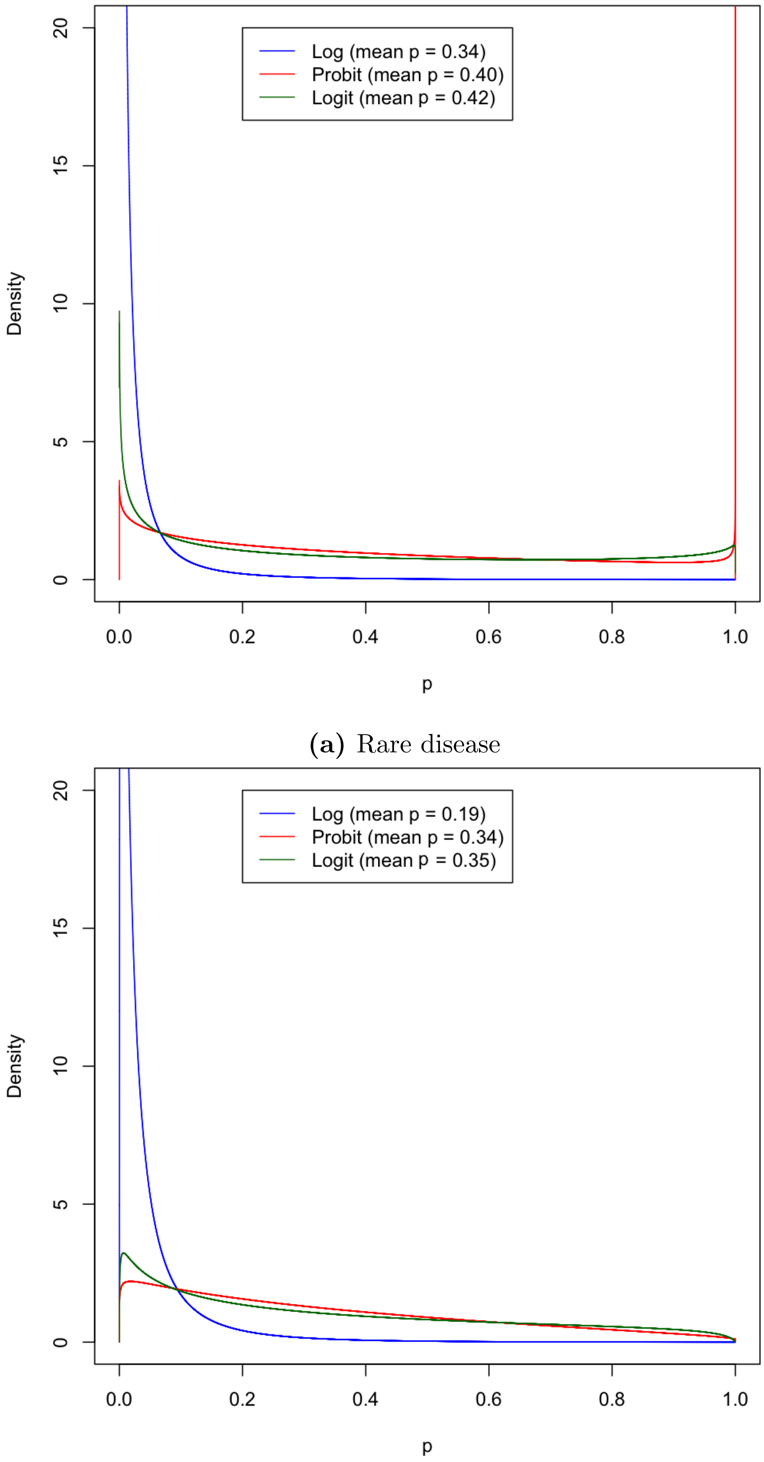
In the previous sections I outlined three continuous models of genetic risk and noted the different assumption that underlie them. In this section I will examine the ways in which these models differ in their predictions about the distribution of genetic risk in the population.

I will look at the predicted distribution of disease probability in cases across different models, and look in more detail at the differences between the logit and probit models. I will then consider the predicted relative recurrence risks and predicted ROC curves for the different models.

2.6.1 Comparing disease probability distributions in cases

Figure 2.8 shows the distribution of p in affected individuals under the three different models. In both cases, the log model produces a smaller mean p and a left-shifted distribution relative to the log and probit models. Additionally, in both scenarios the logit and probit models give relatively similar distributions, with approximately the same mean value of p. Disregarding a sharp peak near $p = 1$ for the probit model in 2.8a, the logit model seems to show slightly more density towards the ends of the distribution, and the probit model shows more density towards the middle.

In these comparisons, the log model stands out as clearly underestimating both the degree of enrichment of genetic risk in cases, predicting very few cases to have a high risk compared to the other two models. On the scale that we have examined, however, the logit and probit models appear similar, and it is difficult to infer the significance of these deviations. We can look at the differences between these two models in more detail by producing values of p given the probit model, and projecting them onto logit space using



(a) Rare disease

(b) Common disease

Figure 2.8: The distribution of genetic disease probabilities in randomly selected cases under the three different risk models, for a relatively rare, highly heritable disease ($K = 0.01$, $\lambda_s = 9$), and a more common, mildly heritable disease ($K = 0.05$, $\lambda_s = 3$). The legend gives the mean value of p in cases.

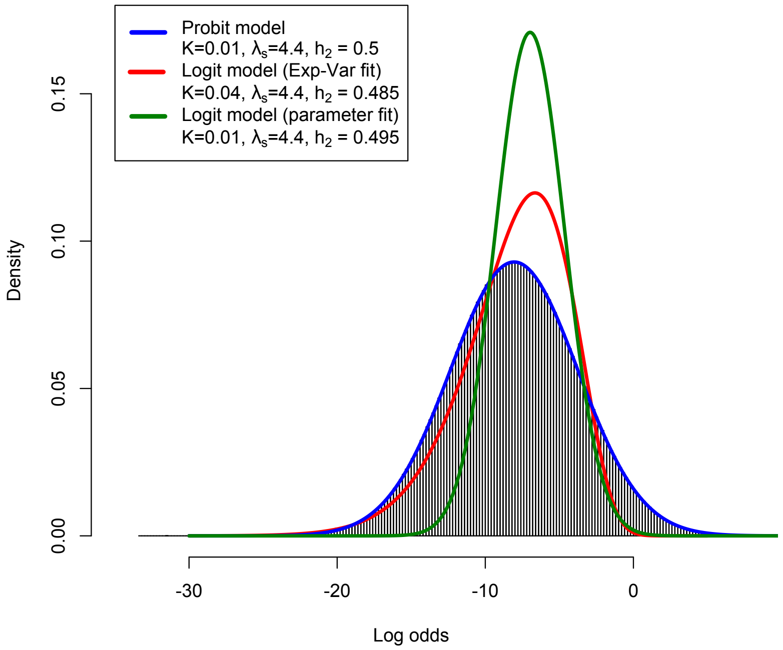


Figure 2.9: Different logistic approximations to a probit distribution. The exact distribution of the logit score under the probit model for $K = 0.01$ and $h^2 = 0.5$ is shown in blue (with bars representing a histogram of samples from the distribution). The red line shows a logistic normal fitted to have the same the mean and variance as the probit model, and the green line shows a logistic normal fitted to have the same K and λ_s values as the probit model.

$$p = \Phi(\eta_{probit}) = (1 + \exp(-\eta_{logit}))^{-1} \tag{2.134}$$

$$\eta_{logit} = log\left(\frac{\Phi(\eta_{probit})}{1 - \Phi(\eta_{probit})}\right). \tag{2.135}$$

Figure 2.9 shows this projection for a probit model with $h^2 = 0.5$ and $K = 0.01$ (bars and blue line). The red and green lines show two logit models: the red line showing the logit model with the same mean and variance on the logit scale as the probit model, and the green line showing the logit model with the same K and λ_s values as the probit model.

We can see that no logit model accurately models the projected probit distribution, due to the high kurtosis of the projection. A model with the same mean and variance, while having similar values of λ_s and h^2, predicts too high a prevalence. The model that has the same K and λ_s also has a similar h^2, but follows a very different curve with a much smaller variance.

This highlights clearly the ambiguity involved in comparing models or results parameterised on these difference scales. Furthermore, we can see that a logit model designed to closely mimic the probit model's risk distribution produces divergent parameters. Despite their superficial similarity, these models cannot be viewed as approximations to each other.

2.6.2 Comparing relative recurrence risk

None of the above distributions reflect any quantities that can be observed in the population. One long measured and studied property in the genetics of disease is the increase in disease risk in relatives of affected individuals, estimations of which are often used to draw conclusions about the genetic architecture of the disease (Compston and Coles, 2008; Sawcer, 2009; Brown et al., 2000).

As we saw in equation 2.58, under the log risk model

$$\lambda_r = \exp(\sigma\rho). \tag{2.136}$$

Substituting $\sigma = 2log(\lambda_s)$ gives

$$\begin{aligned}\lambda_r &= \exp(2log(\lambda_s)\rho) \\ &= \lambda_s^{2\rho}. \end{aligned} \tag{2.137}$$

This means that given the log risk model (and thus also given multiplicative relative risk), the recurrence ratio in relatives λ_r falls off with the logarithm of the coefficient of relatedness ρ. Deviation from this log-linear relationship is often interpreted as evidence of genetic non-additivity (Brown et al., 2000). However, deviations from this relationship could also be evidence that a different model is at play.

Figure 2.10a shows the fall-off in λ_r as a function of ρ for the three models (all with $K = 0.05$ and $\lambda_s = 3$). All models give very similar predictions, though there are slight differences between the models (Figure 2.10b). This includes up to a 6% increase in the risk ratio for probit and logit relative to the log model for highly related individuals ($\rho > 0.5$, including identical twins and siblings of consanguineous parents), and a corresponding decrease in risk for more distance relatives (peaking at a 3% difference at $\rho = 0.25$, or avuncular relationships).

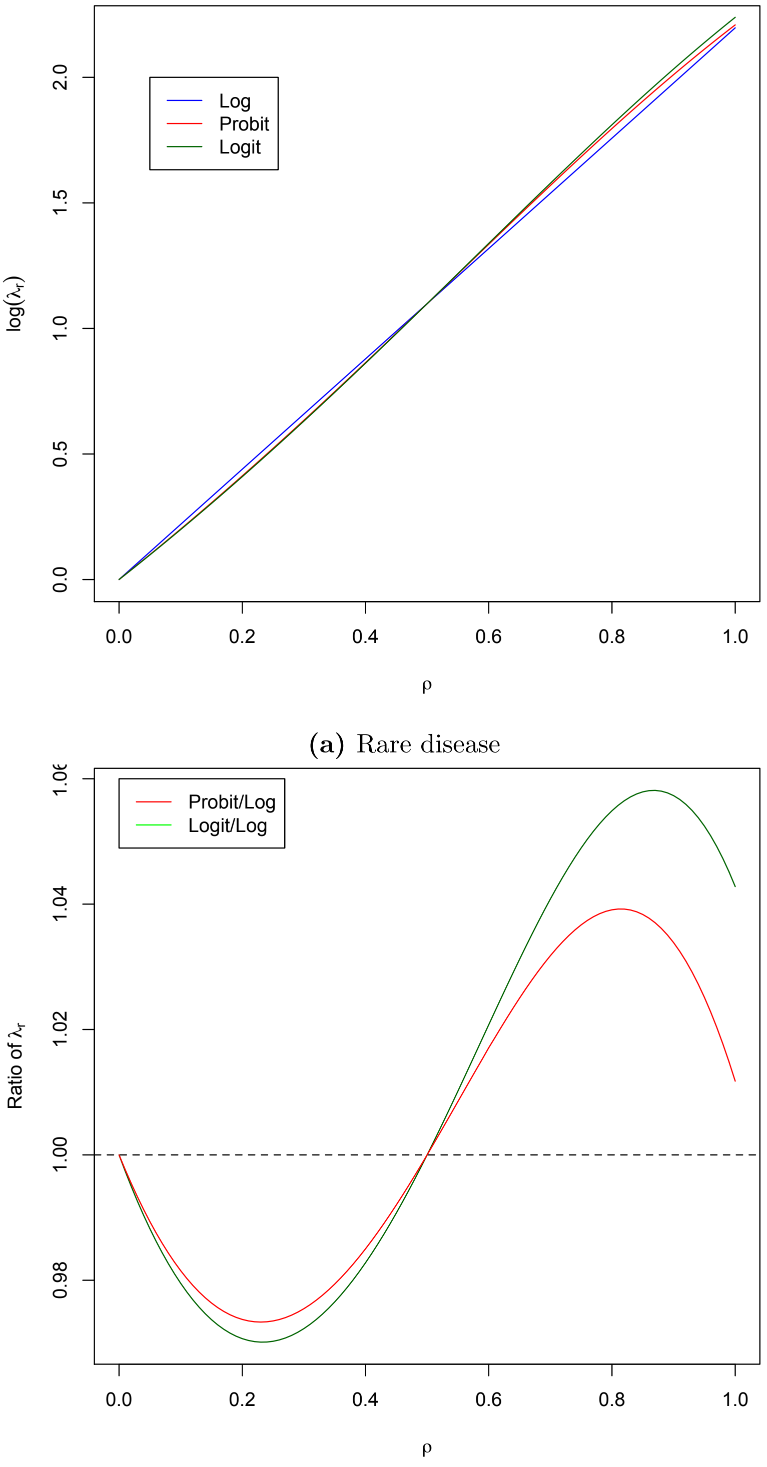
These differences are on the limit of what can be detected in family studies: for 80% power to detect a 3% deviation from $\lambda_s = 3$ at $p < 0.05$ would require over 38 000 avuncular pairs. In addition, even if the log risk model could be rejected, we would not be able to say whether this difference was due to a different additive model applying, or merely a non-additive model. In theory measurements for a large range of different relative types could resolve this question, but in practice an even larger number of relatives would be required. In short, there is no plausible family study that could distinguish between these three models of genetic risk.

2.6.3 Comparing ROC curves for risk prediction

Many authors have attempted to make predictions about how useful genetic risk prediction could be if we managed to account for the total load of genetic risk predicted to exist by family studies. However, the results have been in many cases divergent, even when authors apply their methods to the same datasets. Some authors draw the conclusion that genetic risk prediction is unlikely to ever be of high utility (Clayton, 2009), while others conclude that genetic risk prediction could be of great use (Wray et al., 2010). I discussed the general question of how and when genetic risk prediction could be useful in the introduction, but here I will focus more specifically on how the model used can change your conclusions about the utility of genetic risk prediction.

Figure 2.11 shows the predicted ROC curves for diseases with a prevalence of $K = 1/200$ and $K = 1/20$, and a sibling relative risk of $\lambda_S = 9$ and $\lambda_s = 3$ for the three models. For the rarer disease all the models give divergent answers, with the probit model giving an AUC of 0.98, a logit model an AUC of 0.96, and the log model an AUC of 0.89. For the common disease, the logit and probit models agree on an AUC of 0.93, though with a different sensitivity-specificity trade-off, and the log model gives a much lower AUC of 0.84.

The low predictive accuracies for the log model are probably due to the problems mentioned in Section 2.3.5, and I will disregard these values. It therefore seems like a plausible maximum AUC for rare diseases likely lies between 0.96 and 0.98, and common diseases around 0.93, as predicted by the logit and probit models. However, the significant variability, both in the AUC values and in the shape of the ROC curves, highlights the degree



(a) Rare disease

(b) Common disease

Figure 2.10: a) The log relative risk ($log(\lambda_r)$) under the three models as a function of the coefficient of relatedness ρ, Parameters are $K = 0.05$, $\lambda_s = 3$ b) The ratio of probit and logit λ_r values to log λ_r values, as a function of ρ.

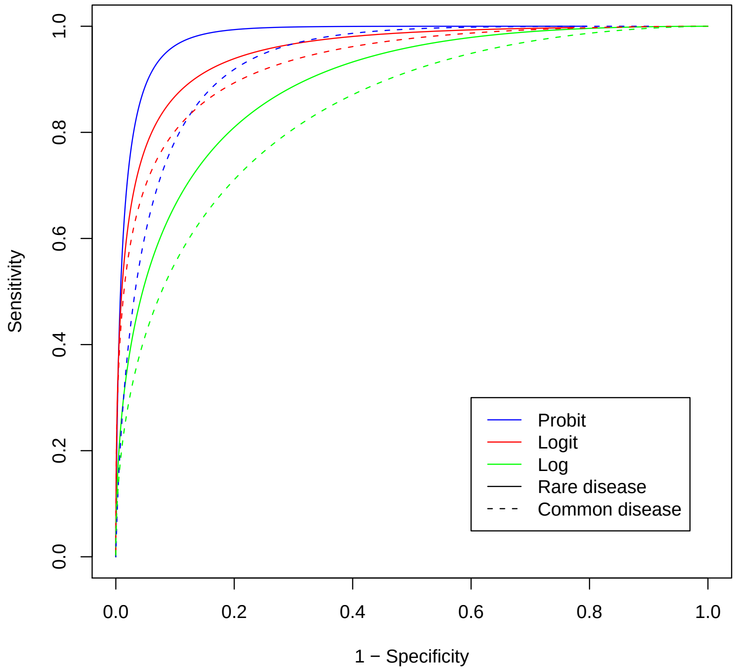


Figure 2.11: The ROC curves for the log, logit and probit models of disease risk for a rare disease with a prevalence $K = 1/200$ and sibling relative risk of $\lambda_S = 9$, and a common disease with $K = 1/20$ and $\lambda_s = 3$, given that all genetic risk has been explained. The corresponding AUCs are 0.89, 0.96 and 0.98 respectively for the rare disease, and 0.84, 0.93 and 0.93 for the common disease.

to which forecasts of the future utility of genetic risk prediction are model specific.

2.7 Conclusion

2.7.1 Summary of models

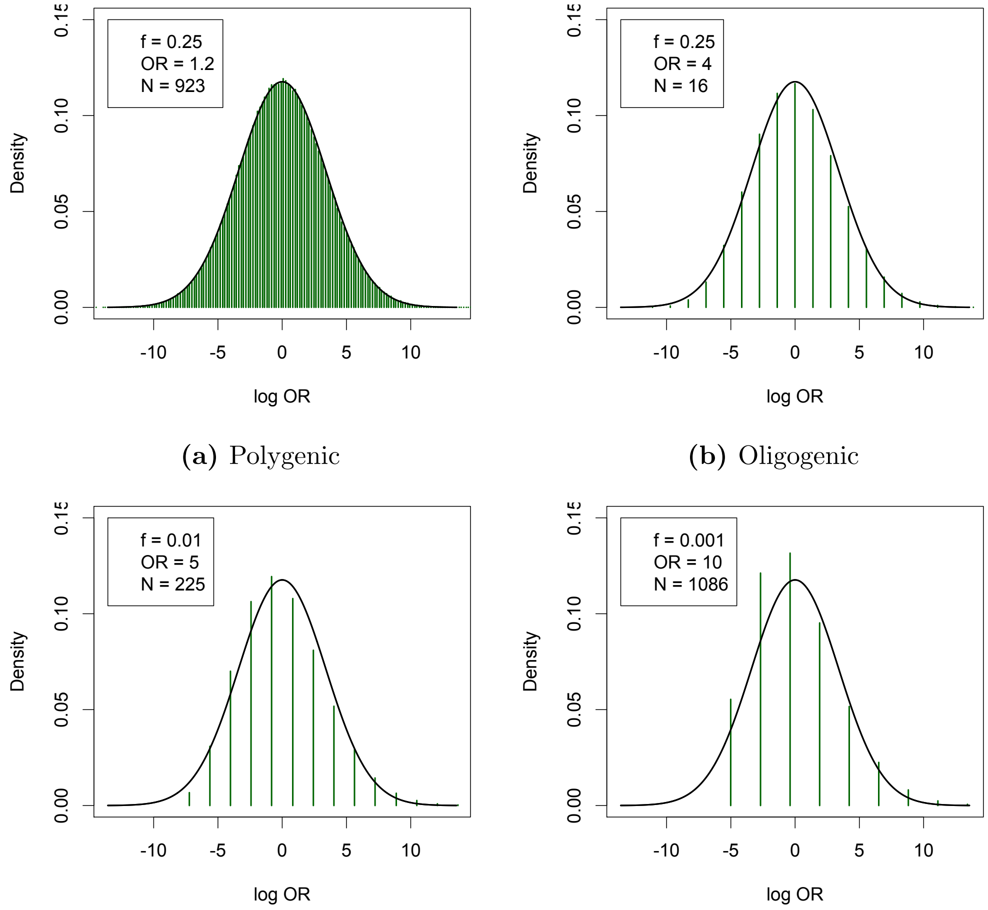
As we have seen, the three models that we have examined can each be seen as the natural result of the assumptions made in one or more major statistical method. We can summarise the three models, and their corresponding methods, using the following table:

Model	Link function	Equivalent models/methods
Log risk	$p = \exp(\eta)$	Risch model, multiplicative relative risk
Probit risk	$p = \Phi(\eta)$	Liability threshold model, latent variable model, probit regression
Logit risk	$p = (1 + \exp(-\eta))^{-1}$	Multiplicative odds ratios, logistic regression

We have seen that these models differ in their predictions about the distribution of risk in populations. Some of these differences are minor (they all have a similar relationship between coefficient of relatedness and relative recurrence risk), but some are large (they give divergent predictions about the maximum utility of genetic risk prediction).

2.7.2 Limitations of this approach

An important caveat is that the analyses of these three models above are all built on two major assumptions. The first is that the risk score η can be approximated by a normal distribution, and the second is that the risk score η is additive.



(a) Polygenic

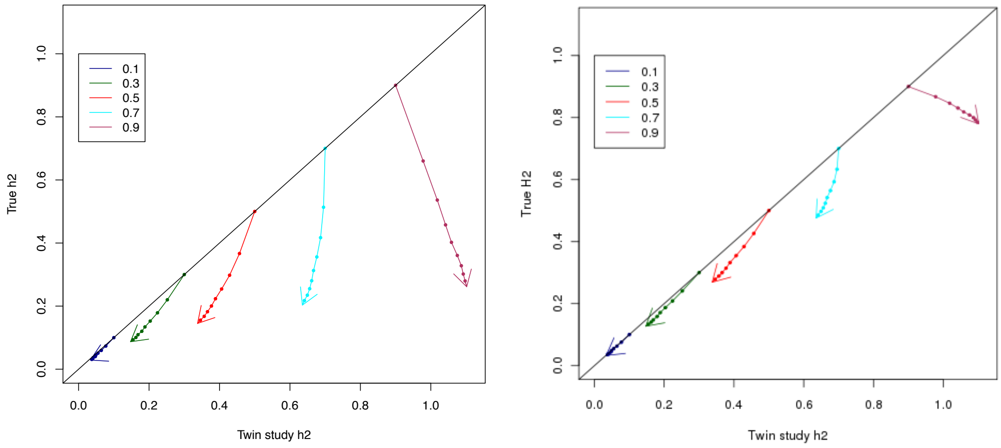
(b) Oligogenic

(c) Low-frequency

(d) Rare

Figure 2.12: The closeness of fit to the normal distribution for variants with different frequencies and odds ratios. The black line represents the normal approximation, and the green bars are odds ratios sampled from the model. The number of variants N is chosen to have $\lambda_s = 3$.

Speaking to the first assumption, Figure 2.12 illustrates how well this approximation holds across different architectures, given the same value of λ_s. In fact, the normal approximation holds for almost all plausible genetic architectures; the approximation is very accurate for polygenic and oligogenic models, and is relatively accurate for low-frequency variants. The approximation becomes significantly less accurate for a disease driven purely by rare,



(a) Narrow sense heritability **(b)** Broad sense heritability

Figure 2.13: The affected of epistasis on heritability estimation from twin studies. The epistasis model used is the multiple threshold model of Zuk et al. (2012), in which the risk score is the minimum of N independent liability scales, each with a heritability h_p^2. The dots represent increasing N (starting with 1, increasing in the direction of the arrow), and the colours represent different values of h_p^2. The first panel shows the overestimation of the narrow-sense heritability, and the second shows the overestimation of the broad-sense heritability.

highly penetrant mutations.

As for the second, non-additivity can alter the models in two ways. Firstly, it can lead to non-normality in the risk score. However, as I mentioned in Section 2.2, it seems likely that most forms of pairwise interaction can be approximated as a normal distribution, and even risk scores based on more detailed forms of epistasis can be modelled as normal (see, for example, Zuk et al. (2012)). Secondly, as we saw for single-locus dominance in section 2.2.3, non-linearity can alter the correlation structure of risk scores in related individuals. Specifically, non-additivity reduces correlation such that $cor[\eta_i, \eta_j] < \rho$. This is turn can lead us to overestimate the heritability of the disease.

We use the model of Zuk et al. (2012) to explore this effect. Figure 2.13

examines how serious this effect will be on our estimation of heritability, and thus the results of our models. Zuk et al. (2012) showed that, under epistasis, the narrow sense heritability (i.e. the correlation in additive risk score) will be greatly overestimated by twin studies (as shown in Figure 2.13a). However, for our purposes we are more interested in the overestimation of the full heritability, which is what determines the univariate distribution of the probit score η. Figure 2.13b shows that this value is significantly less prone to overestimation than the narrow sense heritability, and is only seriously overestimated in cases where $H^2 > 0.8$.

2.7.3 Problems generated by model ambiguity

The use of methods with differing underlying models can itself create ambiguity in results. Suppose we have performed a genome-wide association study of a disease with $K = 0.05$, using logistic regression. We have identified 48 loci, each with an estimated odds ratio of 2 and a frequency of 50%. We wish to compare these results with data from of twin studies, which have found that the disease has a heritability of $h^2 = 0.8$, in order to say what proportion of genetic variance has been explained. There are three ways that we could answer this question

1. Fit the log-normal model from the data using equation 2.131, project the result onto the probit scale using Equation 2.134, calculate the variance and convert to h^2 using Equation 2.75. This gives $h^2 \approx 0.586$

2. Fit the log-normal model, calculate the value of λ_s, and use Equation 2.80 to calculate the corresponding h^2. This gives $h^2 \approx 0.634$

3. Perform probit regression on the original genetic data, and use equation 2.89 to calculate h^2. A simulation of this (generated under the logistic

model) gives $h^2 \approx 0.751$.

The technique used can alter the percentage of heritability explained from 74% to 93%. The smallest answer may lead people to invest further to discover the missing quarter of heritability, while the latter will likely lead to a conclusion that the trait is essentially solved. There is no correct answer, as the question we are asking is inherently problematic: the two results we are comparing were generated under different models. Which of the values is correct (if any) will depend on the true model underlying the genetic risk in the first place, which is unknown.

Chapter 3

Investigating new reference and target sets in genotype imputation

3.1 Introduction

Genome-wide association studies (GWAS) are based on a tag SNP approach. Genotyping arrays use a set of SNPs chosen such that, between them, they are correlated with most of the common variants in the human genome. Any common causal variants will be then be well correlated with at least one SNP on the array, and (providing a large enough sample size is genotyped) such associations can be detected via signals at these tag SNPs.

While tag SNP sets are picked using a high-density reference set, the approach of testing these tag SNPs for association in a GWAS cohort makes no assumptions about what untyped SNPs are being tested. However, it is possible to use the data in the reference set to improve the coverage of the study. The reference set tells us (at least some of) the common SNPs that exist, and allows us to place them together into multi-SNP haplotypes. We can therefore use the tag SNPs we have genotyped to match the haplotypes in our GWAS samples to haplotypes in our reference set, and use this matching to infer these samples' genotypes at other sites. This process is called "genotype imputation", and we refer to the dataset we are predicting genotypes for as the "target set".

Genotype imputation has a number of advantages over tag SNP testing. Firstly, it allows meta-analyses to be performed even when the component studies have been performed using different sets of tag SNPs, by allowing a common set of SNPs to be imputed. Secondly, imputed genotypes, while only probabilistically predicted, are imputed using information from many surrounding SNPs, and thus are often more strongly correlated with the true genotype than any single tag SNP. This gives improved power to detect associations, especially for variants that are not well tagged by the array, and can lead to significant associations being detected that would have been missed otherwise (Huang et al., 2012). Thirdly, it allows test statistics to be produced at all sites in the reference set, which (if the reference set is high enough density) is likely to contain the true causal variant, and thus can allow the function of associated variants to be inspected.

3.1.1 Overview of imputation software and methods

The vast majority of the human genome is diploid, meaning that it is made up two copies. Each copy contains its own set of alleles, which together make up the two multi-marker haplotypes that an individual carries. To perform the haplotype matching that genotype imputation relies on, we first need to reconstruct these two haplotypes from the diploid genotypes produced by genotyping chips, by determining the phase of the alleles at each site (i.e. inferring which alleles are present on the same copy of the chromosome). This process is known as "phasing", and is the most statistically challenging aspect of imputation. The history of imputation is therefore, to a first approximation, a history of phasing techniques.

Experimental and family-based phasing techniques are as old as genetics itself, but statistical phasing techniques began being applied in the 1990s (Browning and Browning, 2011). The first statistical imputation method for unrelated individuals was the Clarke algorithm published in 1990 (Clark, 1990), which inferred the existence of haplotypes based on parsimony. Soon after methods based on Expectation-Maximisation (EM) were developed to estimate haplotype frequencies and phase small numbers of SNPs. Both of these methods are computationally expensive and relatively inaccurate, and thus did not generalise outside of small haplotype blocks (Browning and Browning, 2011). The EM method is still in use, however, for instance in the imputation function of the popular statistical genetics toolkit Plink (Purcell et al., 2007).

Most modern phasing and imputation methods are based on approximate coalescent techniques. Coalescent theory was developed in the 1980s as a way of linking population genetics to genealogy at a single gene or site (Kingman, 1982), and was extended in the 1990s to include recombination (Griffiths and Marjoram, 1996). Because coalescent theory models both polymorphism frequency and stretches of the genome shared by descent it is particularly well suited to modelling haplotype frequencies. While full coalescent theory is computationally difficult to apply in most circumstances, approximate methods have been developed that are computationally tractable (McVean and Cardin, 2005). The most widely used approximation is the Li and Stephens model (Li and Stephens, 2003), which partitions the coalescence likelihood into a series of sequential conditional approximations, which are in turn calculated using a Hidden Markov Model that includes recombination and mutation.

The first piece of software to use the approximate coalescent was PHASE (Stephens et al., 2001). A faster technique, fastPhase (Scheet and Stephens, 2006) (also implemented in BIMBAM (Servin and Stephens, 2007)), was introduced in 2006; this was also the first software to perform genotype imputation *per se*. Other imputation programs using the same approach include IMPUTE (Marchini et al., 2007), IMPUTE2 (Howie et al., 2009), MaCH (Li et al., 2010) and SHAPEIT (Delaneau et al., 2012).

Not all imputation programs are based on an approximate coalescent model. The imputation program Beagle (Browning and Browning, 2007, 2009), while also based on a Hidden Markov Model approach, does not explicitly model mutation and selection, instead using a haplotype clustering model to perform phasing and imputation (Browning, 2006). In contrast, QCall (Le and Durbin, 2011) (an imputation program for sequencing data) performs imputation by directly fitting mutations to a sequence of sampled ancestral recombination graphs.

3.1.2 New reference and target sets in imputation

Imputation methods in GWAS originally used the HapMap phase 2 reference set (Frazer et al., 2007), which contained data on 400 haplotypes from three ethnic groups. This served as a successful reference set for common SNPs in Europeans for the first wave of GWAS, allowing around 75% of common SNPs to be imputed with accuracies of above 98% (Marchini et al., 2007), and allowing meta-analysis of studies from different technologies (Zeggini et al., 2008).

However, in the last five years reference sets have developed substantially. The HapMap phase 3 expanded the dataset in the sample direction to include data from many populations, with five times the total number of samples as HapMap phase 2 (Altshuler et al., 2010). The 1000 Genomes pilot reference set expanded in the marker direction with 16M SNPs, indels and structural variants (Project, 2010), and the 1000 Genomes phase 1 reference set includes an unprecedented 40M SNPs on 1092 samples (Project, 2012), including data from genotyping chips, exome and whole genome sequencing. Many of the newly discovered variants are low frequency (MAF $< 5\%$). We have a far less detailed understanding of how well these new variants can be imputed, and how the changes in reference set will impact imputation.

Likewise, many of the original GWAS that used imputation were carried out on individuals of European descent. However, many important GWAS in recent years have been performed using sample collections from Africa (The MalariaGEN Consortium, 2009; Thye et al., 2012; Akinsheye et al., 2011). As we will discuss later in this chapter, these African populations tend to have a greater diversity (both within and between populations). They also have a lower correlation (linkage disequilibrium, or LD) between markers, and the patterns of LD tend to differ between populations. As a result, genotype imputation in these populations is more complicated and less well understood.

In this chapter I will investigate how changes in reference and target sets impact imputation. This will show how new reference sets allow us to use genotype imputation to fill gaps that old imputation reference sets left. This includes imputing variants at low frequency, and variants from specific functional classes. It will also include imputation into populations where imputation has traditionally had more difficulty, such as African populations.

I will start by studying the impact of sample set and diversity on imputation of common and low frequency variation in Europeans, using HapMap imputation. I will then report two studies of imputation in Africa, including an investigation of HapMap imputation for GWAS meta-analyses, and the use of 1000 Genomes imputation in a single diverse population. Finally, I will discuss how these new imputation reference sets can be used to give us new biological insight into the relationship between variant function and disease association, by allowing us to impute loss-of-function variants into GWAS cohorts.

3.2 The impact of reference set diversity in Europeans

This section describes a study that I carried out and published (Jostins et al., 2011) in the first year of my PhD. The reference sets and software versions used are therefore largely out of date at the time of writing this thesis. However, the broader leasons learned about reference set diverse and genotype imputation are nonetheless still valid.

The HapMap phase 2 reference panel consists of genotype data from three homogeneous populations, with 120 haploid genomes each of European and African origin, and 180 of East Asian origin, genotyped at over 2 million sites. By contrast, the larger HapMap phase 3 (or HapMap3) reference set (Altshuler et al., 2010) is much larger, containing over 1000 samples genotyped at a restricted set of approximately 1.5 million variants. Unlike the HapMap2, this data is drawn from a set of 11 populations, providing a far more diverse dataset. Additionally, the HapMap3 benefits from a more mature genotyping technology, providing higher genotype quality. Taken together, these two HapMap datasets provide a significant and stable set of test data to investigate the impacts of the reference set on imputation quality.

I investigate the relationship between sample size and ancestry and imputation accuracy by comparing results obtained using HapMap2 and HapMap3 as the reference set. My comparative analysis focuses on three areas: (1) what effect does the higher quality of genotyping from HapMap3 compared to HapMap2 have on imputation? (2) what improvements can the large increase in sample size have on imputation accuracy and predicted quality scores, especially for low-frequency SNPs? and (3) what can we infer about the importance of closely matching ancestry of reference and target samples?

Population	Code	HapMap2	HapMap3
African Americans	ASW	0	63
North Europeans	CEU	60	117
Chinese Americans	CHD	0	85
Gujarati	GIH	0	88
Japanese and Chinese	JPT+CHB	90	170
Luhya	LWK	0	90
Mexicans	MEX	0	52
Maasai	MKK	0	143
Toscani	TSI	0	88
Yoruba	YRI	60	155

Table 3.1: A summary of the HapMap sample sets and their sizes in the HapMap2 and HapMap3 datasets. I used release 21 of the phased HapMap2 data, and release 2 of the phased HapMap3 data.

3.2.1 Performing and Scoring Imputation

For the target set, I used 1 374 individuals from the 1958 British Birth Cohort (Power and Elliott, 2006), genotyped on both the Illumina HumanHap550 BeadChip and Affymetrix GeneChip Human Mapping 500k chips as the target set. I used the Illumina data to perform imputation, and checked the answers using the Affymetrix data (Illumina chips having been previously shown to be more powerful for imputation (Anderson et al., 2008)). For the target reference sets, I used the approximately 2.5M polymorphic SNPs of the HapMap2 CEU samples, and various mixtures of HapMap3 samples, with approximately 1.4M polymorphic SNPs. Details on the HapMap reference sets are shown in Table 3.1, and the large-scale genetic relationships between these population (measured by principal component analysis) are shown in Figure 3.1.

To perform the imputation I used the imputation program Beagle (Browning and Browning, 2007) (version 3.0.2). I split the genome up into 500kb chunks, with 250kb buffer region on each side, and ran Beagle for 10 itera-

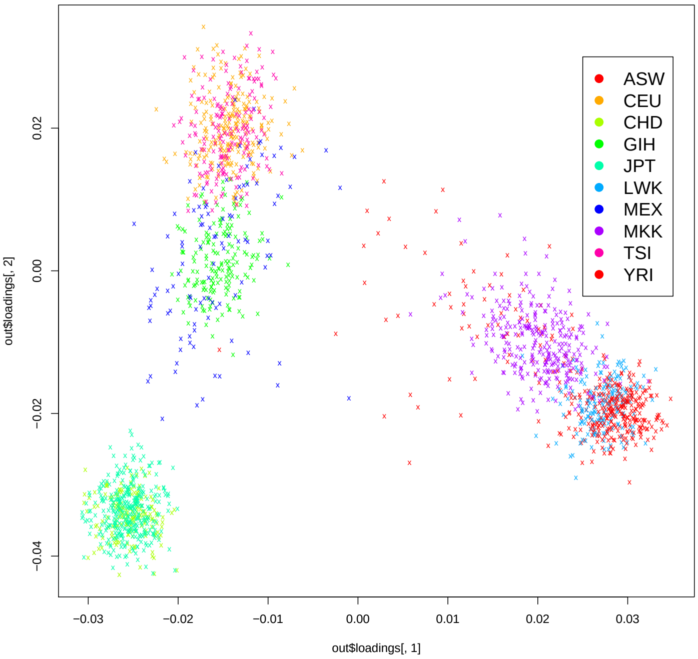


Figure 3.1: The first two principal components for each of the HapMap3 samples, coloured by population. Principal component analysis was performed on all genotypes on chromosome 17, using all founder samples.

tions. To remove poorly imputed SNPs, I applied a filter that removed SNPs with a predicted dosage r^2 of less than 0.9. For several analyses I compare common (MAF $> 5\%$) and low-frequency (MAF $\leq 5\%$) SNPs.

To score the imputation results, I measured both the accuracy of imputation and the usefulness of the predicted quality scores that the imputation method provides. Accuracy was measured using dosage r^2, defined as the square of the Pearson correlation coefficient between the imputed and the actual allele dosage across all imputed samples. The actual dosage is the count of minor alleles for each sample, and the imputed dosage is the expected minor allele count, defined as $2P(aa) + P(Aa)$, where a is the minor allele, and $P(G)$ is the posterior probability of a particular genotype. The dosage r^2 is useful as it is not confounded by minor allele frequency, and thus can be used to compare low-frequency and common SNPs, as well as having a simple relationship to power in a GWAS (Anderson et al., 2008).

For predicted quality scores, most imputation programs (including Beagle) give a predicted dosage r^2 for each SNP, which was evaluated using four criteria: (1) the calibration, or mean difference between predicted and actual dosage r^2 (2) the quality r^2, or the correlation between predicted and actual dosage r^2, (3) the number of overconfident calls, i.e. the number of SNPs that are poorly imputed despite having high predicted dosage r^2, and, vice versa, (4) the number of under-confident calls. I am particularly interested in the number of overconfident SNPs, as these may lead to costly false positives.

3.2.2 Reference Set Quality

While the majority of SNPs in both HapMap2 and HapMap3 are of high quality, the genotyping for a number of previously poorly genotyped SNPs was improved in the development of HapMap3. To investigate whether this increase in reference set quality had a significant effect on imputation, I performed genome-wide imputation on the target set using two 'reduced' HapMap reference sets, and measured differences in dosage r^2. These reduced sets contained only the 56 CEU samples and 1M SNPs that HapMap2 and HapMap3 have in common. I found a small but significant difference due to genotyping quality (mean dosage r^2 0.841 vs 0.845, Figure 3.2), but not enough to explain a meaningful difference in imputation quality between HapMap2 and HapMap3.

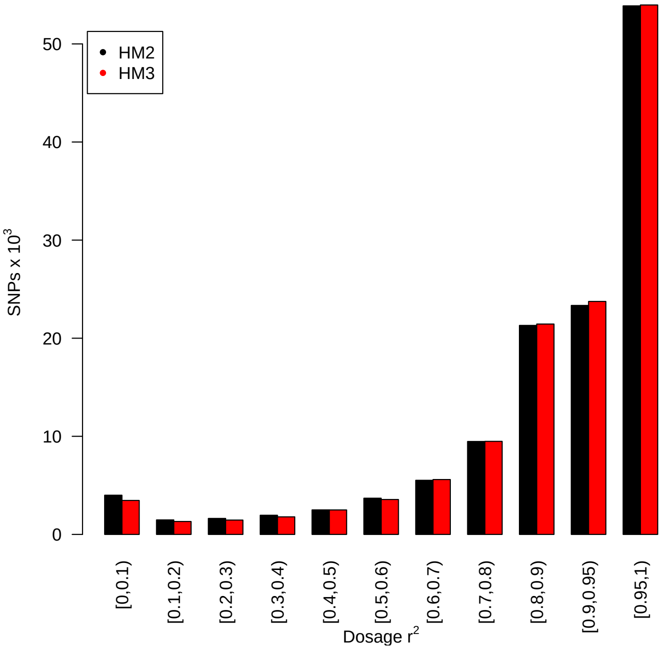


Figure 3.2: A histogram of dosage r^2 for a genome-wide imputation using the reduced HapMap2 and HapMap3 sets, which contain only the 1,069,264 SNPs and 56 CEU samples that both HapMap2 and HapMap3 have genotype information for. The means of the distributions are 0.841 and 0.845, and the difference is significant (t = 7.59, df = 256480, p $<10^{-13}$).

3.2.3 Reference Set Size

To assess the effect of larger HapMap sample sizes, I performed genome-wide imputation on the target set, using five reference sets of increasing size and diversity. I used the HapMap2 and HapMap3 CEU samples (HM2CEU and HM3CEU), which should be the best match to the UK target set, as well as a mixed reference set of HapMap3 European samples (CEU+TSI). To give a large, but still partially matched reference set, I used the HapMap3 European samples mixed with the Indian and Mexican samples (CEU+TSI+GIH+MEX), as these populations cluster together on the first two principal components (see Figure 3.1). Finally, I examined all

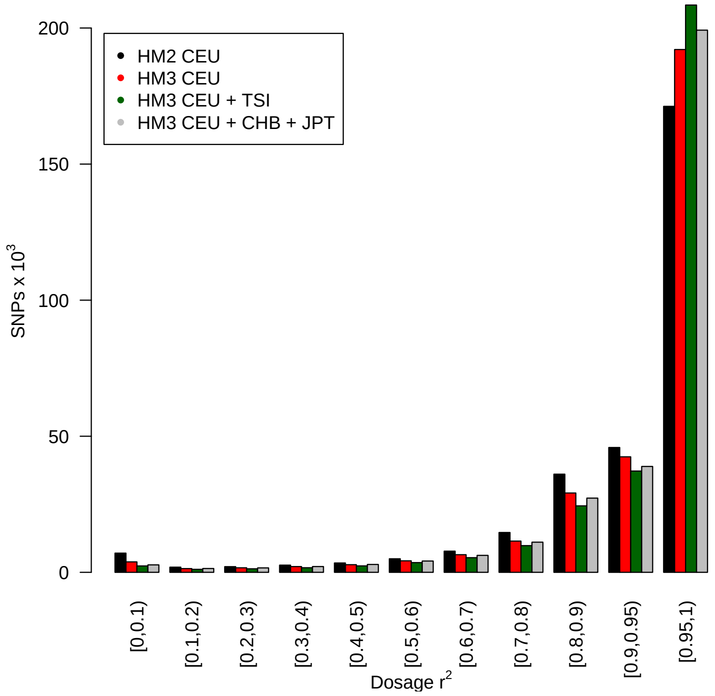


Figure 3.3: The effects of reference set on imputation accuracy. A histogram of dosage r^2 scores genome-wide for samples imputed with HapMap2 and HapMap3 CEU, as well as HapMap3 CEU+TSI, and a reference set consisting of HapMap3 CEU+JPT+CHB of the same size as the CEU+TSI set.

HapMap3 individuals (WORLD), in order to assess a very large and very diverse reference set. Sample sizes are shown in Table 3.2.

I found that HapMap3 yields a substantial increase in imputation accuracy compared to HapMap2, with the number of SNPs in the highest score category ($> 95\%$) increasing, and the number in all lower-scoring categories decreasing (Figure 3.3). A further increase in imputation accuracy is seen when adding the HapMap3 TSI samples. The number of SNPs that pass the filter (have a predicted r^2 greater than 0.9) rises as imputation accuracy increases, although this falls as samples from many populations are added due to a decrease in the imputation software's predicted confidence (see below).

Reference Set	Size	CPU	Passed Filter		Filtered Dosage r^2	
			Common	Low-frequency	Common	Low-frequency
HM2CEU	60	514h[a]	83.7%[b]	52.5%[b]	0.957	0.889
CEU	117	296h	85.1%	59.7%	0.968	0.921
CEU+TSI	205	350h	86.1%	63.1%	0.974	0.934
CEU+TSI +GIH+MEX	345	458h	85.3%	60.3%	0.978	0.957
WORLD	1010	1207h	83.8%	55.5%	0.979	0.968

Table 3.2: Information on Genome-Wide imputation using various reference sets. The CPU columns shows the number of CPU hours used in the imputation, which increases with the size and SNP density of the reference set. The proportion of SNPs that passed the filter (predicted dosage $r^2 \geq 0.9$), and the mean dosage r^2 of those that passed, are shown for common (MAF > 0.05) and low-frequency (MAF ≤ 0.05) SNPs. [a] HM2 has a large SNP set, hence the longer imputation time [b] HM2 has a larger number of SNPs in total

The dosage r^2 of filtered SNPs shows a trend of improved imputation with increasing sample sizes. This increase is statistically significant ($p < 10^{-16}$) for all increases in sample size, with the exception of the WORLD set (Table 3.2). A corresponding increase is seen in computational time, especially for the WORLD set; however, the CEU+TSI+GIH+MEX reference set only takes 55% longer to process than just CEU, despite being nearly 3 times larger.

The improvement for low-frequency SNPs is the most striking. The HM2CEU mean dosage r^2 score is low, especially compared to common SNPs (0.89 vs 0.96). If all samples from all HapMap3 populations are included, this gap nearly disappears (0.96 vs 0.98). In general, fewer low-frequency SNPs pass the imputation quality filter (63% at most), but the accuracy of these imputed low-frequency SNPs can become very high. The improvement in dosage r^2 is inversely proportional to the frequency of the SNP, with the

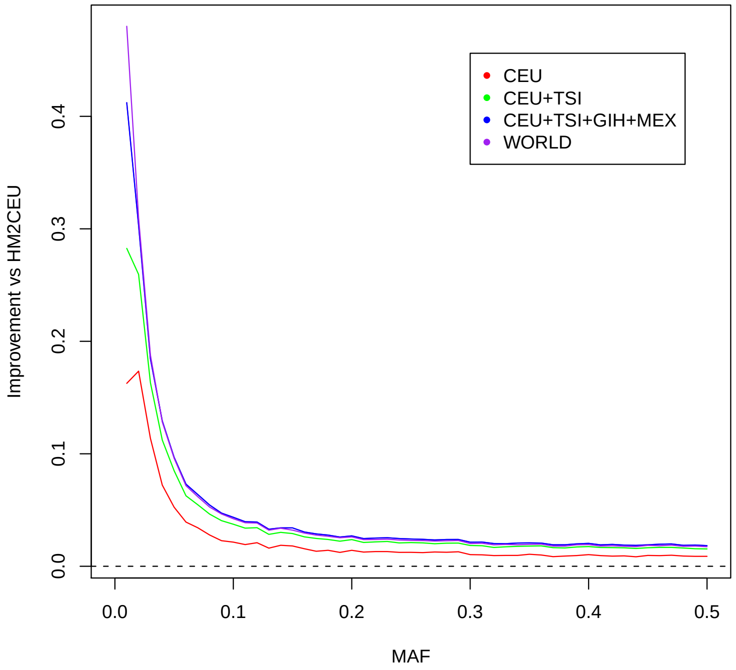


Figure 3.4: The genome-wide increase in dosage r^2 for imputed SNPs relative to HapMap2 CEU, plotted against minor allele frequency, for the four HapMap3 sample mixtures.

greatest improvement observed for the very rarest SNPs (Figure 3.4).

For small reference sets, the calibration of predicted quality scores tends towards overconfidence. As the reference set increases in size, the calibration improves, though very diverse reference sets lead the confidence scores towards under-confidence (Table 3.3). The correlation between predicted and actual dosage r^2 improves, though with a slight decrease for the most diverse sets. These trends are stronger in low-frequency variants than in common ones; low-frequency variants tend to have less well calibrated and correlated predicted quality scores. Larger reference sets decrease the number of overconfident mistakes and the number of under confident mistakes

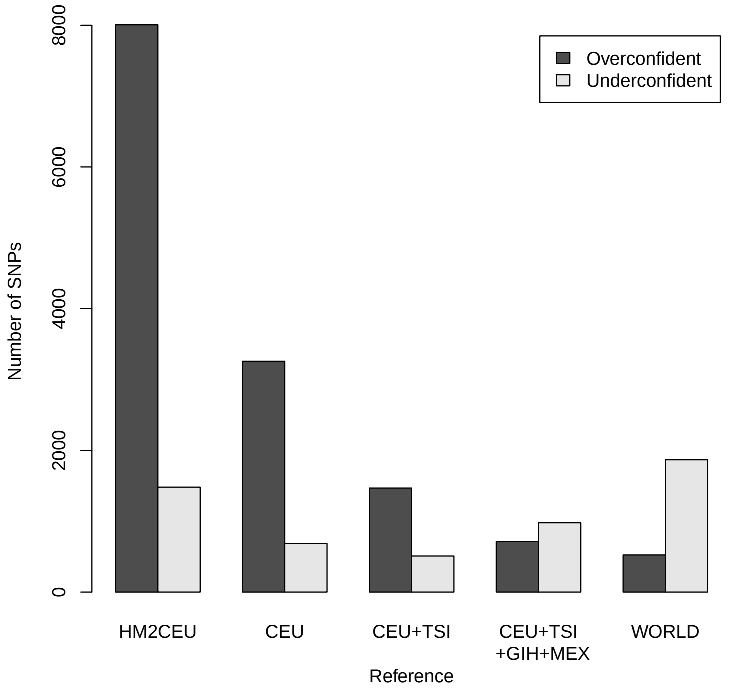


Figure 3.5: The rates of overconfident and under-confident mistakes in imputation, using various reference sets. An overconfident mistake is any SNP that is imputed with a predicted dosage $r^2 > 0.9$, but an actual dosage $r^2 \leq 0.8$, and an under-confident mistake has a predicted dosage $r^2 \leq 0.8$ and an actual dosage r^2 > 0.9.

(with the exception of the WORLD set, which causes a slight inflation in under-confident calls, Figure 3.5).

3.2.4 Reference Set Diversity

I investigated the importance of population matching, independent of sample size, in two ways. Firstly, I compared genome-wide imputation using the HapMap3 CEU+TSI reference set to a CEU+JPT+CHB reference set of the same size and non-CEU proportion. This allows us to investigate the effect of adding poorly matched samples on imputation. Second, I created a num-

Reference Set	Calibration		Quality r^2	
	Common	Low-frequency	Common	Low-frequency
HM2CEU	0.019	0.038	0.78	0.73
CEU	0.008	0.027	0.88	0.76
CEU+TSI	0.002	0.009	0.92	0.79
CEU+TSI +GIH+MEX	-0.006	-0.019	0.93	0.79
WORLD	-0.010	-0.043	0.91	0.76

Table 3.3: Calibration data for Genome-Wide imputation using the five reference sets. Quality calibration is defined as the mean difference between the actual and predicted dosage r^2; a negative value represents conservative quality scores, and a positive value represents liberal quality scores. The quality r^2 is the correlation between the predicted and actual r^2. The SNPs are split into common (MAF $>$ 0.05) and low-frequency (MAF $\leq$ 0.05).

ber of equally sized reference sets for chromosome 17 by combining a range of mixture proportions of either CEU and TSI , or CEU and CHB+JPT. I measured the accuracy of imputation using these reference sets for low-frequency variants. I denote these constant-sized mixed reference sets as CEU/TSI and CEU/CHB+JPT, in order to distinguish between reference sets in which sample size is not held constant (e.g. CEU+TSI).

I found that, while the mismatched CEU+JPT+CHB reference set gives a lower imputation accuracy than CEU+TSI, it still yielded a substantial improvement over the CEU reference set alone. Half of the improvement in imputation accuracy from CEU to CEU+TSI was also gained with the CEU+JPT+CHB reference. This implies that while matching the reference set to the target set is important, even the addition of unrelated samples yields increases in imputation accuracy.

Increased diversity initially correlates with increased imputation accuracy for both CEU/TSI and CEU/CHB+JPT (Figure 3.6), though the former is

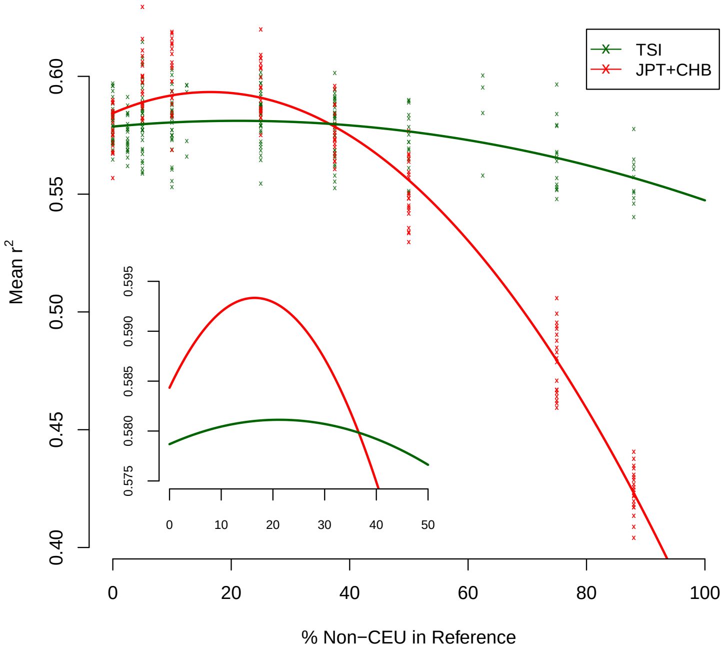


Figure 3.6: The relationship between the dosage r^2 and the proportion of non-CEU samples in a 100-sample reference set. The trend lines are quadratic least squared regression curves, and both explain the data significantly better than a linear relationship (N = 207, p $< 10^{-4}$ and N = 159, p $< 10^{-16}$ for TSI and CHB+JPT respectively). The insert shows an expansion of the trend lines between 0 and 50%.

far less marked than the latter. Beyond a certain proportion of non-CEU samples accuracy starts to fall off as the effect of diversity is outweighed by the effect of mismatching. The optimum population mix is 22% for CEU/TSI, and 17% for CEU/CHB+JPT. It is only above 43% TSI do we see a decrease in imputation accuracy for adding TSI over pure CEU; for CHB+JPT this figure is 33%. This relationship is specific to low-frequency variants.

3.2.5 Discussion

Higher quality reference data and larger sample sizes yield improved imputation accuracy. Using HapMap3 as a reference set compared to using HapMap2 demonstrates this improvement, especially at sites with a low minor allele frequency. While this result was expected I did not anticipate the substantial improvement achieved with large and genetically diverse reference sets. Including samples from such diverse populations as MEX and GIH can provide significant improvement in imputation into UK samples of alleles with a minor allele frequency of less than 5%. Larger reference sets also improve predicted quality scores, with a decrease in overconfident mistakes without inflating under-confident calls.

Overall, an imputation reference set consisting of CEU, TSI, MEX and GIH improves the quality of imputation in all frequency ranges, and greater improvement for very low-frequency SNPs was achieved with very large and highly mixed reference sets. The latter came at the cost of computational power, as well as overly conservative predicted quality scores. The quality scores are likely to be lowered due to the poor match of haplotype frequencies between the reference and target samples, which will in effect decrease the prior on correctly matched haplotypes. Imputation is robust to the precise mix of samples of closely related ancestry (such as CEU/TSI), and small amounts of divergent ancestry can actually improve accuracy (such as CEU/CHB+JPT). However, crude population matching is important, as demonstrated by the reduced accuracy of the CEU+JPT reference compared to CEU+TSI.

My results are consistent with those of Huang et al. (2009), who found that the imputation of Yoruba samples had higher accuracy with a YRI+CHB+JPT HapMap2 reference than with a pure YRI. However, Huang *et al* did not control for reference size, and showed a much smaller improvement compared to my results, probably due to the highly divergent nature of the HapMap2 populations.

These results imply a set of relatively simple rules for picking imputation reference sets: for the best trade-off between accuracy and computation time, the most diverse mixture of populations that still approximately cluster with the target samples of interest on a world-wide PCA plot should be used. However, if imputing genotypes for low-frequency variants with high accuracy is required, all samples available should be used, with the understanding that this will increase computational time, and cause quality scores to be somewhat conservative.

More recent developments in genotype imputation

Since I wrote the above section additional papers have been published by other researchers that shed further light out the relationship between reference set diversity and genotype imputation. Marchini and Howie (2010) performed imputation using HapMap2 data and demonstrated that combining reference haplotypes across continents gives greater imputation accuracy for low-frequency variation regardless of whether IMPUTE2, Beagle or fastPHASE was used, though IMPUTE2 being the most computationally efficient. Similar experiments using 1000 Genomes data carried out by Sung et al. (2012) showed a similar improvement in imputation low-frequency variation with larger and more diverse reference sets, this time while using the MaCH imputation program.

Over the last few years a concensus has emerged that imputation using world-wide datasets (including data from all available populations) is the simplest way of performing high-quality imputation. For instance, Howie et al. (2011) demonstrated that such world-wide datasets give optimal or near optimal imputation results using both cross-validation experiments and imputation into real African GWAS data. The rise of pre-phasing techniques (Howie et al., 2012), which allow fast phasing that is independent of reference set size, has made the use of very large reference sets more computational tractable. The appeal of using world-wide reference sets is that they do not require careful selection of reference haplotypes to match the target panel, and thus can be used out-of-the-box on any set of samples.

3.3 Imputation in African populations

The previous section, and indeed most work on imputation to date, focused on imputing variants into European and East Asian datasets. However, many important GWAS datasets have been generated in African populations, notably studies of malaria (The MalariaGEN Consortium, 2009), tuberculosis (Thye et al., 2012) and sickle cell disease (Akinsheye et al., 2011). Just like European studies, these African studies require imputation, particularly where meta-analyses are performed.

Imputation in Africa provides us with its own unique set of difficulties. African populations show a higher degree of genetic diversity than European populations (both within and between populations (Altshuler et al., 2010)). They show less linkage disequilibrium (Altshuler et al., 2010), and substantial differences in patterns of LD between populations (Teo et al., 2009). Given this, it is unsurprising to note that imputation generally performs less well in African populations (Huang et al., 2009; Altshuler et al., 2010; Howie et al., 2011). However, while imputation is more difficult, the rewards are potentially greater. Good quality imputation can greatly improve power when the causal variant is not well tagged (The MalariaGEN Consortium, 2009), and can also allow well-powered meta-analyses in cases where LD differs between populations (Teo et al., 2010).

In this section I will discuss two studies of imputation in African populations. The first investigates HapMap3-based imputation in a GWAS meta-analyses to discover common associations, and the second looks at using a 1000 Genomes Project high-density reference set to impute into a single, diverse African population.

3.3.1 HapMap-based imputation in a GWAS meta-analyses

Description of the study and data

A large collection of blood samples from individuals diagnosed with severe malaria (including cerebral malaria and severe malarial anaemia), along with matched population controls, have been collected by MalariaGEN consortium partners in 9 African countries. 5425 cases and 6891 controls from three of these collections (Gambia, Malawi and Kenya) were genotyped on three different technologies (Illumina 650K, Illumina 1M and Illumina 2.5M respectively). The aim of the experiment was to identify and investigate genetic loci that correlate with severe malaria, and to investigate changes to standard methodology (including QC, imputation and association techniques) that are required to study these African collections.

Due to the difficulty of taking blood from severely ill children, only a small amount of DNA could be extracted and whole-genome amplification was performed, increasing noise in the genotype data. To produce a robust set of genotype calls, three different calling algorithms were used to process intensity data from the Illumina arrays, separately in each of the three cohorts. A set of consensus calls were obtained by treating as missing any genotype that was discordant among algorithms. SNPs with a missing data rate of $> 2.5\%$ were removed. Sample with outlying missingness of heterozygosity were also removed prior to imputation.

Performing and QCing imputation

Imputation was performed using Impute 2.12, using the phased release 2 of HapMap3 from the Impute website (http://mathgen.stats.ox.ac.uk/impute/). As we saw in section 3.2, a diverse reference set provides maximal imputation accuracy, so I used all HapMap3 haplotypes from all populations (African and non-African) to perform imputation.

The genome was split up into chunks which are either 5Mb, or have 20 000 reference SNPs (whichever is smaller), with an additional 500kb buffer on either side of the segment. I used imputation parameter settings of k = 80 and Ne = 14000. Imputation was performed in parallel for each segment, and segments were reconstructed into chromosomes once all imputations had finished.

To ensure that imputation was performing correctly, I developed a manual imputation QC strategy for examining the output. For each sample cohort I manually examined the following quality-control diagnostic plots to ensure that imputation had performed properly:

(a) a histogram of certainty quality scores across SNPs

(b) a histogram of info quality scores across SNPs

(c) a histogram of per-individual type2 r2 scores, averaged across segments

(d) a histogram of per-segment heterozygous imputation accuracy (proportion of genotyped heterozygous calls that are also confidently imputed as heterozygous)

(e) a plot of per-segment mean type2 r2 scores against the segment's position along the genome

Examples of these plots (taken from the imputation of the Kenya dataset) are shown in Figure 3.7. This imputation run has completed without problems, as the quality scores peak near to 1 (Figures 3.7a and 3.7b), no chunks have abnormally low quality (Figure 3.7d), and the imputation performance shows no significant variation genome-wide (Figure 3.7e). One anomaly is

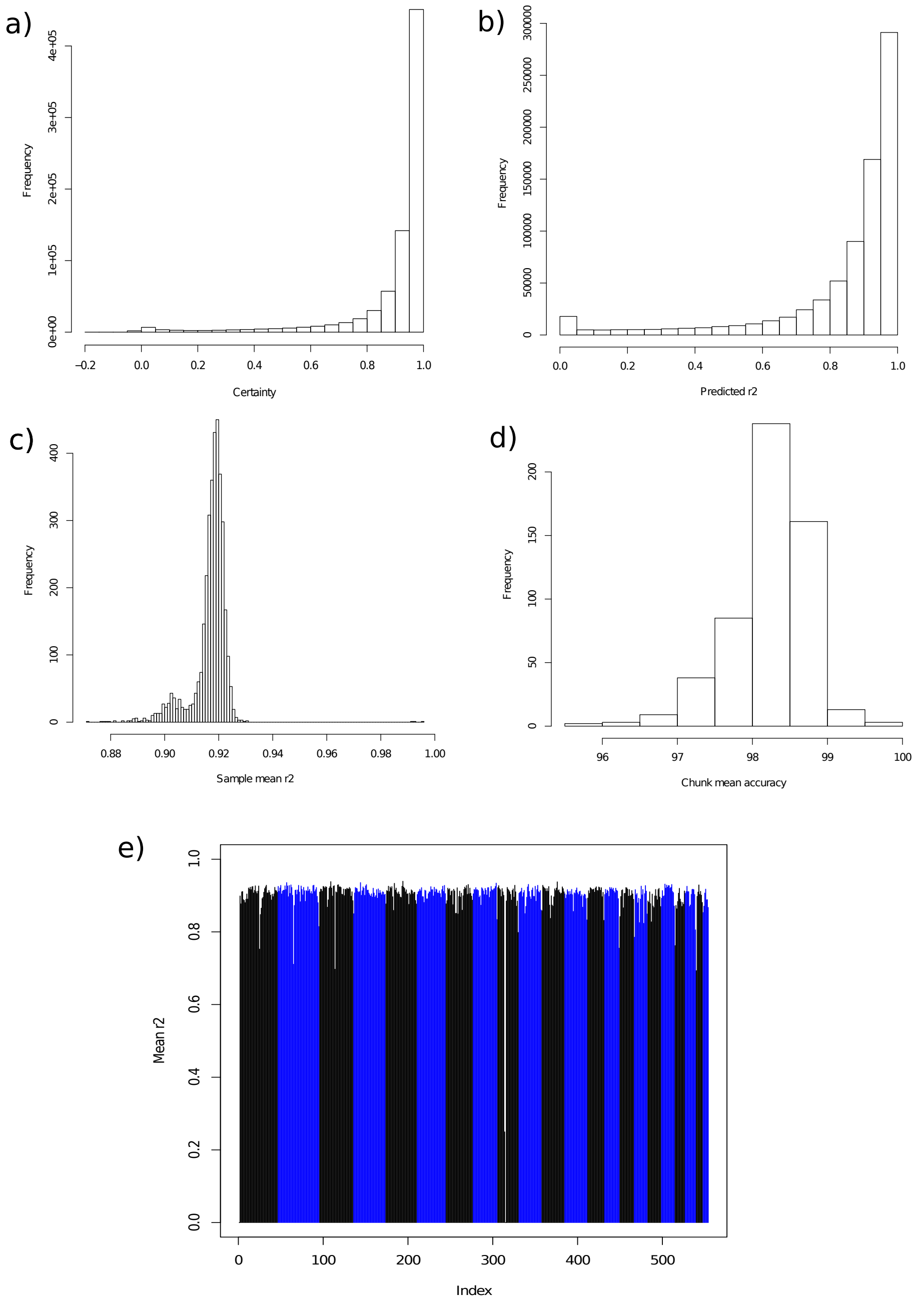


Figure 3.7: Example output from the imputation quality control pipeline for the Kenya imputation. Panels a) and b) show the distribution of two quality scores (certainty and predicted r2) across SNPs, figures c) and d) show the distribution of quality scores across samples and across chunks, and figure e shows the distribution of quality genome-wide (blocks of colour represent chromosomes).

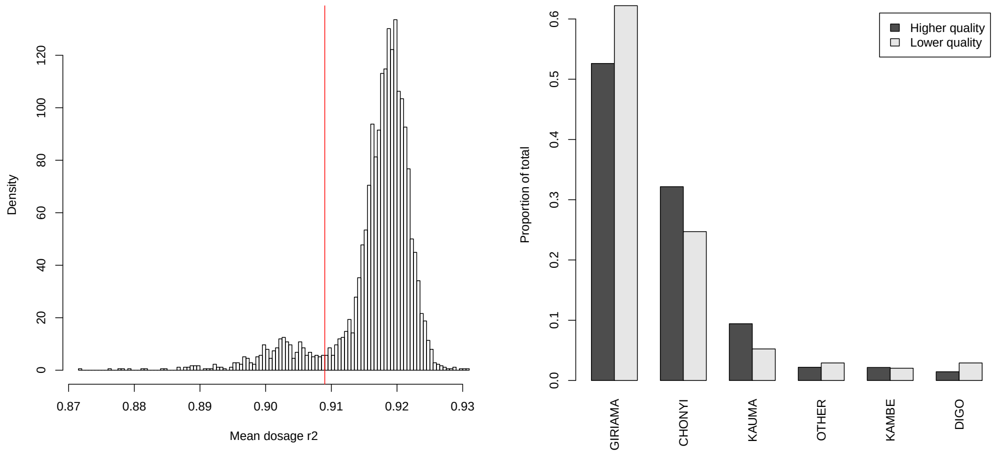


Figure 3.8: a) The distribution of imputation quality (measured by type2 r2) across imputed Kenyan samples. The red line is at r2 = 0.909, and is the minimum between the two peaks. b) The distribution of ethnic groups in the samples in the two peaks. The difference in the two distributions is highly significant (Fisher's exact test, p = 4 x 10^{-4}), suggesting that ethnic differences contribute to the bimodal distribution of imputation quality.

the unusual "bump" in the per-sample imputation plot (Figure 3.7c). Further investigation reveals that this "bump" arises at least in part from ethnic differences within Kenya (Figure 3.8).

Accuracy of imputation across populations

I assessed the accuracy of imputation using the dosage r^2 between imputed and true allele count at directly typed SNPs (This is generated internally by IMPUTE2, and called the type 2 r2). Figure 3.9 shows per-individual dosage r^2 broken down by country. While less accurate than typically achieved in European populations, imputation still captures the majority of common variation in these three populations (a mean dosage r^2 of 0.93 in Malawi, 0.92 in Kenya and 0.87 in Gambia). As in Europeans, common SNPs were better imputed than low-frequency SNPs.

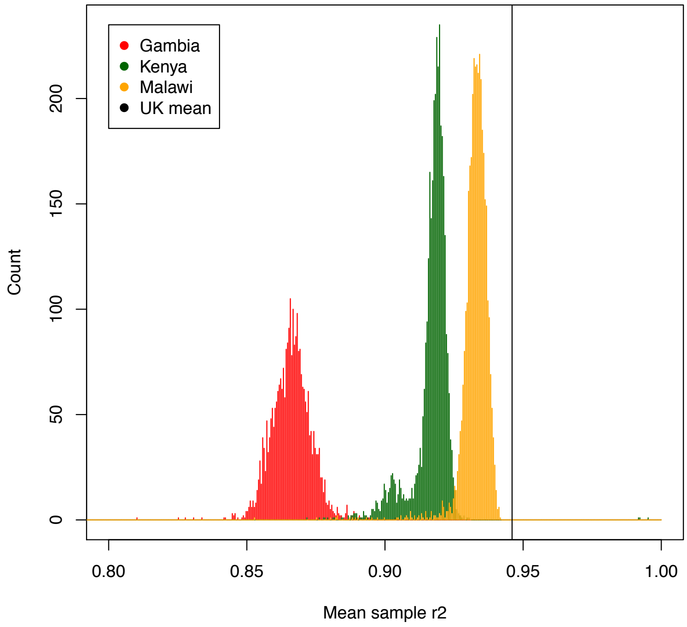


Figure 3.9: Per-sample imputation accuracy measured by dosage r^2, averaged over imputation chunks. Black vertical line shows typical imputation accuracy in a UK population, taken from Section 3.2. Gambian samples (red) perform worst due to the poor coverage of African variation by the Illumina 550K platform, followed by Kenyan samples (green) on the Illumina Omni2.5M, which while dense has limited overlap with our HapMap3 reference, with Malawian samples (yellow) performing best.

As I discussed above, as well as imputation accuracy we are also interested in the numbers of overconfidently and under-confidently imputed SNPs. I evaluated the calibration of the confidence of IMPUTE2 (measured by the info score) against its actual performance at genotyped SNPs. The calibration of confidence was high across our three samples (quality r^2s of 0.93 in Malawi, 0.92 in Kenya, 0.96 in Gambia) but, like overall accuracy, on average worse than in European samples (0.96). I included only SNPs with info score > 0.75 for downstream analyses, leaving a high quality set with mean r2 $>$ 0.9 in all samples, and less than 1% of either very overconfident (predicted r2 > 0.75, actual < 0.6) or very under-confident (predicted < 0.75, actual > 0.9) SNPs. Taken together, these results suggest the underlying model of IMPUTE2, combined with our diverse reference panel, is generally applicable to samples from African populations.

Despite the high performance of imputation overall, I discovered a number of factors that influenced relative imputation performance, including (i) genotyping platform, (ii) ethnic matching of target GWAS samples to the imputation reference panel, and (iii) homogeneity of individual GWAS collections. The Gambian samples (typed on the Illumina 650Y array) show much poorer imputation quality (Figure 3.9) than our Kenyan and Malawian samples (typed on Illumina chips with > 1 million SNPs). While genotyping array represents the single most important factor to imputation accuracy, two aspects of population genetics are also critical: good matching between reference and target samples and homogeneity within a GWAS sample (illustrated by the small number of samples of differential ancestry in Kenya with poorer imputation quality seen in Figure 3.8).

3.3.2 1000 Genomes-based imputation in a single, diverse population

Description of the data

The MalariaGEN Kenya dataset, included in the previously discussed meta-analysis, was genotyped on Illumina's Omni2.5 genotyping chip. This high-density SNP array is the first of a new generation of genotyping chips designed to assay a subset of the large numbers of SNPs discovered by resequencing studies, such as the 1000 Genomes Project. The Kenya malaria dataset is the first of many MalariaGEN datasets that will be genotyped on this chip,

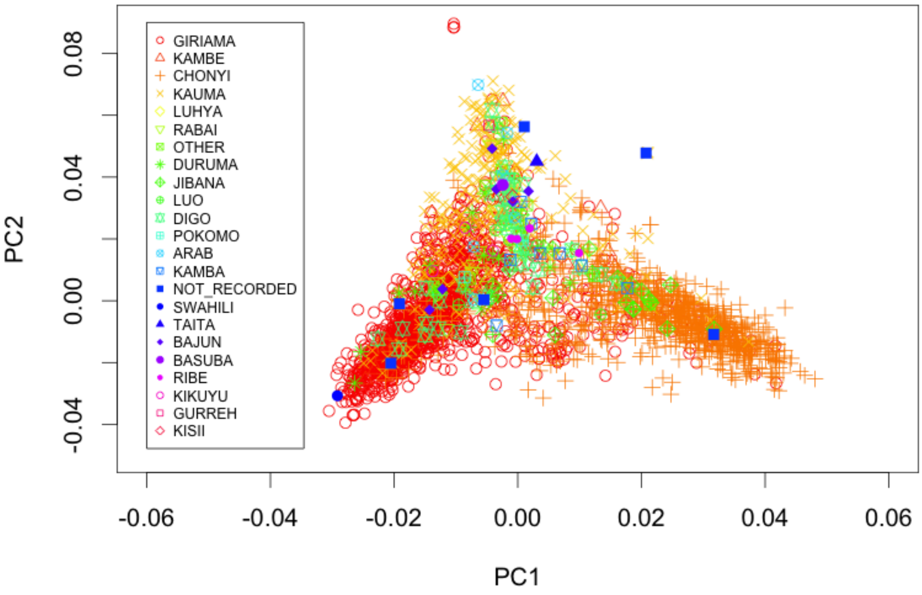


Figure 3.10: A PCA of the 2502 Kenyan samples, coloured by ethnicity.

as it is believed the higher density will allow us to overcome the LD issues that can confound cross-population meta-analysis.

However, this dataset also provides us with an opportunity to make a detailed assessment of the accuracy of high-density imputation into a diverse African population. Two factors make this a particularly good dataset for such assessment. Firstly, the 2502 Kenyan samples are ethnically diverse, as shown by their large number of stated ethnicities, and their significant structure on a principal component plot (both shown in Figure 3.10). We can use this to investigate the impact of target set diversity and structure on imputation accuracy. Secondly, the Omni2.5 is a particularly good system to assess GWAS imputation, as it is built on the backbone of an OmniExpress (a typical, middle cost GWAS chip), with a large number of 1000 Genomes

Reference Set	N. haplotypes	CPU use	Memory use
Pilot Yoruba	120	143hrs	20.5 Gb
Pilot (all samples)	360	163hrs	20.9 Gb
Phase I Yoruba+Luhya	400	165hrs	21.1 Gb
Phase I (all samples)	2420	220hrs	25.4 Gb

Table 3.4: Reference sets used for testing 1000 Genomes imputation, with resources required for imputation.

SNPs added. The OmniExpress backbone, as a model of a GWAS chip, can be imputed into from a high-density dataset, and the additional content can then be used as a validation set.

Performing imputation

Because the Omni2.5 can only be used to assess imputation results for SNPs on that chip, I decided to reduce imputation complexity by only using the Omni2.5 data generated as part of the 1000 Genomes Phase 1. I made a set of four test reference sets from this data, consisting of two 1000 Genomes pilot and two phase 1 datasets, with one containing only African samples, and one containing all samples (Table 3.4).

Imputation was performed only on Chromosome 1, using the Impute2 pipeline described in section 3.3.1. This took between 140 and 220 CPU hours and 20 to 26 CPU Gbs, and was only weakly dependent on reference set size (Table 3.4).

Imputation accuracy was measured using dosage r^2 between imputed and true genotyped at non-OmniExpress SNPs. For per-individual accuracy, I used heterozygous certainty (the mean heterozygous posterior probability at truly heterozygous sites).

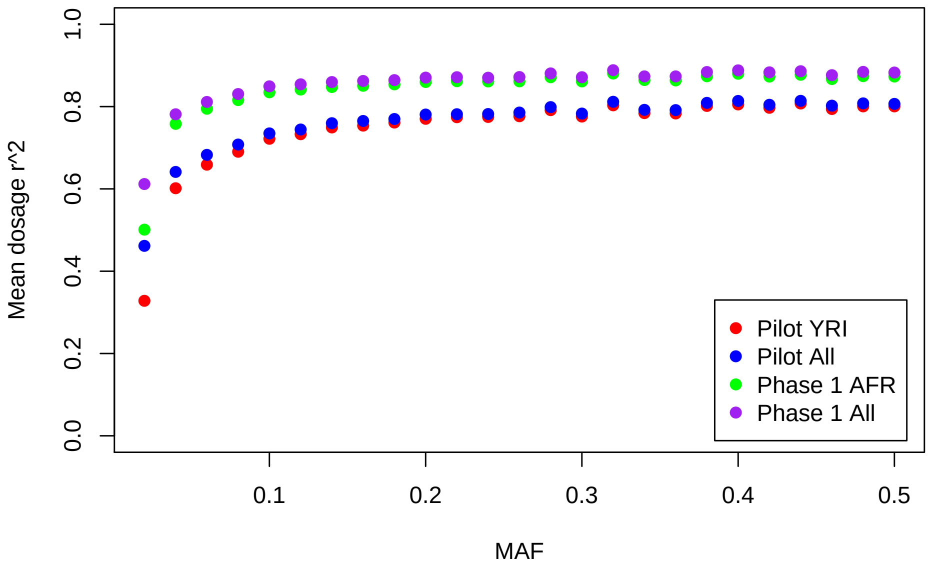


Figure 3.11: The relationship between imputation accuracy and call rate using the various reference sets. YRI=Yoruba, AFR=African. Note that these data has not been filtered by quality score.

Impact of reference set on imputation

Looking first at the pilot data, imputation of 1000 Genomes variants into Kenya performed very badly (Figure 3.11). Even common variants had a mean dosage r^2 of around 0.7. However, going to the Phase 1 data dramatically improved imputation performance, bringing the dosage r^2 up to over 0.8. Interestingly, the non-African haplotypes made almost no improvement to imputation for common SNPs in either the pilot or the phase 1 data. However, for the very low-frequency SNPs (MAF $< 2\%$), introduction of non-African haplotypes dramatically improved imputation, both for the pilot data (0.33 to 0.45) and for the Phase 1 data(0.51 to 0.61). This again reinforces the value of distantly related haplotypes to improve imputation

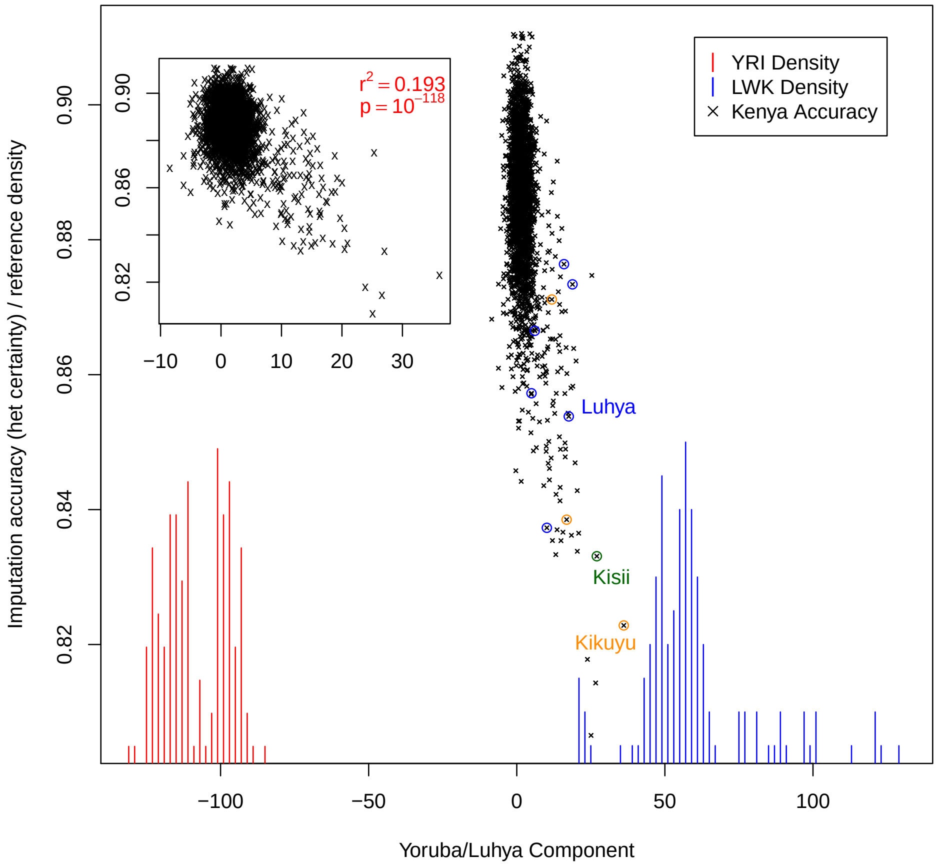


Figure 3.12: Individual variation in imputation accuracy with YRI/LWK principal component. Coloured bars represent the location of reference individuals. A few outlier ethnicities are circled. Inset expands the Kenyan region of the component.

for low-frequency variation.

Impact of target sample on imputation

To investigate the impact of population structure on imputation accuracy, I found the first principal component for the Luhya and Yoruba Phase 1 reference sets, and projected all Kenyan samples onto this axis (using the R

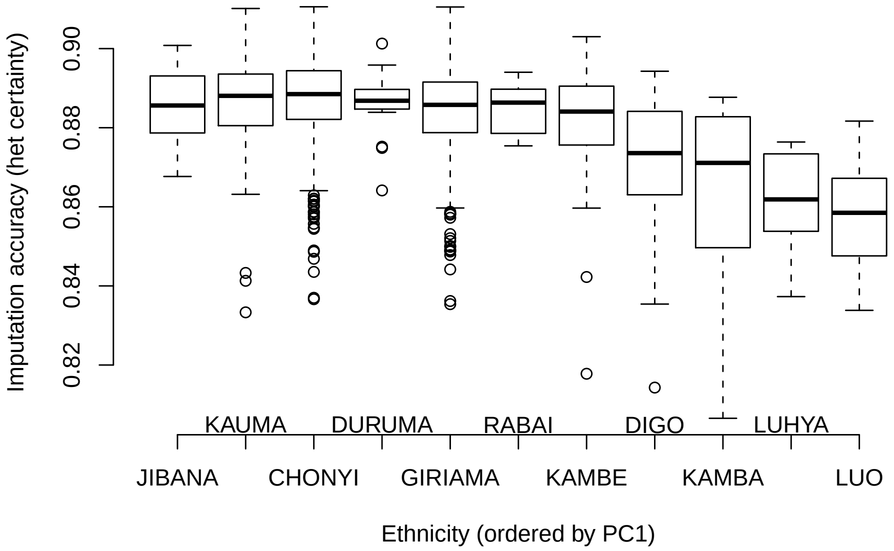


Figure 3.13: The variation in imputation accuracy between the major ethnic groups, ordered by distance from YRI

package snpMatrix). I then correlated this value with the imputation accuracy for the Kenyan samples imputed with the AFR Phase 1 dataset (Figure 3.12). Surprisingly, I found a significant inverse correlation, with samples that lay closer to the Luhya cluster having lower imputation accuracy.

The same relationship appeared to hold if median accuracy across ethnicity was considered, with ethnicities that were genetically more similar to the Luhya having lower median quality (Figure 3.13). However, it also appears that samples that are closest to the Yoruba also show a slight decrease in imputation quality. This suggests that the decrease in quality is in fact due to being ethnic outliers from the main Kenyan cluster, rather than due to similarity to reference populations. This may due to the effect of phasing: IMPUTE2 uses the entire target set to perform phasing, which will lead to samples that are not closely related to the rest of the target set having worse phasing, and thus lower imputation accuracy.

Conclusions

I believe that the results above allow us to draw four conclusions about high-density imputation in diverse populations:

1. The Phase 1 1000 Genomes reference set grants significant improvements in imputation for African populations
2. Low-frequency imputation benefits from extreme diversity, illustrating the need for world-wide genotype reference sets
3. Imputation accuracy in Kenya varies significantly by ethnic group
4. The relationship between accuracy and target/reference match can be complex and counter-intuitive

3.4 Using imputation to explore the impact of loss-of-function variants on complex disease

3.4.1 Loss-of-function variants and the 1000 Genomes project

Loss of function (LoF) variants are SNPs, indels or CNVs where one allele entirely removes the function of one or more genes. These can include SNPs that disrupt a start codon, create a new stop codon or disrupt an essential splice site, indels that create a frame-shift and CNVs that partially or entirely delete a gene. Clearly these mutations are major candidates for having phenotypic effects, and many of the known Mendelian diseases are caused by LoF mutations, but it is also clear that many LoF variants are relatively benign and circulate at high frequency in the population. As part of the 1000 Genomes project, the LoF Group (now the Functional Integration Group) was founded to identify and investigate both common and rare LoF variants.

After extensive filtering, we discovered 1285 high quality LoF mutations in the 1000 Genomes pilot (MacArthur et al., 2012). This was a particularly challenging project, largely due to the high proportion of false positives in this dataset: 1666 putative loss-of-function variants were excluded due to possible mapping artefacts, errors in gene model and systematic sequencing errors. In total, we concluded that the average human genome contains around 100 loss-of-function mutations, with approximately 20 genes homozygously inactivated.

As well as identifying these mutations, an important aim of the project was to shed light on the biology of these mutations. This included identifying differences in the property of genes that harbour common LoF mutations and those where LoF mutations cause Mendelian disease, as well as using RNA-

Seq to study the impact of LoF mutations on gene expression. In this section, I will describe a study that I carried out, using genotype imputation to assess the impact of loss-of-function variants on human complex disease.

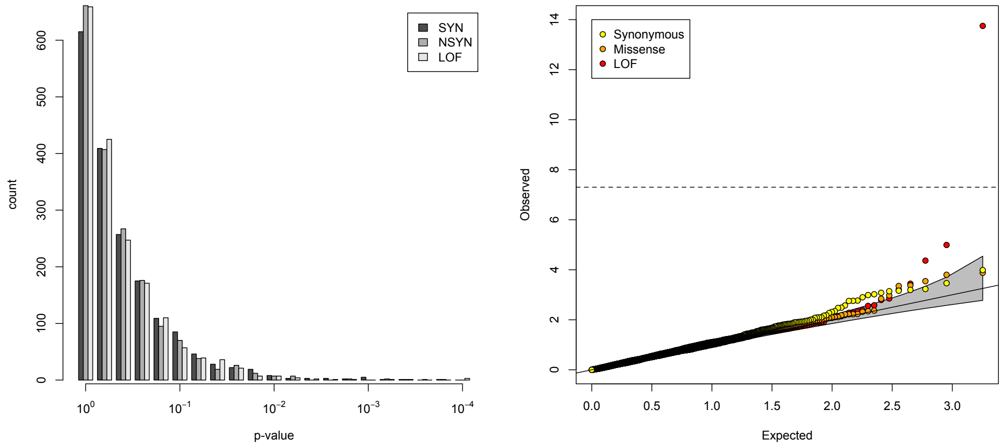
3.4.2 Performing imputation and association analysis

To assess whether LoF variants were enriched for effects on complex disease risk, I imputed all SNPs and indels genotyped in the CEU population in the 1000 Genomes low-coverage pilot (Project, 2010) into the complete Wellcome Trust Case Control Consortium 1 (WTCCC1) dataset (Wellcome Trust Case Control Consortium, 2007), comprising 2,938 controls and 13,241 cases that pass sample QC.

Genotypes for CEU SNPs and indels were obtained from the July 2010 release, and were merged with SNP genotypes from HapMap3 release 2. Imputation of these variants into the WTCCC1 dataset was performed using the IMPUTE2 pipeline described in section 3.3.1.

I investigated potential associations with complex disease risk for 625 high-confidence LoF variants identified as polymorphic in the CEU population. Of these variants, 417 imputed well enough in both controls and at least one cohort to go ahead with association (using an info score threshold of 0.2), resulting in a total of 2901 association tests in the seven disease cohorts. Only 3 variants were close enough to the threshold to be assessed in some cohorts but not others.

I performed a frequentist association analysis using the program SNPTest (Marchini et al., 2007), version 2.2.0. I used an additive model of risk, and a likelihood score test to account for uncertainty in imputed genotypes. Matched synonymous and missense sets were calculated using allele frequencies in controls, taking random draws without replacement of synonymous



(a) Histogram

(b) QQ plot

Figure 3.14: Association of coding variants with complex disease risk. Observed -log10(P) values for disease association in 16,179 individuals from seven complex disease cohorts and a shared control group, following imputation of variants identified by the 1000 Genomes low-coverage pilot, are plotted against the expected null distribution for all LoF variants and frequency-matched missense and synonymous SNPs.

and missense variants from the same 1% frequency bin as each LoF variant. In both cases, five random draws were made.

3.4.3 Results

There were no significant detectable enrichments of associations for LoF variants compared to missense variants at P value thresholds of 10^{-5}, 10^{-4} or 10^{-3} (Fisher's exact P values 0.4994, 0.1245 and 0.8034, respectively), suggesting that common LoF variants are not substantially over-represented among complex disease risk variants compared to other functional coding polymorphisms.

The major caveat of this analysis is that the systematically low frequencies of LoF variants result in a decrease in imputation accuracy, and a subsequent drop in power to detect association. However, note that the *NOD2* frameshift indel, with an allele frequency of <3% and an odds ratio of approximately 3, achieved a P value of 1.78 x 10^{-14} for association with Crohn's disease despite having a low info score for imputation (0.25). This suggests that my analysis would have successfully identified other LoF variants with large effects, even where allele frequency and imputation accuracy was relatively low. Additionally, imputation quality was high for common LoF variants, allowing us to positively rule out a major role of common LoF variants in complex disease.

In addition to the *NOD2* variant that achieved genome-wide significance, two LoF variants achieved Bonferroni-corrected significance: rs16380, a frameshift indel in *ZNF3* (associated in type 1 diabetes), and a novel frameshift indel at chr1:152018423 in the gene *SLC27A3* (associated in hypertension). I pursued the evidence for association for the *ZNF3* variant using data from a meta-analysis of genome- wide association studies of type 1 diabetes incorporating 7,514 cases and 9,045 controls (Barrett et al., 2009a). 3 SNPs were in strong linkage disequilibrium with rs16380 based on 1000 Genomes pilot data that were also examined in the meta-analysis; these showed only nominal significance in the meta-analysis (P = 0.03-0.04), and this association was driven entirely by the samples overlapping with the WTCCC1 analysis: looking only at samples that were not overlapping with WTCCC1, the P value was 0.4012. This suggests that the marginally significant association in the WTCCC1 samples is a chance finding rather than a genuine association.

3.5 Concluding remarks

Throughout this chapter we have seen how new reference sets can add significant value to genome-wide association studies via genotype imputation. This has included allowing assessment of low-frequency variations from both HapMap and 1000 Genomes reference sets, as well as facilitating meta-analysis of diverse African populations and inferring the impact of newly discovered loss-of-function variants in human disease.

However, we have also seen that imputation is most useful when we have access to large, diverse and high-density reference sets. The well-matched but small HapMap2 reference set is not sufficient to allow accurate imputation of low-frequency variation in Europeans (section 3.2). Likewise, despite its high marker density, the 1000 Genomes pilot data is not able to produce accurate imputation in a diverse African population (section 3.3.2). These experiments have shown that to accurately impute all markers down to low frequency, we require sample sizes on the scale of the HapMap3, but with the high-density granted by sequencing.

In essence, this is what has now been achieved by the 1000 Genomes Project Phase 1 release (Project, 2012), which we have seen is capable of imputing low-frequency variation even in a diverse African population (section 3.3.2). This reference set, and subsequent imputation sets from the 1000 Genomes Project and other sequencing projects, presented a new opportunity to extend the reach of genome-wide association studies into new frequency ranges and classes of variation. As such, they represent a valuable, and continually growing, resource for adding value to GWAS.

Chapter 4

Investigating IBD genetics using the Immunochip

4.1 Introduction

This chapter describes a set of studies carried out using a custom genotyping platform named Immunochip. This genotyping chip was collaboratively designed by a large number of researchers in the genetics of complex immune and inflammatory disease, in order to offer an affordable way of performing very large locus discovery and fine-mapping studies. This chapter describes the application of this genotyping chip to the large number of samples collected by the component research groups of the International IBD Genetics Consortium (IIBDGC).

Both the Immunochip in general, and the IIBDGC study in particular, have been very successful in uncovering the genetics of immune-mediated disease. One study described in this chapter increased the number of associated loci known for IBD to 163, which is more than for any other complex disease. The very large number of associations has also necessitated a change in the way we interpret these results, from a locus-by-locus examination of genes to a large-scale bioinformatic interrogation of all loci. Much of this chapter will be dedicated to applying these techniques to the results of the Immunochip studies.

4.1.1 Overview of this chapter

I will begin (Section 4.2) with a discussion of the design of the Immunochip. This section starts with a discussion of the economics and power considerations of large-scale locus discovery and fine-mapping projects. It also contains a brief investigation into the biology of the fine-mapping loci submitted to the Immunochip, and what they tell us about the shared biology of immune-mediated diseases.

Section 4.3 will discuss the IIBDGC Immunochip data itself, and how calling, quality control and association analyses were carried out. It will describe the large number of novel loci this study has uncovered. Section 4.4 describes a detailed set of bioinformatic analyses to transform this locus list into biological insights. These analyses draw on a range of external data, such as associations with other phenotypes, gene networks, gene annotations, population genetic data and expression analyses. This section also sets out the main biological conclusions that can be drawn from these analyses, as I see them.

Finally, I will discuss two smaller studies carried out on this dataset. Section 4.5 discusses an association study of Y chromosome haplogroups in IBD, and reports a novel association with a Northern European Y haplogroup. Section 4.6 discusses a pilot fine-mapping project, investigating coding and non-coding causal variants in the important *NOD2* locus in CD, which will act as a template for larger Immunochip fine-mapping efforts.

Genotyping method	Cost/sample	Number of variants
Sequenom genotyping (1 plex)	£1.25	25-30
Illumina OmniExpress GWAS array	£160	800,000
Agilent and HiSeq targeted sequencing	£90	All in 6Mbp target region
Illumina Infinium iSelect HD custom genotyping	£25[a]	90,000-250,000

Table 4.1: The costs and capacities of various genotyping technologies. All costs are approximate, and assume large order numbers (>5000 individuals). [a] Assuming an order of >100,000 chips.

4.2 An overview of the Immunochip

4.2.1 The economics of the Immunochip

The economics of deep replication

The 30 novel loci discovered by the last International IBD Genetics Consortium's GWAS meta-analysis of Crohn's disease (Franke et al., 2010) have a median odds ratio of 1.1. The total discovery and replication dataset in this study contained 22,441 cases and 29,496 controls, and thus had a 90% power to establish such loci at genome-wide significance ($p < 5 \text{ x } 10^{-8}$), assuming an allele frequency of 0.25 and an additive genetic model. However, a limitation of this study was that the discovery cohort only had ~29% power to detect these signals with a p-value less than the significance threshold required to be taken forward into the replication ($p < 5 \text{ x } 10^{-6}$). This means that we have likely only discovered 29% of the variants that the total collection is well-powered to detect, suggesting another 70 loci that could be discovered. How can we map these loci in an affordable manner?

One option for uncovering some of these associations would be to expand the GWAS collection. Doubling the number of cases on a low-cost genotyping chip such as the Illumina OmniExpress would cost around £160 x 6333 = £1,013,280 (all costs shown in Table 4.1). This would increase the proportion of true associations taken forward for replication to 65%, and would likely result in around 50 new loci for follow-up. Replication on two Sequenom plexes would then cost around £76,370. This would thus involve spending a total of £1,089,650 to discover approximately 33 new loci, at a cost of £29,450 per locus.

Instead of expanding the GWAS collection, we could instead expand the replication genotyping effort (a so-called deep replication experiment). For a replication set containing all SNPs with $p < 10^{-4}$ would contain around 800 SNPs (or 32 Sequenom plexes), and would include 54% of true associations. This would cost £1.25 x 32 x 30,548 = £1,221,920 to uncover approximately 26 loci, or £46,997 per locus.

There is a third option: custom microarray genotyping. Designing a custom genotyping array allows the genotyping of a large number of SNPs at a lower cost than GWAS arrays. For instance, the Illumina iSelect Infinium HD custom genotyping chips can genotype up to 250,000 markers. For small numbers of samples the cost is relatively high (starting at around £100/sample). However, if a very large number of chips are ordered the price can fall substantially, and for orders measured in the hundreds of thousands the price falls to under £25/sample.

At this price, the entire IIBDGC replication cohort can be genotyped for £763,700. Additionally, because tens of thousands of SNPs can be taken forward for replication, we can perform very deep replication. For instance, taking forward the approximately 5000 SNPs that show $p < 10^{-3}$ would allow us to test 76% of true associations. This would allow us to discover 44 new loci at a cost of £17,357 each.

The economics of fine-mapping

Most of the associations that have been established during the IIBDGC meta-analyses are still poorly understood. For all but the most long-established associations the causal variant is unknown, and in many cases the gene or genes that are being acted on are also unknown. Bioinformatic techniques, such as those discussed in section 4.4.3, can shed light some light on the causal genes. However, the gold standard for establishing causation is genetic fine-mapping, i.e. demonstrating that a single variant, and no others, is capable of explaining the observed association.

In general, fine-mapping is not easy to achieve. To take a simple example, consider a common association (allele frequency of 50%) with a small effect size (odds ratio of 1.2), with the lead SNP in high LD ($r^2 = D' = 0.95$) with another variant of the same frequency. To have 80% power to identify the causal variant with high certainty (i.e. posterior > 0.99), we would require genotypes at 20,000 cases and 20,000 controls. In practice, the structure of the genome, and the biases of GWAS detection, will lead to most associations having many variants in high LD. To fine-map these associations we need a large number of samples, genotyped for a large number of SNPs. The IIBDGC cohort, with an effective sample size of around 25,000 cases, has enough power to fine-map a significant fraction of the CD associations detected by GWAS. However, designing this experiment in an affordable manner is difficult.

A basic fine-mapping effort will involve genotyping a limited set of candidate causal variants. If we examine the 40 CD loci that have not been previously fine-mapped with the lowest degree of LD, we find that there are

536 SNPs with $r^2 > 0.8$ to the hit SNP in the 1000 Genomes pilot. This set of SNPs could be genotyped using around 19 Sequenom plexes, and would cost £1,233,504 to genotype the entire IIBDGC cohort. However, if the causal variant has $r^2 < 0.8$ to the hit SNP, we will not find it (and indeed may end up with a false positive causal variant). Additionally, only the primary signal at the locus can be fine-mapped in this fashion.

The ideal fine-mapping experiment involves sequencing entire regions, as this allows us to assay all variants that could drive the association, as well as allowing us to identify new (potentially low-frequency) associations. A pull-down array designed to capture DNA from CD loci, combined with low-cost next-generation sequencing would allow us to perform this. However, while the cost of sequencing is now extremely low, the cost of sample preparation and the pull-down arrays is still relatively high. Even if we restricted sequencing to 20K cases and 20K controls, such a project would still cost in excess of £3,600,000.

Again, a powerful third solution comes in the form of custom genotyping, and in particular via a combined deep replication and fine-mapping array. The same genotyping array that is being used for deep replication (and thus is already being run on a significant fraction of the IIBDGC cohort) can also used to genotype variants in IBD-associated regions taken from the 1000 Genomes project and dbSNP. This allows the primary signal and any secondary signals to be fine-mapped, and also allows any low-frequency variation that is in the SNP databases to be assayed as well. This approach has less full coverage than would be achieved by sequencing, but for common variation the coverage should be nearly as high, at a much lower cost. In essence, the fine-mapping and deep replication efforts are combined on a single chip.

An immune-mediated disease chip

We have seen that custom genotyping is an affordable way to discover and fine-map new loci using existing collections, providing that a large enough purchase is made. If the IIBDGC alone purchased 40,000 chips (enough to genotype all CD and UC cases, and all controls), this would still be too small an order to be cost effective. However, by including deep replication studies from other disease consortia, we can rapidly increase the total number of chip users, and reduce the price to affordable levels.

It was these economic considerations that led to the creation of the Immunochip. This custom genotyping platform was designed for deep replication and fine-mapping in a wide range of studies, with particular focus on immune-mediated diseases (Table 4.2). Along with the reduction in price, there are a number of additional advantages to this cross-consortium collaboration. Firstly, it greatly reduces the costs of control genotyping, as common control sets can be used. Secondly, because there is a high degree of genetic overlap in immune-mediated diseases (see section 4.2.3) a high proportion of deep replication SNPs and fine-mapping regions will be associated to multiple diseases, reducing redundancy and increasing the power to detect new shared associations. Finally, because the chip contains almost all known immune-mediated disease loci at time of creation, and because it is being run on a range of different immune-mediated diseases, it makes a perfect platform for performing cross-phenotype analyses of immune diseases.

4.2.2 The content of the Immunochip

The Immunochip is an Infinium iSelect HD custom genotyping chip, manufactured by Illumina. It contains 196,524 variants (largely SNPs, plus 718 small

Immune-mediated diseases		Other diseases
Autoimmune Thyroid Disease[a] (AITD)		Barrett's oesophagus
Ankylosing Spondylitis (AS)		Bipolar Disease (BD)
Bacteraemia susceptibility (BS)		Glaucoma
Crohn's Disease (CD)		Ischaemic stroke
Coeliac Disease (Coeliac)		Parkinson's Disease
IgA deficiency[a] (IgAD)		Pre-eclampsia
Multiple sclerosis (MS)		Psychosis endophenotypes
Primary Biliary Cirrhosis[a] (PBC)		Statin response
Psoriasis (PS)		Reading and mathematics abilities
Rheumatoid arthritis (RA)		Schizophrenia
Sarcoidosis		
Systemic lupus erythematosus (SLE)		
Type 1 Diabetes (T1D)		
Ulcerative colitis (UC)		
Vasculitis		
Visceral leishmaniasis		

Table 4.2: The diseases involved in the Immunochip design [a]Fine-mapping only, no deep replication.

indels), picked specifically for the purpose of discovering and fine-mapping genetic associations with immune-mediated disease. The variants are selected based on three criteria: deep replication of variants implicated by GWAS, fine-mapping of established disease associations and variants submitted as wildcards. In total, approximately 240,000 SNPs were selected for inclusion, with an assay design success rate of ∼80%.

Deep replication

Approximately 50,000 SNPs are included on the Immunochip as deep replication for the diseases shown in Figure 4.2. These SNPs showed suggestive evidence in GWAS, and are intended to be replicated in a large set of samples in order to discover novel associations. Many of these (including all replication for non-immune-mediated traits) were included as part of the second Wellcome Trust Case Control Consortium project. While these SNPs make up only a quarter of the total, they represent the larger proportion of the genome tagged, as they are largely independent (in contrast to the high level of redundancy in the fine-mapping regions).

Fine-mapping regions

A total of 290 established disease associated loci were included on the Immunochip for fine-mapping. 196 of these came from studies that were submitted, accepted or published when the Immunochip was designed (listed in Table 4.3). An additional 94 loci were included on the basis of personal communication with researchers carrying out GWAS and GWAS meta-analyses that were not yet submitted for publication at the time of chip design (listed in Table 4.4). All but one of these studies have now been published. However, many of the fine-mapping loci included were not included in the final publication for these studies. Some of these loci were subsequently discovered in other studies, but there are still 13 “false” loci that are included on the Immunochip and have never been reported in a publication (Table 4.4). Many of these loci are actually true associations; for instance, three of the four “false” IBD loci are confirmed in the IIBDGC Immunochip data (see section 4.2.2).

Fine-mapping regions were defined by taking 0.2cM on either side of the hit SNP, using the combined HapMap2 genetic map. SNPs for fine-mapping were chosen from the 1000 Genomes pilot 1 two-of-three way SNP site set (dated 10/11/2009), and from dbSNP build 130.

The 290 fine-mapping regions include a high degree of overlap. Exactly

Phenotype	Study	Loci
AITD	Kavvoura et al. (2007)	1
AITD	Brand et al. (2009)	1
AS	Burton et al. (2007)	2
BD	Ferreira et al. (2008)	2
BD	O'Donovan et al. (2008)	1
CD	Barrett et al. (2008)	30
CD	Kugathasan et al. (2008)	2
Coeliac	Hunt et al. (2008)	3
Coeliac	Dubois et al. (2010)	27
IgAD	Ferreira et al. (2010)	1
MS	Booth et al. (2008)	3
MS	De Jager et al. (2009)	5
MS	Bahlo et al. (2009)	1
MS	Esposito et al. (2010)	3
MS	Jakkula et al. (2010)	1
MS	McCauley et al. (2010)	2
MS	Mero et al. (2010)	1
PBC	Hirschfield et al. (2009)	1
PS	Capon et al. (2008)	1
PS	Nair et al. (2009)	6
PS	Zhang et al. (2009)	2
RA	Raychaudhuri et al. (2009b)	23
SLE	Harley et al. (2008)	3
SLE	Kozyrev et al. (2008)	1
SLE	Han et al. (2009)	14
SLE	Gateva et al. (2009)	7
T1D	Cooper et al. (2008)	1
T1D	Smyth et al. (2008)	1
T1D	Barrett et al. (2009a)	34
T1D	Qu et al. (2009)	1
T1D	Wallace et al. (2010)	2
UC	Franke et al. (2008)	1
UC	Kugathasan et al. (2008)	2
UC	Imielinski et al. (2009)	1
UC	Asano et al. (2009)	1
UC	Silverberg et al. (2009)	3
UC	Barrett et al. (2009b)	4
UC	Festen et al. (2009)	1

Table 4.3: Fine-mapping regions included on the Immunochip as a result of studies published or submitted at the time of chip design. The "Loci" column gives the total number of fine-mapping regions on the Immunochip from this study.

Disease	Study	On chip	In study (Confirmed)	“False”
AS	Reveille et al. (2010)	4	3 (1^c)	0
AS	Evans et al. (2011)	2	1 (1^d)	0
CD	Franke et al. (2010)	34	32 (1^a)	1
MS	Sawcer et al. (2011)	11	10	1
PS	Strange et al. (2010)	11	9	2
PS	Stuart et al. (2010)	3	2	1
RA	Stahl et al. (2010)	1	2	1
RA	Freudenberg et al. (2011)	1	1	0
SLE	NA^b	10	0^b (3^e)	7
T1D	Swafford et al. (2011)	1	1	0
T1D	Heinig et al. (2010)	1	1	0
UC	Anderson et al. (2011)	15	13 (2^a)	0

Table 4.4: Fine-mapping regions included on the Immunochip as a result of studies that were not completed at the time of chip design. “On chip” is the total number of loci included on the Immunochip from this study, “In study” is the number of these loci that were subsequently included in the final locus list for that study, “Confirmed” is the number of loci that were not included in the study have subsequently been confirmed elsewhere, and “False” is the number of loci included on the Immunochip from this study that have never been published. ^{a}These loci are confirmed in the study described in this chapter, ^{b}I do not believe that this study has been published yet. ^{c}Confirmed by Evans et al. (2011) ^{d}Confirmed by Danoy et al. (2010) ^{d}Confirmed by Guerra et al. (2012)

how many independent regions exists depends on exactly what parameters are used, but merging any regions with boundaries that lie within 50kb of each other, and excluding the two BD regions, gives 186 separate immune-mediated disease regions.

In addition to the regions included due to established associations, a total of 6378 SNPs from across the MHC were included to allow fine-mapping and imputation of HLA alleles.

Wildcard variants

Many groups with the contributing consortia submitted “wildcard” SNPs. Each contributor was given an allocation of SNPs that could be picked based on criteria not directly related to deep replication or fine-mapping.

Many researchers submitted wildcard variants in candidate genes. For instance, the IBD consortium added three SNPs in the gene *XBP1*, implicated as involved in IBD by a functional and candidate gene study (Kaser et al., 2008). The most associated SNP in the original study, rs35873774, had an odds ratio interval of 0.66-0.84 in 4389 cases and 5322 controls. In the 22,442 cases and 30,837 controls of the IIBDGC Immunochip data, it had an odds ratio interval of 0.92-1.02, suggesting that this association is not real, or at least has been overestimated. A more powerful example is an attempted replication via wildcard genotypes of an association between variants in the gene SIAE and autoimmune disease. The original study that reported the association tested 923 cases and 648 controls (Surolia et al., 2010), but an Immunochip-based study in over 60 thousand individuals failed to replicate the results (Hunt et al., 2012). Often candidate gene studies are expensive to replicate, and many false associations are not disproved. These wildcard replication efforts can allow us to confirm or falsify associations that would not be tested in ordinary circumstances.

Some groups submitted candidate SNPs generated from sequencing experiments. For instance Manny Rivas and colleagues submitted 260 low-frequency SNPs that had been identified through resequencing of IBD GWAS regions, many of which replicated successfully in the IIBDGC Immunochip cohort (Rivas et al., 2011).

Other sets of SNPs were added for other purposes. 100 SNPs within the Killer cell Immunoglobulin-like Receptor (KIR) gene cluster were added, to allow development of techniques to impute KIR serological alleles. 1735 Y chromosome SNPs were included to allow Y haplogroup analyses (discussed in section 4.5 below), and a further 848 SNPs were added from the NHGRI GWAS catalogue to allow testing of GWAS hits from non-immune-mediated diseases.

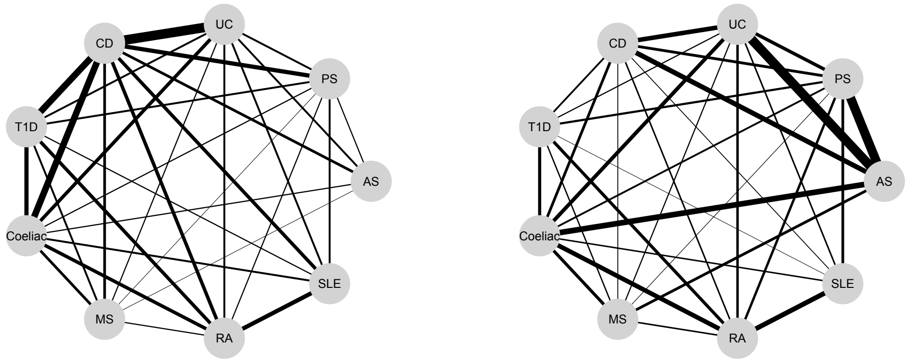
Unpicking “false” IBD fine-mapping regions

There are four IBD fine-mapping regions that were included on the Immunochip despite not appearing in either the Franke et al. (2010) nor Anderson et al. (2011) meta-analysis papers. These include two CD and two UC regions.

In the UC meta-analysis, the first “false” SNP (rs1518070) showed genome-wide significant evidence ($p_{combined}$ = 7.9 x 10^{-9}), leading to its inclusion on the Immunochip. However, final replication did not meet $p_{replication} < 0.05$ due to a high rate of technical failure. The second “false” SNP (rs1569501) showed genome-wide significant evidence of association in the UC GWAS alone, but failed assay design during replication and was thus not included in the final study.

In CD, one “false” SNP (rs1536833) met genome-wide significance in the replication datasets available when the Immunochip was first designed ($p_{combined}$ = 2.6 x 10^{-8}), but dropped just below genome-wide significance when the final replication cohorts were included ($p_{combined}$ = 9.5 x 10^{-8}). The second, rs2098112, showed a significant value of $p_{combined}$ (leading to its inclusion on the Immunochip), but the entire signal was entirely driven by association in the GWAS data, and was excluded from the final list due to lack of signal in the replication.

Of the four IBD fine-mapping regions included “in error”, three were



(a) Unweighted **(b)** Weighted

Figure 4.1: Locus sharing between immune-related diseases, using Immunochip fine-mapping regions. Connecting line width represents number of loci shared, either a) unweighted or b) weighted by square root of the product of the number of associations in both phenotypes.

found to be truly associated in the IIBDGC Immunochip study described in this chapter. The only association that failed to show signal in the Immunochip was rs2098112. Additionally, the improved GWAS imputation described below reduced the association signal from p = 4.5 x 10^{-15} to p = 0.35, showing that this association was driven entirely by poor imputation.

4.2.3 The biology of the Immunochip

The fine-mapping regions on the Immunochip represent a complete survey of the known genetics of immune-mediated disease (or at least, a relatively complete survey of the loci known in mid-2010). What can this list of loci tell us about the shared biology of the diseases that the Immunochip was designed to study?

Locus sharing between immune-mediated diseases

Of the 186 fine-mapping loci, 61 were submitted for more than one disease, including 9 loci shared by at least 4 diseases. Highly shared loci include loci that been traditionally considered important in immunity such as *IL23R/IL12RB2* (5 diseases) and *PTPN22* (4 diseases), and other loci that do not have well-understood roles in immunity such as *KIF21B* (5 diseases).

We can use these shared loci to construct a locus sharing network for 9 autoimmune diseases (excluding diseases with 2 or fewer loci). An unweighted network (Figure 4.1a) shows strong connectivity between CD, UC, T1D and Coeliac. However, these diseases are also those with the largest number of discovered loci, so this connectivity is unsurprising. If we weight the network edges by the geometric mean number of associations in the two diseases, we get a very different network (Figure 4.1b). The strongest connection here is between UC and AS (two comorbid diseases).

Network analyses of Immunochip loci

We can place the Immunochip loci in the context of gene networks, and ask which loci seem to play a central role in these networks. I used two gene network tools (GRAIL and DAPPLE) to construct networks using genes inside Immunochip regions. The first, GRAIL (Raychaudhuri et al., 2009a) (Gene Relationships Across Implicated Loci), is a network connectivity tool that uses text mining to calculate a network distance between genes in different implicated loci. Each gene is measured for enrichment of connectivity to genes in other associated loci, and a p-value is calculated. The second, DAPPLE (Rossin et al., 2011) (Disease Association Protein-Protein Link Evaluator), is a network connectivity tool that uses protein-protein interactions. Each gene is measured for enrichment in either direct or indirect (i.e.

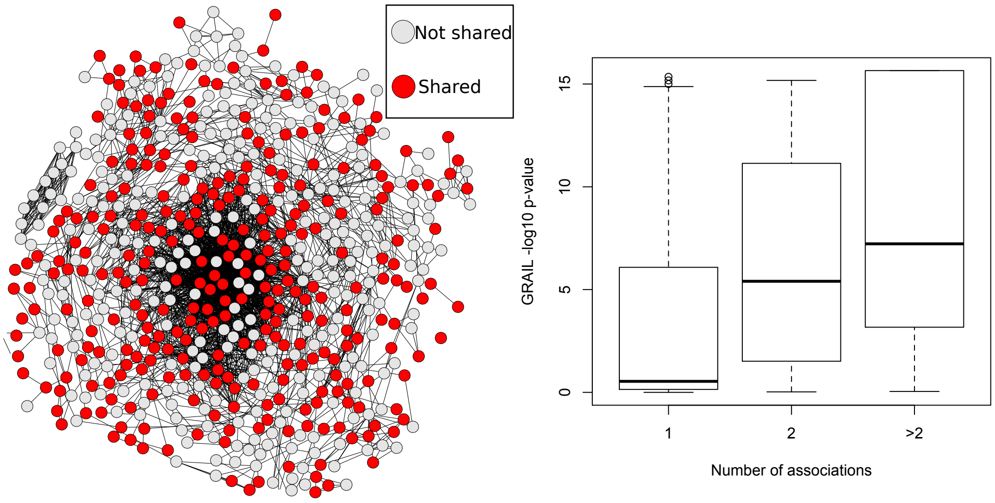


Figure 4.2: The relationship between GRAIL network connectivity and number of associations for Immunochip fine-mapping regions. a) The GRAIL gene network, with genes in shared loci highlighted in red. b) The relationship between GRAIL connectivity p-value and degree of locus sharing

via other proteins) interactions with genes in other loci, and an empirical p-value is calculated by permutation.

Looking at the GRAIL literature network (Figure 4.2), genes that tend to be closest to the centre of the network also tend to be in regions associated with more than one phenotype. In general, there is a correlation between connectivity p-value and number of associations for both GRAIL (Spearman ρ = -0.39, p = 1.45 x 10^{-7}) and DAPPLE (ρ = -0.31, p = 1.15 x 10^{-4}) networks. As intuition might lead us to believe, that loci that play a more central role in the pathways of immune disease are more likely to impact multiple diseases.

The 10 most connected Immunochip fine-mapping loci are shown in Table 4.5. Nine of these regions are associated to more than one disease, though the most significantly connected region, the *TNFSF4* locus, is only associated with SLE. *TNFSF4* (also called *OX40L*) is expressed by dendritic cells and promotes Th2 differentiation and thus humoral immunity, and has

Chrom:Pos (MB)	GRAIL p-value	DAPPLE p-value	Genes	Phenotypes
1:171.4-171.6	3.61 x 10^{-20}	0.23	*TNFSF4*	SLE
1:7.6-8.1	8.32 x 10^{-20}	0.07	*TNFRSF9*	CD, UC, Coeliac
2:204.2-204.5	1.73 x 10^{-19}	<0.002	*ICOS*, *CD28*, *CTLA4*	RA, AITD, T1D, Coeliac
16:28.2-28.9	1.64 x 10^{-18}	<0.002	*IL27*, *NFATC2IP*, *CD19*	CD, T1D
21:44.4-44.5	2.47 x 10^{-18}	0.44	*ICOSLG*	CD, Coeliac
2:191.6-191.7	4.65 x 10^{-18}	<0.002	*STAT4*, *STAT1*	SLE, RA,CD^{a},UC^{a}
1:67.4-67.7	9.97 x 10^{-18}	<0.002	*IL12RB2*, *IL23R*	PS, CD, UC, AS, PBC
20:44.0-44.2	1.11 x 10^{-17}	0.22	*CD40*	RA, CD^{a}, UC^{a}
3:161.1-161.2	3.03 x 10^{-17}	<0.002	*IL12A*	MS, Coeliac
12:54.6-55.1	3.10 x 10^{-17}	<0.002	*IL23A*, *STAT2*	PS, T1D

Table 4.5: The top 10 most connected Immunochip fine-mapping regions, according to a GRAIL network analysis. ^{a}New associations discovered in the IBD Immunochip analysis.

been investigated as a drug target in allergic diseases (Wang and Liu, 2007). If this gene were truly associated only to SLE, and not to other immune-mediated diseases, it would suggest a good starting point for investigating deep-rooted differences between immune diseases. However, we can also use the Immunochip itself to investigate this possibility. The SLE-associated SNP, rs1234315, shows a low but sub-genome-wide-significant signal in the Crohn's disease IIBDGC data (p = 2.03 x 10^{-4}), suggesting that this locus is active in other diseases, but has too small an effect size to be reliably detected in GWAS.

Chrom:Pos (MB)	GRAIL p-value	DAPPLE p-value	Genes[b]	Phenotypes
1:199.1-199.3	0.90	0.63	*KIF21B*	MS, AS, UC, CD, Coeliac
2:162.7-163.1	0.38	0.68[c]	*IFIH1*	IgAD, T1D, PS, CD[a], UC[a]
6:90.9-91.1	0.08	0.96	*BACH2*	T1D, Coeliac, CD, UC[a]

Table 4.6: Immunochip fine-mapping regions associated with at least 3 phenotypes, but with no evidence of connection via either DAPPLE or GRAIL. [a]New associations discovered in the IBD Immunochip analysis. [b]The stated genes are the standard candidate genes given the in the literature [c]*IFIH1* is not included in the protein network used by DAPPLE

As well as highlighting highly connected genes, this analysis can also highlight loci that are associated to many different immune-mediated diseases, but do not show evidence of network centrality. Table 4.6 shows three loci that are associated with at least three diseases, but show $p > 0.05$ in both the GRAIL and DAPPLE analyses. One of these genes, *IFIH1*, was not present in the DAPPLE interaction dataset, so may represent a simple lack of data. One of the others, *KIF21B*, was originally discovered in MS, and was believed to act via its role in neuronal transport (McCauley et al., 2010). However, associations to AS, CD, UC and Coeliac disease suggest a more general role in immunity. All three of these regions are associated in IBD, and two contain candidate genes identified by the IBD-specific gene prioritisation approach described in section 4.4.3. *IFIH1* shows a marginal GRAIL association ($p = 0.032$), and *KIF21B* was prioritised by a gene co-expression network approach.

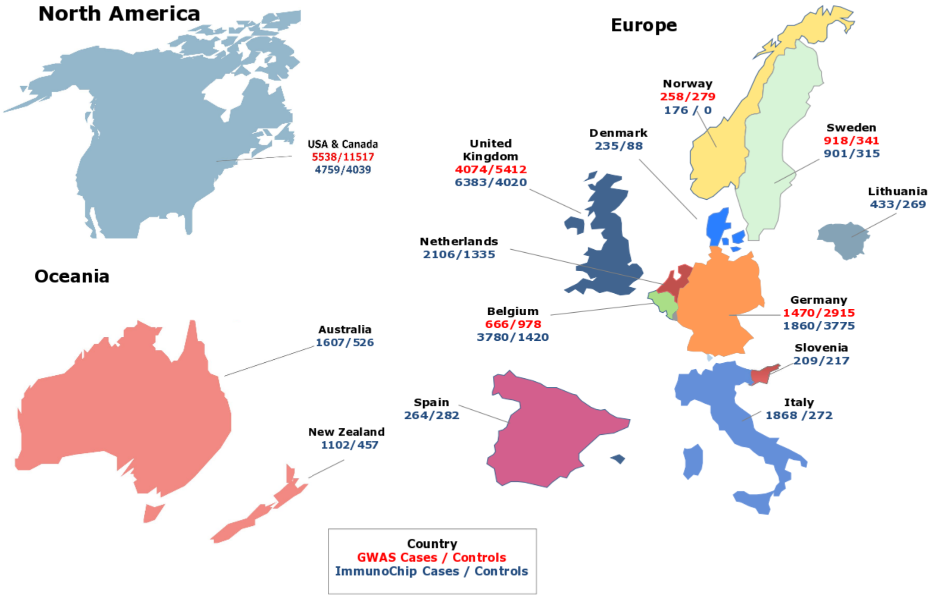


Figure 4.3: Numbers of IBD and control samples passing quality control, from each country participating in this study. The numbers for the Immunochip samples (numbers in blue) only include samples that are not also present in the GWAS (numbers in red).

4.3 QC and association analysis of the IIBDGC Immunochip dataset

4.3.1 The IIBDGC Immunochip dataset

As part of the International IBD Genetics Consortium (IIBDGC), research groups from 15 countries (Figure 4.3) collected Crohn's disease (CD) and Ulcerative colitis (UC) samples and genotyped them using the Immunochip. These data were combined with the GWAS meta-analysis collection to create a large dataset for locus discovery.

The GWAS meta-analysis dataset consists of seven Crohn's disease collections and eight ulcerative colitis collections with genome-wide SNP genotype

Cohort	Countries	Chip	Case / control (unique)
CD cohorts			
BEL1	Belgium, France	ILMN317	513 / 884 (884)
BEL2	Belgium	ILMN317	153 / 94 (94)
CEDARS	USA	ILMN317	835 / 2881 (1364)
CHOP	USA, Canada, Italy, UK	ILMN550	1495 / 6090 (3054)
GERMAN	Germany	ILMN550	480 / 1114 (573)
NIDDK	USA, Canada	ILMN317	759 / 929 (462)
WTCCC	UK	AFFX500	1721 / 2935 (1612)
Total			5956 / 14927 (8043)
UC cohorts			
CEDARS	USA	ILMN317	836 / 2928 (1566)
CHOP	USA, Canada, Italy, Scotland, Canada	ILMN550	664 / 6091 (3038)
GERMANY	Germany	AFFX6	990 / 2915 (2383)
NIDDK1	USA, Canada	ILMN550	498 / 1070 (624)
NIDDK2	USA, Canada	ILMN550	451 / 1428 (1420)
NORWEGIAN	Norway	AFFX6	258 / 279 (279)
SWEDISH	Sweden	ILMN317	918 / 341 (341)
WTCCC	UK	AFFX6	2353 / 5412 (4076)
Total			6968 / 20464 (13727)

Table 4.7: GWAS cohorts, with country of origin, genotyping chip and size. Case and control numbers are after QC, and the number in brackets in the number of unique controls after duplicates between CD and UC have been removed.

data (Table 4.7). The CD cohorts contained a total of 6,299 cases and 15,148 controls, and the UC cohorts contained a total of 7,211 cases and 20,783 controls (the control sets contain largely overlapping samples). Four different chips were used: two produced by Affymetrix (the GeneChip Human Mapping 500K Array and the Genome-Wide Human SNP Array 6.0) and two produced by Illumina (the HumanHap300 BeadChip and the HumanHap550 BeadChip). The majority of these samples were used in the published IIB-

Center	Nationality	CD / UC / control
Bonn	Germany	0 / 0 / 1494
Cedars Sinai	USA	1156 / 822 / 0
Feinstein Institute	Australia	844 / 706 / 464
	Canada	610 / 506 / 305
	New Zealand	422 / 420 / 0
	Netherlands	140 / 157 / 0
	USA	743 / 364 / 2288
	Total	2759 / 2153 / 3057
Kiel	Denmark	66 / 169 / 88
	Germany	1062 / 261 / 1490
	Italy	1273 / 595 / 272
	Lithuania/Baltic	129 / 304 / 269
	New Zealand	260 / 0 / 457
	Norway	122 / 54 / 0
	Spain	264 / 0 / 282
	Sweden	669 / 0 / 0
	Total	3845 / 1383 / 2858
Leuven	Belgium	1434 / 783 / 721
Munich	Germany	0 / 0 / 286
U of Pittsburgh	Australia	0 / 57 / 62
	Canada	0 / 25 / 20
	Germany	0 / 537 / 505
	Netherlands	0 / 327 / 346
	Sweden	0 / 232 / 315
	USA	315 / 218 / 388
	Total	315 / 1396 / 1636
U de Liege	Belgium	1015 / 548 / 699
UMC Groningen	Slovenia	171 / 38 / 217
	Netherlands	1116 / 366 / 989
	Total	1287 / 404 / 1206
UVA	UK	0 / 0 / 2441
Sanger Institute	UK	2952 / 3431 / 1579
Total		14763 / 10920 / 15977

Table 4.8: Immunochip cohorts, broken down by genotyping centre and country of origin. Case and control numbers are after QC, and after samples that overlap the GWAS cohorts have been removed.

DGC meta-analyses (Franke et al., 2010; Anderson et al., 2011).

The Immunochip dataset consists of collections from 15 countries genotyped in 11 different genotyping centres (Table 4.8). Genotyping was performed in 20 batches, with each centre processing between one and three batches. A total of 60,828 samples were genotyped on the Immunochip, comprising 20,076 CD cases, 15,307 UC cases and 25,445 controls. These numbers include many samples that were also present in the GWAS cohorts, which are to be used for fine mapping and not for locus discovery.

Overall, after QC and removing overlapping samples (see below), this dataset has 20,700 CD cases, 17,865 UC cases and 37,747 controls. This is the first time a large meta-analysis has analysed CD and UC together, allowing very high power for variants shared across both phenotypes. For instance, the dataset has an 80% power to detect common IBD associations with an odds ratio greater than 1.06. It is also well-powered to detect low-frequency variants (MAF of 1%) with an odds ratio of >1.35, and rare (MAF = 0.1%) variants with an odds ratio of >2.3.

4.3.2 Genotyping, imputation and quality control

GWAS data

In addition to the quality control performed by individual studies before submission, each GWAS study was subject to the following QC:

1. missing rate per SNP < 0.05
2. missing rate per individual < 0.02
3. heterozygosity per individual ± 0.2
4. missing rate per SNP < 0.02 (after sample removal)

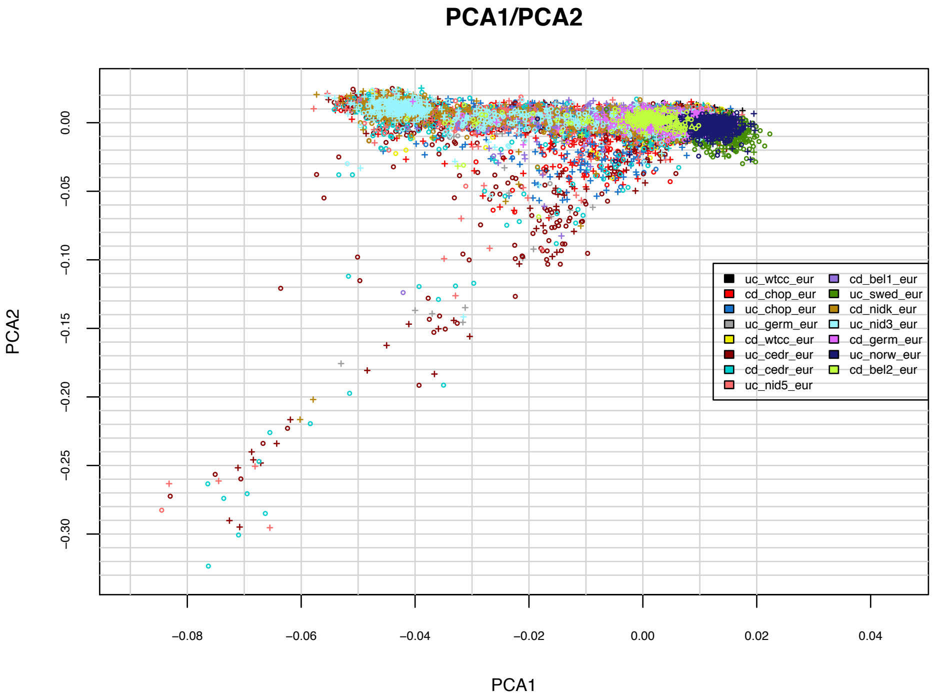


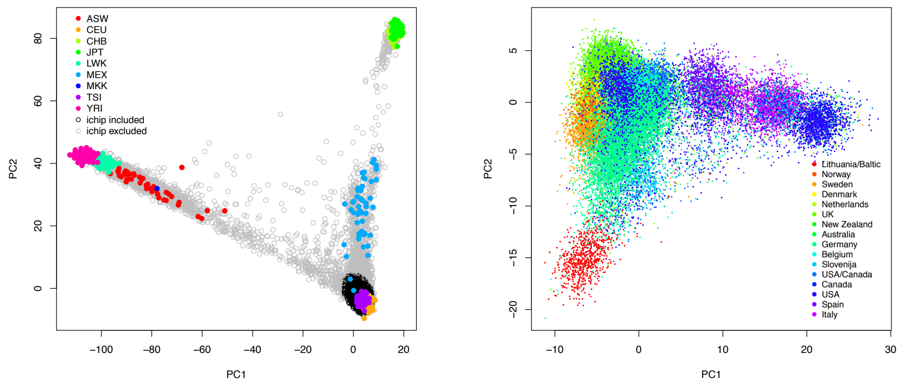
Figure 4.4: All GWAS samples plotted on the first two principal components, coloured by study. Circles are cases, crosses controls

5. missing rate per SNP difference in cases and controls < 0.02

6. Hardy-Weinberg equilibrium (controls) P $< 10^{-6}$

7. Hardy-Weinberg equilibrium (cases) P $< 10^{-10}$.

A set of 17,385 high-frequency SNPs (MAF $> 5\%$) in linkage equilibrium ($r^2 < 0.05$ for all SNP pairs) was generated. Plink was used to calculate relatedness statistics (the estimated coefficient of relatedness $\hat{\pi}$), and individuals with $\hat{\pi} > 0.2$ to another sample were removed. Samples duplicated between CD and UC control datasets were recorded: these samples are kept in for single-phenotype tests, but removed for combined tests. Principal component analysis was performed (Figure 4.4), and principal components that



(a) Continental PCA with ethnic outliers

(b) Within Europe PCA by country

Figure 4.5: a) Principal component projection of Immunochip samples onto a) continental axes fitted from HapMap samples and b) European axes fitted from Immunochip controls

correlated with disease phenotype were recorded for use as covariates.

Imputation of untyped SNPs was performed within each study in batches of 300 individuals. These batches were randomly drawn in order to keep the same case-control ratio as in the total sample from that study. Imputation was performed with the CEU+TSI HapMap3 reference set (containing 1,252,901 polymorphic SNPs), using Beagle 3.13 with a chunk size of 10Mb and default parameters.

Immunochip data

Because many of the variants on Immunochip do not meet the manufacturer's quality standards set for GWAS products, rigorous QC is essential. Furthermore, because samples with poor quality DNA or with other genome-wide problems can adversely affect the genotype calls at high quality samples, I performed a first stage of "coarse" QC on genotypes called using Illumina's GenomeStudio program. I exclude samples with >5% missing data, genome-wide heterozygosity outside a 95% confidence interval in each batch, samples of non-European ancestry (via PCA, see below) or with abnormal mean intensity values from further analysis.

For all remaining samples, I used the optiCall clustering program (Shah et al., 2012) (v0.3.0) to call genotypes, with a no-call cutoff of 0.7 and HWE blanking disabled. I identified duplicate and related samples ($\hat{\pi} > 0.1$) using PLINK with the same set of SNPs used for PCA (details below for details), and removed the duplicate or related sample with the higher missing data rate. I used a set of 692 SNPs present on both the Immunochip and all four GWAS chips to remove Immunochip samples that were also present in the GWAS. I removed samples without a phenotype definition of Crohn's disease, ulcerative colitis or healthy control, and finally removed all samples with > 2% missing data in this improved call-set.

I performed SNP QC in this filtered dataset, removing SNPs with >2% missing data or HWE p-value $< 10^{-10}$ in controls. However, a relatively large number of SNPs still showed poor clustering, driving many false positive associations. To further ensure the quality of genotype calls in our analysis, I selected 3,356 variants for manual inspection, including those with meta analysis p<10^{-5} which fulfilled at least one of the following criteria:

1. Cochran heterogeneity $p < 0.01$ between GWAS and Immunochip (N=871)

2. lie outside fine-mapping regions known to be associated with immune-mediated disease (N=797)

3. are one of the 3 most significantly associated SNPs in a region (N=851)

4. any SNP with $p < 5\text{x}10^{-8}$ which did not fit those criteria (N=195)

5. random SNPs as a comparator (N=642)

I distributed intensity data for these SNPs to 16 members of the IIBDGC for manual inspection. Included was a version of the manual inspection program Evoker (Morris et al., 2010) optimised for multi-cohort inspection, and a document describing the protocol for manual inspection. Each SNP was inspected by three individuals, and was considered to have passed inspection if three individuals passed the SNP, or two passed it and one marked it as a "Maybe". 1015 SNPs were removed during this process. A further 29 SNPs had genotypes manually adjusted (blind to phenotype and association statistics) to correct recoverable errors.

I used principal component analysis to identify ethnic outliers, and to generate covariates to control for population stratification. To identify outliers on the continental scale I constructed a reference set consisting of 662 HapMap founder samples genotyped on the Illumina Human1M, the Affymetrix Human SNP Array 6.0, and the Illumina Omni2.5 for the HapMap3 and 1000 Genomes Projects. This reference set was designed to maximise overlap with the Immunochip, and has a total of 3,268,731 SNPs, of which 83,689 are present on the Immunochip. I used PLINK to LD prune the data such that no pair of SNPs had $r^2 > 0.2$, and I also removed GC/AT SNPs, SNPs within known high LD regions (Price et al., 2008) and SNPs with MAF $< 5\%$. I projected the Immunochip samples on the principal component axes generated using these 17,891 SNPs from the 662 reference samples using the R package snpMatrix (Clayton and Leung, 2007). All samples that did not cluster with the European samples were excluded (Figure 4.5a).

To resolve within-Europe relationships, I performed PCA within the remaining Immunochip samples. LD pruning was performed within European

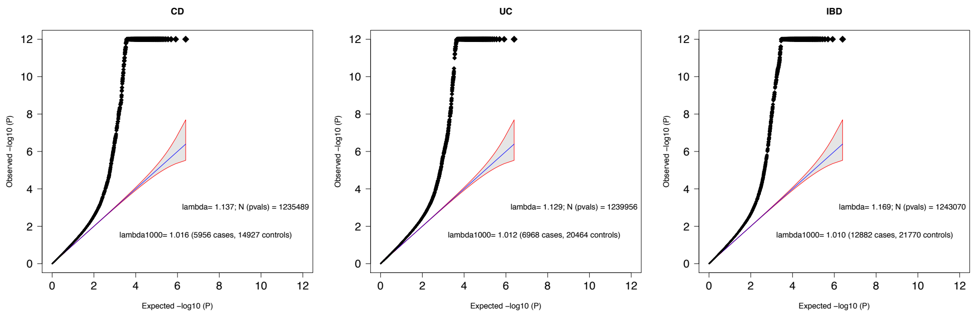


Figure 4.6: QQ plots, λ and λ_{1000} values for the CD, UC and IBD GWAS analyses. Grey shapes show 95% confidence interval under the null.

controls (this was performed three times, to properly break up the LD in fine-mapping regions), and SNPs present in high LD regions or with MAF $< 5\%$ were removed, leaving a total of 19,111 SNPs. I generated principal component axes within the controls, and projected the cases onto these axes to generate PCs for all samples. The first four principal component axes seemed to capture significant population structure (Figure 4.5b), and addition of components beyond the fourth as association covariates in a subset of the Immunochip data did not further reduce the genomic inflation factor.

4.3.3 Association analyses

4.3.4 GWAS and Immunochip analyses

Three association scans were performed for both GWAS and Immunochip. These included a CD analysis (Crohn's disease vs controls), a UC analysis (ulcerative colitis vs controls) and an IBD analysis (combined CD and UC vs controls).

For the GWAS, the CD scan had a total of 5,956 QC+ cases and 14,927 QC+ controls, the UC scan had 6,968 cases and 20,464 controls, and the IBD scan had 12,882 cases and 21,770 controls. For the IBD scan, controls

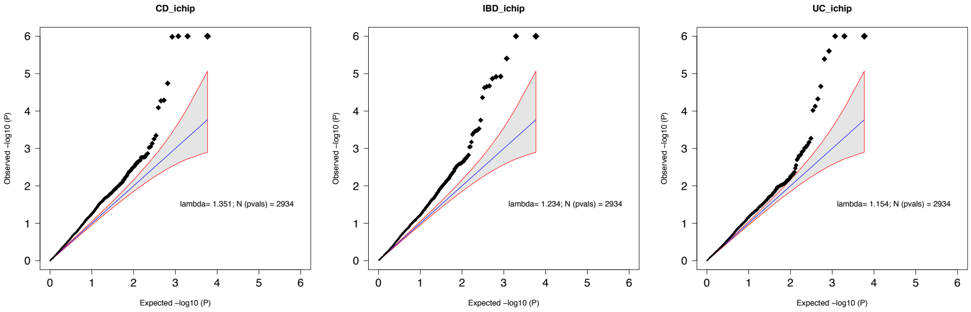


Figure 4.7: QQ plots and λ values for the CD, UC and IBD Immunochip analyses. Grey shapes show 95% confidence interval under the null.

that overlapped between the CD and UC control cohorts were removed from whichever dataset had a greater excess of controls. Association testing was carried out in PLINK, using the dosage data from the imputation and using 10, 7, 15 principal components for CD, UC, IBD respectively as covariates (all PCs that correlated with case-control status). The CD, UC and IBD scans had genomic inflation (λ_{GC}) values of 1.137, 1.129, and 1.169 respectively (Figure 4.6). These inflation figures are substantially lower than the figures for the previous CD and UC meta-analyses.

For the Immunochip analysis, the CD, UC and IBD scans all used the entire control dataset. The CD scan had a total of 14,763 QC+ cases, the UC scan had 10,920 cases, the IBD scan had 25,683 cases, and all scans used the 15,977 QC+ controls. I performed association testing using additive logistic regression in PLINK conditioned on the first four principal components. Test statistic inflation was computed from a set of 3120 SNPs chosen based on GWAS of schizophrenia, psychosis and reading/mathematics ability. Genomic inflation factors were relatively low, given the large sample size and presence of polygenic risk: $\lambda_{GC_{CD}} = 1.353$, $\lambda_{GC_{UC}} = 1.154$, $\lambda_{GC_{IBD}} = 1.234$ (Figure 4.7).

For comparison, I also performed an association test on all IBD samples using the Cochran-Mantel-Haenszel method to stratify by country of origin of the samples. This is one of the standard methods used to analyse GWAS replication data, where population stratification correction via principal components are usually not available. The genomic inflation value for the IBD all analysis was $\lambda_{GC_{IBD}} = 2.00$, showing that without the genome-wide SNP data on the Immunochip this replication analysis would have shown severe inflation.

This also has some worrying implications for the GWAS field, as it suggests that most standard international replication datasets will suffer from test statistic inflation. This in turn could mean that combined GWAS-replication p-values may be too liberal. In the future, it seems prudent that large replication analyses should include a number of ancestry-informative SNPs to control for stratification. Exactly how many such SNPs would be required to reduce inflation is unknown, and the Immunochip provides a platform to investigate this.

4.3.5 Deep replication meta-analysis

A combined analysis was performed using both the GWAS and the Immunochip association results comprising 20,700 Crohn's disease, 17,865 ulcerative colitis cases and 37,747 healthy controls.

All SNPs in GWAS association results with $p < 0.01$ in the CD, UC or IBD scans were selected for replication in the Immunochip dataset (a total of 25,075 SNPs). A fixed-effect meta-analysis was performed using odds ratios and standard errors from the GWAS hit and the Immunochip tag with the highest r^2 to the hit SNP, providing a tag with $r^2 > 0.4$ was available. The Cochran heterogeneity p-value was also calculated (none of the final association signals showed significant heterogeneity after correcting for multiple testing).

SNPs with p $<$ 5 x 10^{-8} in any of the three phenotypes in this analysis were combined into clumps if they had $r^2 > 0.1$. SNPs within these clumps were tested for evidence of association independent of the strongest signal in the clump. Because the tag SNP meta-analysis approach makes standard methods for conditional analysis impossible to carry out, so we used an approximate conditional Z-score

$$Z_i' = Z_i - r_{i,hit} Z_{hit} \tag{4.1}$$

Where Z_i is the Z score of the SNP being tested, Z_{hit} is the Z score of the strongest signal in the clump, and $r_{i,hit}$ is the correlation coefficient between the strongest signal and the SNP being tested. If P($Z_i' > 0$) $<$ 5 x 10^{-8} then this clump is considered to have a secondary signal, and the SNP with the Z_i' largest in magnitude is recorded as a secondary signal in this clump. All other SNPs in the clump are then tested for a tertiary signal independent of the first two, using

$$Z_i' = Z_i - r_{i,hit} Z_{hit} - r_{i,2nd} Z_{2nd} \tag{4.2}$$

We do not test for additional signals after the third. Theoretically, this could be extended to an arbitrary number of signals, but the approximation will become less accurate as additional signals are tested for.

This approach yielded 193 genome-wide significant independent signals of association. None of these signals had significant heterogeneity of effect size, and all had their Immunochip intensity cluster plots manually inspected to ensure that they were well clustered.

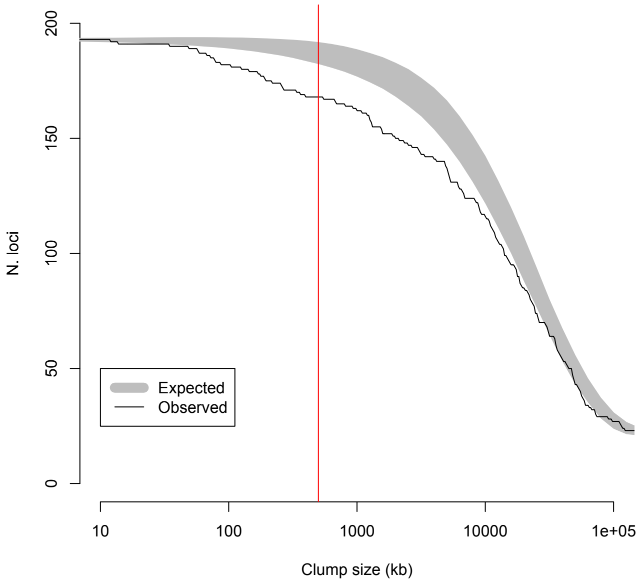


Figure 4.8: The results of a null simulation of association clumping. The x-axis shows varying thresholds of proximity for two statistically independent signals to be considered in the same locus. The y-axis shows the number of loci for a particular threshold, from 193 (the total number of independent signals) at the left when no signals are combined to fewer than 50 when even extremely distant signals 100Mb apart are combined. The grey shaded area shows the 95% confidence interval from simulations of 193 random signals, and the black line shows the true number of loci for a given clumping value. The red line is 500kb, the actual clumping distance we used.

4.3.6 Combining signals into loci

The large number of independent signals (193) makes categorising them into functionally separate loci problematic. We conventionally define signals as coming from the same locus if their lead SNPs lie within a certain physical or genetic distance of each other. However if this physical distance parameter is too large functionally independent signals that are adjacent by chance may be incorrectly combined. Conversely, selecting too small a distance parameter could cause variants that act relatively proximately on the same gene to be split into independent loci.

To test the effect of this distance parameter on classifying signals into loci, I performed a null simulation. I selected randomly from the PCA SNPs to simulate null signals, and examined what proportion of signals are incorrectly merged together for a given distance parameter value. Based on this, I decided to define a locus as a 500kb unit: 250kb on either side of the hit SNP. This results in between 95% and 99% of null loci being correctly separated (Figure 4.8).

Each independent signal had a region defined around it, which was 250kb on either side of the hit SNP, or the extent of LD (defined as the positions of the furthest up-and-downstream variants with $r^2 > 0.5$ to the hit SNP). Overlapping regions were merged together, providing that they were associated to compatible phenotypes under the likelihood analysis (see below); i.e. loci were not merged if one was uniquely associated with CD, and the other uniquely associated with UC. The final merged regions were defined as loci, with their extents being the maximum extent of their component signals. A total of 163 independent loci were thus defined (Table 4.9).

4.3.7 Crohn's disease/Ulcerative colitis likelihood modelling

We used a likelihood modelling approach to classify signals into four categories according to their relative strength of association to CD and UC. We used a multinomial logistic regression model with additive log-odds ratio parameters β_{CD} and β_{UC}. The model was fitted to the Immunochip genotypes using the `mlogit` package in R.

We fit this model with four sets of parameter constraints:

1. CD-specific model: $\beta_{UC} = 0$ (i.e. UC cases and controls have the same frequency), β_{CD} fitted by maximum likelihood

2. UC-specific model: $\beta_{CD} = 0$, β_{UC} fitted by maximum likelihood

3. IBD unsaturated (same-effect size) model: $\beta_{CD} = \beta_{UC} = \beta_{IBD}$ (i.e. frequency is the same in CD and UC cases), β_{IBD} fitted by maximum likelihood

4. IBD saturated (different effect sizes) model: β_{CD} and β_{UC} both fitted by maximum likelihood

Note that models 1-3 are all constrained versions (1 d.f.) of model 4 (2 d.f.).

We calculated likelihoods for each model, and performed a likelihood ratio test of each of models 1-3 against model 4. If the likelihood ratio test had p < 0.05 for all 3 models (the 2 d.f. model is nominally significantly a better fit than any of the 1 d.f. models), we classified the signal as "saturated" (i.e. associated to both CD and UC, but with evidence of different effect sizes). Otherwise, we classified the signal according to which of the first three models had the largest likelihood. Note that being classified as IBD unsaturated should be interpreted as "associated to both CD and UC, without significance evidence of differing effect sizes".

In Table 4.9 below, the "IBD" section contains all loci where the main signal was classified as IBD unsaturated or IBD saturated. An exception was made for the CD associations at *PTPN22* and *NOD2*, where the correct model was "IBD saturated", as there were significant UC associations that went in the opposite direction to the CD effect.

Even within these classifications there is a significant variation in the balance of CD and UC effect sizes (Figure 4.9). To capture this we also used polar-transformed log odds ratios as a continuous measure of CD vs UC effect size balance. This is defined as θ = atan2(log(ORCD), log(ORUC)). Large values of θ correspond to associations with a stronger UC component, smaller values correspond to a stronger CD component.

4.3.8 Comparison of this locus list to previous CD and UC lists

Because this study has access to raw genotype data from both CD and UC for the first time, it has allowed us to clarify several aspects of the 99 previously reported associations:

- While previously suspected, we have confirmed that the associations in the MHC are distinct for CD and UC, and therefore should be split into two phenotype specific associations, rather than a single IBD locus.
- Conversely our improved imputation has re-localised the CD association previously reported as *VAMP3* to be the same effect as the adjacent previous UC association to *TNFRSF9*, making this a single IBD locus.
- Two previously independent associations on chromosome 2 near 102Mb (one CD, one UC) have both been shown to be IBD, and accordingly have been merged into independent effects in a single IBD locus. Similarly, a previous CD SNP (chromosome 2 near 198Mb), which is now associated to UC as well, was incorporated into a new nearby UC locus.
- Five previous associations (Chr2@198Mb, Chr5@36Mb, Chr6@3Mb, Chr6@44Mb, Chr13@42Mb) are no longer genome-wide significant. In four cases, our improved PCA-corrected analysis is >2 orders of magnitude less significant than the previous country-stratified analysis, suggesting that these associations may have been driven in part by uncor-

rected population structure. In the final instance the key SNP failed Immunochip design.

Thus, from 99 previously reported loci, one was split, three were merged and five were lost, leaving 92 established and 71 novel loci. This highlights both the overall robustness of our previous analyses as well as potential pitfalls in small-scale replication genotyping, for which correction for population stratification is difficult.

We also compared the total phenotypic variance of CD and UC explained by our loci compared to previously published estimates. In ulcerative colitis we improved from 3.9% of phenotypic variance explained by known loci to 7.0% explained by our 193 signals. For Crohn's disease we improved from 7.6% to 12.0%. Two additional comments are necessary: first, I have decided here to report phenotypic variance explained, rather than heritability, due to the difficulties in measuring narrow-sense heritability discussed in Chapter 2. Second, the odds ratios estimated from the Immunochip are smaller than previous estimates for several key loci in CD, including *NOD2*, *IL23R* and *ATG16L1*. This difference was not explained by an abnormal degree of stratification or differential ancestry at these sites. Our new odds ratios are estimated in replication samples in this project, so this effect may reflect less severe disease than the samples previously collected for GWAS.

Table 4.9: The 163 inflammatory bowel disease loci, split into Crohn's Disease specific, Ulcerative Colitis specific, and shared across Inflammatory Bowel Disease. Key genes are those identified by one of the candidate gene prioritisation analyses described in the text, and bold genes are identified by more than one bioinformatic approach. Loci shaded grey are newly identified in this study. SNP IDs marked * denote the presence of a second genome-wide significant alternative signal at this locus, and ** denotes the presence of two or more additional signals. Odds ratios marked with a †show evidence of heterogeneity of effect size between CD and UC.

Chrom:Pos (Mb)	SNP	P-value	RAF	OR	Key Genes (+N additional in locus)
Crohn's Disease					
1:78.37-78.87	rs17391694	2.96 x 10^{-9}	0.889	1.134	(5)
1:114.05-114.55	rs6679677	2.03 x 10^{-15}	0.907	1.196†	*PTPN22*, *DCLRE1B*, (7)
1:120.2-120.7	rs3897478	1.97 x 10^{-11}	0.891	1.161	*ADAM30*, (5)
1:172.6-173.1	rs9286879	5.53 x 10^{-22}	0.249	1.125	***FASLG***, *TNFSF18*, (0)
2:27.38-27.88	rs1728918	4.86 x 10^{-16}	0.299	1.123	*UCN*, (23)
2:62.3-62.8	rs10865331	9.77 x 10^{-10}	0.396	1.098	(3)
2:230.84-231.34	rs6716753	1.17 x 10^{-16}	0.196	1.134	*SP140*, (5)
2:233.87-234.42	rs12994997	4.14 x 10^{-70}	0.523	1.233	***ATG16L1***, *INPP5D*, (7)
4:48.11-48.61	rs6837335	1.75 x 10^{-8}	0.647	1.086	*TXK*, *TEC*, *SLC10A4*, (3)
4:102.61-103.11	rs13126505	1.84 x 10^{-12}	0.096	1.172	(1)
5:55.18-55.68	rs10065637	3.68 x 10^{-12}	0.773	1.123	***IL6ST***, *IL31RA*, (1)
5:72.29-72.79	rs7702331	5.63 x 10^{-10}	0.621	1.088	(4)

5:173.09-173.59	rs17695092	4.68 x 10^{-9}	0.703	1.095	*CPEB4*, (2)
6:21.17-21.67	rs12663356	4.01 x 10^{-12}	0.533	1.095	(3)
6:31.02-31.52	rs9264942	4.96 x 10^{-28}	0.378	1.145	*HLA-C*, *PSORS1C1*, *NFKBIL1*, (19)
6:127.2-127.7	rs9491697	3.79 x 10^{-10}	0.439	1.077	(3)
6:127.99-128.49	rs13204742	8.38 x 10^{-15}	0.124	1.173	(2)
6:159.24-159.74	rs212388	3.04 x 10^{-14}	0.410	1.105	*TAGAP*, (5)
7:26.63-27.13	rs10486483	3.48 x 10^{-8}	0.247	1.089	(2)
7:27.92-28.42	rs864745	3.65 x 10^{-9}	0.497	1.087	*CREB5*, *JAZF1*, (1)
8:90.62-91.12	rs7015630	1.42 x 10^{-8}	0.739	1.075	*RIPK2*, (4)
8:129.31-129.81	rs6651252	1.45 x 10^{-16}	0.865	1.185	(0)
13:44.2-44.7	rs3764147	2.19 x 10^{-21}	0.248	1.155	*LACC1*, (3)
15:38.64-39.14	rs16967103	3.88 x 10^{-9}	0.203	1.088	***RASGRP1***, *SPRED1*, (2)
16:50.31-51	rs2066847**	5.86 x 10^{-209}	0.024	3.103†	***NOD2***, *ADCY7*, (5)
17:25.59-26.09	rs2945412	8.68 x 10^{-17}	0.587	1.137	*LGALS9*, *NOS2*, (3)
19:0.87-1.37	rs2024092	8.26 x 10^{-22}	0.215	1.156	*GPX4*, *HMHA1*, (20)
19:46.6-47.1	rs4802307	2 x 10^{-10}	0.706	1.099	(9)
19:48.95-49.45	rs516246	1 x 10^{-15}	0.483	1.107	*DBP*, *SPHK2*, *IZUMO1*, *FUT2*, (22)
21:34.52-35.02	rs2284553	2.14 x 10^{-16}	0.599	1.123	***IFNGR2***, ***IFNAR1***, *IFNAR2*, *IL10RB*, (9)

Ulcerative Colitis					
1:2.25-2.75	rs10797432	2.62 x 10^{-12}	0.522	1.078	***TNFRSF14***, *MMEL1*, *PLCH2*, (8)
1:19.88-20.42	rs6426833**	2.39 x 10^{-68}	0.542	1.265	(9)
1:199.84-200.34	rs2816958	1.98 x 10^{-17}	0.887	1.23	(3)
2:198.18-199.12	rs1016883	2.87 x 10^{-8}	0.817	1.1	*RFTN2*, *PLCL1*, (7)
2:199.27-200.12	rs17229285*	1.73 x 10^{-13}	0.496	1.117	(0)
3:52.8-53.3	rs9847710	1.05 x 10^{-8}	0.416	1.064	*PRKCD*, *ITIH4*, (8)
4:103.26-103.76	rs3774959	3.66 x 10^{-12}	0.358	1.118	***NFKB1***, *MANBA*, (2)
5:0.34-0.84	rs11739663	1.81 x 10^{-8}	0.760	1.071	*SLC9A3*, (8)
5:134.19-134.69	rs254560	2.55 x 10^{-9}	0.397	1.056	(6)
6:32.33-32.86	rs6927022	4.71 x 10^{-133}	0.535	1.444	***HLA-DQB1***, *-DRB1*, -DQA1 (13)
7:2.53-3.03	rs798502	6.09 x 10^{-17}	0.709	1.127	***CARD11***, *GNA12*, *TTYH3*, (4)
7:26.97-27.47	rs4722672	2.06 x 10^{-8}	0.183	1.091	(14)
7:107.18-107.72	rs4380874*	2.07 x 10^{-26}	0.405	1.137	*DLD*, (9)
7:128.32-128.82	rs4728142	4.37 x 10^{-14}	0.444	1.104	***IRF5***, *TNPO3*, *TSPAN33*, (11)
11:95.77-96.27	rs483905	1.21 x 10^{-8}	0.292	1.056	*JRKL*, *MAML2*, (2)

11:114.13-114.63	rs561722	5.15 x 10^{-17}	0.663	1.12	*FAM55A*, *FAM55D*, (5)
15:41.29-41.81	rs28374715	2.43 x 10^{-8}	0.738	1.082	*ITPKA*, *NDUFAF1*, *NUSAP1* ,(8)
16:30.22-30.72	rs11150589	6.04 x 10^{-10}	0.463	1.09	***ITGAL***, (20)
16:68.33-68.83	rs1728785	3.71 x 10^{-8}	0.767	1.075	*ZFP90*, (6)
17:70.39-70.89	rs7210086	1.89 x 10^{-9}	0.797	1.111	(3)
19:46.87-47.37	rs1126510	1.55 x 10^{-9}	0.363	1.075	*CALM3*, (14)
20:33.55-34.05	rs6088765	2.21 x 10^{-8}	0.437	1.079	*PROCR*, *UQCC*, *CEP250*, (8)
20:42.81-43.31	rs6017342	1.43 x 10^{-43}	0.530	1.228	*ADA*, *HNF4A*, (9)
Inflammatory Bowel Disease					
1:0.99-1.49	rs12103	7.66 x 10^{-13}	0.182	1.099	*TNFRSF18*, *TNFRSF4*, (30)
1:7.77-8.27	rs35675666	1.12 x 10^{-15}	0.838	1.112	*TNFRSF9*, (6)
1:22.45-22.95	rs12568930	1.26 x 10^{-17}	0.821	1.095†	(3)
1:67.4-67.95	rs11209026**	8.12 x 10^{-161}	0.933	2.013†	***IL23R***, *IL12RB2*, (4)
1:70.74-71.24	rs2651244	2.29 x 10^{-8}	0.599	1.015†	(3)
1:151.54-152.04	rs4845604	3.52 x 10^{-16}	0.857	1.144†	*RORC*,(14)
1:155.22-156.12	rs670523	5.79 x 10^{-11}	0.324	1.06†	*UBQLN4*, *IT1*, *STO1*,(28)
1:160.6-161.1	rs4656958	6.8 x 10^{-9}	0.686	1.061	***CD48***, *SLAMF1*, *ITLN1*, *CD244*, (12)

1:161.22-161.72	rs1801274	2.12 x 10^{-38}	0.509	1.124†	***FCGR2A***, ***FCGR2B***, ***FCGR3A***, **HSPA6** (11)
1:197.33-197.87	rs2488389	8.45 x 10^{-13}	0.220	1.115	*C1orf53*,(2)
1:200.62-201.12	rs7554511	1.24 x 10^{-32}	0.725	1.164	*KIF21B*,(6)
1:206.68-207.18	rs3024505	6.66 x 10^{-42}	0.160	1.208†	***IL10***, ***IL20***, ***IL19***, ***IL24***, (7)
2:24.87-25.37	rs6545800	6.14 x 10^{-16}	0.445	1.109†	*ADCY3*, (6)
2:28.36-28.86	rs925255	2.67 x 10^{-15}	0.557	1.092†	*FOSL2*, *BRE*, (1)
2:43.56-44.06	rs10495903	8.03 x 10^{-12}	0.130	1.086†	(5)
2:60.95-61.45	rs7608910	8.65 x 10^{-32}	0.394	1.138	***REL***, *C2orf74*, *KIAA1841*, *AHSA2*, (6)
2:65.42-65.92	rs6740462	2.35 x 10^{-8}	0.739	1.081	*SPRED2*, (1)
2:102.41-103.31	rs917997*	3.12 x 10^{-20}	0.231	1.103†	*IL1R2*, *IL18RAP*, *IL18R1*, *IL1R1*, (5)
2:162.85-163.35	rs2111485	1.93 x 10^{-8}	0.404	1.066	*IFIH1*, (5)
2:191.67-192.17	rs1517352	3.28 x 10^{-11}	0.600	1.077	***STAT1***, *STAT4*, (2)
2:218.89-219.39	rs2382817	3.7 x 10^{-12}	0.408	1.073	***SLC11A1***, ***CXCR2***, ***CXCR1***, *PNKD*, (11)
2:241.31-241.83	rs3749171*	3.07 x 10^{-21}	0.167	1.135†	***GPR35***, (12)
3:18.51-19.01	rs4256159	9 x 10^{-15}	0.140	1.107†	(0)

3:47.96-49.96	rs3197999**	1.01 x 10^{-47}	0.296	1.18	***MST1***, ***PFKFB4***, *MST1R*, *UCN2*, (61)
4:74.6-75.1	rs2472649	2.57 x 10^{-8}	0.824	1.095†	***CXCL5***, ***CXCL1***, ***CXCL3***, *IL8*, (7)
4:122.91-123.53	rs7657746	2.76 x 10^{-13}	0.753	1.116	***IL2***, *IL21*, (2)
5:10.44-10.94	rs2930047	1.03 x 10^{-8}	0.382	1.065	***DAP***, (2)
5:40.02-40.74	rs11742570**	1.81 x 10^{-82}	0.605	1.198†	*PTGER4*, (1)
5:95.99-96.49	rs1363907	5.62 x 10^{-13}	0.411	1.068	*ERAP2*, *ERAP1*, *LNPEP*, (2)
5:129.75-130.26	rs4836519	4.24 x 10^{-10}	0.768	1.072†	(1)
5:130.36-132.01	rs2188962*	1.35 x 10^{-52}	0.425	1.158†	***IRF1***, ***IL13***, ***CSF2***, ***SLC22A4***, (14)
5:141.26-141.76	rs6863411	3.59 x 10^{-14}	0.630	1.089†	*SPRY4*, *NDFIP1*, (5)
5:150.02-150.52	rs11741861	2.94 x 10^{-37}	0.093	1.249†	*TNIP1*, *IRGM*, *ZNF300P1*, (8)
5:158.53-159.07	rs6871626**	1.43 x 10^{-42}	0.337	1.181†	***IL12B***, (3)
5:176.54-177.04	rs12654812	1.68 x 10^{-8}	0.335	1.068	***DOK3***, (17)
6:14.46-14.96	rs17119	3.08 x 10^{-11}	0.786	1.071	(0)
6:20.47-21.06	rs9358372*	8.66 x 10^{-14}	0.379	1.089†	(2)
6:90.71-91.21	rs1847472	1.57 x 10^{-10}	0.655	1.06	(1)
6:106.18-106.68	rs6568421	8.24 x 10^{-20}	0.301	1.108†	(2)

6:111.55-112.09	rs3851228	1.08 x 10^{-13}	0.073	1.153	***TRAF3IP2***, *FYN*, *REV3L*, (2)
6:137.75-138.25	rs6920220	1.4 x 10^{-21}	0.206	1.102†	*TNFAIP3* ,(1)
6:143.65-144.15	rs12199775	1.99 x 10^{-8}	0.929	1.129	*PHACTR2*, (5)
6:167.12-167.62	rs1819333	6.76 x 10^{-21}	0.523	1.081†	***CCR6***, ***RPS6KA2***, *RNASET2*, (3)
7:49.94-50.55	rs1456896*	7.28 x 10^{-15}	0.688	1.088	*ZPBP*, *IKZF1*, (4)
7:98.5-99	rs9297145	8.21 x 10^{-12}	0.265	1.082	*SMURF1*, (6)
7:100.06-100.61	rs1734907	1.67 x 10^{-13}	0.149	1.114†	***EPO***, (21)
7:116.64-117.14	rs38904	1.31 x 10^{-8}	0.532	1.054†	(6)
8:126.28-126.78	rs921720	8.3 x 10^{-20}	0.609	1.081†	*TRIB1*, (1)
8:130.37-130.87	rs1991866	1.65 x 10^{-9}	0.422	1.054	(2)
9:4.73-5.23	rs10758669	7.88 x 10^{-45}	0.349	1.174	***JAK2***, (4)
9:93.67-94.17	rs4743820	3.6 x 10^{-9}	0.702	1.056†	*NFIL3*, (2)
9:117.3-117.89	rs4246905**	2.8 x 10^{-32}	0.709	1.142	*TNFSF8*, *TNFSF15*, *TNC*, (2)
9:138.99-139.64	rs10781499*	4.38 x 10^{-56}	0.412	1.188†	***CARD9***, *PMPCA*, *SDCCAG3*, (19)
10:5.83-6.33	rs12722515	3.76 x 10^{-10}	0.849	1.102†	***IL2RA***, ***IL15RA***, (6)
10:30.47-30.97	rs1042058	5.93 x 10^{-11}	0.592	1.075†	***MAP3K8***, (3)
10:35.04-35.55	rs11010067	2.49 x 10^{-25}	0.346	1.115†	***CREM***, (3)
10:59.74-60.24	rs2790216	8.07 x 10^{-9}	0.778	1.066	*CISD1*, *IPMK*, (2)
10:64.12-64.89	rs10761659**	6.37 x 10^{-46}	0.543	1.166†	(3)
10:75.42-75.92	rs2227564	6.75 x 10^{-10}	0.770	1.082†	(13)

10:80.78-81.28	rs1250546	3.15 x 10^{-18}	0.604	1.096†	(5)
10:82-82.5	rs6586030	9.24 x 10^{-16}	0.847	1.115†	*TSPAN14*, *10orf58*, (4)
10:94.18-94.68	rs7911264	2.98 x 10^{-8}	0.519	1.066	(4)
10:101.03-101.53	rs4409764	1.03 x 10^{-54}	0.491	1.182	*NKX2-3*, (6)
11:1.62-2.12	rs907611	2.7 x 10^{-10}	0.315	1.068	***TNNI2***, *LSP1*, (17)
11:58.08-58.58	rs10896794	6.8 x 10^{-10}	0.762	1.08	*CNTF*, *LPXN*, (8)
11:60.52-61.02	rs11230563	9.03 x 10^{-13}	0.654	1.085	*CD6*, *CD5*, *PTGDR2*, (12)
11:61.31-61.81	rs4246215	1.93 x 10^{-15}	0.338	1.079†	*C11orf9*, *FADS1*, *FADS2*, (12)
11:63.85-64.39	rs559928	4.19 x 10^{-11}	0.821	1.101	***CCDC88B***, *RPS6KA4*,(20)
11:65.4-65.9	rs2231884	2.91 x 10^{-10}	0.157	1.083†	*RELA*, *FOSL1*, *CTSW*, *SNX32*, (22)
11:76.04-76.54	rs2155219	4.24 x 10^{-36}	0.509	1.151†	(5)
11:86.87-87.37	rs6592362	2.32 x 10^{-8}	0.248	1.083	(1)
11:118.49-118.99	rs630923	7.07 x 10^{-9}	0.846	1.074†	*CXCR5*, (17)
12:12.4-12.9	rs11612508	1.06 x 10^{-8}	0.267	1.058†	*LOH12CR1*, (8)
12:40.5-41.03	rs11564258*	6.38 x 10^{-29}	0.025	1.334†	*MUC19*, (1)
12:47.95-48.45	rs11168249	7.78 x 10^{-9}	0.467	1.054†	*VDR*, (8)
12:68.24-68.74	rs7134599	8.51 x 10^{-32}	0.378	1.096†	***IFNG***, *IL26*, *IL22*, (1)
13:27.27-27.77	rs17085007	2.79 x 10^{-19}	0.183	1.106†	(2)
13:40.45-41.26	rs941823**	2.07 x 10^{-14}	0.758	1.071†	(3)
13:99.7-100.2	rs9557195	2.37 x 10^{-14}	0.772	1.112	*GPR183*, *GPR18*,(6)

14:69.02-69.52	rs194749	2.7 x 10^{-10}	0.226	1.075†	*ZFP36L1*, (4)
14:75.45-75.95	rs4899554	2.71 x 10^{-8}	0.819	1.083†	***FOS***, *MLH3*, (6)
14:88.22-88.72	rs8005161	2.35 x 10^{-14}	0.089	1.153	***GPR65***, *GALC*, (1)
15:67.18-67.68	rs17293632	5.97 x 10^{-16}	0.235	1.067†	*SMAD3*, (2)
15:90.92-91.42	rs7495132	9.48 x 10^{-11}	0.891	1.134	*CRTC3*, (3)
16:11.12-11.95	rs529866*	1.73 x 10^{-16}	0.803	1.124†	***SOCS1***, ***LITAF***, *RMI2*, (10)
16:23.61-24.11	rs7404095	9.68 x 10^{-10}	0.572	1.06	***PRKCB***, (5)
16:28.26-28.93	rs26528	9.65 x 10^{-22}	0.451	1.099†	*RABEP2*, *IL27*, *EIF3C*, *SULT1A1*, (11)
16:85.75-86.25	rs10521318	1.41 x 10^{-9}	0.915	1.155†	*IRF8*, (4)
17:32.34-32.84	rs3091316	1.22 x 10^{-26}	0.722	1.122†	***CCL13***, ***CCL2***, *CCL11*, (4)
17:37.66-38.16	rs12946510	4.1 x 10^{-38}	0.465	1.157	*IKZF3*, *ZPBP2*, *GSDMB*, *ORMDL3*, (13)
17:40.28-40.78	rs12942547	5.51 x 10^{-22}	0.580	1.103†	***STAT3***, *STAT5B*, *STAT5A*, (13)
17:57.71-58.21	rs1292053	8.85 x 10^{-13}	0.446	1.076†	*TUBD1*, *RPS6KB1*, (9)
18:12.55-13.05	rs1893217	3.05 x 10^{-26}	0.157	1.171†	(6)
18:46.14-46.64	rs7240004	1.31 x 10^{-9}	0.616	1.057†	*SMAD7*, (2)
18:67.28-67.78	rs727088	4.65 x 10^{-9}	0.484	1.077	***CD226***, (2)
19:10.22-10.76	rs11879191*	2.04 x 10^{-18}	0.797	1.136	***TYK2***, *PPAN-P2RY11*, *ICAM1*, (25)
19:33.48-33.98	rs17694108	5.85 x 10^{-15}	0.282	1.1	*CEBPG*, (8)

19:55.13-55.63	rs11672983	6.5 x 10^{-11}	0.392	1.087	*NLRP7*, *NLRP2*, *KIR2DL1*, *LILRB4*, (15)
20:30.47-31.03	rs6142618	6.05 x 10^{-10}	0.564	1.072†	*HCK*, (10)
20:31.12-31.62	rs4911259	1.2 x 10^{-9}	0.383	1.075	*DNMT3B*, (8)
20:44.49-44.99	rs1569723	9.95 x 10^{-14}	0.259	1.091†	***CD40***, *MMP9*, *PLTP*, (11)
20:48.7-49.2	rs913678	4.59 x 10^{-8}	0.662	1.056	*CEBPB*, (5)
20:57.57-58.07	rs259964	1.01 x 10^{-12}	0.464	1.085	*ZNF831*, *CTSZ*, (5)
20:62.09-62.59	rs6062504	1.09 x 10^{-23}	0.684	1.104	*TNFRSF6B*, *LIME1*, *SLC2A4RG*, (24)
21:16.56-17.06	rs2823286	9.28 x 10^{-30}	0.708	1.157†	(0)
21:40.21-40.71	rs2836878	4.62 x 10^{-48}	0.733	1.18†	(3)
21:45.37-45.87	rs7282490	2.35 x 10^{-26}	0.391	1.105	*ICOSLG*, (9)
22:21.67-22.17	rs2266959	1.39 x 10^{-16}	0.186	1.105	*MAPK1*, *YDJC*, *UBE2L3*, *RIMBP3*, (9)
22:30.12-30.73	rs2412970	2.7 x 10^{-14}	0.457	1.08	***LIF***, ***OSM***, *MTMR3*, (8)
22:39.4-39.97	rs2413583*	4.4 x 10^{-33}	0.833	1.209†	*ATF4*, TAB1, APOBEC3G, (16)

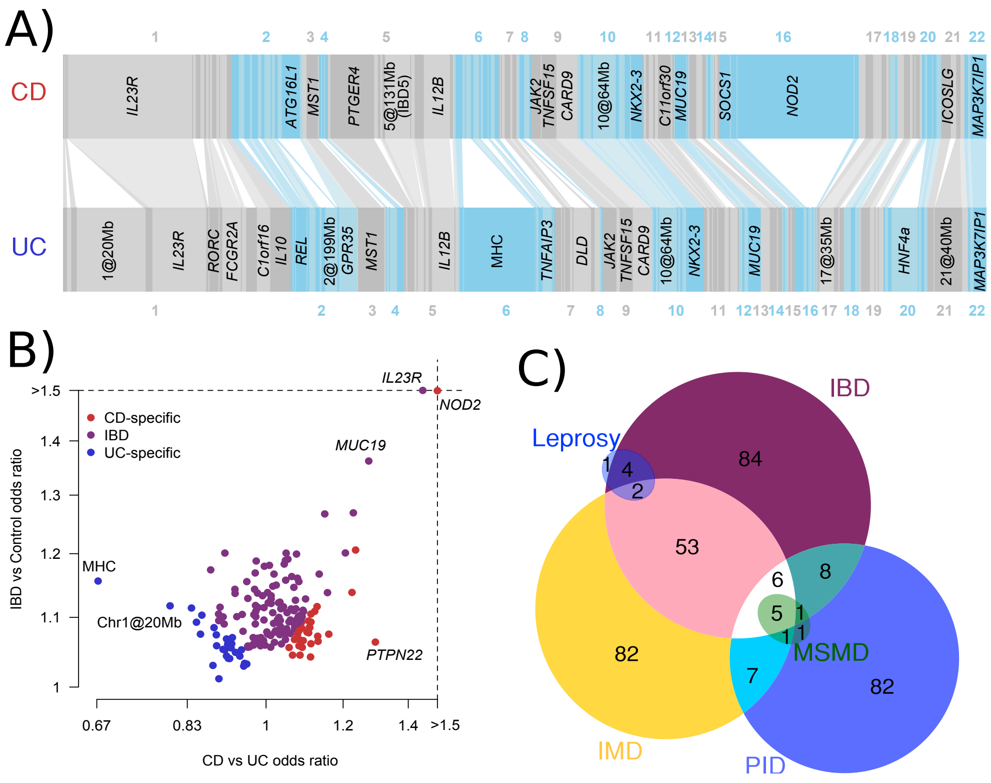


Figure 4.9: The IBD genome. A) The 163 IBD loci identified in this study. Each bar, ordered by genomic position, represents an independent locus, and the width of the bar is proportional to the variance explained by that locus in CD and UC. Bars are connected together if they are identified as being associated with both phenotypes. Loci are labelled if they explain more than 1% of the total variance explained by all loci for that phenotype. B) The 193 independent signals, plotted by total IBD odds ratio and phenotype specificity (measured by the odds ratio of CD relative to UC), and coloured by their IBD phenotype classification from Table 1. C) Number of overlapping IBD loci with other immune-mediated diseases (IMD), leprosy, and Mendelian primary immunodeficiencies (PID). Within PID, we highlight Mendelian susceptibility to mycobacterial disease (MSMD).

4.4 Biological and bioinformatic interpretation of 163 IBD loci

Our meta-analysis of the GWAS and Immunochip data identified 193 statistically independent signals of association at genome-wide significance ($P < 5 \times 10^{-8}$) in at least one of the three phenotypes (CD, UC, IBD). These signals were merged into 163 regions, of which 71 have never been reported before (Table 4.9). This is more loci than has ever been recorded for a complex disease, and the number of loci, and the large number of genes they contain, make a locus-by-locus interpretation difficult. To go from a list of regions to a set of specific biological hypotheses we have to use computational techniques, and make use of external datasets.

In this section I will discuss a number of ways in which this can be achieved, starting with a brief overview of the IBD loci. I will go on to use genetic data from other disease loci (both complex and Mendelian) to place IBD genetics in the context of other immune diseases. Next I will describe a range of methods for identifying candidate causal genes within the identified loci using gene networks and functional information. I will then describe a detailed analysis of the identified candidate genes in terms of Gene Ontology (GO) terms and canonical pathways, followed by an analysis of the IBD loci in the context of natural selection. Finally, I will describe a number of functional analyses based on gene expression data carried out by other members of the IIBDGC Immunochip analysis group.

4.4.1 Global patterns in the "IBD genome"

A traditional Manhattan plot of this study does not provide much information, due to the large number of peaks and high variation in significance between them. Instead, I have developed an alternative visualisation, which I call the "Belgravia plot" (by analogy with the flat, regular Regency terraces in Belgravia, London). This plot (Figure 4.9A) shows the relative contribution of each locus to the total variance explained in UC and CD using width, rather than height. This gives us an intuitive overview of the importance of the global structure of IBD. For instance, CD is more dominated by the two major loci (*NOD2* and *IL23R*), with UC having a more even distribution.

The likelihood-based model selection analysis described in Section 4.3.7 gives us information on the global level of genetic sharing between the two IBD phenotypes. The vast majority of loci (110) are associated with both disease phenotypes, of which 62 have an indistinguishable effect size in UC and CD, while 48 show evidence of heterogeneous effects (highlighted in Table 4.9). Of the remaining loci, 30 are classified as specific for CD and 23 for UC, but notably, 43 of these 53 show the same direction of effect in the non-associated disease (overall P = 2.8 x 10^{-6}), suggesting that only a few of the loci may be truly disease specific.

While likelihood-based approaches for the classification of IBD loci are instructive, it should be noted that there is continuous variability in the CD-UC balance of effect sizes among loci (Figure 4.9B). While locus sharing is almost exclusively in the same direction, risk alleles at two CD loci, *PTPN22* and *NOD2*, show significant ($P < 0.005$) protective effects in UC, highlighting them as particularly informative about biological differences between these related diseases.

4.4.2 IBD genetics in the context of autoimmunity and infection

To place the IBD loci in the context of other immune-related diseases, I generated lists of associations with other immune-related disease. I included complex autoimmune and immune-mediated and diseases (IMD), and autosomal dominant or recessive primary immunodeficiencies (PID).

I took autosomal dominant and recessive genes identified as causing PID from Notarangelo et al. (2009). Genes that lie within 250kb of each other were merged together into regions, giving 135 genes across 121 independent

Disease	Locus overlap	Fold-enrichment	Enrichment OR	95% CI	P-value
PS	14	13.5	14.71	8.5-25.5	4.15 x 10^{-12}
AS	8	12.56	13.18	6.5-26.8	3.22 x 10^{-7}
AD[a]	3	12.1	12.32	3.9-38.6	0.002
PBC	13	10.99	11.88	6.7-21.0	3.12 x 10^{-10}
PSC[b]	1	10.93	11.00	1.5-78.6	0.085
RA	12	10.92	11.74	6.5-21.1	1.64 x 10^{-9}
Celiac	16	10.57	11.64	7.0-19.5	4.56 x 10^{-12}
T1D	20	9.99	11.28	7.1-19.0	2.35 x 10^{-14}
SLE	12	9.75	10.47	5.8-18.9	5.87 x 10^{-9}
All AI	66	8.62	13.94	10.2-19.1	5.15 x 10^{-44}
MS	17	8.19	9.06	5.5-15.0	5.11 x 10^{-11}
Asthma	7	7.61	7.91	3.7-16.9	4.90 x 10^{-5}
All PID	20	4.85	5.42	3.4-8.7	8.52 x 10^{-9}

Table 4.10: Enrichment in overlap between IBD loci and loci for other immune-mediated diseases. The enrichment OR is measured on the logistic scale (as described in section 4.4.4). [a]Atopic dermititus. [b]Primary sclerosing cholangitis

regions. I took associated regions for the IMD list from the NHGRI GWAS catalogue, and included the following diseases: Primary sclerosing cholangitis, primary biliary cirrhosis, rheumatoid arthritis, type 1 diabetes, multiple sclerosis, celiac disease, atopic dermatitis, psoriasis, ankylosing spondylitis, asthma and systemic lupus erythematosus. All SNPs in the catalogue with p $< 5 \times 10^{-8}$ were included. As with the IBD loci, I defined a region as 250kb on either side of the hit SNP, and overlapping regions were merged together into loci. This generated a total of 156 independent IMD loci. I assessed overlap between lists (IBD, PID and IMD) using the method described in Section 4.4.4.

A large proportion (113 of 163) of the IBD loci are shared with other complex diseases or traits. Sixty-six of these 113 are among the 154 loci previously associated with other immune-mediated diseases (Hindorff et al., 2009), which is 8.6 times the number that would be expected by chance (Figure 4.9C, $P < 10^{-16}$). Comparing overlaps with specific diseases (Table 4.10) is confounded by the differential power in studies of different diseases. For instance, while type 1 diabetes (T1D) shares the largest number of loci (20/39, 10-fold enrichment), this is partially driven by the large number of known T1D associations. Indeed, seven other immune-mediated diseases show stronger enrichment of overlap, with the largest being ankylosing spondylitis (8/11, 14-fold) and psoriasis (14/17, 13-fold).

In addition to this well-established genetic overlap between IBD and other complex immune mediated diseases, we can now show that IBD loci are also markedly enriched (4.9-fold, $P < 10^{-4}$) in genes involved in primary immunodeficiencies (PIDs, Figure 2C). Consistent with an important role for T-cells in IBD, most of the PIDs overlapping with IBD loci are characterised by reductions in levels of circulating T-cells (*ADA*, *CD40*, *TAP1/2*, *NBS1*, *BLM*, *DNMT3B*), levels of Th17 (*STAT3*), memory T-cells (*SP110*) or regulatory T-cells (*STAT5B*), rather than reduced levels of circulating B-cells (cell count characteristics taken from Notarangelo et al. (2009)).

Compared to the overlap between PID genes and IBD loci, the subset of PIDs leading to Mendelian susceptibility to mycobacterial infection (MSMD) (Notarangelo et al., 2009; Bustamante et al., 2011; Patel et al., 2008) are enriched still further. Of the eight known autosomal genes that increase susceptibility to MSMD, six are located within IBD loci (46-fold enrichment, $P = 1.3 \text{ x } 10^{-6}$), and a seventh, *IFNGR1*, narrowly missed genome-wide significance ($P = 6 \text{ x } 10^{-8}$). A further relationship to MSMD is seen in the new association near the gene *CD40*, which is involved in MSMD induced by mutations in the X chromosome gene *NEMO* (Filipe-Santos et al., 2006). Furthermore, genetic defects in *STAT3* (Holland et al., 2007; Minegishi et al.,

2007) and *CARD9* (Glocker et al., 2009b), also within IBD loci, lead to PIDs involving skin infections with staphylococcus and candidiasis, respectively, further underscoring the importance of host-microbe interactions in IBD.

This mycobacterial disease overlap is not limited to Mendelian susceptibility. We also find IBD associations in 7/8 loci known to be associated with complex susceptibility to leprosy GWAS (Zhang et al., 2011), including 6 cases where the same SNP is implicated (Figure 4.9C).

There appears to be a shared biology underlying these all these overlapping mycobacterium susceptibility loci. All of the MSMD mutations that overlap with IBD cause defects in interferon signalling pathways, which are known to be important in mycobacterium infection (Flynn et al., 1993). Additionally, the six MSMD genes, four of the leprosy genes and *CD40* fit together into a single well-connected subnetwork within the GRAIL and DAPPLE networks described below (Figure 4.10A). This subnetwork also includes many important signalling proteins involved in IBD and bacterial defence, including *IFNG*, *IL10* and *NFKB1*.

4.4.3 Prioritising candidate genes in IBD loci

We used various methods to reduce the 1438 genes in our locus list to a more limited list of candidate variants. We used both gene network analyses, and analyses of SNP function, to implicate candidate genes.

We used GRAIL and DAPPLE (discussed in Section 4.2.3) to prioritise genes based on network connectivity. In both cases, we removed the HLA region (due to its large size), and fixed four well-established IBD genes as causal (*NOD2*, *IL23R*, *ATG16L1* and *PTPN22*). We took any gene with $p < 0.05$ as implicated. We also included genes from the gene expression network discussed in section 4.4.6.

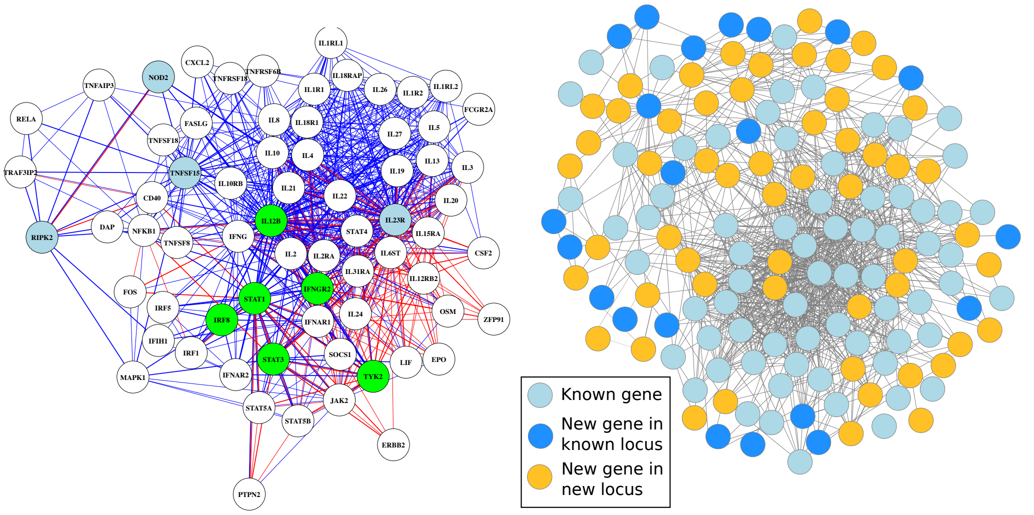


Figure 4.10: a) A combined network graph including GRAIL (blue lines) and DAPPLE (red lines) connections, consisting of all genes connected to MSMD or leprosy genes (highlighted in green and teal respectively) b) The GRAIL network for all genes with GRAIL P < 0.05. Genes included in our previous GRAIL networks in CD and UC are shown in light blue, newly connected genes in previously identified loci in dark blue, and genes from newly associated loci in gold.

Compared to previous analyses that identified candidate genes in 35% of loci (Anderson et al., 2011; Franke et al., 2010) our updated GRAIL-connectivity network identifies candidates in 53% of loci, including increased statistical significance for 58 of the 73 candidates from previous analyses. The new candidates come not only from genes within newly identified loci, but also integrate additional genes from previously established loci (Figure 4.10B). The joint-IBD loci are more likely to contain GRAIL connected genes than CD- or UC-specific loci (P = 0.005), pointing to the shared core of genetic risk between the two diseases.

We also used existing annotations of variant function to search for likely causal mechanisms. We used SeattleSeq to annotate all variants in high LD ($r^2 > 0.8$) with missense or nonsense SNPs, producing 29 IBD associations that caused a protein code change. We also collected eQTL data from a range of studies, including lymphoblastoid cell lines of asthmatic children (Dixon et al., 2007), various tissues from obese patients (Greenawalt et al., 2011), and a collection of eQTL studies from the Chicago eQTL browser. We found evidence that 64 IBD associations altered the expression of at least one gene.

Overall, our network analyses and functional annotations highlighted a total of 300 candidate genes in 125 loci, of which 37 contained a single gene supported by two or more methods.

4.4.4 Testing for enrichment of functional terms within IBD loci

Gene Ontology (GO) terms and canonical pathways are a natural way to ask questions about the function of the genes in the identified IBD loci. We can ask whether there is an enrichment of certain functional terms in IBD loci, as well as whether these functional loci are associated with a particular type of locus (e.g. CD loci). Below I outline a method for performing tests for enrichment and association of functional terms. I then go on to apply this to the IBD data, to find functional terms associated with IBD, as well as identifying terms associated with CD-UC balance, and are more strongly enriched in IBD relative to other immune-mediated diseases. Finally, I use this methodology to investigate potential functional biases introduced by the structure of Immunochip, and by using genes identified by the prioritisation approaches described above.

A methodology for testing functional enrichment in IBD loci

Basic framework

We wish to assess the enrichment of a particular functional term (e.g. a GO term) in causal IBD genes. Given a list of causal genes, we could easily calculate an enrichment odds ratio λ_i of a functional term i in IBD genes relative to the genome as a whole, and perform a statistical test of $\lambda_i = 1$ vs $\lambda_i > 1$. However, we do not know the causal variant for most IBD regions, and most IBD regions contain multiple genes. To compensate for this, we use an extension of the standard odds ratio method that takes into account the presence of non-causal genes.

Assume that we have M loci, designated by $j = (1, ..., M)$ each of which contains N_j genes. For each associated locus j we set an indicator variable δ_{ij} to 1 if the functional term i is present in locus j, and 0 otherwise. We also calculate a genome-wide frequency for term f_i that is equal to the proportion of all genes that contain the term i.

We calculate g_i, the frequency of term i in causal genes, given an enrichment odds ratio λ_i as

$$g_i = \left(1 + \frac{1 - f_i}{\lambda_i f_i}\right)^{-1}. \tag{4.3}$$

We then assume that all other genes have a frequency of the term f_i. Assuming that there is exactly one causal gene in the region, the log likelihood L_i is given by

$$L_i = \sum_j \delta_{ij} log\left(1 - (1 - f_i)^{N_j}(1 - g_i)\right) + \sum_j (1 - \delta_{ij}) log\left((1 - f_i)^{N_j}(1 - g_i)\right). \tag{4.4}$$

We fit the parameter λ_i by maximum likelihood using the Nelder-Mead optimisation method, implemented in the statistical package R. We assess the significance of the parameter λ_i by performing a likelihood ratio test of $\lambda_i = 1$ vs $\lambda_i \neq 1$.

Extension to arbitrary predictors

We can extend the method above to include arbitrary per-locus predictors $X = \{x_{jk}\}$ that correlate with level of enrichment of a function term. We can extend the definition of g_i to take the form of a generalised logistic model

$$g_i = \left(1 + \frac{1-f_i}{f_i} exp(-\beta_0 - \vec{\beta}X)\right)^{-1}. \tag{4.5}$$

We keep the enrichment odds ratio (in this case as $\lambda_i = exp(\beta_0)$), but also include regression coefficients for the other predictors $\vec{\beta}$. The predictors X can be discrete (e.g. $x_{jk} = 0$ if locus j is a UC locus, and $x_{jk} = 1$ if it is a CD locus), or continuous (e.g. $x_{jk} = \theta_j$, where θ_j is the polar-transformed odds ratio described in section 4.3.7). The model is fitted by maximum likelihood in the same way as the simple enrichment model, and likelihood ratio tests for $\beta_k = 0$ can be used to assess the significances of the parameters.

Extension to interval overlap

We can extend the above methodology to look for an enrichment in overlap between a set of genomic intervals (e.g. a set of wide linkage peaks) and our IBD loci. We assume that we have a set of genomic intervals $k = 1.,,.R$, each of length l_k. We will also assume that the length of each locus is l_j and the length of the whole genome is l_g. We can thus modify equations 4.3, 4.4 and 4.5 above by setting

$$f_i = \frac{1}{l_g}\sum_k (l_k + l_i). \tag{4.6}$$

It was this extension that enabled me to evaluate the significance of overlap between our IBD loci and GWAS associations discussed in section 4.4.2.

Functional term associations in the IBD loci

I tested the 300 genes prioritised in section 4.4.3 for enrichment in 15,526 human GO terms (27/02/2012 release) and 833 canonical pathways (taken from KEGG, Reactome and Biocarta). I identified 286 GO terms and 48 pathways demonstrating significant enrichment in genes contained within IBD loci. The top associations are shown in Table 4.11, though the large number makes interpreting the entire list difficult.

We can use the hierarchical nature of the GO terms to bring some order to these terms. For instance, cytokine production is the most significantly enriched term, but within that four child terms drive this: IFNγ, IL12, TNFα and IL10. These cytokines are all produced by the cells of the innate immune system (including macrophages, dendritic cells and natural killer cells) in response to bacterial stimulation. This immediately suggests that the IBD risk alleles are, as a whole, interfering with the correct response to bacteria by altering the resulting rates of cytokine production.

The second strongest enrichment was in immune system processes (P = 2.6 x 10^{-23}), with activation of T-, B-, and NK-cells being the strongest contributors to this signal (P = 1.6 x 10^{-22}). Strong enrichment was also seen for response to molecules of bacterial origin (P = 9.6 x 10^{-20}), further evidence for a close relationship between IBD risk and bacterial exposure.

We can test whether any of these enriched functional terms show evidence of differential enrichment between CD and UC phenotypes, both by using the

Term	Description	Loci	P-value
GO:0002376	immune system process	69	3.45 x 10^{-26}
GO:0002682	regulation of immune system process	60	2.61 x 10^{-25}
GO:0001817	regulation of cytokine production	39	2.65 x 10^{-24}
GO:0046649	lymphocyte activation	32	1.77 x 10^{-23}
GO:0031347	regulation of defence response	39	4.78 x 10^{-23}
GO:0048518	positive regulation of biological process	90	3.23 x 10^{-22}
GO:0050865	regulation of cell activation	36	1.63 x 10^{-21}
GO:0045321	leukocyte activation	33	1.84 x 10^{-21}
GO:0048522	positive regulation of cellular process	83	9.27 x 10^{-21}
GO:0002237	response to molecule of bacterial origin	28	2.41 x 10^{-20}
GO:0050776	regulation of immune response	46	2.90 x 10^{-20}
GO:0002684	positive regulation of immune system process	45	3.05 x 10^{-20}
GO:0042110	T cell activation	24	1.56 x 10^{-19}
GO:0006955	immune response	51	1.76 x 10^{-19}
GO:0002694	regulation of leukocyte activation	33	3.09 x 10^{-19}
GO:0001775	cell activation	38	3.40 x 10^{-19}
GO:0032496	response to lipopolysaccharide	26	5.36 x 10^{-19}
GO:0051249	regulation of lymphocyte activation	31	8.13 x 10^{-19}
GO:0070663	regulation of leukocyte proliferation	24	8.67 x 10^{-19}
GO:0080134	regulation of response to stress	43	1.55 x 10^{-18}
KO:04630	Jak-STAT signalling pathway	20	4.80 x 10^{-15}
KO:05140	Leishmania infection	16	3.89 x 10^{-14}
KO:04060	Cytokine-cytokine receptor interaction	25	1.66 x 10^{-13}
BI	Th1/Th2 differentiation	10	1.64 x 10^{-12}
BI	NO2-dependent IL12 pathway	7	3.25 x 10^{-10}
RE:690 0	Signalling in immune system	24	3.35 x 10^{-10}
KO:04062	Chemokine signalling pathway	16	1.10 x 10^{-9}
BI	IL12-dependent signalling pathway	7	7.73 x 10^{-9}
KO:05330	Allograft rejection	9	2.34 x 10^{-8}
KO:04660	T-cell receptor signalling pathway	13	2.49 x 10^{-8}

Table 4.11: The top 20 most enriched GO terms, and top 10 canonical pathways, in IBD loci. Terms starting "GO" are Gene Ontology terms, those starting "KO" are KEGG pathways, "RE" are Reactome pathways and "BC" are Biocarta pathways

Term	Description	Direction	p_θ	$p_{CD/UC}$
GO:0007243	intracellular protein kinase cascade	CD	0.0046	0.0005
GO:0051241	negative regulation of multicellular organismal process	UC	0.0796	0.0039
GO:0000165	MAPK cascade	CD	0.0058	0.0086
GO:0002237	response to molecule of bacterial origin	CD	0.0099	0.0140

Table 4.12: Pathways that show evidence of differential enrichment ($p < 0.01$) in CD and UC. The "direction" shows which phenotype has the higher enrichment of this term. p_θ is the evidence of association between functional term and CD-UC balance parameter θ. $p_{CD/UC}$ is evidence of differential enrichment in CD and UC loci (as defined in Table 4.9)

Term	Description	p_{IMD}	p_{PID}	p_{axis}
KO:04350	TGFβ signalling pathway	0.015	0.004	0.001
BI	Erythropoietin signalling pathway	0.03	0.04	0.004

Table 4.13: Pathways that show evidence of enrichment ($p_{axis} < 0.01$) in IBD loci relative to other immune-mediated disease loci. p_{IMD} and p_{PID} is the enrichment p-value relative to complex immune-mediated diseases and Mendelian primary immunodeficiencies respectively, and p_{axis} is the enrichment p-value of IBD relative to both IMD and PID.

phenotype-specific loci defined in Table 4.9, and using the continuous CD-UC balance parameter θ defined in section 4.3.7. Neither analysis produced any results that met Bonferroni-corrected significance, but results that showed nominal ($p < 0.01$) significance are shown in Table 4.12. Perhaps the most interesting is the evidence that CD may have a larger enrichment of terms involved in response to bacterial products, as this reinforces the opposite direction of effect we see at the *NOD2* locus (itself responsible for responding to the bacterial product MDP).

We can perform a similar analysis comparing IBD to the set of immune-mediated complex diseases and primary immunodeficiencies described in sec-

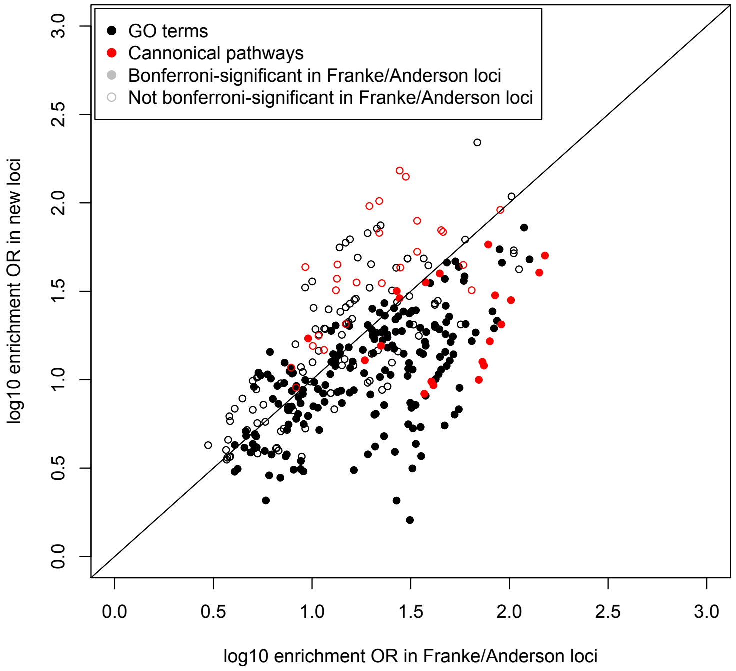


Figure 4.11: GO enrichment in known vs. new loci. The enrichment odds ratios for enriched GO terms are plotted for loci discovered via GWAS and for new loci identified in the current Immunochip analysis. The circled are filled if they were significant in the GWAS loci, and empty if they are only significant when all loci are combined.

tion 4.4.2. Again, no functional term produced a Bonferroni-significant result, but the strongest enrichment was in the TGFβ signalling pathway (Table 4.13). TGFβ is a widely expressed protein that has been implicated in many diseases. However, knock-out mice develop colorectal cancer, potentially suggesting a particular role for TGFβ in the intestinal immune system (Sterner-Kock et al., 2002).

Testing for functional biases in Immunochip genes

The Immunochip was constructed using variant lists submitted by immune-related disease association consortia. We may therefore expect there to be a bias towards discovering loci that are associated to both IBD and other immune-related diseases. This would, in turn, cause an artificial inflation in enrichment of immune-related GO terms. To test this hypothesis, I re-calculated enrichment odds ratios for the 286 enriched GO terms and 48 canonical pathways in two non-overlapping subsets of the 163 loci: (i) the 92 loci described in our previous meta-analyses, and (ii) the 71 newly discovered loci. If our analysis for identifying new IBD loci were biased (via the Immunochip design) toward loci shared across autoimmune diseases we would expect larger enrichment odds ratios in set (ii) compared to (i). Figure 4.11 shows that in fact, the opposite is true: the previous loci are, on average, slightly more strongly enriched than our new loci ($p = 2.2\text{x}10^{-9}$). This difference might suggest that the strongest IBD loci (i.e. those already known) play a more central role in key immune functions than our new discoveries.

This lack of observable bias, while initially surprising, can largely be explained by our experimental design, and the specifics of the SNP selection process for the Immunochip. As part of that design we included the top 2000 most associated SNPs each from the earlier CD and UC GWAS meta-analyses regardless of function or association with other phenotype (corresponding to $p < 0.0009$ for CD and $p < 0.0004$ for UC). This subset of SNPs therefore represents a functionally unbiased, genome-wide replication set that includes 147 (55 new, 92 known) of our 163 reported loci. Therefore the non-IBD part of the Immunochip contributed to only 16 of our loci, of which only 8 are known to be also associated with another immune-mediated disease. This number is too small to strongly bias enrichment analyses, as demonstrated

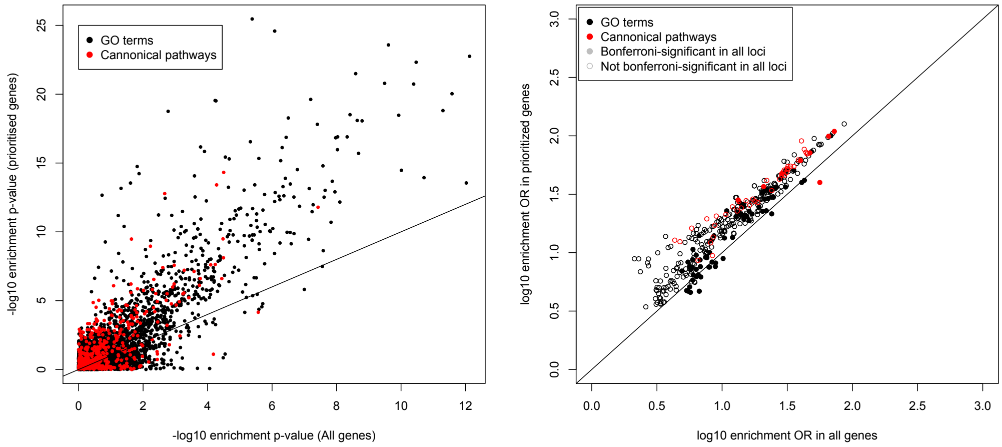


Figure 4.12: Enrichment p-values (a) and odds ratios (b) for GO terms (black dots) and canonical pathways (red dots) calculated on all 1438 genes in IBD loci (x-axis) and just the 300 prioritised genes (y-axis).

above.

Another potential source of bias is the use of the 300 genes selected by our gene prioritisation procedure. There is good reason to use these genes, as doing so grants a large increase in power to detect associations for both GO terms and canonical pathways (Figure 4.12a). However, this procedure is also likely to produce a bias towards the classes of genes and pathways that can be easily detected using gene prioritisation methods. To measure this effect, I calculated enrichment odds ratios for the selected GO term and canonical pathways using just the prioritised genes, and then using the entire set of genes inside the loci. Figure 4.12b shows that the estimated odds ratios are indeed higher when using the prioritised genes, suggesting that this introduces a detectable bias towards the detection of well-studied pathways. However, this bias is relatively small.

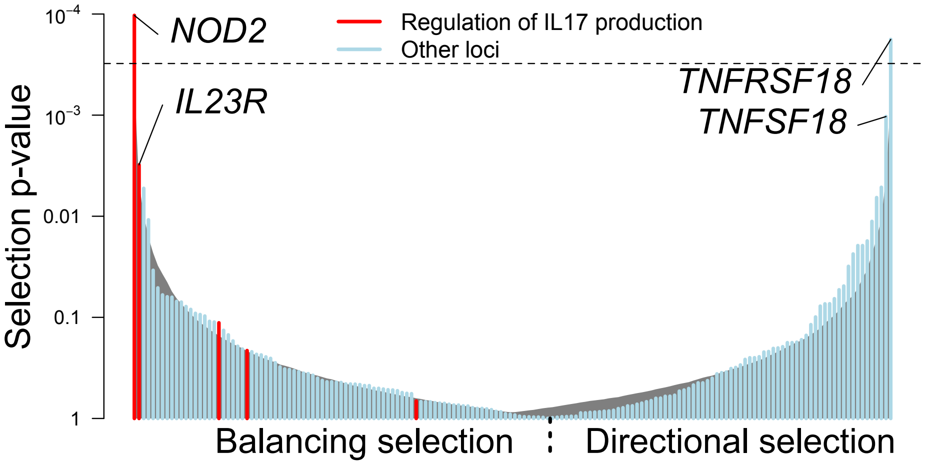


Figure 4.13: Signals of selection at IBD SNPs, from strongest balancing on the left to strongest directional on the right. The grey curve shows the 95% confidence interval for randomly chosen frequency-matched SNPs, illustrating our overall enrichment ($p = 5.5 \times 10^{-6}$), while the dashed line represents the Bonferroni significance threshold. SNPs highlighted in red are annotated as involved in regulation of IL17 production, a key IBD functional term related to bacterial defence, and are enriched for balancing selection.

4.4.5 Natural selection in IBD loci

Infectious organisms are known to be among the strongest agents of natural selection (Lederberg, 1997). It seems logical to ask whether the strong genetic relationship between infection and IBD that emerges from the above analyses also suggests a role for natural selection in the evolutionary history of IBD susceptibility. There are many plausible types of selection that may be acting on IBD risk variants. The risk alleles may be under directional selection, either positive (if the decrease in fitness due to infection outweighs the increase in fitness due to inflammation), and negative (if vice versa). They may also be under balancing selection, indicative of an allele frequency dependent scenario typified by host-microbe co-evolution, as can be observed with parasites (Lederberg, 1997).

To test selection on IBD loci, I used data, provided by Joe Pickrell, generated using the TreeMix method developed by Pickrell and Pritchard (Pickrell and Pritchard, 2012) They constructed population trees from the Human Genetic Diversity Panel (HGDP) data (Li et al., 2008), and produced a per-variant score that measures the extent to which population allele frequencies at that site are over-dispersed relative to this tree. The most over-dispersed sites are likely to have been subjected to directional (positive or negative) selection, whereas those that match the tree most closely are likely to have been subjected to balancing selection.

I picked the best HDGP proxy SNP for each of our associated variants (picking only the UC associated variant from the HLA), and extracted the scores for these variants. Because the score is confounded with allele frequency, I calculated empirical p-values for each variant as follows: pick all variants with an allele frequency within 1 percentage point of the hit variant's allele frequency, and measure the proportion of variants with a score greater than the score of the hit variant. I calculated p-values for directional selection (the proportion of variants with a score higher than the hit variant), and p-values for balancing selection (the proportion with scores lower than the hit variant), as well as two-tailed p-values.

Two SNPs show Bonferroni-significant selection: the most significant signal, in *NOD2*, is under balancing selection (P = 5.2 x 10^{-5}), and the second most significant, in the receptor *TNFRSF18*, showed directional selection (P = 8.9 x 10^{-5}). The next most significant variants were in the ligand of that receptor, *TNFSF18* (directional, P = 5.2 x 10^{-4}), and *IL23R* (balancing, P = 1.5 x 10^{-3}). As a group, the IBD variants show significant enrichment in

Term	Description	Direction	P-value
GO:0032660	regulation of interleukin-17 production	Balancing	0.00014*
GO:00327	positive regulation of interleukin-17 production	Balancing	0.00020
GO:0009897	external side of plasma membrane	Directional	0.0018
GO:0008283	cell proliferation	Directional	0.0020
GO:0032653	regulation of interleukin-10 production	Balancing	0.0020

Table 4.14: Top 5 pathways that show evidence of natural selection in the IBD loci. *Significant after Bonferroni-correction for 334 enriched GO terms and pathways.

selection (Figure 4.13) of both types (P = 5.5 x 10^{-6}).

In order to assess whether extent or direction of selection was correlated with specific functions, I used the GO term enrichment method described above. I converted the selection p-values to Z scores using an inverse normal transformation, and tested for association between these scores and GO terms. The top five associations are shown in Table 4.14. The top result was the GO term “regulation of interleukin-17 production”, which met Bonferroni-corrected significance (Figure 4.13). The important role of IL17 in both bacterial defence and autoimmunity suggests a key role for balancing selection in maintaining the genetic relationship between inflammation and infection, and this is reinforced by a nominal enrichment of balancing selection in loci annotated with the broader GO term “defence response to bacterium” (p = 0.007).

4.4.6 Gene expression analyses of IBD loci

Gene expression datasets provide a powerful resource to interpret GWAS results. Two other groups within the IIBDGC Immunochip analysis group

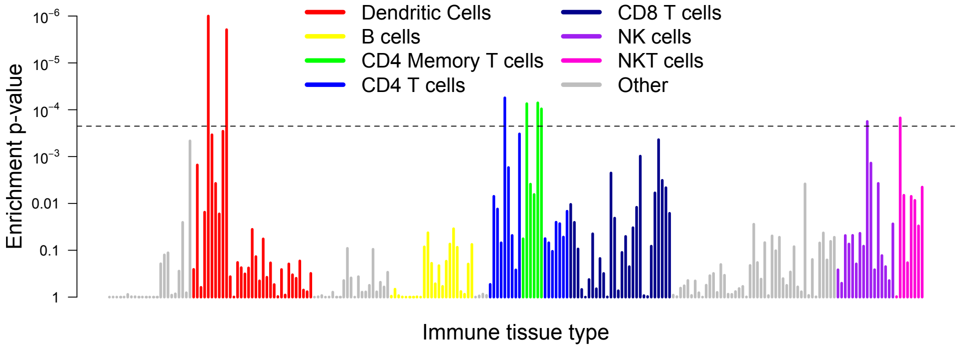


Figure 4.14: Evidence of enrichment in IBD loci of differentially expressed genes from various immune tissues. Each bar represents the empirical P-value in a single tissue, and the colours represent different cell type groupings. The dashed line is Bonferroni-corrected significance for the number of tissues tested.

used gene expression to investigate the new IBD locus list.

Xinli Hu and Soumya Raychaudhuri used enrichment of cell-type specific genes to study the cell types implicated by our locus list, using a previously described method (Hu et al., 2011). They tested for enrichment of cell-type expression specificity of genes in IBD loci in 223 distinct sets of sorted, mouse-derived immune cells from the Immunological Genome Consortium (Hyatt et al., 2006). Dendritic cells showed the strongest enrichment, followed by weaker signals that support the GO analysis, including CD4+ T, NK and NKT cells (Figure 4.14). Notably, several of these cell types express genes near our IBD associations much more specifically when stimulated; our strongest signal, a lung-derived dendritic cell, had $p_{stimulated} < 10^{-6}$ compared with $p_{unstimulated} = 0.0015$, consistent with an important role for cell activation.

Ken Hui and Eric Shadt used gene expression networks and eQTL data to infer causality in IBD associations. They screened genes in IBD loci against 211 co-expressed "modules" (sets of genes) previously identified by weighted

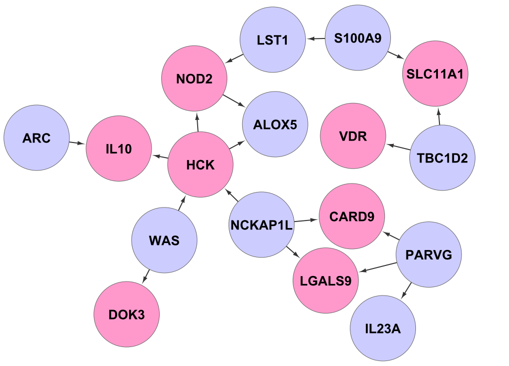


Figure 4.15: *NOD2*-focused cluster of the IBD causal subnetwork. Pink genes are in IBD associated loci, blue are not. Arrows indicate inferred causal direction of expression regulation.

gene co-expression network analyses (Zhang and Horvath, 2005) performed on multiple tissues (Greenawalt et al., 2011; Emilsson et al., 2008; Schadt et al., 2008), and identified a significantly enriched module in omental adipose tissue from obese patients ($p < 10^{-12}$). They then used gene expression and genotype data from these patients to construct a causal network using a Bayesian network reconstruction algorithm (Zhu et al., 2007). To illustrate the power of this approach, Figure 4.15 shows a small subset of this network around the gene *NOD2*, which also contains many important bacterial interaction genes including *IL10* and *CARD9*. This network implicates a number of new IBD genes as playing a part in response to bacteria, and in particular highlights the new IBD gene *HCK* as a potential regulator of the important IBD genes *NOD2* and *IL10*.

4.4.7 Take home messages about the biology of IBD

We have used a range of bioinformatic analyses to attempt to extract biological insight from the 163 loci and 1438 genes implicated by the Immunochip analysis. This has in turn produced a large amount of data, which itself needs to be interpreted. Below I will distil what I believe to be the major biological lessons that these analyses have taught us about the aetiology of IBD.

CD and UC show a very high degree of genetic overlap, with almost all of the 163 loci showing some degree of association to both. Likewise, there does not appear to be any strong differences in the function of CD and UC specific loci. However, many loci show a significant heterogeneity of odds ratio between the two phenotypes, with many having differing (or occasionally opposite) effects on CD and UC risk. Perhaps in the future we need to think about genetic differences between CD and UC not in terms of different loci, but as differently weighted combinations of the same loci. The same property may apply to subphenotypes of IBD (such as ileal verses colonic disease), and possibly even to the relationship between IBD and other immune-mediated diseases.

IBD shows genetic overlap with almost all diseases of immunity. However, **there is a startling overlap between IBD and susceptibility to both complex and Mendelian mycobacterial disease**. This is further highlighted by the large number of loci that contain genes involved in interferon gamma, including both the *IFNG* gene itself and its receptor *IFNGR2*, which is known to play a vital role in defence against *Mycobacterium tuberculosis* (Flynn et al., 1993). This relationship appears to have led to **significant natural selection on IBD alleles during human history**, and in particular balancing selection on the regulation of pro- and anti-inflammatory cytokines IL17 and IL10.

Cell types of both the innate and adaptive immune system play an important role in IBD. Gene expression data implicated dendritic and natural killer cells on the innate side, and CD4+ T-cells on the adaptive side. The

Gene Ontology analysis, however, implies a different mode of action of these two cell types. **IBD risk alleles seem to lead to defects of bacteria-induced cytokine production in innate immunity and defects of cell activation and signal transduction in the adaptive immune system**. This is not an exclusive relationship (one innate immune cell type has an activation-related GO term, “regulation of natural killer cell activation”), but it does seem to hold as a rule of thumb.

4.5 IBD and Y haplogroups

There is much suggestive evidence of a relationship between sex chromosomes and immunity. Most autoimmune diseases are more prevalent in females than in males (Whitacre, 2001), and individuals with Turner syndrome (a partial or total absence of one sex chromosome) are at higher risk of developing various immune-related diseases (Lleo et al., 2012). There is also evidence that the progression of HIV infection can vary between carriers of different Y haplogroups (Sezgin et al., 2009). However, large systematic studies of the effect of Y chromosome variation on human immune disease are relatively rare.

As mentioned in the introduction, 1735 Y chromosome variants were placed on the Immunochip for the purpose of assigning Y haplogroups. This gives us an opportunity to make a detailed and well powered study of the relationship between IBD risk and Y haplogroups. In this section I will describe the analysis of these variants in the IIBDGC Immunochip dataset.

4.5.1 Calling Y SNPs and assigning haplogroups

I selected males from the QC+ Immunochip sample set based on their mean normalised intensity at Y chromosome sites. There were a total of 22,129 males available, with 9,811 controls, 6,204 CD cases and 6,114 UC cases.

Because (at the time this study was carried out) the optiCall method used for genotype calling on the autosomes had not yet been adapted to run on sex chromosomes, I used the calling software Illuminus (Teo et al., 2007). The calls were generally of low quality, so I selected 150 haplogroup informative marker (Karafet et al., 2008) and manually inspected and fixed clusters using the program Evoker (Morris et al., 2010).

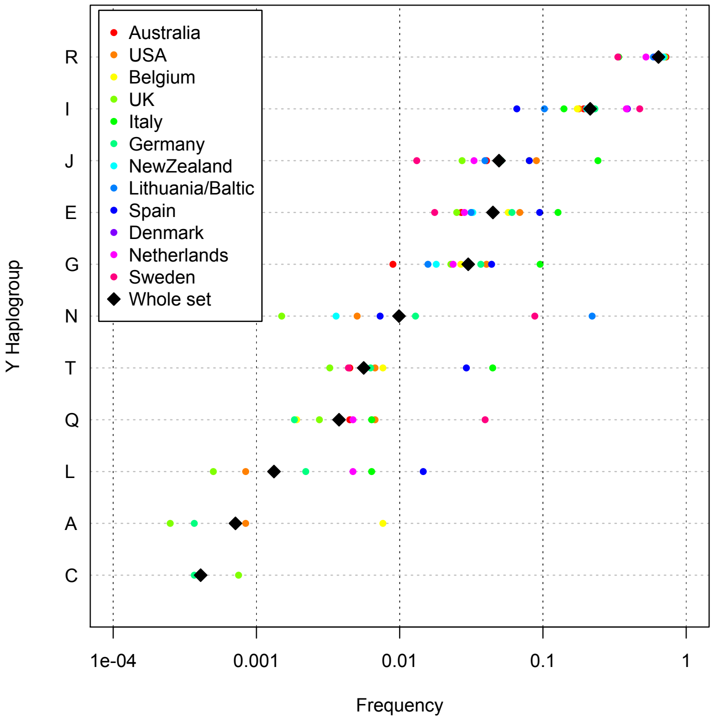


Figure 4.16: Y haplogroup frequencies in controls across the IIBDGC Immunochip dataset.

I developed a novel maximum likelihood method to assign haplogroups to these individuals (implemented in C++ as the program YFitter (Luke Jostins, 2011)). All but 9 males could be unambiguously assigned to a major haplogroup. The dataset contained samples from 10 major haplogroups, including 6 haplogroups with a frequency of greater than 1% (Figure 4.16).

4.5.2 Association analyses and controlling for stratification

I used logistic regression to assess association across these 6 common major haplogroups. The frequency spectrum differs between IBD cases and controls, even after including country-of-origin, sample collection and four autosomal principal components as covariates ($\chi^2 = 14.2$, df = 5, p = 0.014). The per-

Haplogroup	OR (95% CI)	P-value
E	1.07 (0.92 - 1.24)	0.393
G	1.20 (0.99 - 1.20)	0.059
I	1.00 (0.93 - 1.10)	0.837
J	0.85 (0.76 - 1.03)	0.112
N	1.53 (1.12 - 2.07)	0.006
R	0.96 (0.89 - 1.03)	0.229

Table 4.15: Association statistics for the Y chromosome haplogroups

haplogroup results show that this association is largely driven by a strong association between haplogroup N and IBD (Table 4.15).

Haplogroup N shows significant variation in frequency between European populations (Figure 4.16). This may lead us to suspect that the association results are due to population stratification. There are two major sources of stratification in IBD: a higher incidence in Ashkenazi Jewish, and an increasing incidence in Northern Europe compared to Southern Europe (Shivananda et al., 1996). We can rule out the former as haplogroup N has a lower frequency in Ashkenazim (Behar et al., 2004), which would produce the opposite direction of association to that observed. However, haplogroup N is at a significantly higher frequency in Northern Europe, so this is a plausible source of stratification. While I conditioned on country of origin and principal components, it is possible that additional stratification is driving the haplogroup N association.

To attempt to remove such stratification, I selected two homogeneous cohorts with over 10% frequency of haplogroup N (one Swedish and one Lithuanian). To ensure the population was homogeneous, I used principal component analysis to remove 136 individuals that lay more than two standard deviations from the mean on any of the first four PCs. Even within these two highly homogeneous populations, the results were very similar to

Collection	Cases/controls	OR (95% CI)
Sweden	165/193	1.19 (0.60 - 2.37)
Sweden (PC corrected)		1.27 (0.62 - 2.58)
Lithuania	192/109	1.94 (1.13 -3.33)
Lithuania (PC corrected)		1.81 (1.02 - 3.19)
Meta-analysis	228/228[a]	1.61 (1.05 - 2.47)
Meta-analysis (PC corrected)		1.57 (1.01 - 2.45)

Table 4.16: Association of haplogroup N with IBD in two homogenous populations. Studies were combined using variance-weighted fixed-effect meta-analysis. [a]Effective sample size

the across-Europe results (Table 4.16).

In these homogeneous groups, case-control status was correlated with principal components, weakly in Sweden (omnibus p = 0.050) and strongly in Lithuania (p = 3.6 x 10^{-4}). Equally, haplogroup N shows evidence of population stratification via a correlation between the haplogroup and principal components (p = 0.032 and p = 0.014). However, conditioning on the first four principal components within these countries does not significantly alter the results (Table 4.16), providing further evidence that this association is not driven by stratification.

4.5.3 Identifying candidate causal variants

Because the Y chromosome does not undergo recombination, the haplogroup association does not implicate a genomic region in the same way as an autosomal association does, and thus does not immediately suggest candidate genes or mutations.

To understand potential biological underpinnings of the haplogroup N enrichment in IBD, I used sequence data from the 1000 Genomes Project (specifically, from the Complete Genomics high-coverage sequencing) to iden-

Gene	Number	Location
AMELY	1	Intron
CD24	2	Upstream of TSS
KDM5D	3	CDS (missense)
NLGN4Y	8	Intron
PRKY	1	Intron
RPS4Y1	1	Intron
RPS4Y2	1	CDS (synonymous)
TBL1Y	5	Intron
TTTY10	2	Intron
TTTY14	2	Intron
TTTY15	3	Intron
USP9Y	4	Intron
UTY	13	Intron
ZFY	2	Intron

Table 4.17: Candidate genic mutations that may underlie the haplogroup N IBD association.

tify Y chromosome mutations specific to that haplogroup. A total of 379 mutations lie on or within the N haplogroup branch. 50 of these were present in or near genes, implicated 15 candidate genes (Table 4.17). These included a mutation 3kb upstream of *CD24*, a cell adhesion gene known to be up-regulated in inflammatory bowel disease, and a missense mutation in *KDM5D*, which encodes for a MHC antigen known to be involved in male-to-female tissue graft rejection.

4.6 Fine-mapping the *NOD2* locus

The IIBDGC has an ongoing project to fine-map IBD loci using the Immunochip. This project uses both the large European dataset discussed above, and an additional set of approximately 12,000 transethnic samples. It also aims to incorporate functional information from external datasets, such as gene expression and functional sequencing. It is aimed at both establishing the causal risk variants that underlie GWAS associations, and investigating the biological mechanisms through which these risk variants act. Calling and analysis of these datasets are currently ongoing.

In this section, I will describe the results of a pilot project carried out to investigate the methods and resources that could be used in such a fine-mapping project. This pilot project was focused on a single Crohn's disease fine-mapping region, the long-established *NOD2* locus. I will show how the Immunochip data can be used to infer new biological insights on both coding and non-coding associations at this locus.

4.6.1 Characterising coding mutations in *NOD2*

There are 24 polymorphic missense mutations in *NOD2* on the Immunochip. Six of these have been previously established as associated with IBD (Rivas et al., 2011). By performing stepwise logistic regression, I found that eight of these coding mutations show independent associations that are significant after correcting for the number of coding variants tested (i.e. $p < 0.002$), including the six known mutations and two that have not been reported before (Asn289Ser and Ala918Asp). With 8/24 mutations showing association, it is clear that a large proportion of the *NOD2* mutation space is associated with IBD. However, the Immunochip data can allow us to investigate in more

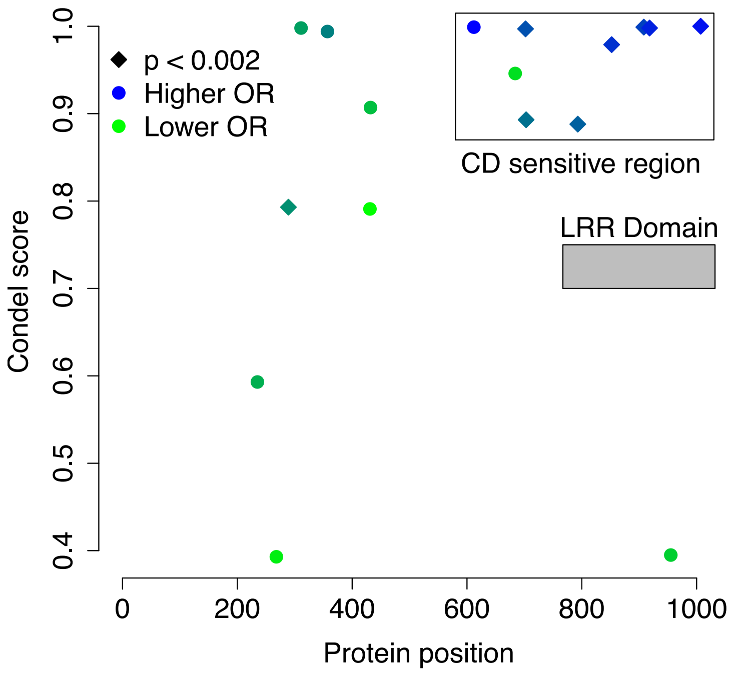


Figure 4.17: Functional characterisation of coding signals at NOD2. 17 coding variants are shown on a plot of their position in the protein and their Condel score, with colouring used to show their odds ratio. The LRR domain (responsible for bacterial sensing) is also shown.

detail what drives certain mutations to increase CD risk, while others appear to be benign from the point of view of IBD.

I took 17 of the highest frequency (MAF $>$ 0.0005) non-synonymous variants and calculated independent odds ratios for each (conditioning on the six established *NOD2* coding mutations, plus the common regulatory signal discussed below). I also produced a Condel score (Gonzalez-Perez and Lopez-Bigas, 2011) for each mutation, which combines various measures of conservation and protein structure to estimate the probability that the mutation is pathogenic. Figure 4.17 shows the relationship between Condel score, position in the protein, and odds ratio. We can see a striking relationship:

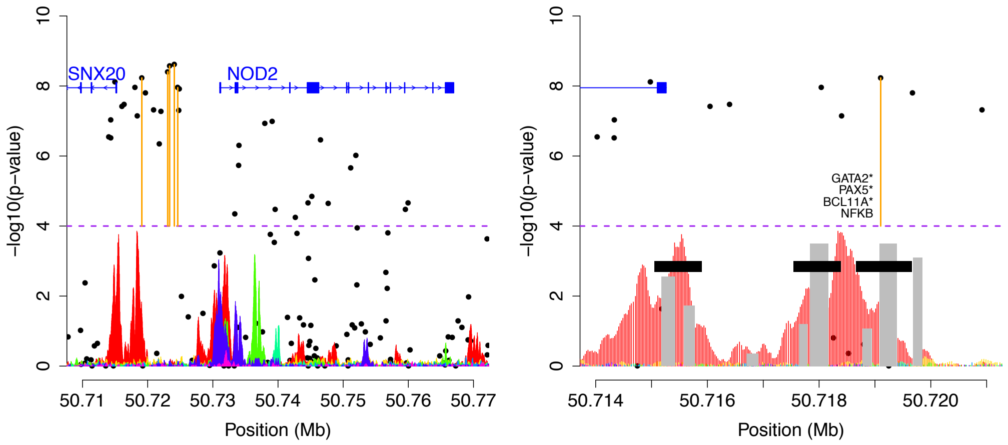


Figure 4.18: Fine-mapping and functional characterisation of a common regulatory signal at *NOD2*. Variants in orange are candidate causal variants. The coloured spikes under at the bottom of the plot show H3K27Ac histone modification levels in various tissues, with red being lymphoblastoid cell lines. Grey blocks are open chromatin and black blocks are transcription factor binding sites, with binding sites within 20bp of the candidate causal variant named in panel b).

mutations with a high Condel score, towards the end of the protein, almost invariably increase the risk of IBD. However, variants towards the start of the protein, or with a low Condel score, are rarely associated. It is likely that this "CD sensitive region" of *NOD2* represents mutations that disrupt the Leucine-Rich Repeat (LRR) domain. The LRR domain is responsible for detecting the bacterial product MDP, and is known to play a key role in Crohn's disease risk (Abraham and Cho, 2006).

4.6.2 Characterising a common regulatory signal at *NOD2*

Once we condition on the coding mutations mentioned above, a genome-wide significant signal remains around 50kb upstream of *NOD2* (Figure 4.18a). This signal is the same signal (but in the opposite direction) as the common *NOD2* association with leprosy susceptibility (Zhang et al., 2011), and is also associated with expression of both *NOD2* and *SNX20* in monocytes (Zeller et al., 2010). However, the association with IBD has not been reported before, as it can only be detected at genome-wide significance after conditioning on the coding variants.

Again, we are interested in the function of this association. The first step is to establish the set of SNPs that could plausibly be causal. To do this, we test the association statistics for the hit SNP conditional on each variant in LD with it, and rule out all SNPs that still show conditional association ($p < 0.01$). After performing this test, a total of 5 SNPs remain that could plausibly be the causal variant.

The next step is to establish what functional impact these candidate causal variants may have. Establishing the function of non-coding variants is difficult, but we can make some headway by using epigenetic sequencing data from the Encyclopaedia of DNA elements (ENCODE) (The ENCODE Project Consortium, 2012; Myers et al., 2011). In Figure 4.18b, I have overlaid H3K27Ac histone modification levels in various tissues: this is known to be an indicator of active enhancers (Creyghton et al., 2010). We can see that one of the candidate causal variants overlaps a peak that is specific to the lymphoblastoid cell line, suggesting an immune-tissue specific enhancer region. Looking closer at this region, we can see multiple sites of open chromatin and transcription factor binding (Figure 4.18b), with the candidate variant lying within one of these. The variant is nearby to binding sites for transcription factors involved in erythropoiesis (GATA2, PAX5 and BCL11A), as well as the protein NFκB, which regulates inflammation.

Taken together, this evidence points towards the existence of a common Crohn's disease risk variant in an upstream enhancer of *NOD2*. The upstream enhancer is active only in immune tissues, and appears to regulate expression of both *NOD2* and the neighbouring gene *SNX20*. This risk variant may act by interfering with a transcription factor binding, possibly a transcription factor involved in haemopoiesis.

4.7 Conclusions

The majority of this chapter has been focused on the use of the Immunochip to discovery new IBD loci in Europeans. The scale of the project has necessitated new approaches to both data handling and results interpretation, requiring a greater range of both techniques and expertise than previous IIBDGC analyses. Overall this has been a successful project, delivering both many new loci and biological information.

However, this project is only the first of many Immunochip analyses. At the time of writing we have just produced a new release of IBD Immunochip data, including over 86 thousand samples from both European and East Asian sample collections. This dataset will be used in a range of projects, including those that will fine-map existing loci, to study the contribution of IBD loci across different populations, and investigate the genetics of IBD subphenotypes. It will also be combined with Immunochip datasets from other diseases, in order to make a detailed investigation into the shared genetics of immune-mediated disease.

The results described in this chapter have taught us a number of lessons that will aid these future projects. Some of these are important, but perhaps uninteresting matters of quality control and data handling. For instance, the large manual inspection effort described in section 4.3.2 has given us many insights into the behaviour of Immunochip intensity readings, as well as setting up a framework for large, collaborative cluster plot inspection. Other lessons will have wider ranging consequences. For instance, the joint analysis of CD and UC demonstrated that two diseases can have an extremely high degree of genetic overlap (110 of 163 loci shared), but remain genetically distinct due to a high degree of effect size heterogeneity. We have learned that the relative balance of contribution of each locus can be just as important as the overall degree of locus sharing.

One of the strongest lessons to emerge from this analysis is the potential for integrating functional datasets into genetic studies. Gene expression, protein-protein interaction, canonical pathways and literature networks all added a great degree of value to the locus discovery effort. Most striking, the *NOD2* pilot fine-mapping project demonstrated the power of functional sequencing data in aiding the identification and understanding of non-coding causal variants. As a result of these successes, ongoing Immunochip projects are integrating, and in some cases specifically generating functional datasets as an integral part of their respective studies.

Chapter 5

High-throughput genomic studies of multiplex families

5.1 Introduction

The previous two chapters have discussed methods for mapping and interpreting disease associations in unrelated case/control cohorts. This has proven extremely successful at discovering common risk loci, including a large number of risk alleles for inflammatory bowel disease (IBD). However, case-control studies, using genotyping chips, are far from the only method of studying genetic risk.

As I discussed in the introduction, there are many types of risk variant that case-control GWAS studies are not well suited to study. In particular, the tag SNP approach is poorly powered to detect very low frequency variants, even if they have large effect sizes. The rise of next-generation sequencing, however, gives us the opportunity to directly assay such variants via whole genome or whole-exome sequencing. The question is how to distinguish the (very small) number of causal risk variants from the (very large) number of low-frequency variants that have no effect on disease.

One potentially powerful tool is the study of multiplex (or multiply affected) families. Multiplex families have long been a staple of human disease genetics, and are the starting point for both the heritability and linkage studies that underlie much of our knowledge of complex disease. In recent years family studies have fallen out of favour in complex disease genetics as a result of the relatively poor performance of linkage studies and the success of GWAS. However, as we will see, multiplex families are more likely to harbour rare, high penetrance causal variants than unrelated cases. Furthermore, the fact that these variants are shared across multiple affected individuals gives us information that can allow us to whittle down the list of candidate variants by focusing only on those that are shared by many affecteds within the family.

I will start this chapter with a brief discussion of the history of multiplex family studies in complex disease (section 5.2). This section will also outline the approach to studying multiplex families that I describe in this chapter, in the context of the studies that have come before. I will then introduce some statistical models for analysing multiplex families in terms of high penetrance and polygenic risk factors (section 5.3). This will lead to the introduction of a new method for prioritising multiplex families that are most likely to carry a high penetrance mutation, using GWAS risk variants.

Section 5.4 will discuss a large multiplex family with over 40 family members suffering from IBD, collected with the aim of identifying rare causal mutations. We have performed a detailed genetic investigation into this family, using targeting and whole-genome genotyping, as well as whole-exome and whole-genome sequencing. I will discuss the known risk in this family, and explore the linkage and haplotype evidence for association. I will then describe the analysis of the sequencing data, calling SNPs, indels and structural variants, and combining them with the linkage information. Finally, I will describe a filtering procedure designed to identify candidate causal variants on the basis of their frequency, function and segregation within the family. This identifies a total of 120 candidate variants, including coding and regulatory SNPs and indels, and structural variants.

In the final section (section 5.5) I will describe a validation and replication experiment designed to discover which of these candidate variants may be causal. I will describe the error modes that can create false candidates, and how they can be counteracted. Finally I will describe three methods for genetically replicating these associations, including using case-control cohorts, unaffected siblings and other multiplex families, and explore the power of these approaches.

5.2 A history of multiplex family studies in complex disease

It was the existence of multiply affected families that first led scholars to begin investigating what we now call disease genetics. At the turn of the 19th century John C. Otto published an extensive pedigree analysis of a haemophilic family in New Hampshire, tracing it back for three generations (Raabe, 2008). He also hypothesised that haemophilia may be traceable to only a few pilgrim families, the first description of what would now be called a founder effect. It is this very concept of family and population specific causal mutations that underlies the research in this chapter.

Studies of disease families were a focus of many Victorian scientists. Both French physician Paul Broca, and the English surgeon James Paget documented many multiplex cancer families, leading to the first studies into familial aggregation in what is now called complex disease (Schneider et al., 1986). While we may see the roots of the modern concept of family history in these developments, in other fields the recognition of familial clustering took a darker and more ideologically driven form. For instance, the hereditary degeneracy theories of psychiatric disease in late 19th century France fed rapidly into contemporary prejudices about the mentally ill that lay far from modern concepts of medical care (Dowbiggin, 1991).

Paget specifically argued that these cancer families were the result of a hereditary factor, but both he and Broca noted that the high (and not precisely known) prevalence of cancer made it difficult to rule out these families as merely chance occurrences. The reality of familial clustering was only established with the rise of systematic epidemiological studies, and the statistical frameworks required to analyse them, at the start of the 20th century.

As in the 19th century, studies of cancer lead the way (Schneider et al., 1986), and these studies came of age when the pioneering epidemiological studies of Janet Lane-Claypon (Lane-Claypon et al., 1926) conclusively demonstrated an enrichment of familial clustering in cancer. Even at this stage the genetic studies were informing biological knowledge: familial clustering was shown to occur strongly in cancer at a single location (in particular breast cancer), but only weakly in cancers from distinct locations, highlighting the importance of considering cancers of different tissues as distinct diseases.

Despite having (as we know now) a higher heritability, the study of multiplex families in inflammatory bowel disease developed later. This is partly because the current diagnostic landscape of IBD solidified later: while diagnoses of IBD stretch back to the 19th century, the distinct diagnoses of ulcerative and Crohn's colitis emerged only at the beginning of the 20th century (Kirsner, 1995). The existence of families with multiple affected individuals was noted from 1906, and nuclear families with three or more affecteds were documented from the 1930s (Kirsner, 1995). However, it was not until the 1960s and the advent of twin studies (Kirsner, 1973) that a hereditary role for IBD was widely accepted. Around this time the existence of very large IBD families in the Jewish population began to be noted, with a particularly striking family with seven affected members being reported in 1963 (Sherlock et al., 1963).

In the latter half of the 20th century, family history was recognised as the single strongest known predictor of IBD (Satsangi et al., 1997). Many collections of multiplex families were made during this time: in 2004 Russell and Satsangi (2004) reviewed studies of 19 distinct multiplex IBD family collections. These studies were important in establishing the broad strokes of IBD genetics. They gave the first indication that CD and UC were genetically distinct, yet related, diseases. Furthermore, they hinted at significant substructure within IBD aetiology, by demonstrating a genetic effect on disease location, and suggesting a genetic role in disease progression. Overall, family studies established IBD as a complex genetic disease, comparable in heritability to other immune-mediated diseases such as type 1 diabetes and multiple sclerosis.

Around the turn of the 21st century, linkage studies of multiplex IBD families led to the identification of the major IBD susceptibility loci *NOD2* (Hampe et al., 1999; Hugot et al., 2001) (in CD) and HLA (Williams et al., 2002) (in UC). However, large meta-analyses of linkage studies, including nearly 2000 families, failed to identify further genome-wide significant loci (van Heel et al., 2004), and even had difficulty consistently replicating the (by then fine-mapped) *NOD2* locus. This ultimately led to the replacement of family-based methods with genome-wide association studies (a phenomenon reviewed in Chapter 1).

The failure of linkage meta-analysis in IBD showed that IBD is not caused solely by high penetrance alleles at a small number of loci. However, it does not imply that high penetrance alleles do not exist; only that, if they do exist, they are individually at low frequency and are located in a number of different loci (so-called locus heterogeneity). Indeed, many of these multiplex families are likely to harbour high penetrance mutations, which can potentially be detected via their co-segregation with disease status within that family. It was this approach that identified mutations in the IL10 receptor subunits as an important contributor to early onset IBD (Glocker et al., 2009a).

Recent developments in whole-genome and whole-exome sequencing have opened up new avenues for the discovery of high penetrance causal variants. The power of this approach was demonstrated with the discovery of the gene underlying the previously unsolved Mendelian disease Miller syndrome (Ng et al., 2010). This study used whole-exome sequencing of four patients, combined with filtering based on databases of common variation and software for predicting the severity of coding mutations, and identified candidate causal mutations in the gene *DHODH*. Over recent years, this approach has become the dominant means of solving Mendelian diseases (Bamshad et al., 2011), and has even been used to identify mutations that underlie syndromic forms of IBD (Worthey et al., 2011a; Fiskerstrand et al., 2012).

Given the success of this sequencing approach, we would like to also use it to identify penetrant mutations in multiplex families with complex IBD. However, there are a number of challenges in generalising this approach. Firstly, there is no guarantee that any given affected individual, and even any given multiplex family, will carry a penetrant mutation. Ideally we would like to sequence families that are likely to carry such mutations, and thus we require methods to decide which families to select for study. Secondly, even if a causal mutation is present in a family it is unlikely to be fully penetrant. Likewise, because the disease is relatively common compared to Mendelian diseases some family members may have the disease despite not carrying the mutation (so-called "phenocopies"). We thus need methods that can discover such mutations in families that may include both affected non-carriers and unaffected carriers. Finally, as we saw in Chapter 4 many common IBD risk variants lie in regulatory rather than coding regions, and it is possible that this will also be true for rare risk variants. We would thus like to generalise the variant prioritisation procedure to include potential non-coding candidate risk variants.

5.3 Modelling and controlling polygenic risk in multiplex families

There are many potential factors that can lead to familial aggregation in a disease without leading to families suitable for locus mapping. An obvious reason (and one that has been discussed since the 19th century) is chance co-occurrence: the large number of families in the world makes it likely that there exist families that have a large number of affecteds despite the absence of an underlying genetic risk factor. This effect can be additionally confounded by uncertainty in the prevalence, or population stratification, both of which could inflate the chance of seeing multiplex families by chance. For instance, the higher prevalence of IBD in individuals of Ashkenazi Jewish individuals will lead to a larger number of multiplex families in the Jewish population, even if the increased risk in this group was entirely due to environment.

Additionally, a shared exposure to an environmental risk factor can lead a family to develop a higher incidence than would be expected by chance. Diagnostic bias can also lead to familial clustering, as a strong family history may lead to more vigilant screening or overdiagnosis (this is particularly likely to occur for diseases with a high rate of undiagnosed cases, such as prostate cancer (Fleshner, 1995)). These non-genetic causes all highlight the importance of careful screening of multiplex families to establish a genetic cause.

Furthermore, for the purposes of mapping loci an excess of familial aggregation as a result of genetics may not be enough to make a family useful for study. It is now becoming clear that a substantial portion of the heritability of complex traits is due to highly polygenic risk. Williams et al. (2002)

estimated the contribution of polygenic risk in three complex diseases in the Wellcome Trust Case-Control Consortium data, by applying a linear mixed-model method. This gave lower bounds on the liability-scale variance due to polygenic risk from common loci of 22% for Crohn's disease, 31% for Type I Diabetes and 38% for Bipolar Disorder. In many cases a significant minority of this polygenic risk has already been characterised, for example via the 193 independent IBD risk factors identified via the IIBDGC Immunochip study (see Chapter 4), but much still remains undiscovered.

The risk variants that make up this polygenic risk each have a small effect size, and thus are unlikely to individually co-segregate with affection status in multiplex families. They are therefore outside of the scope of what can be studied by sequencing families. However, it will contribute to familial aggregation of cases within multiplex families, creating another class of families that need to be excluded from family sequencing studies.

A good first stage in understanding the impact of polygenic and penetrant risk on multiplex families is to construct and examine theoretical models of risk in families. Recent theoretical studies have investigated models of high penetrance mutations (Al-Chalabi and Lewis, 2011), as well as models of continuous polygenic risk (Yang et al., 2010) in multiplex families. However, to answer questions about the relative contribution of penetrant and polygenic risk, we need to construct a model that contained both elements.

In this section, I will develop a model of genetic risk that combines a polygenic risk with the presence of dominant, high penetrance alleles, and study how different parameterisations of this model (corresponding to different heritabilities, prevalence and balances of polygenic/penetrant risk) alter the distribution of affecteds in multiplex families. I will also develop and test a method for performing genetic risk prediction in a partially genotyped pedigree, and using such risk prediction to prioritise multiplex families that are likely to carry high penetrance mutations over those that are likely to carry only polygenic risk.

5.3.1 A combined polygenic/penetrant model of multiplex families

To describe the combined polygenic/penetrant model of genetic risk, I will first lay out the two components: a liability threshold model for polygenic risk due to common variants of low effect, and a dominant Mendelian model for higher penetrance variants. I will then combine these two models together to produce a general model of which both component models are special cases.

Throughout this section I will consider a nuclear family, with two parents denoted by subscripts m and f (for mother and father, treated as interchangeable), and O offspring denoted by subscripts $c_i : i = 1, ..., O$. I will use indicator variables d_i to denote the affection status of individuals. I use a parameter K to denote the disease prevalence in the population.

The polygenic model

We model the polygenic component of the disease using a liability threshold model (as described in Chapter 2). To recap, each individual in the family is given a liability $L_i = A_i + E_i$, where the genetic liability $A_i \sim N(0, h^2)$ is an additive polygenic component of risk, and the environmental liability $E_i \sim N(0, 1 - h^2)$ is an (individual-specific) environmental component. h^2 is called the heritability of liability, and measures the proportion of liability that is shared by identical twins: as this model assumes additive polygenic risk, h^2 is also the narrow-sense heritability. An individual is affected (i.e.

$d_i = 1$) if $L_i > T$, where T is the liability threshold $T = \Phi^{-1}(1-K)$ and Φ is the cumulative distribution function of the standard normal distribution.

The liabilities for each family member are

$$\begin{aligned} L_m &= A_m + E_m && (5.1)\\ L_f &= A_f + E_f && (5.2)\\ L_{c_i} &= A_{c_i} + E_{c_i} = \frac{1}{2}(A_m + A_f) + M_{c_i} + E_{c_i}, && (5.3) \end{aligned}$$

where $M_{c_i} \sim N(0, h^2/2)$ is a Mendelian segregation term. We can reformulate these equations in terms of $4 + O$ standard normal variables Z_i,

$$\begin{aligned} L_m &= hZ_1 + \sqrt{1-h^2}Z_3 && (5.4)\\ L_f &= hZ_2 + \sqrt{1-h^2}Z_4 && (5.5)\\ L_{c_i} &= \frac{h}{2}(Z_1 + Z_2) + \sqrt{1 - \frac{h^2}{2}}Z_{c_i}, && (5.6) \end{aligned}$$

The probability of an individual having disease state d_i given a genetic liability a_i is given by

$$P(d_i|A_i) = \begin{cases} \Phi(\frac{T-A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1; \\ 1 - \Phi(\frac{T-A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 0 \end{cases} . \tag{5.7}$$

We can write down a similar expression conditional on parental genetic liabilities

$$P(d_c|A_m, A_f) = \begin{cases} \Phi(\frac{T-(A_m+A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } d_i = 1; \\ 1 - \Phi(\frac{T-(A_m+A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } d_i = 0 \end{cases} . \tag{5.8}$$

The probably mass function for a set of affection statuses $\vec{d} = (d_m, d_f, d_{c_1}, ..., d_{c_O})$ is thus given by

$$\begin{aligned} P(\vec{d}) &= \iint_{-\infty}^{\infty} P(d_m|hz_1)P(d_f|hz_2)\phi(z_1)\phi(z_2) \\ &\times \prod_{i=1}^{O} P(d_{c_i}|hz_1, hz_2)dz_1dz_2. \end{aligned} \tag{5.9}$$

Because siblings are interchangeable and independent conditional on parental genetic liabilities, we can model the number of affected offspring using a binomial distribution. The joint probability of observing parent genotypes (d_m, d_f), and also observing a total of y_c affected offspring is thus

$$\begin{gathered} P(d_m, d_f, \sum d_{c_i} = y_c) = \\ \iint_{-\infty}^{\infty} P(d_m|hz_1)P(d_f|hz_2)\phi(z_1)\phi(z_2)\binom{O}{y_c} \\ \times P(d = 1|hz_1, hz_2)^{y_c} P(d = 0|hz_1, hz_2)^{O-y_c} dz_1dz_2. \end{gathered} \tag{5.10}$$

Finally, because parents are interchangeable, we can write down the probability of observing y total affecteds in the family (including parents and children) as

$$\begin{aligned} P(\sum d = y) = \quad & P(d_m = 1, d_f = 1, \sum d_{c_i} = y - 2) \\ & +2P(d_m = 1, d_f = 0, \sum d_{c_i} = y - 1) \\ & +P(d_m = 0, d_f = 0, \sum d_{c_i} = y). \end{aligned} \tag{5.11}$$

The dominant penetrant model

The dominant penetrant model assumes that a large number of individually rare variants exist in the population, each of which has a dominant effect with intermediate penetrance. Certain diseases are known to show such a heterogeneity of genetic architecture, for instance in diabetes (Molven and Njølstad, 2011) and breast cancer (Chen and Parmigiani, 2007), and it is possible this is true for other diseases.

This model assumes that a proportion R of cases have a dominant mutation with a penetrance of $\pi > K$. The total combined frequency of these mutations is thus KR/π (and therefore $\pi/R > K$): note that this is the proportion of people who carry at least one mutation, not the allele frequency. We will use the indicator variable $r_i = 1$ to denote that individual i carries a mutation, and assume that each individual carries at most one mutation.

The disease probabilities, conditional on genotype, are given by

$$P(d_i = 1|r_i = 1) = \pi \tag{5.12}$$

$$P(d_I = 1|r_i = 0) = \frac{K(1-R)}{1-KR/\pi}, \tag{5.13}$$

and transmission probabilities from parents to child are given by

$$P(r_{c_i} = 1|r_m = 0, r_f = 0) = 0 \tag{5.14}$$

$$P(r_{c_i} = 1|r_m = 1, r_f = 0) = \frac{1}{2} \tag{5.15}$$

$$P(r_{c_i} = 1|r_m = 1, r_f = 1) = \frac{3}{4}. \tag{5.16}$$

We can combine these two together to give disease probabilities conditional on parental genotype

$$P(d_{c_i} = 1 | r_m = 0, r_f = 0) = \frac{K - KR}{1 - KR/\pi} \tag{5.17}$$

$$P(d_{c_i} = 1 | r_m = 1, r_f = 0) = \frac{K + \pi - 2KR}{2(1 - KR/\pi)} \tag{5.18}$$

$$P(d_{c_i} = 1 | r_m = 1, r_f = 1) = \frac{K + 3\pi - 4KR}{4(1 - KR/\pi)}. \tag{5.19}$$

As with the polygenic model, offspring are interchangeable and independent conditional on parental genotype, so again we model the number of affected offspring binomially:

$$\begin{gathered} P(\textstyle\sum d_{c_i} = y_c | r_m = 1, r_f = 0) = \\ \binom{O}{y_c} \left[P(d_{c_i} = 1 | r_m, r_f)\right]^{y_c} \left[1 - P(d_{c_i} = 1 | r_m, r_f)\right]^{O - y_c}. \end{gathered} \tag{5.20}$$

We can then incorporate parental affection status, conditional on genotype, into the total count of affecteds y

$$\begin{gathered} P(\textstyle\sum d_i = y | r_m, r_f) = \\ P(d_m = 1, d_m = 1 | r_m, r_f) P(\textstyle\sum d_{c_i} = y - 2 | r_m, r_f) \\ + P(d_m = 1 \text{ or } d_f = 1 | r_m, r_f) P(\textstyle\sum d_{c_i} = y - 1 | r_m, r_f) \\ + (1 - P(d_m = 1 | r_m))(1 - P(d_f = 1 | r_m)) P(\textstyle\sum d_{c_i} = y | r_m, r_f), \end{gathered} \tag{5.21}$$

where

$$
\begin{aligned}
&P(d_m = 1 \text{ or } d_f = 1 | r_m, r_f) = \\
&P(d_m = 1 | r_m) + P(d_f = 1 | r_f) - 2P(d_m = 1 | r_m) P(d_f = 1 | r_f). \qquad (5.22)
\end{aligned}
$$

Finally we marginalize out parental genotypes using the population frequency

$$
\begin{aligned}
P(\sum d_i = y) = \quad & \left(\tfrac{KR}{\pi}\right)^2 P(\sum d_i = y | r_m = 1, r_f = 1) \\
& + 2\tfrac{KR}{\pi}(1 - \tfrac{KR}{\pi}) P(\sum d_i = y | r_m = 1, r_f = 0) \\
& + (1 - \tfrac{KR}{\pi})^2 P(\sum d_i = y | r_m = 0, r_f = 0). \qquad (5.23)
\end{aligned}
$$

The combined polygenic/dominant penetrant model

The combined model takes into account both polygenic risk and the presence of penetrant dominant risk alleles. To do this we set two thresholds, one for non-carriers for the dominant risk alleles $T_{wt} = \Phi^{-1}\left(1 - \frac{K - KR}{1 - KR/\pi}\right)$, and one for carriers $T_{dom} = \Phi^{-1}(1 - \pi)$. We then model transmission of both the penetrant risk alleles and a continuous liability.

The continuous liability is again given as $L_i = A_i + E_i$, where the genetic liability $A_i \sim N(0, h_d^2)$ only includes heritability due to common variants, excluding the rare penetrant mutations. This polygenetic heritability is given by $h_p^2 = h^2 - h_d^2$, where $h_d^2 = \frac{\sigma_d^2}{1 + \sigma_d^2}$ is the variance explained on the liability scale by the penetrant risk alleles, where

$$
\sigma_d^2 = \frac{KR}{\pi} \left[T_{dom} - T\right]^2 + (1 - \frac{KR}{\pi}) \left[T_{wt} - T\right]^2. \qquad (5.24)
$$

Note that $h_d^2 \to 1$ as $\pi \to 1$ and as $R \to 1$.

We now specify the disease probability conditional on both the polygenetic liability (A_i) and the presence of absence of a penetrant mutation (r_i)

$$P(d_i = 1|A_i, r_i) = \begin{cases} \Phi(\frac{T_{dom} - A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1 \text{ and } r = 1; \\ \Phi(\frac{T_{wt} - A_i}{\sqrt{1-h^2}}) & \text{if } d_i = 1 \text{ and } r = 0 \end{cases} . \tag{5.25}$$

Again we can give a child's disease probability conditional on the genetic liability and presence of penetrant mutations in the parents, by taking into account the multiple thresholds with different transmission probabilities

$$P(d_{c_i} = 1|A_m, A_f, r_m, r_f) = \\ \begin{cases} \Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 0 \text{ and } r_f = 0; \\ \frac{1}{2}\Phi(\frac{T_{dom} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) + \frac{1}{2}\Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 1 \text{ xor } r_f = 1; \\ \frac{3}{4}\Phi(\frac{T_{dom} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) + \frac{1}{4}\Phi(\frac{T_{wt} - (A_m + A_f)/2}{\sqrt{1-h^2/2}}) & \text{if } r_m = 1 \text{ and } r_f = 1 \end{cases} \tag{5.26}$$

As before, we can write down the probability of observing y_c affected offspring given parental genotypes by modelling the number of affecteds as a binomial

$$\begin{gathered} P(d_m, d_f, \sum d_{c_i} = y_c|r_m, r_f) = \\ \iint_{-\infty}^{\infty} P(d_m|h_p z_1, r_m) P(d_f|h_p z_2, r_f) \phi(z_1) \phi(z_2) \\ \times \binom{O}{y_c} P(d = 1|h_p z_1, h_p z_2, r_m, r_f)^{y_c} \\ \times (1 - P(d = 1|h_p z_1, h_p z_2, r_m, r_f)^{y_c})^{O - y_c} dz_1 dz_2. \end{gathered} \tag{5.27}$$

We then include parental affection status to give the probability mass function for the total number of affecteds y given parental genotypes

$$\begin{aligned} P(\sum d = y|r_m, r_f) = \quad & P(d_m = 1, d_f = 1, \sum d_{c_i} = y - 2|r_m, r_f) \\ & +2P(d_m = 1, d_f = 0, \sum d_{c_i} = y - 1|r_m, r_f) \\ & +P(d_m = 0, d_f = 0, \sum d_{c_i} = y|r_m, r_f), \end{aligned} \tag{5.28}$$

and finally we marginalize out parental genotypes using the population frequency to give the final probability mass function

$$\begin{aligned} P(\sum d_i = y) = \quad & \left(\tfrac{KR}{\pi}\right)^2 P(\sum d_i = y|r_m = 1, r_f = 1) \\ & +2\tfrac{KR}{\pi}(1 - \tfrac{KR}{\pi})P(\sum d_i = y|r_m = 1, r_f = 0) \\ & +(1 - \tfrac{KR}{\pi})^2 P(\sum d_i = y|r_m = 0, r_f = 0). \end{aligned} \tag{5.29}$$

Results

I have implemented the above combined model using R, and used it to explore how the expected number of affecteds in multiplex families for a relatively uncommon disease ($K = 0.01$) varies depending on model and model parameters.

Figures 5.1a and 5.1b show the results of this multiplex model to families of 8 ($O = 6$), with dominant penetrance of $\pi = 0.5$. The solid lines give the purely polygenic model $R = 0$, the black lines give the purely penetrant model $h^2 = 0$, and other lines give various parameterisations of the combined model.

The first thing to note is that multiplex nuclear families can be very common given only a moderate degree of polygenic risk. Families with 5 or

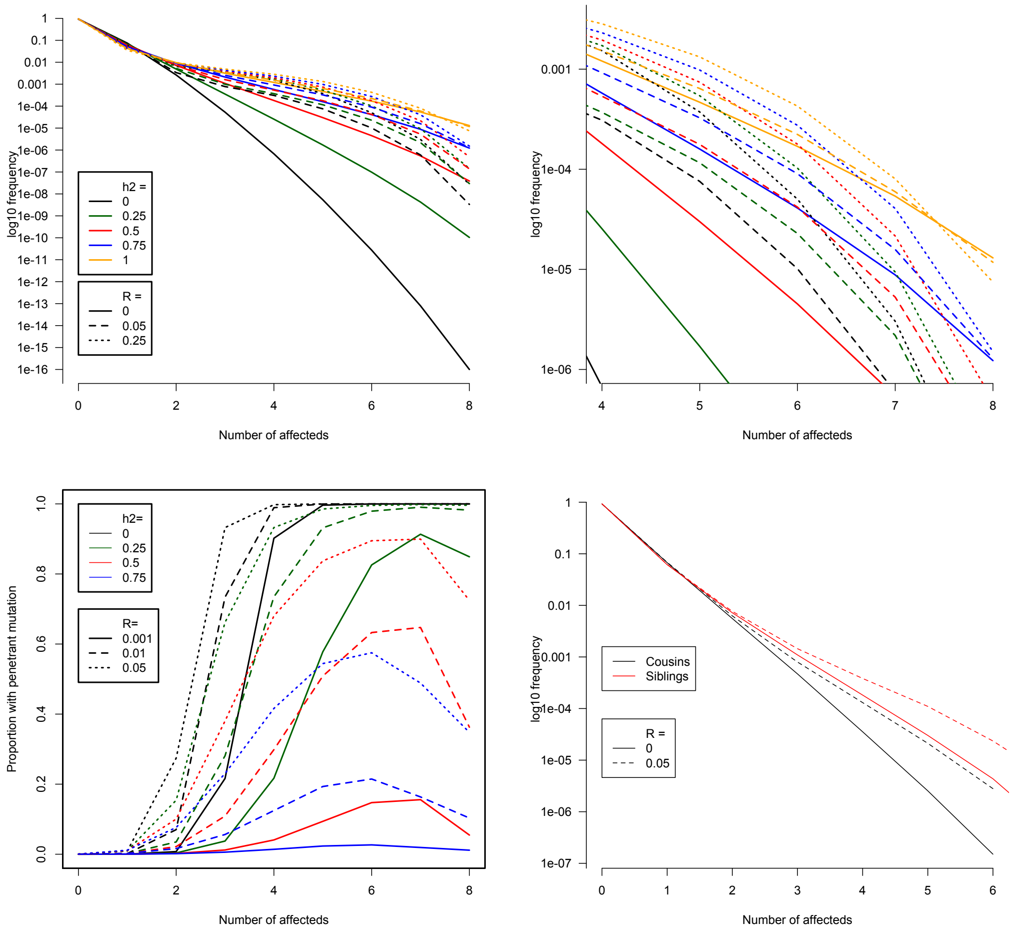


Figure 5.1: The results of the combined multiplex family model. a) Distribution of number of affecteds per nuclear family with 6 children, for different values of h^2 and R. b) A zoomed in view of the same model c) The probability that a nuclear family harbours a penetrant mutation, for different values of h^2 and R d) Comparison of sibships ($O = 6$) and cousinships ($k = 2, O_1 = 2, O_2 = 2$), with $h^2 = 0.5$ (generated by simulation). In both cases I used $K = 0.01$ and $\pi = 0.5$.

more affecteds, an occurrence that is virtually impossible under the null (less than one family in 200 million) become possible (one family in 500 thousand) for a moderate polygenic heritability of 0.25 and positively common (one family in 30 thousand) for strong heritability of 0.5. Multiplex families are likely to be relatively common, even without high penetrance mutations.

However, the flip-side of this is that a high degree of familial aggregation can be seen even for not particularly heritable diseases given a small contribution of dominant alleles. A disease with no polygenic liability, but 5% of cases caused by penetrant mutations, will show as many families with 4 affecteds as a disease with 50% heritability (despite the former case having a heritability of less than 1%). This seems to lead to the somewhat counterintuitive conclusion that families multiply affected by a weakly heritable disease will be easier to map than equivalent families with a strongly heritable disease, though this may be confounded by correlations between polygenic and penetrant heritabilities.

We can turn these results around and instead ask what proportion of multiplex families of a certain size harbour a penetrant mutation (Figure 5.1c). In the absence of polygenic risk, the vast majority of nuclear families with more than 4 affecteds harbour a penetrant mutation, even if such mutations explain a very small proportion of the total disease burden ($R > 0.001$). However, this becomes progressively less true as the heritability rises, and for highly heritable diseases penetrant mutations only become common in multiplex nuclear families if they already explain a non-trivial amount of all cases to start with (greater than 1% for $h^2 = 0.5$, and greater than 5% for $h^2 = 0.75$).

Figure 5.1d compares the results of the combined model for nuclear and extended families of the same size. Specifically, I have compared a nuclear family with eight individuals (two parents and six children, i.e. $O = 6$ offspring) to an extended family with eight individuals (two siblings, their partners, and their two children each, i.e. $k = 2, O_1 = O_2 = 2$, with grandparental state disregarded). I consider only a heritability of $h^2 = 0.5$.

As we have already seen, under the polygenic model with $h^2 = 0.5$, observing five or more affected nuclear family members is not unlikely (1 in 30 thousand). However, Figure 5.1d shows that for an extended family of the same size this is a relatively rare even (1 in 400 thousand). This gap between nuclear and extended families is reduced if the presence of high penetrance mutations is considered. Introducing a small number of penetrant mutations ($R = 0.05$, $\pi = 0.5$) increases the number of families with at least 5 affecteds 9-fold for the cousinship (to 1 in 42 thousand), but only 4-fold for the nuclear family (to 1 in 7200). This corresponds to a 93% of cousinships with 5 affecteds carrying a penetrant mutation, compared to 83% for nuclear families.

From these analyses we can draw a number of lessons for studying multiplex families

- Even if only a minority of variance is explained by rare variants, these rare variants can still result in the occurrence of a relatively large number of multiplex families.

- However, relatively large numbers of multiplex families are also expected given the levels of polygenic risk ($h^2 = 0.2 - 0.5$) that have been shown to exist for many complex diseases. Thus the presence of strong familial clustering is not alone evidence of a penetrant mutation.

- Extended multiplex families, with aggregation occurring across cousins as well as siblings, is stronger evidence of a penetrant mutation.

- Using additional methods to decrease or factor out the contribution of polygenic risk will be valuable in identifying families that are likely to harbour penetrant risk variants

5.3.2 Risk prediction in multiplex families

An outline of risk prediction in families

As we have seen in the above section, the presence of polygenic risk can lead to a high frequency of multiplex families even in the absence of penetrant mutations. However, for many diseases we already have a grasp on this polygenic variation via the results of GWAS. For instance, the 193 independent associations to Crohn's Disease explain 12.7% of variance in disease liability (see Chapter 4). Using the upper bound of 84% (calculated in Chapter 1) and a lower bound of 22% (from Williams et al. (2002)), we know that we have discovered somewhere between 15% and 58% of the polygenic risk for Crohn's disease.

We can use these GWAS loci to produce estimates of polygenic risk, and use this polygenic risk to prioritise those families that are more likely to harbour penetrant mutations. Assume that a given family has N members, of whom y are affected. We wish to select families for which y is significantly larger than what would be expected given the observed genotypes, G, i.e. those that minimize:

$$P(\hat{y} > y|N, G) \tag{5.30}$$

If G is known for all family members then disease probabilities for each individual can be calculated directly from the odds ratios as described in Chapter 2, and then used to calculate equation (5.30) by sampling. However, most family based experiments will not generate genotype data across all members of the pedigree for a variety of reasons, including cost, DNA availability, consent, or death. A solution is to sample disease status as in the complete information case, conditional on a set of unobserved genotypes G_{unobs} that are themselves sampled from the conditional distribution

$$P(G_{unobs}|\mathbf{f}, T, G_{obs}), \tag{5.31}$$

where $\mathbf{f}$ is the population allele frequency, T is the family structure, and G_{obs} are the known genotypes. Sampling from this distribution is not trivial, but is possible via a modified Inside-Outside algorithm (Baker, 1979) (itself a generalisation of the forward-backwards algorithm used in Hidden Markov Models). The Inside-Outside is used for inference on tree-like data structures, and has been applied to certain multiple sequence alignment problems (Durbin, 1998). Here, we instead use Inside-Outside to sample from the posterior distribution of genotypes across a family. Briefly, we decompose the marginal genotype posteriors into inside and outside probabilities, similar to the forward and backward probabilities from an HMM. The inside probability accounts for information from each individual and their descendants, whereas the outside probability accounts for the individual's other relatives (including ancestors, siblings and cousins).

These values can be computed recursively via the standard Inside-Outside approach (Section 5.3.2), which enables the sampling of one individual's genotypes. When sampling an entire family, however, we must sample down the tree from the root, with each individual's genotypes conditioned on their parents' sampled genotypes (Section 5.3.2). We accomplish this by modifying the outside probability to include parental genotypes (Section 5.3.2).

Description of the Inside-Outside algorithm in trees

Definitions

The Inside-Outside algorithm is a generalisation of the Forward Backward algorithm, originally designed to extend parameter estimation from Hidden Markov Models to stochastic context-free grammars (Baker, 1979). Here we reformulate the Inside Outside algorithm as a method of performing parameter estimation and sampling on a directed tree.

A directed tree is a directed acyclic graph in which all nodes have a unique path originating from a single node. We will denote nodes by subscripts i, j, k. Each node i may have a parent p_i, offspring o_i and/or siblings s_i. A node without parents is called a "root node" or "root", and a node without children is called a "leaf node" or "leaf".

Each node i has an associated emission d_i (e.g, an observed genotype), as well as a hidden state x_i (e.g. an unobserved genotype) with statespace S_i. The values of hidden states will be denoted a, b, c etc, e.g. $(x_i = a)$ denotes that node i has hidden state value a.

The tree defines a graphical model that specifies the probability density functions for all the variables (hidden states and emissions) as conditional probabilities. Specifically, the probability density function of emission d_i is specified conditional on hidden state x_i taking value a by the likelihood

$$L_i(a) = P(d_i|x_i = a). \tag{5.32}$$

The probability density function for a non-root hidden state variable x_i taking on value b is specified conditional on the parent's hidden state x_{p_i} taking on value a by the transition probability

$$T_i(b|a) = P(x_i = b|x_{p_i} = a). \tag{5.33}$$

The probability distribution of the hidden state associated with the root x_{root} is given by the root prior

$$\pi(a) = P(x_{root} = a). \tag{5.34}$$

We will refer to all emissions associated with node i and nodes descended from node i as D_i, and all emissions not associated with node i or its descendants as $D_{!i}$. Note that these can both be expressed recursively

$$D_i = \{d_i, D_{o_i}\} \tag{5.35}$$

for non-leaves and $D_i = d_i$ for leaves, and

$$D_{!i} = \{D_{s_i}, D_{!p_i}, d_{p_i}\} \tag{5.36}$$

for non-roots and $D_{!i} = \emptyset$ for the root. All emissions associated with all nodes can be expressed as D, and $D = \{D_i, D_{!i}\}$ for any i.

We will use the Inside-Outside algorithm to deduce the probability density functions of hidden states x_i conditional on observed emissions associated with all nodes D.

The Inside Probability

The inside probability $\alpha_i(a)$ is defined as the probability of observing emission associated with node i and all its descendants, given that the hidden state x_i takes on value a

$$\alpha_i(a) = P(D_i|x_i = a). \tag{5.37}$$

For leaves, $D_i = d_i$, and hence $\alpha_i(a) = L_i(a)$. For non-leaves we have

$$\begin{aligned}
\alpha_i(a) &= P(D_i|x_i = a) \\
&= P(d_i|x_i = a) \prod_{j \in o_i} P(D_j|x_i = a) \\
&= P(d_i|x_i = a) \prod_{j \in o_i} \sum_{b \in S_j} P(D_j|x_i = b) P(x_j = b|x_i = a) \\
&= L_i(a) \prod_{j \in o_i} \sum_{b \in S_j} \alpha_j(b) T_j(b|a). \qquad (5.38)
\end{aligned}$$

Because we require the inside probabilities of all offspring of a node to calculate its own inside probability we calculate the inside probabilities first for the leaves, and then propagate them recursively up the tree. The overall likelihood of all emissions D is

$$P(D) = \sum_{a \in S_{root}} \alpha_{root}(a) \pi(a). \tag{5.39}$$

The Outside Probability

The outside probability $\beta_i(a)$ is defined as the joint probability of observing emissions not associated with node i and its descendants, and the node i being in hidden state $x_i = a$ is

$$\beta_i(a) = P(D_{!i}, x_i = a). \tag{5.40}$$

For the root node, $D_{!i} = \emptyset$, so $\beta_{root}(a) = P(x_{root} = a) = \pi(a)$. For non-root nodes, we can calculate the outside probability recursively as

$$\begin{aligned}
\beta_i(a) &= P(D_{!i}, x_i = a) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, x_i = a, D_{!i}) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, x_i = a, D_{!p_i}) P(d_i | x_{p_i} = c) \prod_{j \in s_i} P(D_j | x_{p_i} = c) \\
&= \sum_{c \in S_{p_i}} P(x_{p_i} = c, D_{!p_i}) P(x_i = a | x_{p_i} = c) P(d_i | x_{p_i} = c) \\
&\quad \times \prod_{j \in s_i} \sum_{b \in S_j} P(D_j | x_j = b) P(x_j = b | x_{p_i} = c) \\
&= \sum_{c \in S_{p_i}} \beta_{p_i}(c) T_i(a|c) L_{p_i}(c) \prod_{j \in s_i} \sum_{b \in S_j} \alpha_j(b) T_j(b|c). \qquad (5.41)
\end{aligned}$$

The outside probability for each node requires the outside probability of the node's parent. We thus calculate it first for the root, and then propagate recursively down the tree. The outside probabilities are also dependent on the inside probabilities, which are therefore calculated first.

Conditional sampling across the tree

We can calculate the posterior distribution of hidden state x_i conditional on all emissions D in terms of the inside and outside probabilities as

$$P(x_i = a | D) = \frac{\alpha_i(a) \beta_i(a)}{P(D)}. \qquad (5.42)$$

We can sample from this posterior distribution for each node. However, this approach cannot jointly sample hidden states across the entire tree. To do this we need to propagate sampled states down the tree, starting with the root. The hidden state for the root can be sampled from the posterior distribution

$$P(x_{root} = a|D) = \frac{\alpha_{root}(a)\pi(a)}{P(D)}. \tag{5.43}$$

To sample non-roots, we must first calculate the partial outside variable, which includes the hidden state c of the parent, and can be calculated as

$$\begin{aligned} \beta_i^p(a,c) &= P(x_i = a, x_{p_i} = c, D_{!i}) \\ &= \beta_{p_i}(c)T_i(a|c)L_{p_i}(c)\prod_{j\in s_i}\sum_{b\in S_j}\alpha_j(b)T_{jp_i}(b|c). \end{aligned} \tag{5.44}$$

The hidden state of node i can then be sampled from the posterior conditional on the sampled state of the parent c

$$P(x_i = a|D, x_{p_i} = c) = \frac{\beta_i^p(a,c)\alpha_i(a)}{\sum_{a\in S_i}\beta_i^p(a,c)\alpha_i(a)}. \tag{5.45}$$

Like the calculation of the outside probabilities, the samples are propagated down the tree.

Application of the Inside-Outside algorithm to family trees

A family is not strictly a directed tree, due to the addition of new founders (via marriage) in each generation. However, we can make a family into a directed tree by treating parent couples as a single node, consisting of a founder and a non-founder individual. The root node of this directed family tree consists of the top pair of founders. While I have currently only used this method for family trees with only one founder-founder couple, in fact any family relationships that do not include inbreeding (i.e. any that take the form of a polytree) can be modelled if the polytree is transformed to a directed tree by reversing the transition matrix (using $T_{p_i}(a|b) = T_i(b|a)P(x_{p_i})/P(x_i)$).

We use the Inside-Outside algorithm to sample unobserved genotypes conditional on all other genotypes for a single biallelic polymorphism with allele frequency f (although this is readily generalised to an arbitrary number of independent polymorphisms). We model individuals as nodes, and genotypes as hidden states for each node. For non-parent couples the state-space is

$$x_i \in S_i = \{AA, AB, BB\}, \tag{5.46}$$

and for parent couples it is

$$x_i = (x_i^f, x_i^{nf}) \in \{AA, AB, BB\}^2, \tag{5.47}$$

where x_i^f is the founder's genotype state and x_i^{nf} is the non-founder's genotype.

Genotype calls for each individual are modelled as emissions, and we assume that these genotypes are certain and thus for genotyped individuals x_i and d_i are identical (though genotype error can be included by modifying the likelihoods below). Genotypes can also be missing (N). Thus the emissions for a non-parent couple node is

$$d_i = g_i, \tag{5.48}$$

and for parent couples is

$$d_i = \{g_i^f, g_i^{nf}\}. \tag{5.49}$$

Likelihoods for non-parent couples are

$$L_i(a) = \begin{cases} 1 & \text{if } a = g_i \text{ or } g_i = N; \\ 0 & \text{otherwise.} \end{cases} , \tag{5.50}$$

and for parent couples are

$$L_i(a) = \begin{cases} 1 & \text{if } a_i^f = g_i^f \text{ and } a_i^{nf} = g_i^{nf}; \\ 1 & \text{if } a_i^f = g_i^f \text{ and } g_i^{nf} = N \text{ or } a_i^{nf} = g_i^{nf} \text{ and } g_i^f = N; \\ 1 & \text{if } g_i^f = g_i^{nf} = N; \\ 0 & \text{otherwise.} \end{cases} . \tag{5.51}$$

Transitions can only occur from a parent couple to a non-parent couple, or from a parent couple to a parent couple. For a parent couple to a non-parent couple, transmission is simple Mendelian inheritance

$$T_{ij}(a|b) = P(C = a|P1 = b^f, P2 = b^{nf}), \tag{5.52}$$

where C is the child's genotype, and $P1$ and $P2$ are parental genotypes. For parent couple to parent couple transmission, we need to include the probability density on the founder genotype

$$T_i(a|b) = P(C = a^{nf}|P1 = b^f, P2 = b^{nf})P(a^f|f), \tag{5.53}$$

where $P(a^f|f)$ is the population frequency of the founder's genotype, assuming Hardy-Weinberg equilibrium. Finally, the prior on the root node is given by the population frequency

$$\pi(a) = P(a^f|f)P(a^{nf}|f). \tag{5.54}$$

Using this formulation, marginal posteriors can be calculated for each unobserved genotype, and joint genotypes for the entire family can be sampled from the joint posterior distribution.

Mangrove: An R package for risk prediction in families

To summarise the above approach, we can calculate the probability of seeing at least y affected families members in a family given known GWAS risk loci $P(y|G_{obs}, \beta, f)$ using the following process:

1. Convert the family tree with genotype data into a true directed tree with emissions as described in section 5.3.2
2. Calculate α_i, β_i and β_i^p statistics using the Inside-Outside algorithm as described in section 5.3.2
3. Sample N sets of genotypes for ungenotyped family members using the method in 5.3.2
4. Sample affection status for each individual conditional on samples genotyped, using standard risk prediction (Chapter 2)
5. Count the number of families with more than y affected family members

These stages have all been implemented in the R package Mangrove, which is available from the Comprehensive R Archive Network (CRAN). Mangrove is specifically designed to use genetic risk prediction to prioritise individuals or families for sequencing. As well as risk prediction in families, Mangrove can also perform both risk prediction and quantitative trait prediction in unrelated individuals. I have provided detailed documentation, and a vignette containing usage examples for both families and unrelated individuals, with the package.

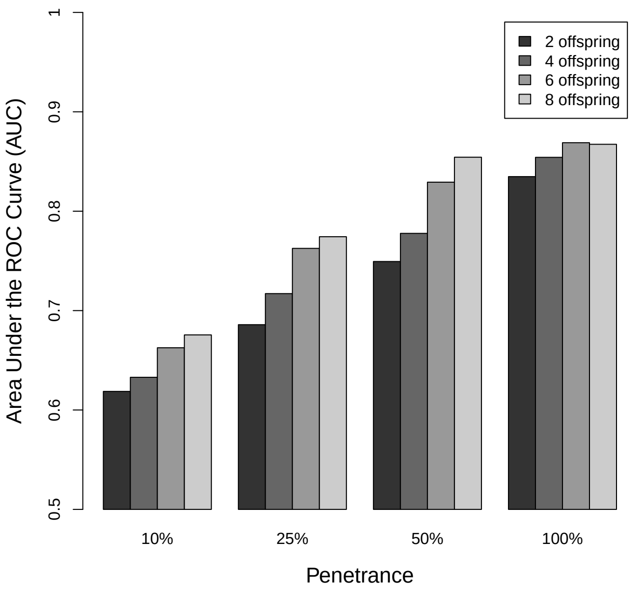


Figure 5.2: Ability to predict the presence of a high penetrance mutation (measured by AUC) in multiplex families using a polygenic risk score. We assume a disease with a prevalence of 1%, a heritability of 50%, and a genetic risk score that captures 12.5% of variance. All families have three affected individuals, and the AUC is shown for families of different total size and dominant mutations of varying penetrance.

Assessing the efficacy of risk prediction in families in prioritising penetrant mutations

The aim of the risk prediction prioritisation described above is to increase the chance that a family selected for sequencing carries a high penetrance mutation. To investigate how powerful this approach is I performed simulations of families with and without a high penetrance mutation.

Consider two families both subject to polygenic risk for a disease and one additionally containing a high penetrance dominant mutation. We would like to be able to identify the latter family for the type of family sequencing experiment described above. To evaluate the ability of the above method to identify families containing such high penetrance mutations I simulated nuclear families with between 2 and 8 offspring, where three total family members were affected by a disease having 1% prevalence and heritability of 50% (these values correspond approximately to immune mediated diseases such as Crohn's disease). Half the families contained a dominant mutation with a penetrance from 10–100%, and the other half arose simply from polygenic risk and chance.

For each family, we computed the value of equation (5.30) based on a GWAS risk predictor explaining 25% of heritability (again by analogy to Crohn's disease). Figure 5.2 shows the area under the ROC curve (AUC), which in this instance can be interpreted as the probability of correctly distinguishing between one family with a penetrant mutation and one without. For a low-penetrance mutation in a small family AUC is only $\sim$0.6, but for a medium-penetrance mutation in a large family, AUC is $\sim$0.85, which would provide a substantial advantage over simply selecting the family with the largest number of affected individuals.

5.4 Linkage and sequence analysis of a multiplex IBD family

We have seen how multiplex families are likely to show an enrichment for rare, high penetrance risk variants. This is particularly true for multiplex families that span extended pedigrees, and in pedigrees with a low predicted risk given common variants. Via linkage and haplotyping methods, these families can also be analysed for candidate regions that may harbour such mutations. The falling cost of sequencing means that whole-exome or whole-genome sequencing can then be used to attempt to identify causal candidates in the family using linkage data and functional information.

To attempt to discovery such high penetrance mutations, we collected samples from extended families with multiple members affected by inflammatory bowel disease (IBD). Here I discuss the analysis of one such family.

Note that some non-important details of the family have been altered in this chapter to ensure anonymity. These include the gender of subjects, the number of offspring and the details of family relationships. In no case does this affect the conclusions drawn, though it may lead to small inconsistencies in the precise details of results.

5.4.1 Description of the family

The family comprises over 800 individuals of Ashkenazi Jewish descent, spanning four generations connected via a founding couple born at the turn of the 20th century (Figure 5.3). The family is characterised by its large number of offspring per parental couple, with an average of 9. The founding couple had seven offspring (including two identical twins), six of these have at least two descendants with IBD.

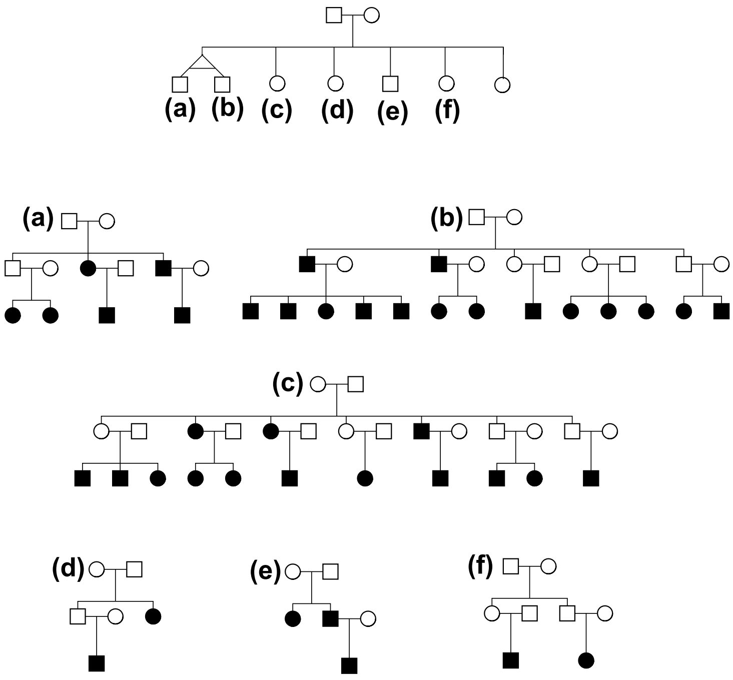


Figure 5.3: A pedigree for the family under study, showing affecteds and parents of affecteds. The top figure shows how the founders of the six subpedigrees (a-f) are related. The founders of subpedigrees a) and b) are identical twins.

A total of 41 individuals have been diagnosed with IBD, including 35 with a diagnosis of Crohn's disease and 7 with a diagnosis of ulcerative colitis. We were able to independently confirm the diagnosis via medical records in all but five cases. The location of disease in the bowel was variable. The average age of onset was 18.8 years (95% CI: 16-22, n=30) and at the time of sample collection, one-quarter of the patients had undergone surgical resections.

This family is a good candidate for discovering a high penetrance mutation. They have a wide geographic distribution, with affected individuals present in seven cities around the world, making an environmental cause of the disease less likely. Additionally, because the affecteds are spread across first and second cousins, polygenic risk is far less likely to explain the large number of affecteds.

5.4.2 Segregation analysis

Before looking at any genetic data, we can use the structure of the family to make a plausible guess at what sort of genetic risk factors we may be looking for. We will look specifically at subfamilies (a) and (b) as the identical twin founders make the analysis significantly easier.

Suppose we take the most optimistic view of the genetics of this family, i.e. that all cases are explained by a single dominant mutation. Together, the two identical twins have 18 offspring, of which 10 are either affected, or have affected children (or both). The most favourable model would be to suppose that these 10 individuals all inherited a causal mutation from these identical twins, and the rest did not. Furthermore, we will assume that all affected family members carry this mutation.

Under this favourable model, 9 parents and 18 affected children, as well as approximately half of their 66 unaffected siblings, will carry the mutation, of which 21 have the disease. This gives a penetrance of 35% (21 out of 60). In fact, as discussed in section 5.5.2, unaffected siblings are less likely to inherit a causal mutation. If we correct for this, the estimated penetrance in the highly favourable model is 41%, with a 95% confidence interval of 24-48%.

This model is almost certainly overly optimistic, as in a family of this size many of the cases are likely to be phenocopies, and likewise causal mutations may be segregating in parts of the family with no affecteds. It is also possible that the mutation is recessive, interacts with another risk factor (either genetic or environmental), or is only one of many undiscovered risk factors in the family. However, the model does illustrate how, even in the best-case scenario, we are looking for a mutation with incomplete penetrance (<50%).

Family	N	y	$E(y\|K)$	$E(y\|G)$	$\frac{y}{E(y\|G)}$	$P(y\|G)$
Whole family	806	41	6.04 (1 - 11)	10.24	4.00	$< 10^{-4}$
Subfamily (a)	112	6	0.84 (0-3)	1.02 (0-4)	5.90	0.0012
Subfamily (b)	112	15	0.84 (0 - 3)	0.97 (0-4)	15.42	$<10^{-4}$
Subfamily (c)	140	14	1.05 (0 - 3)	1.56 (0 - 5)	8.97	$<10^{-4}$
Subfamily (d)	147	2	1.10 (0-3)	1.24 (0-4)	1.62	0.352
Subfamily (e)	81	3	0.61 (0 - 2)	1.63 (0 - 5)	1.84	0.243
Subfamily (f)	138	2	1.04 (0 - 3)	3.11 (0 - 7)	0.74	0.706

Table 5.1: A Mangrove analysis of the IBD family, including analyses of the six subfamilies. N is the total number of individuals in this subfamily, y is the number of affected individuals, $E(y|K)$ is the expected number of affected given the prevalence alone, $E(y|G)$ is the expected number given genotyped common variants. $\frac{y}{E(y|G)}$ is the enrichment of cases over that predicted by common variants, and $P(y|G)$ is the probability of observing y or more affected in this pedigree given common variation. Numbers in brackets are 95% confidence intervals.

5.4.3 Known IBD risk variants in the family

We successfully genotyped 38 CD and UC risk variants in 152 family members across the entire family in order to assess the extent to which the increased incidence may be explained by known genetic risk factors. I used odds ratios and frequencies taken from the IIBDGC GWAS meta-analysis data (using only Jewish samples), except for the 3 *NOD2* variants for which I used the Immunochip data (described in Chapter 4). Together, these variants explain 7.8% of variance in CD liability and 2.0% in UC liability.

I used the R package Mangrove (described in Section 5.3.2) to assess the number of cases we would expect in the family given these common variants. I used population prevalence of CD and UC of 0.6% and 0.15%, collected by Adam Levine from Jewish patients in GP surgeries in North London (personal communication).

Compared to the baseline prevalence, the family shows a 6.8-fold enrichment in IBD. While the family does show a marked increase in risk (1.7-fold)

SNP	OR_{het}	OR_{hom}	P-value
rs2066844	1.83	8.65	1.6 x 10^{-12}
rs2066845	1.90	11.61	2.1 x 10^{-4}
rs2066847	2.56	29.8	1.8 x 10^{-16}
Compound heterozygous (Excess odds ratio over additivity)	x3.46	-	2.0 x 10^{-55}

Table 5.2: Odds ratios for *NOD2* mutations under a non-additive model, fitted from the IIBDGC Immunochip data described in Chapter 4. The p-values give the significance of the full model compared to a model with this term replicated with a purely additive term.

due to common risk variants, there is still a 4-fold enrichment in IBD even given these common variants (Table 5.1).

We can further break this down by subfamily (Table 5.1). Subfamilies (d)-(f) show a particularly marked enrichment in common risk variants, which would predict a 2.2-fold increase in prevalence. The expected number of affected given common risk variants (5.98) is remarkable close to the observed number (7), suggesting that there is unlikely to be any high penetrance mutations in this area of the family. By contrast, subfamilies (a)-(c) show a very large gap between the predicted and actual number of affecteds (9.9 times that predicted by common variants), suggesting that these subfamilies are good candidates for harbouring high penetrance mutations.

Modelling non-additivity in NOD2 risk variants

One complication is that the above analysis assumes an additivity genetic architecture. While this model fits most of the IBD risk variants well, it does not accurately model the *NOD2* risk variants, which show significant evidence of both recessive effects at single coding variants and epistatic interaction between coding variants (Table 5.2).

In subfamilies (a) and (b) *NOD2* mutations are relatively uncommon,

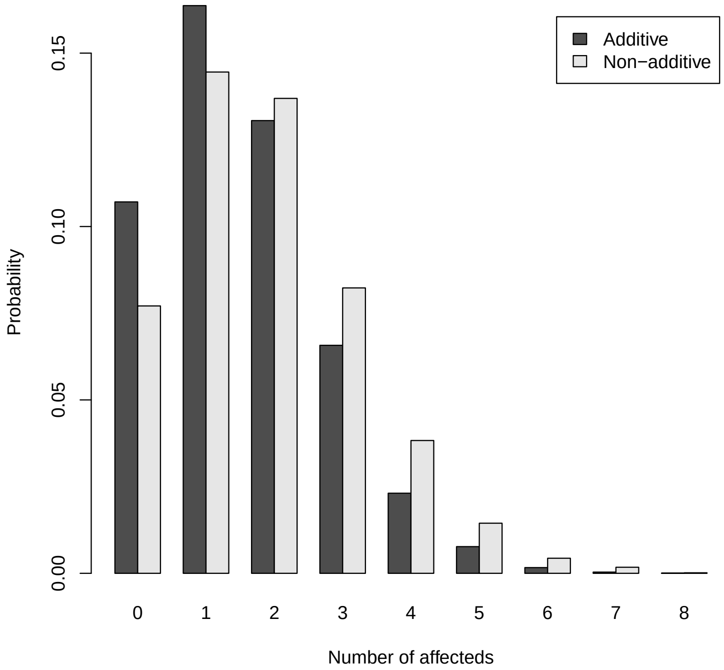


Figure 5.4: The distributed of cases expected in subfamily (c) under an additive and a non-additive model of NOD2 risk.

and no individuals were homozygous or compound heterozygous for *NOD2*, suggesting that the non-additive model will only decrease the total number of predicted affecteds. However, in subfamily (c) seven individuals are either homozygous or compound heterozygous for one of the three classical *NOD2* mutations, suggesting that the contribution of known genetics in this family could be larger than an additive analysis suggests.

I used data from the IIBDGC Immunochip dataset (described in Chapter 4) to fit a non-additive *NOD2* model by logistic regression (Table 5.2), and used the Mangrove method to perform risk prediction in subfamily (c) using this model. Non-additivity increases the expected number of affecteds slightly, from 1.56 to 1.88 ($p = 5.5 \text{ x } 10^{-11}$). However, the real increase is on the extremes (Figure 5.4), where the probability of seeing 6 or more affecteds increases by a factor of three (from 0.4% to 1.3%). Despite this increase, the

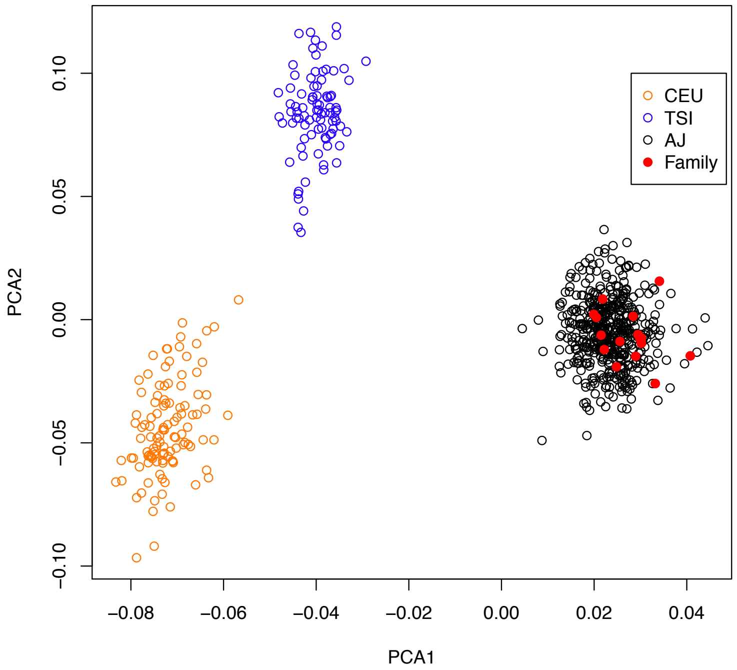


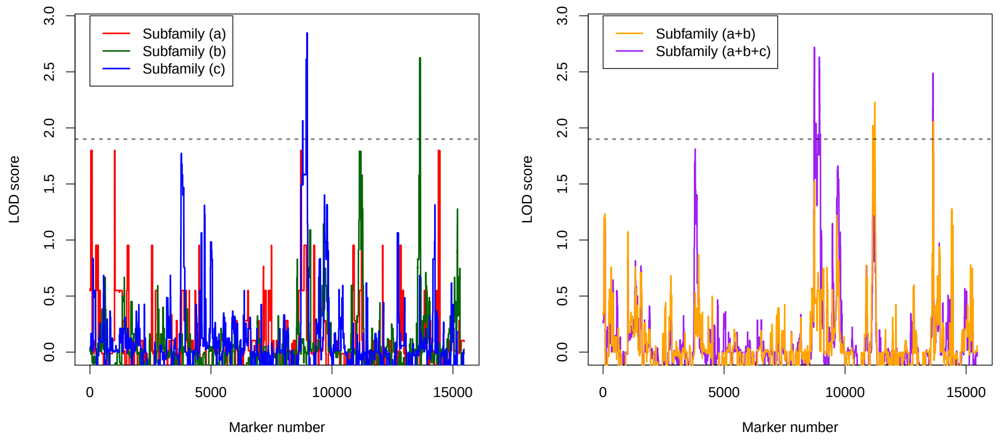
Figure 5.5: A principal component analysis of the family, using HapMap populations (TSI=Italian, CEU=Northern European) and Ashkenazi Jewish (AJ) reference populations.

probability of seeing 14 affecteds in subfamily (c) given common variation remains very small ($<< 10^{-4}$).

5.4.4 Linkage and haplotype analysis of the family

Genotyping data

A total of 60 individuals (30 affected and 30 unaffected) from subfamilies (a)-(c) were genotyped on an Illumina CytoSNP 12 BeadChip array. Genotypes were called using BeadStudio. Genotypes inconsistent with Mendelian segregation were set to missing, and SNPs with greater than 1% missingness, minor allele frequency less than 1% in founders or Hardy-Weinberg Equilibrium p-value less than 10^{-5} in founders were removed.



(a) Within subfamilies

(b) Across subfamilies

Figure 5.6: Non-parametric linkage results for the family.

As a reference population, we used genetic data from a study of 471 Ashkenazi Jewish individuals genotyped on the Affymetrix Human SNP Array 6.0 (Bray et al., 2010), obtained via the NCBI's Gene Expression Omnibus (GEO) database (Barrett et al., 2011). Principal component analysis confirmed that the family members clustered with the Ashkenazi reference population (Figure 5.5).

We created a 1cM maximally informative genetic map by taking all SNPs present in both the reference set and the family, and for which there was no missing data in the family. We performed LD thinning in the reference dataset (such that $r^2 < 0.2$ for all SNPs). We then selected the SNP with the highest heterozygosity in the family founders in every 1cM block. Allele frequencies for these SNPs were calculated from the reference set.

Linkage analysis

We performed non-parametric linkage using Merlin (Abecasis et al., 2002) (v1.1.2). As we expect large increases in allele sharing due to high penetrance mutations, the standard linear approximation used by Merlin is too conservative, so we used the more accurate Kong and Cox exponential model (Kong and Cox, 1997). We used the maximally informative map and allele frequencies described above.

We ran linkage separately on the three subfamilies (a)-(c). We also used Fisher's method to combine the results for subfamilies (a)-(b) (i.e. the offspring of the identical twins), and for all subfamilies (a)-(c). The results are shown in Figure 5.6. None of the results meet the criteria for genome-wide significance (a LOD score of 3.3 (Lander and Kruglyak, 1995)). A number of linkage peaks reached the level of significance that Lander and Kruglyak (1995) suggest can be interpreted as "suggestive evidence" (a LOD score of 1.9). These are shown in Table 5.3.

The linkage peaks inferred are broad, and contain many genes. Even if we reduce this down to genes that are expressed in the immune or digestive systems, there are still between 7 and 89 genes in each linkage peak (Table 5.3). Low-throughput sequencing of exons in some of these candidates did not produce any likely candidate causal variants.

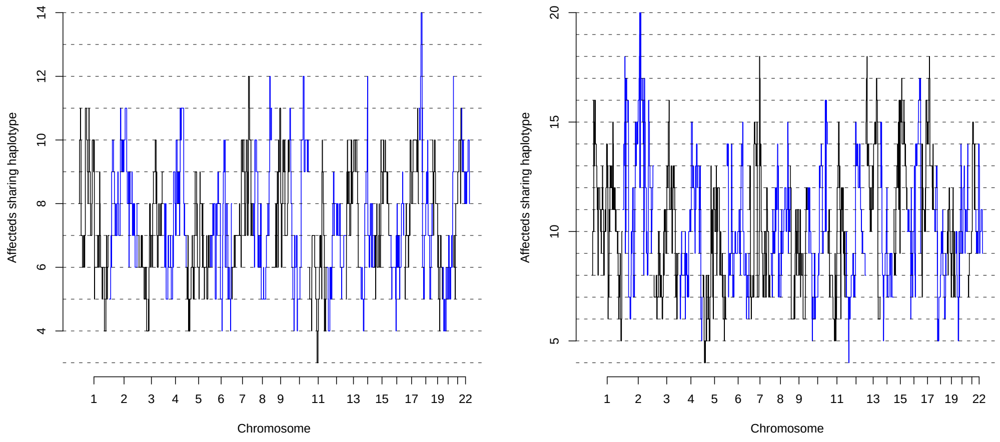
Haplotype analysis

As well as using the genotype data to find evidence of significant linkage, we can also use it for the related purpose of inferring the flow of haplotypes within the family. This can allow us to identify regions of the genome that are widely shared across subfamilies, and identify which family members do and do not share a candidate mutation on a particular haplotype. It can be used to inform the analysis of sequence data.

The computing resources required to carry out a full haplotype analysis grows exponentially with the number of samples. As a result, directly infer-

Chr	Pos in Mb	LOD score (subfamilies)	P-value	Genes (expressed)
Subfamily (b)				
18	6.98-9.71	2.62	2.54 x 10^{-4}	10 (7)
Subfamily (c)				
10	72.59-82.39	2.81	1.62 x 10^{-4}	81 (23)
Subfamily (a)+(b)				
13	89.61-96.75	2.23 (0.95, 1.58)	6.78 x 10^{-4}	24 (8)
18	6.98-9.71	2.05 (0.01, 2.62)	1.05 x 10^{-3}	10 (7)
Subfamily (a)-(c)				
10	19.17-81.96	2.72 (1.80, 0.12, 1.49)	2.01 x 10^{-4}	256 (89)
18	6.99-9.71	2.49 (0.01, 2.62, 0.69)	3.57 x 10^{-4}	10 (7)

Table 5.3: Suggestive linkage peaks (LOD > 1.9) in the family. Positions are given as the region in which markers have LOD > MAXLOD - 1. Numbers in brackets are LOD scores of the individual subfamilies that went into the analysis. The number of genes expressed in either the immune or digestive systems in the linkage peak is calculated from the expression datasets described in section 5.4.7



(a) Subfamilies (a)+(b)

(b) Subfamilies (a)-(c)

Figure 5.7: Haplotype sharing in affecteds across the genome for subfamilies (a)+(b) (of 18 total) and subfamilies (a)-(c) (of 31 total).

ring haplotypes across subfamilies using Merlin was not possible. Instead, we developed a method for parallelising the calculating of haplotypes across subfamilies, involving the following steps:

1. Perform haplotype analysis in two subfamilies separately
2. For each pair of individuals across the two subfamilies, produce a small pedigree consisting of siblings of these individuals, and ancestors that connect them together. Use this to perform genome-wide identity-by-descent estimation in these two individuals.
3. For every possible set of haplotype assignments at every point in the genome, calculate the difference between the calculated identity-by-descent value and the value predicted by the haplotypes generated in step 1, summed across all pairs of individuals.
4. At each position in the genome, pick the haplotype assignment that minimises this value

We carried out this analysis on subfamilies (a)+(b) using this method, and on subfamilies (a)-(c) by then matching up haplotypes between subfamilies (a)+(b) and (c).

Haplotype sharing in subfamilies (a) and (b)

The maximum number of affected family members sharing the same haplotype across the genome for subfamilies (a) and (b) is shown in Figure 5.7a. The most widely shared haplotype is on chromosome 18 (corresponding to the suggestive linkage peak in Table 5.3), and is shared by 14 of the 18 genotyped affecteds. This haplotype is present in all five affected nuclear families in subfamily (b), and two of the four in subfamily (a).

Using the same approach as described in section 5.4.2, we can use this haplotype information to estimate the potential penetrance of a dominant mutation that lies on this haplotype. This model produces an estimate of the penetrance of 39% (95% CI 27-56%). It also implies between 4 and 7 phenocopies, corresponding to a phenocopy rate of 2.6% (95% CI 1.0-6.3%). While this is elevated compared to the population prevalence, this may be partly explained by ascertainment bias: this family, and in particular this subfamily, was selected for investigation due to the large number of affecteds, and this is likely to slightly inflate the number of affecteds due to winner's curse.

Haplotype sharing in subfamilies (a)-(c)

The maximum degree of haplotype sharing in subfamilies (a)-(c) is found on chromosome 2 (between 13.3Mb and 14.3Mb). This does not correspond to any of the suggestive peaks in the linkage analysis. This haplotype is shared across 10 of the 16 affected nuclear families, and affects 20 of the 31 genotyped affecteds in this part of the family.

A dominant causal mutation on this haplotype could have a relatively high penetrance (48%, 95% CI 36-64%). However, it would also imply between 11 and 14 phenocopies, corresponding to a phenocopy rate of 4.2% (95% CI: 2.4%-7.1%). This is more than 5-fold higher than the population prevalence, and 4-fold higher than the rate predicted from common risk variants in this part of the family, suggesting that a dominant mutation on this haplotype alone would be insufficient to explain the incidence of IBD in this family.

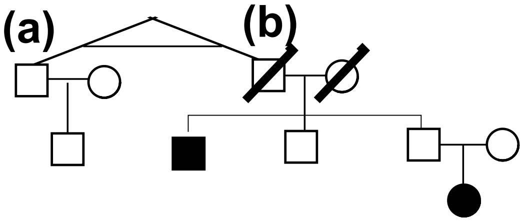


Figure 5.8: The founder subpedigree used for whole-genome sequencing.

5.4.5 Whole-genome sequencing in the family

Samples chosen for sequencing

Whole-genome sequencing allows a complete survey of variation within a family. It allows us to characterise structural variation, as well as SNPs and indels in non-coding DNA that may have a regulatory function. However, the cost is substantially higher, and thus we can only perform sequencing on a limited number of individuals. We decided to concentrate on subfamilies (a) and (b), as they are descended from two identical twins. This both increases the chance that a shared mutation is acting in both families, and reduces the cost of sequencing (because two founders can be sequenced for the price of one).

Figure 5.8 shows the 8 samples that we decided to sequence. These samples have been picked to capture the shared haplotypes introduced by the identical twins who founded subfamilies (a) and (b). Additionally, we included enough offspring to allow us to assign mutations to haplotypes, and thus allow us to impute variants on shared haplotypes into all affected members of subfamilies (a) and (b).

Generating and quality controlling raw sequence

We performed whole-genome sequencing using the Illumina HiSeq 2000, generating 2x100bp reads. A total of 407.5Gb of sequence was generated, and

Call set	SNPs	% dbSNP	Ts/Tv
Union	7.46M	79.5%	1.76
Intersection	6.09M	90.2%	2.07
VQSR (99%)	5.86M	92.2%	2.04
VQSR (90%)	5.16M	94.7%	2.12

Table 5.4: Summary statistics for various whole-genome sequencing call sets

aligned to build 37 of the human genome using BWA (Li and Durbin, 2009) v0.5.9. The mapping rate was 95.49% (range 94.07-96.35%), and the average coverage across the eight individuals was 16.1X (range 12.3 - 23.6X).

QC of the sequence data was performed using the BAMCheck pipeline developed by Petr Danecek, and all sequencing lanes passed. Samples were checked against their CytoSNP12 genotyping data (described above) to assure that samples swaps had not occurred. GATK (McKenna et al., 2010) v1.2 was used to perform local realignment around known indels, and to recalibrate base pair quality scores.

Calling SNPs and indels

Raw lists of SNPs and indels were generated using the GATK UnifiedGenotyper and samtools mpileup (Li et al., 2009) (v0.1.17). A total of 7.46M SNPs and 1.50M indels were called, of which 82% and 53% respectively were called by both approaches. This union SNP set is relatively poor: over 20% of SNPs are not seen in dbSNP, and the transition to transversion ratio (which should be above 2) is only 1.76 (Table 5.4). To improve the dataset, we carried out Variant Quality Score Recalibration (VQSR) using GATK. This technique fits a mixture model of true and false positive variants using QC metrics and a truth set of known polymorphic variants, and uses this to produce a calibrated quality score (the VQSLOD) for each variant.

Statistic	Call sets	Description
QD	SNP/Indel	Variant quality divided by depth
HaplotypeScore	SNP/Indel	Data consistency with exactly two haplotypes per individual
MQ	SNP	RMS mapping quality of reads mapping to site
MQRankSum	SNP	Test statistic for bias in MQ
DP	SNP	Total depth of reads at site
FS	SNP/Indel	Test statistic for bias in strand
ReadPosRankSum	SNP/Indel	Test statistic for bias in position in read

Table 5.5: QC statistics used for VQSR. In all cases "bias" refers to a difference in reference and non-reference reads. RMS stands for "root-mean-square", i.e. $\sqrt{\frac{1}{N}\sum_i^N x_i^2}$.

We used a variety of QC statistics as input for VQSR (Table 5.5). For SNP truth datasets, we used HapMap3 and 1000 Genomes Omni2.5 polymorphic sites, and for an indel truth dataset we used indels observed twice in the Mills and Devine (Mills et al., 2011a) dataset. A total of 5.86M SNPs and 1.22M indels passed the basic VQSR filter (VSQR99, equivalent to VQSLOD $>$ 2.52 for SNPs and $>$ 0.13 for indels), and these call sets had very favourable statistics (Table 5.4). A more stringent level of filtering (VQSR90, equivalent to VQSLOD $>$ 5.18 for SNPs and VQSLOD $>$ 3.20 for indels) provides a very high quality dataset at the expense of calling fewer variants.

We can use the CytoSNP 12 genotype data to test the sensitivity of the SNP call sets. Figure 5.9 shows this sensitivity as a function of non-reference allele count. As well as showing good quality statistics, the VQSR datasets have a very high sensitivity: the basic VQSR99 set has a 99.7% sensitivity for variants present in at least two individuals, and the stringent VQSR90 set, while less sensitive, still has a very high sensitivity (99.0%). A caveat to this analysis is that the CytoSNP 12 was designed late in the Illumina BeadChip

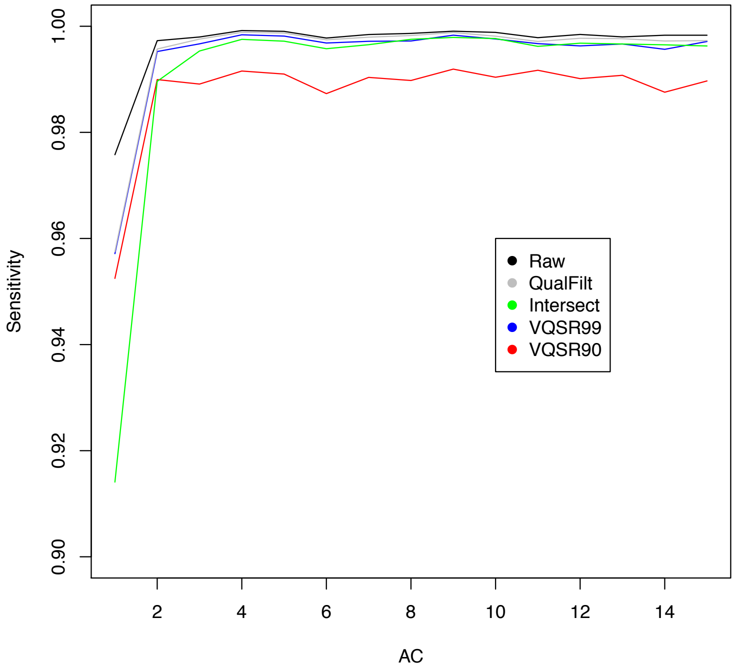


Figure 5.9: The sensitivity of the various WGS call-sets compared to array genotyping, as a function of non-reference allele count (AC).

line (in 2008) in order to genotype low concentrations of DNA, and as such is strongly biased towards "genotypeable" (i.e. complex, well-behaved) SNPs. The sensitivity values should thus be considered the sensitivity to detect "easy" SNPs.

Calling structural variants

Unlike for SNP and indel calling, there is no single well-established method for calling structural variants (SVs) from sequence data. Instead, most SV calling efforts combine information from a range of different complementary calling methods (Mills et al., 2011b).

To call SVs from the whole-genome sequencing data I used six different

Method	Insertions	Deletions	Inversions	Complex
BreakDancer	0	4630	517	0
CNVnator	2816	17371	0	0
Pindel	2573	2574	165433	0
RDXplorer	491	335	0	0
SECluster	1347	0	0	0
Genome STRiP	0	1377	0	0
SVMerge confirmed	814	3519	19184	8355

Table 5.6: Summary statistics for the different whole-genome sequencing structural variant callsets, along with the combined SVMerge set

calling methods to generate candidates. These included two methods that call SVs based on read-depth (RDXplorer (Yoon et al., 2009) and CNVnator (Abyzov et al., 2011)), two that call based on paired end reads (BreakDancer (Chen et al., 2009) and SECluster (Wong et al., 2010)), one that uses a combined read-length and paired-end method (Genome STRiP (Handsaker et al., 2011)) and one that calls based on split reads (Pindel (Wong et al., 2010)). We used the program SVMerge to combine these candidates together into a single set. We used the recommended SVMerge settings for filtering candidate sets, and removed calls that overlapped centromeres, teleomeres or gaps in the reference. The merged list of variants was then checked by local assembly (using the assembly program Velvet (Zerbino and Birney, 2008)) to confirm breakpoints. A breakdown of the number of variants called is shown in Table 5.6. Note that a very large number of inversions and complex events are called, coming almost exclusively from Pindel. As Pindel already uses local realignment, the 19,184 inversions could not actually be confirmed by an independent method, and should thus be considered suspect.

A total of 1210 SVs had at least a 50% reciprocal overlap with known structural variants (taken from Zhang et al. (2006), Conrad et al. (2010) and Mills et al. (2011b)). Of these, 179 of the 814 insertions had been previously

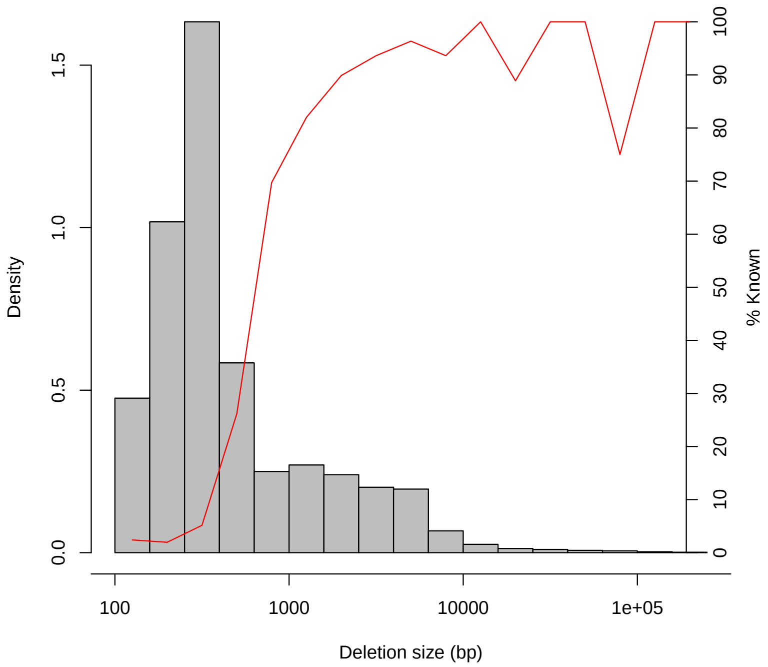


Figure 5.10: The distribution of deletion size in our call set, combined with the proportion observed (with >50% reciprocal overlap) in at least one of the three external datasets.

discovered, and 968 of the 3519 deletions. However, only 23 of the inversions and 37 of the complex events were previously known, suggesting again that these classes of variants are unreliably called. We decided that the likely very high false positive rate in inversions and complex events made them unreliable, and discarded them.

Looking in detail at the deletions, the number of called mutations also seen in the databases varies widely with the size of deletion (Figure 5.10) . 88.8% of deletions sized between 100 and 1000bp are novel, compared to only 9.6% of deletions greater than 1000bp. This likely represents a combination of false negatives in the database (for instance, Conrad et al. (2010) only examined SVs larger than 443bp), and false positives in our call set.

5.4.6 Whole-exome sequencing in the family

Samples chosen for sequencing

Whole-exome sequencing is a more limited approach than whole-genome sequencing, and only allows the assessment of small-scale variation in coding regions. However, the substantially lower cost means that many more samples can be sequenced, potentially allowing a far more extensive study of coding variation than can be afforded by whole-genome sequencing.

All affected individuals from the family with DNA available (a total of 40) were sequenced, along with 13 unaffected family members to allow phasing. Additionally, we sequenced 26 control exomes, taken from unaffected members from the same ethnic group and geographic region as the family, to allow us to identify population-specific variation that may otherwise be mistaken for risk variants.

Processing of whole-exome sequencing data

We performed whole-exome sequencing, using a SureSelect Human All Exon 50 Mb kit for target enrichment and the Illumina HiSeq 2000 for sequencing. We used the same pipeline for quality control, mapping, realignment, recalibration and variant calling that was developed for the whole-genome sequencing (sections 5.4.5 and 5.4.5). The samples had a mean coverage of 154.0X in the target region (range 131.2X - 186.4X).

The VQSR99 set contained 128410 SNPs (87% known, Ts/Ts = 2.84), of which 105243 were also in the VQSR90 set (89% known, Ts/Tv = 2.96). The indel dataset contained too few indels to apply VQSR, so instead we used the default GATK hard-filters to create a high-quality indel set. This included 9906 indels (52% known).

5.4.7 Identifying candidate variants in the family

Identifying candidate mutations

Between the whole-genome and whole-exome sequencing we have called over 7.5 million SNPs, indels and structural variants. Given the analyses reported above, we can be nearly certain that there exists, somewhere in this list, at least one mutation that causes a substantial increase in risk of inflammatory bowel disease. To identify such mutations, we need to filter out the vast majority of variants that do not contribute to IBD risk.

We have developed separate filtering procedures for the three different classes of variants: coding SNPs and indels, non-coding SNPs and indels and structural variants (laid out in detail in Table 5.7). Each filtering procedure begins with a platform-specific quality filter to remove poorly performing variants, followed by the removal of high-frequency variants using various databases of common variation.

Our next stage is to filter out any variants that are not present in at least half of the family members being considered. In the case of the data deriving from whole-genome sequencing we infer this from the haplotype flow information discussed in Section 5.4.4. For the SNPs and indels, we examine the consistency of the genotypes with what would be expected if the variant lay on the maximimally shared haplotype at that point in the genome. If the genotypes are consistent with this haplotype (given at most one genotyping error), and the haplotype is shared by at least half of affecteds in subfamilies (a)+(b), we include the variant. This approach has the notable

Filter	Description
Filters for coding variants:	
High quality	Genotype quality > 10 in at least 60% of samples
Uncommon coding variant	Frequency <2.5% in ESP^a, and less than <5% in our 26 AJ controls (annotated using ANNOVAR^b)
Affected sharing	Is shared by at least 50% of sequenced affecteds in either subfamilies (a)+(b), subfamily (c), or the entire family
Coding consequence	Is a missense, nonsense, essential splice, stop or frameshift mutation (annotated using Ensembl VEP^c)
Deleteriousness	Predicted to be deleterious to protein function (measured using Condel^d).
Filters for non-coding variants:	
Haplotype consistency	Genotypes are consistent with maximally shared haplotype in linkage data (given at most one genotyping error).
Uncommon variant	Has an non-reference allele frequency <2.5% in 1000 Genomes Phase 1 Europeans^e
Haplotype sharing	Variant is predicted to lie on a haplotype shared by at least 9 affected members of subfamilies (a) and (b)
Conserved	GERP^f score > 2 or phastCons^g score > 0.5, using UCSC vertebrate alignments^h
Regulatory function	Within an Ensembl regulatory region (via VEP^c) or within both a transcription factor binding site (TFBS) and a region of open chromatic (DNase1) in at least one ENCODE cell line^i (via UCSC^j)
Filters for structural variants:	
Novel	Does not have >50% reciprocal overlap with a variant in Conrad *et al*^{k}, 1000 Genomes^l or HGV^m.
Not a CNV region	Overlaps no more than 5 variants in HGV^m
Haplotype sharing	Variant overlaps a haplotype shared by at least 9 affected members of subfamilies (a) and (b)
Potential function	Overlaps at least one coding base
Filters for all variants:	
Genic variant	Overlaps a gene region in GenCode release 7^n
Expressed gene	Gene is expressed in at least one immune or gut tissue type, either in the Gene Expression Barcode^o or our gene expression datasets.

Table 5.7: Filters used to identify candidate causal variants. [a]NHLBI GO Exome Sequencing Project (ESP) (2012). [b]Wang et al. (2010) [c]McLaren et al. (2010) [d]Gonzalez-Perez and Lopez-Bigas (2011) [e]Project (2012) [f]Davydov et al. (2010) [g]Siepel et al. (2005) [h]Dreszer et al. (2012) [i]The ENCODE Project Consortium (2012) [j]Rosenbloom et al. (2012) [k]Conrad et al. (2010) [l]Mills et al. (2011b) [m]Zhang et al. (2006) [n]Harrow et al. (2006) [o]McCall et al. (2011)

advantage of allowing us to assess the variant in more affecteds than were sequenced. However, if a causal variant has been introduced to the family multiple times on separate haplotypes, this variant will be missed (in the family this is true for the *NOD2* mutations, for example). Thus for the exome sequencing, where data is available for nearly all affecteds, we did not use the haplotype information, instead directly counting the number of affected individuals carrying each haplotype.

The next stage involves removing variants that are unlikely to have a functional impact. Coding SNPs and indels are filtered based on their predicted impact on protein function. Non-coding SNPs and indels from the whole-genome sequence are filtered based on their level of evolutionary conservation and their presence in putative regulatory features. Structural variants are filtered based on whether they delete coding sequence.

The final stage is to remove variants that, while possibly functional, are unlikely to be functionally relevant to IBD risk. We use two sets of gene expression data (one public reference set, one dataset generated by us) to identify genes that are expressed in tissues relevant to IBD (tissues of the immune or digestive systems). All mutations are filtered out if they do not overlap a gene identified as expressed in a relevant tissue.

In the next three sections I will describe the results of this filtering on the three different classes of variant, and discuss some of the candidate variants that this analysis uncovers.

Coding SNPs and indels

Across the entire family there were 7,626 protein-changing mutations that are at low frequency in the general population. Of these, 223 were shared by at least 50% of affecteds in at least one subfamily, and 36 were implicated as

Filter	SNPs	Indels
Low frequency protein mutations	7462	164
Shared by 50% of a subfamily	220	3
Deleterious	72	3
Expressed	35	1

Table 5.8: Summary of the filtering procedure for exome variants

functional in a relevant tissue (Table 5.8).

Ordering by the maximum frequency in affecteds in either subfamily, or across the entire family, the *NOD2* frameshift mutation ranks second in the list of candidates (Table 5.9). This mutation is a the strongest known risk factor for Crohn's disease, and acts as a reassuring positive control, demonstrating that this method can prioritise mutations with relatively low penetrance. This is particularly reassuring as the *NOD2* region was not identified as a suggestive linkage peak or widely shared haplotype, due to it being introduced by multiple founders: this shows that the sequencing and prioritisation approach can identify true associations that the linkage approaches cannot.

The most widely shared novel candidate mutation across the family was a missense mutation in the gene *PDE4FIP*, encoding the protein Myomegalin. This gene has not previously been implicated as having a role in immunity. Next down, a mutation in the gene *PIK3C2A* was found to be widely shared in subfamily (c): this gene is relatively poorly understood, but may play a role in autophagy (Vanhaesebroeck et al., 2010). Towards the top of the list we also find a missense variant in *NLRP2* (a protein known to regulate inflammation in macrophages (Fontalba et al., 2007)) that is shared across subfamilies.

Chr:Pos	Alleles	Affected carriers			Gene	Mutation
		(a+b)	(c)	All		
1:144871738	C/A	16	11	27	*PDE4DIP*	Aka1742Ser (0.73)
16:50763778	G/GC	0	10	15	*NOD2*	Leu1007Fs
11:17191207	T/C	0	10	10	*PIK3C2A*	Lys28Glu (0.55)
11:64527189	C/T	14	0	14	*PYGM*	Arg61His (0.82)
19:55481394	C/T	4	9	13	*NLRP2*	Ser4Leu (0.74)
3:148601439	G/C	1	9	11	*CPA3*	Arg273Pro (0.70)
11:5536759	G/A	0	9	10	*UBQLNL*	Gln305X
3:136664737	C/T	13	1	15	*NCK1*	Ala180Val (0.50)
11:5424701	T/C	5	8	15	*OR51B5*	Ile292Thr (0.86)
11:64854223	C/A	0	8	8	*ZFPL1*	Pro147His (0.55)

Table 5.9: Top 10 SNP protein coding candidate mutations. The number after the amino acid change is the Condel score on the canonical transcript.

Filter	SNPs	Indels
Low frequency mutations on maximal haplotype	125189	38290
Shared by at least 9 affecteds	26993	8501
Conserved base	3143	584
Regulatory function	110	12
Expressed in relevant tissue	74	7

Table 5.10: Summary of the filter procedure for non-coding variants

Non-coding SNPs and indels

A total of 35,494 SNPs and indels were at low frequency in the population, and were shared by at least 9 affecteds in subfamilies (a)+(b) (Table 5.10). Further filtering produced 81 candidate variants, which were both conserved and lay in putative regulatory regions (Table 5.11).

Chrom:Pos	Alleles	Affected	Conservation	Gene	Regulatory features
18:9380839	G/A	14	1.04	*TWSG1*	Ensembl
14:61835929	C/T	11	2.98	*PRKCH*	Ensembl, TFBS(EBF), DNase1
4:169330682	T/C	11	2.11	*DDX60L*	Ensembl, TFBS(c-Fos, junD), DNAse1
1:7018684	G/A	10	4.30	*CAMTA1*	Ensembl, TFBS(KAP1,p300,CEBPB), DNase1
9:112012029	G/C	10	4.17	*EPB41L4B*	TFBS(CTCF,Rad21), DNase1
10:247426	G/A	10	3.58	*ZMYND11*	Ensembl, TFBS(c-Fos,STAT3,c-Jun), DNase1
1:7717739	G/C	10	3.47	*CAMTA1*	Ensembl
1:108439711	A/C	10	3.12	*VAV3*	TFBS(EBF,EBF1), DNase1
22:19347974	C/T	10	2.27	*HIRA*	Ensembl, TFBS(CTCF,Pol2)
2:103039025	T/A	10	2.05	*IL18RAP*	TFBS (GATA2, STAT2, others) DNase1
17:62181435	C/CA	10	0.98	*ENR1*	TFBS(c-Fos,STAT3,others), DNase1
13:98919267	C/CAA	10	0.91	*FARP1*	Ensembl
10:99441651	G/GA	10	0.78	*AVPI1*	Ensembl
1:52150585	C/CT	10	0.61	*OSBPL9*	Ensembl, TFBS(p300), DNase1

Table 5.11: Top 10 SNP candidates, and all 4 indel candidates, shared by at least 10 individuals. SNP conservation is given as the GERP score, and indel conservation is given by the phastCons score.

Filter	SNPs	Indels
Novel insertions or deletions >100bp	2332	262
Shared by at least 9 affecteds	645	80
Delete coding sequence	5	0
Expressed	3	0

Table 5.12: Summary of the filtering procedure for structural variants

No single candidate stood out as both clearly functional and widely shared. Only one potential regulatory mutation was on the maximally shared haplotype (i.e. shared by 14 individuals). This was a novel mutation in a putative regulatory region of *TWSG1* (a gene implicated in BMP signalling and B cell differentiation). However, of the 4 cell lines the regulatory feature was detected as active in, none was related to the immune or digestive system, and there was no clear evidence of transcription factor binding at this position.

There were some promising candidate mutations that were shared by a reduced number of affecteds. A mutation in a B- and T-cell active regulatory region near *PRKCH* (involved in T-cell activation (Fu et al., 2011)) is shared by eleven affecteds. This gene has previously been implicated in susceptibility to atrophic gastritis by a candidate gene study (Goto et al., 2010). Another strong candidate is *IL18RAP*, a receptor for interleukin-18 (known to be important in Crohn's disease (Maerten et al., 2004)), and a candidate causal gene in the IIBDGC Immunochip analysis (Chapter 4). The mutation itself is in a binding site for STAT2, a transcription factor known to be downregulated in IBD (Mudter et al., 2005), though only 10 individuals share this mutation.

Chrom:Pos	Alleles	Affected	Gene (bases deleted)
7:142494034-142495142	1108bp deletion	12	*TRBJ2* (50bp) to *TRBJ6* (48bp)
13:95363645-95363829	184bp deletion	10	*SOX21* (184bp)
4:84221936-84222193	257bp deletion	9	*HSPE* (77bp)

Table 5.13: The three candidate structural variants

Structural variants

725 novel structural variants lay within regions of the genome with haplotypes shared by at least 9 individuals (Table 5.12). Because of the difficulty in genotyping structural variants we were not able to test whether these variants fell on the maximally shared haplotype. Of the 725 mutations, 5 deleted coding sequence, and 3 of these lay within genes expressed in the digestive or immune system (Table 5.13).

The functional structural variant that is most widely shared lies in the T-cell receptor β (TCRB) locus, and appears to delete seven TCRBJ genes (including all the most commonly used ones (Freeman et al., 2009)). At first glance, this makes it an excellent candidate. However, the TCRB region undergoes VDJ recombination during T cell development, and the deletion may well have occurred during normal somatic development. Furthermore, parts of the TCRB region are known to be copy number variable in healthy individuals (Mackelprang et al., 2002), meaning that even a germ-line mutation may be benign. The other two candidate SVs are not particularly widely shared, and do not lie in any obvious candidate genes.

New *NOD2* variants

I mentioned above that the well-established *NOD2* frameshift mutation ranked second in the list of coding candidate variants in the family. The importance of this mutation in the genetics of Crohn's disease led us to specifically investigate *NOD2* variants that our above prioritisation analysis may have missed. Doing so uncovered two new *NOD2* mutations that are likely increasing the risk of IBD in this family.

One mutation was carried in a heterozygous state by one of the spouses that underwent whole-genome sequencing. This mutation (Arg791Gln) is present in dbSNP (rs104895464), but is at very low frequency in the general population (0.1%). It has a high Condel score (0.997) and lies in the middle of the LRR domain: this places it in the "CD sensitive region" described in Chapter 4, section 6.1, and thus is very likely to increase the risk of Crohn's disease. However, the mutation is not very common in the family: It was observed once in the sequencing, and from the haplotype flow we can infer that it was only passed on to one affected offspring.

A second novel *NOD2* mutation is found in the exome sequencing, and occurs at the same base pair as one of the traditional *NOD2* mutations (Gly908Arg). This mutation, Gly908Cys, is not present in 1000 Genomes or ESP datasets, though it has been observed twice (in 662 individuals) in the NIH ClinSeq project (Biesecker et al., 2009; Biesecker, 2012). This allele has an even higher Condel score than the established variant (0.999 vs 0.997), suggesting that it too will increase the risk of Crohn's disease. This mutation was introduced by a spouse, and was passed on to two affected children (both of whom are thus discovered to be compound heterozygous for this and a second *NOD2* mutation). Because of the striking nature of this mutation and the fact that it had (at the time) never been reported before, we performed capillary sequence validation to confirm its existence.

While both of these mutations likely increase the risk of IBD, both were introduced by spouses and thus are not carried on a haplotype shared across nuclear families. Additionally, as they are together only carried by three affected individuals, they can only explain only a small fraction of the affecteds in the family.

Sample set	N	Reason
Affecteds and parents	74	Validation of sites and genotypes
Jewish controls	~100	Validation of allele frequency
Case/control cohort	~600	Replication via association
Unaffected siblings	~250	Replication via transmission
Other multiplex families	~200	Replication via additional families
Total	~1200	

Table 5.14: Summary of the samples used for in the replication and validation effort. The columns give the name of the sample set, the number of samples included, the reason for their inclusion.

5.5 Follow-up of candidate causal variants

In the previous section I described a number of candidate variants (120) that could be driving the prevalence of IBD in a multiplex family. The vast majority of these are not associated with IBD: instead, they are likely to be a combination of technical errors and variants that have risen to high frequency in the family by chance.

To reduce the number of candidate causal variants, we have designed a validation and replication exercise to identify erroneous and non-associated candidates. This involves genotyping approximately 1200 samples using 8 Sequenom plexes (around 220 variants). These will consist of candidates from the IBD family discussed above, as well as candidate variants from other families and other important known risk variants (such as the *NOD2* mutations). The samples to be used, as well as the reasons that they are included, are shown in Table 5.14. In this section I will describe the intended validation and replication tests, and discuss their power to confirm or falsify candidate causal variants.

Once these tests have been carried out, and if candidate variants still remain, these variants will be carried forward into functional studies to identify likely causal mechanisms. I will not discuss these functional experiments here.

5.5.1 Technical validation of causal variants

During the 1000 Genomes loss-of-function project described in chapter 3, we learned that LoF variants are greatly enriched for technical errors compared to other classes of variations (MacArthur et al., 2012). This was not due to any particular property of the variants themselves, but instead due to the fact that loss-of-function variants are extremely rare. In essence, because the number of true loss-of-function variants is depleted relative to other categories, while the number of technical errors is approximately constant regardless of functional category, the proportion of errors is much higher.

The list of candidate variants from the family suffers from a similar effect. We have picked these candidates based on a number of criteria that will diminish the pool of true variants and increase the relative number of errors. The classes of functional variants that we have selected for are known to be under negative selection: coding SNPs predicted to be damaging to protein structure are under strong negative selection (Barreiro et al., 2008), and mutations inside non-coding regulatory regions are also known to be rarer than in the genome as a whole (The ENCODE Project Consortium, 2012). We have also selected variants that are common within the family, but rare in the general population, which itself will inflate the error rate.

In this section I will discuss some sources of technical error in the candidates, and discuss validation strategies that can overcome these problems.

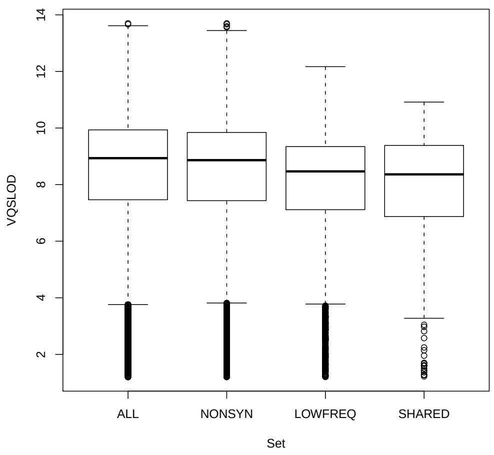


Figure 5.11: The VQSLOD scores for the exome call sets after various sequential filtering steps.

False positive variants

Some of the candidate variants will be false, the result of systematic errors in sequencing. The VQSR calibration will have given us a degree of robustness to such errors, but it is likely that at least some will remain. Figure 5.11 show the VQSLOD score for the exome variants after various stages of filtering. There is a difference in score of approximately 0.8 between the entire exome dataset and the shared, low-frequency coding variants. This shows that systematic errors of the type measured by VQSR are more common in our datasets. More specifically, it corresponds to an estimated 2.2-fold increase in false positive rate in filtered variants (95% CI: 1.6-3.1).

Ideally, all candidate variants should be validated using an independent technology. Capillary resequencing is perhaps the most accurate form of validation (for example, we use this technology to validate the novel *NOD2* variant discussed above), but it is low throughput. PCR amplification is another low-throughput method that can be used to validate structural variants. A more high-throughput validation technology is the Sequenom (Bradic et al., 2011) mass spectrometry method (the main method used for validation of LoF variants in the 1000 Genomes project). This requires processing a large number of samples to accurately validate sites, but can be combined with the various genotyping efforts described below.

Poorly genotyped variants

Another potential source of false candidates is genotype error. A variant may be real, and present in the family, but some samples have been assigned the wrong genotypes. This can lead a variant that is present only in a small number of individuals to seem to be present in a larger number. This is particularly likely to be a problem in the whole-genome sequencing data, where the coverage is much lower, and incorrect genotypes in a small number of individuals can lead to a variant being incorrectly inferred to lie on a shared haplotype. Again, the most reliable method of detecting these problems is to perform genotyping on the same samples using an independent technology. This can be combined with the site validation described in the previous section.

Common variants

Some of the candidate variants may in fact be at high frequency in the general population. While we have filtered these datasets based on population frequency, there are two factors that may lead a high-frequency variant to remain in the list. Firstly, the variant may be absent from the reference set used, either because it was not detected in the original call list, or was filtered out as poorly performing. Secondly, the variant may be at high frequency exclusively in the Ashkenazi Jewish population. For instance, of the

35,191 exome variants that were below 2.5% in Americans of both European and African descent, 222 were detected at above 10% in the Ashkenazi Jewish control exomes. For the whole-genome sequencing no Jewish controls were available, meaning many of our non-coding candidates may be at high frequency in the Jewish population.

The solution to this problem is to genotype all candidate variants in a control population taken from the same ethnic group and geographic region as the family.

5.5.2 Independent replication of causal variants

Even if the variant is real, is truly low frequency and has been correctly genotyped, it still may be present in a large number of affected family members merely by chance. This is especially true in our case, where we know that this family does not show a genome-wide significance linkage peak, and many of our candidate variants do not lie within even suggestive linkage peaks. To demonstrate that a variant is causal, we need to provide independent replication of the association. In this section I will discuss three different methods of replicating a candidate mutation by genotyping in further samples.

Validation in a case-control cohort

While we filtered out variants with an allele frequency of above 2.5%, many of our candidate variants are still polymorphic in the general population. Such variants may well not be well tagged in GWAS, but case-control cohorts well powered to detect them if genotyped directly. For variants at intermediate frequency (between 0.1% and 1%) we can attempt to replicate these variants in standard case-control cohorts of IBD.

Assuming a risk allele frequency of f, a prevalence of K and a dominant

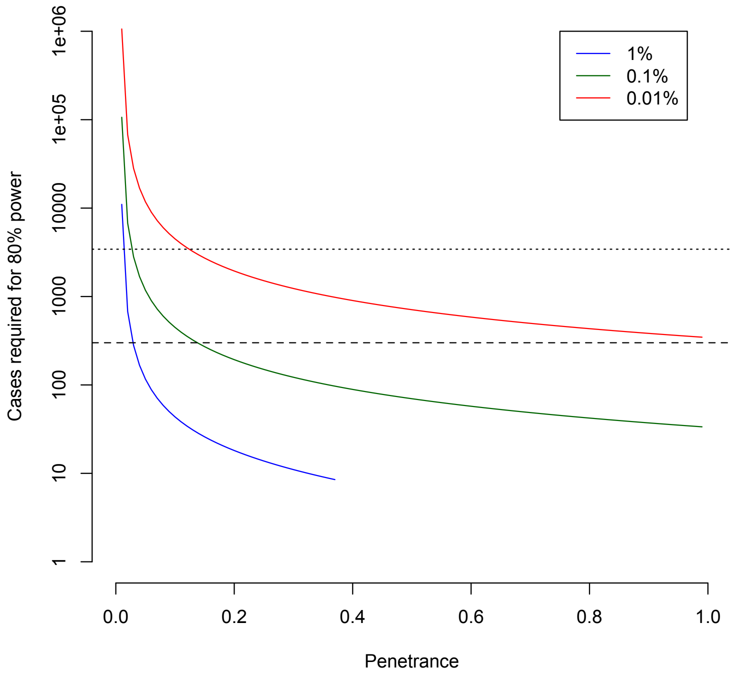


Figure 5.12: The same size required to have 80% power to replicate a mutation with a given penetrance with $p < 0.01$, assuming a prevalence of $K = 0.0075$. The colours of the lines represent the allele frequency in the general population. The dashed line represents a small replication effort (300 cases and 300 controls), and the dotted line represents a large effort (3000 cases and 3000 controls).

penetrance of π (such that $\pi < \frac{K}{f(2-f)}$), the proportion of affecteds in the general population who carry this mutation is

$$\begin{aligned} P(r = 1|d = 1) &= \frac{P(d = 1|r = 1)P(r = 1)}{P(d = 1)} && (5.55) \\ &= f(2 - f)\frac{\pi}{K}. && (5.56) \end{aligned}$$

Similarly, the proportion of unaffected carriers is

$$P(r = 1|d = 0) = f(2 - f)\frac{1 - \pi}{1 - K}. \tag{5.57}$$

The sample size required to detect a difference in the number of carriers between cases and controls, for a given penetrance and allele frequency, is shown in Figure 5.12. A small genotyping effort (300 cases and 300 controls) is well powered to detect (and therefore also to rule out) medium penetrance mutations (>10%) with an allele frequency of greater than 0.1%. A large genotyping effort (such as the whole-genome sequencing experiment described in Chapter 6) would have a power to detect and rule out medium penetrance mutations with a population frequency of greater than 0.01%.

Replicating truly rare mutations is extremely difficult using case-control cohorts, though datasets on the scale of the International IBD Genetics Consortium's replication cohort (discussed in chapter 4) would be well powered to replicate intermediate penetrance mutations with allele frequencies as low as 1 in 200,000.

Validating using unaffected siblings

A standard way to validate a potential causal variant is to track its co-segregation with affection status within the family it was discovered in. In the approach described above, we have prioritised variants for follow-up based on their presence in a large number of affecteds. However, the unaffected siblings of these affected individuals have not been tested, and these unaffected individuals can provide an additional validation set. Where a parent is heterozygous for the candidate mutation, we can test for evidence of causality by testing whether it is transmitted to less than half of unaffected children. Here I will consider what allele frequencies we expect in unaffected siblings as a function of penetrance, and what power these unaffected siblings can

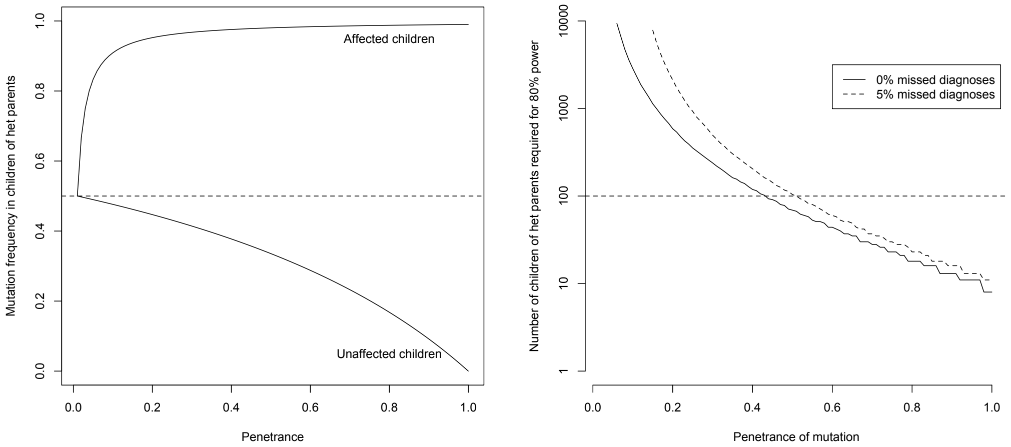


Figure 5.13: a) The frequency of a dominant mutation in affected and unaffected children of an individual heterozygous for this mutation. b) The number of unaffected children of parents heterozygous for the mutation required to validate causality with $p < 0.01$ by a binomial hypothesis test, as a function of the penetrance of the mutation. The solid line represents the case where all unaffected individuals are correctly diagnosed, whereas the dashed line represents a scenario in which 5% of unaffected siblings in fact are (or will become) affected. In both cases I assume $K = 0.0075$.

provide to validate causality.

We will assume that one parent caries the mutation, and there is therefore an even chance that a child will inherit it, i.e. $P(r = 1) = P(r = 0) = \frac{1}{2}$. The overall disease prevalence in the children is thus

$$\begin{aligned} P(d = 1) &= P(d = 1|r = 0)P(r = 0) + P(d = 0|r = 0)P(r = 1) \\ &= \frac{K + \pi}{2}. \qquad (5.58) \end{aligned}$$

The proportion of unaffected children who are wild-type is thus

$$P(r = 0|d = 0) = \frac{P(d = 0|r = 0)P(r = 0)}{P(d = 1)} \tag{5.59}$$

$$= \frac{1 - K}{(1 - K) + (1 - \pi)}. \tag{5.60}$$

We can calculate the same value for affected children

$$P(r = 0|d = 1) = \frac{P(d = 1|r = 0)P(r = 0)}{P(d)} \tag{5.61}$$

$$= \frac{K}{\pi + K}. \tag{5.62}$$

These two equations are plotted as a function of π (for a fixed $K = 0.0075$) in Figure 5.13a. While the mutation frequency in affected children rises very rapidly with the penetrance, the corresponding frequency in unaffecteds falls much more slowly. Figure 5.13b shows the number of unaffected children of heterozygous parents required to validate a candidate mutation at $p < 0.05$. For high penetrance mutations ($\pi > 0.7$) validation can be performed in a modest number of unaffected siblings ($N < 30$), though for intermediate penetrance mutations ($\pi > 0.4$) larger number of unaffecteds are required ($N \sim 100$).

This analysis assumes that all individuals who are currently believed to be unaffected are truly unaffected. However, a proportion of these individuals are likely to have the disease but not yet have been diagnosed, or will go on to develop the disease later in life. This could seriously increase the frequency of the mutation in unaffecteds.

To model this, we assume that a proportion α of the unaffected siblings are in fact cases. We will denote the true affection status with d^T, such that

$P(d^T = 1|d = 0) = \alpha$. The proportion of individuals classified as unaffected who are wild-type is given by

$$\begin{aligned} P(r = 0|d = 0) &= P(r = 0|d^T = 0)P(d^T = 0|d = 0) \\ &\quad + P(r = 0|d^T = 1)P(d^T = 1|d = 0) \qquad (5.63) \\ &= (1-\alpha)\frac{1-K}{(1-K)+(1-\pi)} + \alpha\frac{K}{K+\pi}. \qquad (5.64) \end{aligned}$$

This diagnostic uncertainty can seriously reduce the power of validation using unaffected siblings. The dashed line in Figure 5.13b shows how many more siblings are needed to account for this diagnostic uncertainty. For instance, to validate a mutation with a penetrance of $\pi = 0.4$ requires $N = 115$ siblings under perfect diagnostic conditions, but $N = 190$ when there is a 5% underdiagnosis rate.

For the candidate variants in the family we are studying, the number of unaffected offspring of carrier parents varies from 50 to 250, depending on the number of subfamilies the mutation is segregating in. We thus will have power to replicate mutations with a high penetrance (>60%) for most mutations, down to about 30% for more widely shared mutations.

Replication in other multiplex families

Perhaps the gold standard for replicating a causal mutation found in a family is to show that it segregates with disease status it in a second family. As we saw in section 5.3.1, multiplex families are more likely to carry more penetrant mutations, and thus screening a large enough number of multiplex families is likely to turn up other instances of the mutation even if the allele frequency in the population is low. For instance, a 0.1% variant with a penetrance of 50% will be present in 1.3% of cases, but will be present in approximately 10% of patients with at least 2 affected first degree relatives (calculated using the model described in section 5.3). Once such families are identified, affected children of mutation carriers can be tested for an over-inheritance of the mutation.

As we saw in Figure 5.13a, providing that penetrance is above around 20%, affected children of heterozygous parents should be carriers at least 95% of the time. If 8 such affected children can be collected from additional families, and the mutation is causal, more often than not (>65% of the time) all will carry the mutation. However, if the mutation is not causal, there is only a 0.4% chance of all children carrying this mutation. Even for a disease with a 10% penetrance, only 12 children are required to produce the same effect. Thus, identifying less than a dozen affected children in families carrying the mutation is often sufficient to demonstrate causality.

5.6 Conclusions

Discovering high penetrance mutations in multiplex families is, unsurprisingly, a more complex endeavour for complex diseases than for Mendelian disease. We have seen how a large number of multiplex families can arise as a result of polygenic risk alone, and great care must be taken to select families that are likely to carry penetrant mutations. Even if an affected family is detected, a combination of phenocopies, incomplete penetrance and less obviously severe mutations can make correct identification of the causal variants difficult.

Given this, it is not surprising that the above approach did not produce the single, clearly highly damaging mutation shared by all affecteds that would be expected from a Mendelian disease family. Instead, a detailed genotyping, sequencing and filtering experiment produced a series of over a hundred plausible candidates. One of the most valuable resources in the identification of these variants has been tools for inferring both coding and non-coding function, including variant effect prediction, information on regulatory regions, and tissue specific gene expression data. This has allowed us to drastically reduce the list of candidates on the basis of putative function.

A list of multiple candidate variants is likely to be the standard output for family sequencing studies in complex disease. As has been the case with common associations, the key to turning these candidate variants into established associations will be independent replication. I have shown, for certain variants there is potential for replication with unaffected siblings, and within case-control cohorts. However, the most valuable form of replication is likely to be the detection of evidence of co-segregation with affection status in other multiplex families. This highlights the value of collecting samples from many multiplex families, and of collaboration between different research groups studying multiplex families.

From these observations, I believe that we can identify the two most important developments that will drive forward the study of multiplex families in coming years. The first will be the integration of increasingly detailed functional datasets, and in particular datasets that can assess regulatory function. The second will be collaboration, and in particular reciprocal replication, between research groups in order to establish causal variants in multiple families.

Chapter 6

Conclusions

6.1 Connections and themes

The projects described in this thesis, while all focused towards locus discovery, have been more or less distinct. I have investigated the historical and statistical foundations of complex disease genetics in chapters 1 and 2, studied the utility of genotype imputation in Chapter 3, described the discovery of new inflammatory bowel disease (IBD) risk loci via custom genotyping in Chapter 4, and investigated genetic risk factors in a multiplex IBD family in Chapter 5. Each chapter contained a unique dataset, and in each case I investigated this dataset using the methods most relevant to that data type.

Despite this, certain topics have come up multiple times throughout the chapters. For instance, the importance of the Crohn's disease *NOD2* locus has come up in almost every chapter: as an important development in the history of disease genetics in Chapter 1, one of the few common loss-of-function risk variants in Chapter 3, as a pilot locus for fine-mapping in Chapter 4, and as an important contribution of risk in the family discussed in Chapter 5.

However, beyond specific topics like this, certain wider themes have emerged as relevant to all the individual projects, and possibly to the field of complex disease genetics as a whole.

One of the major themes has been the economics of experimental design in disease genetics. In this field we do not design the "ideal" experiment, we design the experiment that has the highest power or utility given the availability of datasets, technology, samples and statistical methods. Sometimes the driving considerations have been explicitly financial: the design of the Immunochip was born from economic arguments about the relationships between sample size, power, unit cost and bulk buying. Other experiments have been driven by the exploitation of unique sample resources, such as the relatively low-cost sequencing of very large multiplex families. Still others have been about leveraging external datasets: genotype imputation using sequencing datasets is a prime example of statistical methods, combined with external datasets, adding substantial value to existing studies. Success in complex disease genetics is largely dependent on being able to recognise the potential for such "high-value" studies as new resources become available.

A related theme is the appearance of "next-generation" datasets that can add value to the genetic data used in locus discovery. The studies in Chapters 3 and 5 would have been essentially impossible without the use of genome-wide external datasets of population sequencing and genome function respectively. While the list of 163 loci in Chapter 4 would probably have been obtainable without external datasets, the transformation of this locus list into biological hypotheses would have been essentially impossible. In essence, these datasets allow a quantitative understanding of biological function throughout the genome, and as they improve and become more integrated with the genetic datasets, our ability to discover and characterise disease associations can only grow.

Another theme has been the role of theory in complex disease genetics. The statistical techniques described in Chapter 2 have been used throughout this thesis. The scale of data generated in modern complex disease genetics means that these techniques are the only way to analyse this data, making understanding the assumptions and models implied by these methods especially important. However, we have also seen the role of biological theory in the analysis of this data. In Chapter 5, it was knowledge of biological function (both genome function and gene expression) that allowed us to reduce the number of candidate causal mutations to a manageable number. This union of statistical and biological theory was particularly important in Chapter 4, where both bioinformatic interrogation and biological insight were required to transform a long list of loci into a set of more concrete biological hypotheses. I believe that this integration between biological and statistical theory will be increasingly important as the field moves forward, as I will discuss later in this chapter.

Finally, a continued theme throughout this thesis has been the historical trajectory of locus discovery. In Chapter 1 I took an unashamedly teleological view of the field, describing how experiments (if not discoveries) are often predicted far in advance. While I do not believe that science in general is an onwards march of progress, there have always been certain discoveries that have been foreseen long before they came pass. Given sufficient time, the source of the Nile will be discovered, the moon will be walked on, and the genetic differences that lead to disease susceptibility will be identified. The question is not if but when.

In this spirit of teleology, I will spend the rest of this chapter considering the future of locus discovery, looking ahead to the next crop of experiments that are already underway, and those that will appear in the more distant future. We will see the same themes I have discussed above appear again, in many cases becoming more important as the scale of the data and the level of data integration in human disease genetics increases.

6.2 A next-generation GWAS using low-coverage sequencing

As I have discussed in this thesis, we can attempt to map low-frequency and rare disease alleles in a variety of ways. I have presented projects that use imputation, custom genotyping within known loci and sequencing in multiplex families to identify low-frequency risk variants. As discussed in the introduction, other groups are also using exome sequencing, targeting sequencing of candidate loci and custom genotyping of low-frequency coding variation to the same end. All of these experiments look at a restricted class of variants or samples. Sequencing in families can only identify variants present in those families, and imputation can only identify variants that are in the reference sets and can be inferred from common variation. Targeted sequencing or genotyping is only as powerful as the selection of targets, and any "surprising" variants (e.g. large effect size regulatory variation) will be missed.

A more "hypothesis-free" way of discovering low-frequency risk variation is to extend the genome-wide approach that has been so successful in GWAS using next-generation sequencing. Complete (high-coverage) sequencing is currently too expensive to produce the sample sizes required, but the imputation framework described in Chapter 3 can be applied to incomplete (or low-coverage) sequencing data to allow us to infer the genotypes of nearly all variants in the genome at a fraction of the cost of complete sequencing. This opens up the possibility of affordable whole-genome sequencing of case and control cohorts with sample sizes large enough to detect low-frequency associations.

In order to put this technique into practice, two research groups from the

Wellcome Trust Sanger Institute, in collaboration with the UK IBD Genetics Consortium (UKIBDGC), designed a large low-coverage sequencing project of IBD cases. The project is funded by the Wellcome Trust and the MRC, and will sequence 5000 cases (3000 CD and 2000 UC). This data will be combined with the 4000 UK10K cohort controls to produce (to my knowledge) the largest whole-genome sequencing case-control dataset ever produced.

The cases picked for sequencing underwent a detailed selection procedure to maximise the likelihood of detecting associations. Samples were selected for sequencing on the basis of family history (at least three affected family members, or one affected first-degree relative), age of onset (diagnosed before age 17) and severity of disease (more than three surgical interventions). Other samples were prioritised on the basis of having attached functional data, including gene expression and epigenetic assays, in order to allow functional studies to be performed using the whole-genome sequencing data.

The aim of the dataset is to identify suggestive ($p < 10^{-5}$) associations to replicate in a larger cohort. The experiment will have 77% and 55% power to detect low-frequency (MAF of 1%) associations of intermediate effect size ($OR = 2$) that are unique to CD or UC respectively. For shared associations this power rises to 85%, with 26% power for risk alleles with a frequency below 0.5%. There are more than 7000 additional UKIBDGC cases (2500 CD and 4500 UC) ready for use in replication, with others to come, which if combined with a large number of controls will have high power to replicate associations down to at least 0.5%.

While sequencing for the final dataset is not yet complete, we have run a small pilot project to test and refine the methodology. This involved sequence data from 4249 samples with a mean coverage of 3.7X, and focused on a 40Mb region of chromosome 16 that contained the *NOD2* locus (a positive

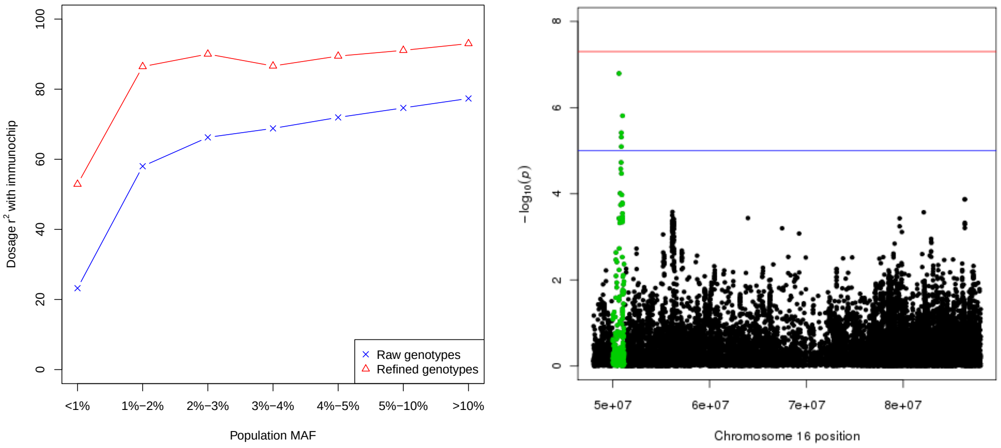


Figure 6.1: a) The correlation between allele dosage as calculated from the sequencing data and from the Immunochip data, before and after imputation genotype improvement. b) Manhattan plot of variants with MAF between 1% and 5% for the last 40Mb of chromosome 16 after extensive QC. Association testing was carried out using SNPTest on the imputation posteriors. The green dots are variants in the *NOD2* region.

control for low-frequency association). Samples were genotyped, and SNPs and indels called, using the pipeline described in Chapter 5. We then used the imputation program Beagle (Browning and Browning, 2007) to refine the genotype likelihoods, which substantially improved the accuracy of the calls when measured by concordance with Immunochip data on the same samples (Figure 6.1a). Overall, the refined genotypes had an r^2 of approximately 87% with the true genotypes at sites with a minor allele frequency of between 1% and 2%. Substantial further QC on both SNPs and samples was required to produce a clean enough dataset to allow association testing. After filtering, association tests at low-frequency variants (MAF of between 1% and 5%) yielded a clean Manhattan plot with the *NOD2* region showing clear evidence of association (Figure 6.1b). This demonstrates that the approach is sound, that specific low-frequency variants can be detected and that with enough filtering false positives rates can be controlled

The analysis of the full dataset will involve numerous methodological challenges, in addition to the significant computation burden. The probabilistic genotypes need to be well-calibrated to allow association testing, and false positive associations generated by countless new error modes will need to be identified. Standard tests, such as burden tests, will need to be redeveloped to deal with the uncertainty in the data. However, if these problems can be overcome, this experiment will allow the first well-powered whole-genome survey of low-frequency IBD risk variation to date.

6.3 Towards the ideal locus discovery experiment

I said in Chapter 1 that human disease genetics is a field where future experiments are anticipated relatively far in advance. The genome-wide linkage studies of the late 1990s and early 2000s were anticipated since at least the development of RFLP linkage in the 1970s. Likewise, the power of GWAS was foreseen since at least the 1990s. In both cases the conceptual framework was present, waiting for the technology and the sample collections to make them a reality. In the same way, we can now anticipate what the next (and possibly final) locus discovery experiment may look like in the future.

Much as the original set of GWAS experiments were followed by meta-analyses and international replication experiments, it seems reasonable that in the next few years the various international disease genetics consortia will combine their sequencing data into meta-analyses. These are likely to be very heterogeneous analyses, combining information from targeted, exome and whole-genome sequencing across a range of technologies. Doubtless this will then be followed by replication in tens of thousands of samples. The power of these projects will depend on the coverage of their component studies, but it is likely that a large number of low-frequency and rare associations will be identified at this point.

Beyond this, we start to move towards what disease geneticists refer to as the “right” or “ideal” locus discovery experiment. The cost of sequencing has fallen dramatically in recent years, and the speed and ease of sample preparation seems set to rise dramatically. We are on the verge of the $1000 genome, and it seems likely that the next decade will bring the cost of whole-genome sequencing down to $100 a sample or below. Within 20 years it is likely that a WGS experiment including hundreds of thousands of samples would have a price tag measured in the low millions of pounds, and be as technically feasible as GWAS.

Even in the absence of a concerted effort from researchers, it seems likely that such datasets will become available eventually. Cheap and readily available genome sequencing is already being used in clinical genetics practice to diagnose genetic disease (Worthey et al., 2011b; Rios et al., 2010), to guide cancer treatment (Link et al., 2011), and as a cost-effective form of carrier testing (Bell et al., 2011). It is likely that a relatively large proportion of patients will undergo routine whole-genome sequencing in the not-too-distance future, and many of these patients will consent to their data being used for research. The cost of the “ideal” WGS experiment may well end up being covered by the budgets of public healthcare services and private insurance companies.

Let us imagine that, sooner or later, researchers will have access to high-quality genome sequencing from 100,000 IBD cases (around a third of the patients in the UK), including sporadic and familial cases, as well as sequence data from their parents and an arbitrarily large number of other healthy controls. What could this ideal dataset tell us about the genetics of IBD?

Firstly, as discussed in Chapter 4, this dataset would have a very high power to fine-map associations with odds ratios larger than 1.1. If any IBD loci exist that have not yet been fine-mapped by other projects, a dataset of this size and completeness would allow the vast majority to have the causal variant determined.

Secondly, this dataset would allow us to characterise a large proportion of the common, low-effect size variants that contribute to polygenic risk, detecting most common risk variants with odds ratios > 1.03. Distinguishing these variants from the effects of very subtle population stratification may be difficult, but sequence data is also available on the parents this can be easily overcome. These polygenic risk loci are likely to cover a significant proportion of the genome, and be extremely difficult to fine-map, making them hard to interpret. However, they could be combined with other external datasets to perform detailed network and pathway analyses, in the same manner as the Immunochip loci were used in Chapter 4.

Thirdly, almost all low-frequency risk variants (MAF $> 1\%$) with an odds ratio of greater than 1.15, and all the rare (MAF $> 0.01\%$) mutations with odds ratios greater than 3, could be identified via this dataset. Aggregation tests for rare variants (Neale et al., 2011) would also allow us to identify genes or other functional units that carry extremely low-frequency risk mutations. This high power and completeness of data would allow us to ask questions about the biological properties that lead to some genes, parts of genes or classes of variation to carry risk variants, while others do not (using the techniques described for loss-of-function variants in Chapter 3 and *NOD2* coding variants in Chapter 4).

Fourthly, the family data would allow us to identify high penetrance familial mutations in IBD. The sequencing of parents would allow us to detect the contribution of de novo point, indel and structural mutations to IBD (in a similar fashion to recent studies of autism (Neale et al., 2012)). It would also let us identify extremely rare near-Mendelian mutations that are shared only by a handful of families, using the techniques described in Chapter 5.

Finally, this dataset would allow the full power of genetic risk prediction to be utilised, via standard risk prediction using established variants, and via identity-by-state and identity-by-descent methods. Assuming 50% of liability variance is captured by the risk score, we would be able to define a "high-risk" group who are more likely to than not to develop the disease (which would catch 3% of cases), and a "medium-risk" group that have a 1 in 6 chance of developing the disease (catching another 28% of cases). Even in these extreme cases genetic risk prediction would not give any guarantees, and some 10% of cases will still have a lower genetic risk than the population mean. However, the information provided may still be a significant aid to diagnosis and prevention, particularly in combination with non-genetic risk prediction.

While we are dreaming up ideal datasets, we can also imagine what other functional information that may come attached to our ideal genetic dataset. Today many IBD patients undergo measurement of certain cytokine and antibody concentrations, and the battery of tests is always increasing. It has recently been shown that gene expression data from CD8+ T-cells can be used to predict disease prognosis in IBD (Lee et al., 2011), and it stands to reason that such assays will become standard procedure in the future. Sequencing assays of epigenetic data, such as of methylation, transcription factor binding or open chromatin, are also rapidly becoming important in research, and may eventually become clinical tools. It seems likely that our ideal dataset will be accompanied by at least some data on gene expression, epigenetic marks and other relevant biological quantities. Combining the WGS data with this functional data will both allow us to reconstruct how the genetic risk factors act biologically.

6.4 Beyond locus discovery

This thesis has largely been focused on methods for discovering disease-associated loci. In this chapter I have discussed even more ways that we can discover risk loci, both in the near and more distance future. However, as I discussed in Chapter 1, risk loci in and of themselves are not inherently valuable to science or society. Despite all the debate about missing heritability, very few scientists have a deep and abiding desire to increase the "heritability explained" counter up to 100%. It is only when these loci can improve our understanding of disease biology, or directly impact patient care, that they really start earning the investment put into finding them. It is in the follow-up of these disease-associated loci that the real biological discovery starts to take shape.

In Chapter 2 I described the field of complex disease genetics as inherently statistical, and I stand by that statement. However, after disease loci have been identified, the task of following them up has historically been passed on to our less statistically, and more experimentally minded colleagues. For instance, the discovery of the *NOD2* loci via linkage was followed up by a decade of experimental work, establishing and investigating the biological links between *NOD2*, IBD and immunity (Shaw et al., 2011).

However, this is not a sustainable approach. The GWAS era ended the days when disease loci could be counted on one hand (and numbered in a universally recognised fashion: "IBD5", "IDDM2", "BRCA1"). Now risk loci are numbered in the hundreds, encompassing thousands of genes. Understanding the function of these loci has moved from something that can be established in the lab, and into the domain of statistical genetics described in Chapter 2. This mirrors developments in other fields, such as the rise of gene expression profiling and functional sequencing assays that have made functional biology into a high-throughput science (The ENCODE Project Consortium, 2012).

Many of the more striking discoveries that I have described in this thesis have come from the integration of genetic and functional data. As I hinted at in the previous section, future genetic studies are likely to become increasingly tied in with functional assays, allowing GWAS-style studies at all levels of disease biology. This will also be of benefit to purely experimental scientists following up these experiments, as the unit of follow-up will change from a gene name to a more detailed biological mechanism, pathway or hypothesis.

The next great challenge of statistical genetics in the coming decade will be to take the techniques and philosophy that have driven locus discovery, and turn them to the task of understanding the biological mechanisms of disease risk. This will require new models and new methodology, but perhaps more importantly it will require statistical geneticists to engage with disease biology, and experimental biologists to engage with statistical models. As I have seen throughout my thesis, complex disease genetics is an inherently statistical field, but it is also an inherently biological one.

References

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet., 30(1): 97–101, Jan 2002.

C. Abraham and J. H. Cho. Functional consequences of NOD2 (CARD15) mutations. Inflamm. Bowel Dis., 12(7):641–650, Jul 2006.

A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res., 21(6):974–984, Jun 2011.

I. Akinsheye, A. Alsultan, N. Solovieff, et al. Fetal hemoglobin in sickle cell anemia. Blood, 118(1):19–27, Jul 2011.

A. Al-Chalabi and C. M. Lewis. Modelling the effects of penetrance and family size on rates of sporadic and familial disease. Hum. Hered., 71(4):281–288, 2011.

T. J. Albert, M. N. Molla, D. M. Muzny, et al. Direct selection of human genomic loci by microarray hybridization. Nat. Methods, 4(11):903–905, Nov 2007.

D. Altshuler, J. N. Hirschhorn, M. Klannemark, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet., 26(1):76–80, Sep 2000.

D. M. Altshuler, R. A. Gibbs, L. Peltonen, et al. Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311):52–58, Sep 2010.

C. A. Anderson, G. Boucher, C. W. Lees, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet., 43(3):246–252, Mar 2011.

C.A. Anderson, F.H. Pettersson, J.C. Barrett, et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. The American Journal of Human Genetics, 83(1):112–119, 2008.

K. Asano, T. Matsushita, J. Umeno, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat. Genet., 41(12):1325–1329, Dec 2009.

O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type II. J. Exp. Med., 79(2):137–158, Feb 1944.

S.C. Bagley, H. White, and B.A. Golomb. Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. Journal of clinical epidemiology, 54(10):979–985, 2001.

M. Bahlo, D. R. Booth, S. A. Broadley, et al. Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat. Genet., 41(7):824–828, Jul 2009.

S. C. Bain, J. B. Prins, C. M. Hearne, et al. Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. Nat. Genet., 2(3):212–215, Nov 1992.

J.K. Baker. Trainable grammars for speech recognition. The Journal of the Acoustical Society of America, 65:S132, 1979.

M. J. Bamshad, S. B. Ng, A. W. Bigham, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet., 12(11):745–755, Nov 2011.

L. B. Barreiro, G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. Natural selection has driven population differentiation in modern humans. Nat. Genet., 40(3):340–345, Mar 2008.

J. C. Barrett and L. R. Cardon. Evaluating coverage of genome-wide association studies. Nat. Genet., 38(6):659–662, Jun 2006.

J. C. Barrett, D. G. Clayton, P. Concannon, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet., 41(6):703–707, Jun 2009a.

J. C. Barrett, S. Hansoul, D. L. Nicolae, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet., 40(8): 955–962, Aug 2008.

J. C. Barrett, J. C. Lee, C. W. Lees, et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat. Genet., 41(12):1330–1334, Dec 2009b.

T. Barrett, D. B. Troup, S. E. Wilhite, et al. NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res., 39(Database issue):D1005–1010, Jan 2011.

A. Barton, W. Thomson, X. Ke, et al. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nature genetics, 40(10):1156–1159, 2008.

T. C. Bates, J. F. Price, S. E. Harris, et al. Association of KIBRA and memory. Neurosci. Lett., 458(3):140–143, Jul 2009.

D. E. Bauer and S. H. Orkin. Update on fetal hemoglobin gene regulation in hemoglobinopathies. Curr. Opin. Pediatr., 23(1):1–8, Feb 2011.

D. C. Baumgart and S. R. Carding. Inflammatory bowel disease: cause and immunobiology. Lancet, 369(9573):1627–1640, May 2007.

G. W. Beadle and E. L. Tatum. Genetic Control of Biochemical Reactions in Neurospora. Proc. Natl. Acad. Sci. U.S.A., 27(11):499–506, Nov 1941.

A. B. Begovich, V. E. Carlton, L. A. Honigberg, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am. J. Hum. Genet., 75(2):330–337, Aug 2004.

D. M. Behar, D. Garrigan, M. E. Kaplan, et al. Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum. Genet., 114(4):354–365, Mar 2004.

C.J. Bell, D.L. Dinwiddie, N.A. Miller, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med., 3(65):65ra4, 2011.

G. I. Bell, S. Horita, and J. H. Karam. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes, 33(2):176–183, Feb 1984.

S. T. Bennett, A. M. Lucassen, S. C. Gough, et al. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat. Genet., 9(3):284–292, Mar 1995.

J. Berkson. Application of the logistic function to bio-assay. Journal of the American Statistical Association, 39(227):357–365, 1944.

J. Berkson. Why I prefer logits to probits. Biometrics, 7(4):327–339, 1951.

C.N. Bernstein. The natural history of inflammatory bowel disease. Crohn's Disease and Ulcerative Colitis: From Epidemiology and Immunobiology to a Rational Diagnostic and Therapeutic Approach, page 343, 2011.

L. G. Biesecker. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. Genet. Med., 14(4):393–398, Apr 2012.

L. G. Biesecker, J. C. Mullikin, F. M. Facio, et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res., 19(9):1665–1674, Sep 2009.

J.M. Bland and D.G. Altman. Statistics notes: the odds ratio. BMJ: British Medical Journal, 320(7247):1468, 2000.

C.I. Bliss. The calculation of the dosage-mortality curve. Annals of Applied Biology, 22(1):134–167, 1935.

D. Booth, R. Heard, G. Stewart, et al. Refining genetic associations in multiple sclerosis. Lancet Neurol, 7(7):567–569, Jul 2008.

F. J. Bosker, C. A. Hartman, I. M. Nolte, et al. Poor replication of candidate genes for major depressive disorder using genome-wide association data. Mol. Psychiatry, 16(5):516–532, May 2011.

D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet., 32(3):314–331, May 1980.

M. Bradic, J. Costa, and I. M. Chelo. Genotyping with Sequenom. Methods Mol. Biol., 772:193–210, 2011.

O. J. Brand, J. C. Barrett, M. J. Simmonds, et al. Association of the thyroid stimulating hormone receptor gene (TSHR) with Graves' disease. Hum. Mol. Genet., 18(9):1704–1713, May 2009.

S. R. Brant. Update on the heritability of inflammatory bowel disease: the importance of twin studies. Inflamm. Bowel Dis., 17(1):1–5, Jan 2011.

S. M. Bray, J. G. Mulle, A. F. Dodd, et al. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc. Natl. Acad. Sci. U.S.A., 107(37):16222–16227, Sep 2010.

NE Breslow and BE Storer. General relative risk functions for case-control studies. American journal of epidemiology, 122(1):149–162, 1985.

M. A. Brown, S. H. Laval, S. Brophy, and A. Calin. Recurrence risk modelling of the genetic susceptibility to ankylosing spondylitis. Ann. Rheum. Dis., 59(11): 883–886, Nov 2000.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet., 84(2):210–223, Feb 2009.

S. R. Browning. Multilocus association mapping using variable-length Markov chains. Am. J. Hum. Genet., 78(6):903–913, Jun 2006.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet., 81(5):1084–1097, Nov 2007.

S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet., 12(10):703–714, Oct 2011.

B. Buijsse, R.K. Simmons, S.J. Griffin, and M.B. Schulze. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiol. Rev., 33(1):46–62, 2011.

P. R. Burton, D. G. Clayton, L. R. Cardon, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat. Genet., 39(11):1329–1337, Nov 2007.

J. Bustamante, C. Picard, S. Boisson-Dupuis, L. Abel, and J.L. Casanova. Genetic lessons learned from x-linked mendelian susceptibility to mycobacterial diseases. Annals of the New York Academy of Sciences, 1246:92–101, 2011.

F. Capon, M. J. Bijlmakers, N. Wolf, et al. Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene. Hum. Mol. Genet., 17(13):1938–1945, Jul 2008.

A. Cassidy, J.P. Myles, M. van Tongeren, et al. The LLP risk model: an individual risk prediction model for lung cancer. Brit. J. Cancer, 98(2):270–276, 2008.

N. Chatterjee, J. H. Park, N. Caporaso, and M. H. Gail. Predicting the future of genetic risk prediction. Cancer Epidemiol. Biomarkers Prev., 20(1):3–8, Jan 2011.

K. Chen, J. W. Wallis, M. D. McLellan, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods, 6(9): 677–681, Sep 2009.

S. Chen and G. Parmigiani. Meta-analysis of BRCA1 and BRCA2 penetrance. J. Clin. Oncol., 25(11):1329–1333, Apr 2007.

W. Chen, D. Stambolian, A. O. Edwards, et al. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. Proc. Natl. Acad. Sci. U.S.A., 107:7401–7406, 2010.

A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol., 7(2):111–122, Mar 1990.

D. Clayton. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. Genet. Epidemiol., 36(4):409–418, May 2012.

D. Clayton and H. T. Leung. An R package for analysis of whole-genome association studies. Hum. Hered., 64(1):45–51, 2007.

D. G. Clayton. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet., 5(7):e1000540, Jul 2009.

F. S. Collins. Shattuck lecture–medical and societal consequences of the Human Genome Project. N. Engl. J. Med., 341(1):28–37, Jul 1999.

D. Commenges, H. Jacqmin, L. Letenneur, and C. M. Van Duijn. Score test for familial aggregation in probands studies: application to Alzheimer's disease. Biometrics, 51(2):542–551, Jun 1995.

O. Companioni, F. Rodríguez Esparragón, A.M. Fernndez-Aceituno, and J.C. Rodrguez Prez. Genetic Variants, Cardiovascular Risk and Genome-Wide Association Studies. Rev. Esp. Cardio., 64(6):509–514, 2011.

A. Compston and A. Coles. Multiple sclerosis. Lancet, 372(9648):1502–1517, Oct 2008.

P. Concannon, W. M. Chen, C. Julier, et al. Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. Diabetes, 58(4):1018–1022, Apr 2009.

D. F. Conrad, D. Pinto, R. Redon, et al. Origins and functional impact of copy number variation in the human genome. Nature, 464(7289):704–712, Apr 2010.

J. D. Cooper, D. J. Smyth, A. M. Smiles, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet., 40(12):1399–1401, Dec 2008.

J.J. Corneveaux, A.J. Myers, A.N. Allen, et al. Association of CR1, CLU and PICALMwith alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. Hum. Mol. Genet., 19(16):3295–3301, 2010.

A. Cortes and M. A. Brown. Promise and pitfalls of the Immunochip. Arthritis Res. Ther., 13(1):101, 2011.

C. Cotsapas, B. F. Voight, E. Rossin, et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet., 7(8):e1002254, Aug 2011.

J.S. Cramer. The origins and development of the logit model. Logit Models from Economics and Other Fields, pages 149–158, 2003.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U.S.A., 107(50):21931–21936, Dec 2010.

F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. Nature, 192:1227–1232, Dec 1961.

A. G. Cudworth and H. Festenstein. HLA genetic heterogeneity in diabetes mellitus. Br. Med. Bull., 34(3):285–289, Sep 1978.

M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. Nat. Genet., 29(2):229–232, Oct 2001.

P. Danoy, K. Pryce, J. Hadler, et al. Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn's disease. PLoS Genet., 6(12):e1001195, 2010.

E. V. Davydov, D. L. Goode, M. Sirota, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol., 6(12):e1001025, 2010.

P. L. De Jager, X. Jia, J. Wang, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet., 41(7):776–782, Jul 2009.

deCODEme. Genetic risk calculation, July 2012. URL `http://demo.decodeme.com/health-watch-information/risk-calculation`.

A. Deisseroth, A. Nienhuis, J. Lawrence, et al. Chromosomal localization of human beta globin gene on human chromosome 11 in somatic cell hybrids. Proc. Natl. Acad. Sci. U.S.A., 75(3):1456–1460, Mar 1978.

O. Delaneau, J. Marchini, and J. F. Zagury. A linear complexity phasing method for thousands of genomes. Nat. Methods, 9(2):179–181, Feb 2012.

A. Dewan, M. Liu, S. Hartman, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science, 314(5801):989–992, Nov 2006.

H. M. Dick. HLA and disease. Introductory review. Br. Med. Bull., 34(3):271–274, Sep 1978.

C. Dina, D. Meyre, C. Samson, et al. Comment on "A common genetic variant is associated with adult and childhood obesity". Science, 315(5809):187; author reply 187, Jan 2007.

A. L. Dixon, L. Liang, M. F. Moffatt, et al. A genome-wide association study of global gene expression. Nat. Genet., 39(10):1202–1207, Oct 2007.

N. J. Dovichi. DNA sequencing by capillary electrophoresis. Electrophoresis, 18 (12-13):2393–2399, Nov 1997.

I.R. Dowbiggin. Inheriting madness: professionalization and psychiatric knowledge in nineteenth-century France, volume 4. Univ of California Pr on Demand, 1991.

T. R. Dreszer, D. Karolchik, A. S. Zweig, et al. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res., 40(Database issue): D918–923, Jan 2012.

P. C. Dubois, G. Trynka, L. Franke, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet., 42(4):295–302, Apr 2010.

R. H. Duerr, K. D. Taylor, S. R. Brant, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science, 314(5804): 1461–1463, Dec 2006.

R. Durbin. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univeristy Press, 1998.

D. Duricova, N. Pedersen, M. Elkjaer, et al. Overall and cause-specific mortality in Crohn's disease: a meta-analysis of population-based studies. Inflamm. Bowel Dis., 16(2):347–353, Feb 2010.

R.A. Eeles, Z. Kote-Jarai, A.A. Al Olama, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat. Genet., 41(10):1116–1121, 2009.

M. Eisenstein. Oxford nanopore announcement sets sequencing sector abuzz. Nature Biotechnology, 30(4):295–296, 2012.

R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. Hum. Hered., 21(6):523–542, 1971.

V. Emilsson, G. Thorleifsson, B. Zhang, et al. Genetics of gene expression and its effect on disease. Nature, 452(7186):423–428, Mar 2008.

F. Esposito, N. A. Patsopoulos, S. Cepok, et al. IL12A, MPHOSPH9/CDK2AP1 and RGS1 are novel multiple sclerosis susceptibility loci. Genes Immun., 11(5): 397–405, Jul 2010.

D. M. Evans, C. C. Spencer, J. J. Pointon, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat. Genet., 43(8):761–767, Aug 2011.

D.S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics, 29(1):51–76, 1965.

D.S. Falconer and T.F.C Mackay. Introduction to quantitative genetics. Longman., 4th edition, 1996.

J.B. Fan, A. Oliphant, R. Shen, et al. Highly parallel SNP genotyping. In Cold Spring Harbor symposia on quantitative biology, volume 68, pages 69–78. Cold Spring Harbor Laboratory Press, 2003.

M. A. Ferreira, M. C. O'Donovan, Y. A. Meng, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat. Genet., 40(9):1056–1058, Sep 2008.

R. C. Ferreira, Q. Pan-Hammarstrom, R. R. Graham, et al. Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. Nat. Genet., 42(9):777–780, Sep 2010.

E. A. Festen, P. Goyette, R. Scott, et al. Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut, 58(6):799–804, Jun 2009.

O. Filipe-Santos, J. Bustamante, M.H. Haverkamp, et al. X-linked susceptibility to mycobacteria is caused by mutations in NEMO impairing CD40-dependent IL-12 production. The Journal of experimental medicine, 203(7):1745–59, 2006.

R.A. Fisher. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52:399–433, 1918.

R.A. Fisher. The detection of linkage with "dominant" abnormalities. Annals of Human Genetics, 6(2):187–201, 1935.

T. Fiskerstrand, N. Arshad, B. I. Haukanes, et al. Familial diarrhea syndrome caused by an activating GUCY2C mutation. N. Engl. J. Med., 366(17):1586–1595, Apr 2012.

N. Fleshner. Re: Familial clustering of breast and prostate cancers and risk of postmenopausal breast cancer. Journal of the National Cancer Institute, 87(7): 536–537, 1995.

J. L. Flynn, J. Chan, K. J. Triebold, et al. An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection. J. Exp. Med., 178(6): 2249–2254, Dec 1993.

S. P. Fodor, J. L. Read, M. C. Pirrung, et al. Light-directed, spatially addressable parallel chemical synthesis. Science, 251(4995):767–773, Feb 1991.

A. Fontalba, O. Gutierrez, and J. L. Fernandez-Luna. NLRP2, an inhibitor of the NF-kappaB pathway, is transcriptionally activated by NF-kappaB and exhibits a nonfunctional allelic variant. J. Immunol., 179(12):8519–8524, Dec 2007.

A. Franke, T. Balschun, T. H. Karlsen, et al. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat. Genet., 40(11):1319–1323, Nov 2008.

A. Franke, D. P. McGovern, J. C. Barrett, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet., 42(12):1118–1125, Dec 2010.

K.A. Frazer, D.G. Ballinger, D.R. Cox, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164):851–861, 2007.

P. Frederic and F. Lad. A technical note on the logitnormal distribution. Research Report, Mathematics and Statistics department at Canterbury University (NZ), 2003.

J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res., 19(10):1817–1824, Oct 2009.

J. Freudenberg, H. S. Lee, B. G. Han, et al. Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. Arthritis Rheum., 63(4):884–893, Apr 2011.

G. Fu, J. Hu, N. Niederberger-Magnenat, et al. Protein kinase CH is required for T cell activation and homeostatic proliferation. Sci Signal, 4(202):ra84, 2011.

A.E. Garrod. The incidence of alkaptonuria: a study in chemical individuality. Lancet, 2(24):1616–1620, 1902.

V. Gateva, J. K. Sandling, G. Hom, et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat. Genet., 41(11):1228–1233, Nov 2009.

E. O. Glocker, D. Kotlarz, K. Boztug, et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. N. Engl. J. Med., 361(21):2033–2045, Nov 2009a.

E.O. Glocker, A. Hennigs, M. Nabavi, et al. A homozygous card9 mutation in a family with susceptibility to fungal infections. The New England journal of medicine, 361(18):1727–35, 2009b.

A. Goate, M. C. Chartier-Harlin, M. Mullan, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature, 349(6311):704–706, Feb 1991.

A. Gonzalez-Perez and N. Lopez-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet., 88(4):440–449, Apr 2011.

E.L. Goode, G. Chenevix-Trench, H. Song, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat. Genet., 42(10):874–879, 2010.

Y. Goto, A. Hishida, K. Matsuo, et al. PRKCH gene polymorphism is associated with the risk of severe gastric atrophy. Gastric Cancer, 13(2):90–94, Jun 2010.

D. M. Greenawalt, R. Dobrin, E. Chudin, et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res., 21(7):1008–1016, Jul 2011.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol., 3(4):479–502, 1996.

W. Guan, A. Pluzhnikov, N. J. Cox, et al. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. Hum. Hered., 66(1):35–49, 2008.

S. G. Guerra, T. J. Vyse, and D. S. Cunninghame Graham. The genetics of lupus: a functional perspective. Arthritis Res Ther, 14(3):211, May 2012.

J. Gulcher and K. Stefansson. Genetic risk information for common diseases may indeed be already useful for prevention and early detection. Eur. J. Clin. Invest., 40(1):56–63, Jan 2010.

DG Haegert. Analysis of the threshold liability model provides new understanding of causation in autoimmune diseases. Medical hypotheses, 63(2):257–261, 2004.

J.B.S. Haldane. Methods for the detection of autosomal linkage in man. Annals of Human Genetics, 6(1):26–65, 1934.

W. D. Hall, R. Mathews, and K. I. Morley. Being more realistic about the public health impact of genomic medicine. PLoS Med., 7(10), Oct 2010.

J. Hampe, S. Schreiber, S. H. Shaw, et al. A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. Am. J. Hum. Genet., 64(3):808–816, Mar 1999.

J. W. Han, H. F. Zheng, Y. Cui, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet., 41(11):1234–1237, Nov 2009.

R. E. Handsaker, J. M. Korn, J. Nemesh, and S. A. McCarroll. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet., 43(3):269–276, Mar 2011.

J. B. Harley, M. E. Alarcon-Riquelme, L. A. Criswell, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat. Genet., 40(2):204–210, Feb 2008.

D. Harold, R. Abraham, P. Hollingworth, et al. Genome-wide association study identifies variants at CLU and PICALM associated with alzheimer's disease. Nat. Genet., 41(10):1088–1093, 2009.

J. Harrow, F. Denoeud, A. Frankish, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol., 7 Suppl 1:1–9, 2006.

M. Heinig, E. Petretto, C. Wallace, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature, 467(7314):460–464, Sep 2010.

A. Herbert, N. P. Gerry, M. B. McQueen, et al. A common genetic variant is associated with adult and childhood obesity. Science, 312(5771):279–283, Apr 2006.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A., 106(23):9362–9367, Jun 2009.

D. A. Hinds, L. L. Stuve, G. B. Nilsen, et al. Whole-genome patterns of common DNA variation in three human populations. Science, 307(5712):1072–1079, Feb 2005.

A. D. Hingorani, J. P. Casas, D. I. Swerdlow, et al. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet, 379(9822):1214–1224, Mar 2012.

G. M. Hirschfield, X. Liu, C. Xu, et al. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. N. Engl. J. Med., 360(24):2544–2555, Jun 2009.

S. M. Holland, F. R. DeLeo, H. Z. Elloumi, et al. STAT3 mutations in the hyper-IgE syndrome. N. Engl. J. Med., 357(16):1608–1619, Oct 2007.

R.S. Houlston, E. Webb, P. Broderick, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet., 40(12):1426–1435, 2008.

B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet., 44(8):955–959, Aug 2012.

B. Howie, J. Marchini, and M. Stephens. Genotype imputation with thousands of genomes. G3 (Bethesda), 1(6):457–470, Nov 2011.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet., 5(6):e1000529, Jun 2009.

X. Hu, H. Kim, E. Stahl, et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet., 89(4):496–506, Oct 2011.

J. Huang, D. Ellinghaus, A. Franke, B. Howie, and Y. Li. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur. J. Hum. Genet., 20(7):801–805, Jul 2012.

L. Huang, Y. Li, A.B. Singleton, et al. Genotype-imputation accuracy across worldwide human populations. The American Journal of Human Genetics, 84 (2):235–250, 2009.

J. P. Hugot, M. Chamaillard, H. Zouali, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature, 411(6837):599–603, May 2001.

K. A. Hunt, D. J. Smyth, T. Balschun, et al. Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. Nat. Genet., 44(1):3–5, Jan 2012.

K. A. Hunt, A. Zhernakova, G. Turner, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet., 40(4):395–402, Apr 2008.

FB Hutt, WR LAMOREUX, et al. Genetics of the fowl. 11. a linkage map for six chromosomes. Journal of Heredity, 31:231–235, 1940.

G. Hyatt, R. Melamed, R. Park, et al. Gene expression microarrays: glimpses of the immunological genome. Nat. Immunol., 7(7):686–691, Jul 2006.

M. Imielinski, R. N. Baldassano, A. Griffiths, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat. Genet., 41 (12):1335–1340, Dec 2009.

V. M. Ingram. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature, 180(4581):326–328, Aug 1957.

International HapMap Consortium. The International HapMap Project. Nature, 426(6968):789–796, Dec 2003.

International HapMap Consortium. A haplotype map of the human genome. Nature, 437(7063):1299–1320, Oct 2005.

International Parkinson's Disease Genomics Consortium and Wellcome Trust Case Control Consortium 2. A two-stage meta-analysis identifies several new Loci for Parkinson's disease. PLoS Genet., 7:e1002142, 2011.

J. P. Ioannidis, R. Tarone, and J. K. McLaughlin. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology, 22(4):450–456, Jul 2011.

D. A. Jackson, R. H. Symons, and P. Berg. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proc. Natl. Acad. Sci. U.S.A., 69(10):2904–2909, Oct 1972.

E. Jakkula, V. Leppa, A. M. Sulonen, et al. Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am. J. Hum. Genet., 86(2):285–291, Feb 2010.

D. Jawaheer, M. F. Seldin, C. I. Amos, et al. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. Arthritis Rheum., 48(4):906–916, Apr 2003.

G. A. Jervis. Phenylpyruvic oligophrenia deficiency of phenylalanine-oxidizing system. Proc. Soc. Exp. Biol. Med., 82(3):514–515, Mar 1953.

T. Jess, M. Gamborg, P. Munkholm, and T. I. S?rensen. Overall and cause-specific mortality in ulcerative colitis: meta-analysis of population-based inception cohort studies. Am. J. Gastroenterol., 102(3):609–617, Mar 2007.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. Continuous Univariate Distributions, volume I. John Wiley & Sons Inc., New York, 2nd edition, 1994. ISBN 0-471-58495-9.

L. Jostins, K. I. Morley, and J. C. Barrett. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur. J. Hum. Genet., 19(6):662–666, Jun 2011.

T. M. Karafet, F. L. Mendez, M. B. Meilerman, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res., 18(5):830–838, May 2008.

A. Kaser, A. H. Lee, A. Franke, et al. XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. Cell, 134(5):743–756, Sep 2008.

F. K. Kavvoura, T. Akamizu, T. Awata, et al. Cytotoxic T-lymphocyte associated antigen 4 gene polymorphisms and autoimmune thyroid disease: a meta-analysis. J. Clin. Endocrinol. Metab., 92(8):3162–3170, Aug 2007.

J. Keane, S. Gershon, R. P. Wise, et al. Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. N. Engl. J. Med., 345(15): 1098–1104, Oct 2001.

B. Kerem, J. M. Rommens, J. A. Buchanan, et al. Identification of the cystic fibrosis gene: genetic analysis. Science, 245(4922):1073–1080, Sep 1989.

JFC Kingman. On the genealogy of large populations. Journal of Applied Probability, pages 27–43, 1982.

JB Kirsner. Genetic aspects of inflammatory bowel disease. Clin Gastroenterol, 2 (557):76, 1973.

J.B. Kirsner. The historical basis of the idiopathic inflammatory bowel diseases. Inflammatory Bowel Diseases, 1(1):2–26, 1995.

G. D. Kitsios, N. Tangri, P. J. Castaldi, and J. P. Ioannidis. Laboratory mouse models for the human genome-wide associations. PLoS ONE, 5(11):e13782, 2010.

R. J. Klein, C. Zeiss, E. Y. Chew, et al. Complement factor H polymorphism in age-related macular degeneration. Science, 308(5720):385–389, Apr 2005.

R. G. Knowlton, O. Cohen-Haguenauer, N. Van Cong, et al. A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. Nature, 318(6044): 380–382, 1985.

A. Kong and N. J. Cox. Allele-sharing models: LOD scores and accurate linkage tests. Am. J. Hum. Genet., 61(5):1179–1188, Nov 1997.

S. V. Kozyrev, A. K. Abelson, J. Wojcik, et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. Nat. Genet., 40(2):211–216, Feb 2008.

P. Kraft and D. J. Hunter. Genetic risk prediction–are we there yet? N. Engl. J. Med., 360(17):1701–1703, Apr 2009.

L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet., 58(6):1347–1363, Jun 1996.

S. Kugathasan, R. N. Baldassano, J. P. Bradfield, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nat. Genet., 40 (10):1211–1215, Oct 2008.

E. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet., 11(3):241–247, Nov 1995.

E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. U.S.A., 84(8):2363–2367, Apr 1987.

E. S. Lander, L. M. Linton, B. Birren, et al. Initial sequencing and analysis of the human genome. Nature, 409(6822):860–921, Feb 2001.

J.E. Lane-Claypon et al. A further report on cancer of the breast with special reference to its associated antecedent conditions. Ministry of Health. Reports on Public Health and Medical Subjects., 32, 1926.

T. Langaee and M. Ronaghi. Genetic variation analyses by Pyrosequencing. Mutat. Res., 573(1-2):96–102, Jun 2005.

R. M. Lawn, E. F. Fritsch, R. C. Parker, G. Blake, and T. Maniatis. The isolation and characterization of linked delta- and beta-globin genes from a cloned library of human DNA. Cell, 15(4):1157–1174, Dec 1978.

S. Q. Le and R. Durbin. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res., 21(6):952–960, Jun 2011.

J. Lederberg. Infectious disease as an evolutionary paradigm. Emerging Infect. Dis., 3(4):417–423, 1997.

J. C. Lee, P. A. Lyons, E. F. McKinney, et al. Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. J. Clin. Invest., 121(10):4170–4179, Oct 2011.

L. G. Lee, C. R. Connell, and W. Bloch. Allelic discrimination by nick-translation PCR with fluorogenic probes. Nucleic Acids Res., 21(16):3761–3766, Aug 1993.

C. W. Lees, J. C. Barrett, M. Parkes, and J. Satsangi. New IBD genetics: common pathways with other diseases. Gut, 60(12):1739–1753, Dec 2011.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14):1754–1760, Jul 2009.

H. Li, B. Handsaker, A. Wysoker, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–2079, Aug 2009.

J. Li and W.B. Wang. Tag snp selection. Analysis of Complex Disease Association Studies: A Practical Guide, page 49, 2010.

J. Z. Li, D. M. Absher, H. Tang, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science, 319(5866):1100–1104, Feb 2008.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, 165(4): 2213–2233, Dec 2003.

Y. Li, C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res., 21(6):940–951, Jun 2011.

Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol., 34(8):816–834, Dec 2010.

D.C. Link, L.G. Schuettpelz, D. Shen, et al. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA : the journal of the American Medical Association, 305(15):1568–76, 2011.

P. Linsel-Nitschke, A. Gotz, J. Erdmann, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease–a Mendelian Randomisation study. PLoS ONE, 3(8):e2986, 2008.

R. J. Lipshutz, D. Morris, M. Chee, et al. Using oligonucleotide probe arrays to access genetic diversity. BioTechniques, 19(3):442–447, Sep 1995.

Q. R. Liu, T. Drgon, C. Johnson, et al. Addiction molecular genetics: 639,401 SNP whole genome association identifies many ”cell adhesion” genes. Am. J. Med. Genet. B Neuropsychiatr. Genet., 141B(8):918–925, Dec 2006.

A. Lleo, L. Moroni, L. Caliari, and P. Invernizzi. Autoimmunity and Turner's syndrome. Autoimmun Rev, 11(6-7):A538–543, May 2012.

D.M. Lloyd-Jones, K. Liu, L. Tian, and P. Greenland. Narrative review: Assessment of C-reactive protein in risk prediction for cardiovascular disease. Ann. Intern. Med., 145(1):35–42, 2006.

R. J. Loos, I. Barroso, S. O'rahilly, and N. J. Wareham. Comment on "A common genetic variant is associated with adult and childhood obesity". Science, 315 (5809):187; author reply 187, Jan 2007.

Q. Lu and R. C. Elston. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am. J. Hum. Genet., 82(3):641–651, Mar 2008.

Luke Jostins. YFitter: a program for assigning haplogroups using maximum likelihood, 2011. URL `http://sourceforge.net/projects/yfitter/`. v0.2.

V. Lyssenko, A. Jonsson, P. Almgren, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N. Engl. J. Med., 359(21):2220–2232, 2008.

D. G. MacArthur, S. Balasubramanian, A. Frankish, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science, 335(6070): 823–828, Feb 2012.

R. Mackelprang, C. S. Carlson, L. Subrahmanyan, et al. Sequence variation in the human T-cell receptor loci. Immunol. Rev., 190:26–39, Dec 2002.

M. Macpherson, B. Naughton, A. Hsu, and J. Mountain. Estimating genotype-specific incidence for one or several loci, November 2007. URL `https://www.23andme.com/for/scientists/`.

P. Maerten, C. Shen, S. Colpaert, et al. Involvement of interleukin 18 in Crohn's disease: evidence from in vitro analysis of human gut inflammatory cells and from experimental colitis models. Clin. Exp. Immunol., 135(2):310–317, Feb 2004.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. Nature Reviews Genetics, 11(7):499–511, 2010.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet., 39(7):906–913, Jul 2007.

E. R. Mardis. The impact of next-generation sequencing technology on genetics. Trends Genet., 24(3):133–141, Mar 2008.

P. Margaritte, C. Bonaiti-Pellie, M. C. King, and F. Clerget-Darpoux. Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease. Am. J. Hum. Genet., 50(6):1231–1234, Jun 1992.

E. J. Masicampo and D. R. Lalande. A peculiar prevalence of p values just below .05. Q J Exp Psychol (Hove), Aug 2012.

H. Matsuzaki, S. Dong, H. Loi, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat. Methods, 1(2):109–111, Nov 2004a.

H. Matsuzaki, H. Loi, S. Dong, et al. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res., 14(3):414–425, Mar 2004b.

M. N. McCall, K. Uppal, H. A. Jaffee, M. J. Zilliox, and R. A. Irizarry. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res., 39(Database issue): D1011–1015, Jan 2011.

M. I. McCarthy and E. Zeggini. Genome-wide association studies in type 2 diabetes. Curr. Diab. Rep., 9(2):164–171, Apr 2009.

J. L. McCauley, R. L. Zuvich, A. L. Beecham, et al. Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum. Mol. Genet., 19(5):953–962, Mar 2010.

D. McGonagle, A. Aziz, L. J. Dickie, and M. F. McDermott. An integrated classification of pediatric inflammatory diseases, based on the concepts of autoinflammation and the immunological disease continuum. Pediatr. Res., 65(5 Pt 2):38R–45R, May 2009.

A. McKenna, M. Hanna, E. Banks, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20(9):1297–1303, Sep 2010.

W. McLaren, B. Pritchard, D. Rios, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26 (16):2069–2070, Aug 2010.

G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 360(1459):1387–1393, Jul 2005.

G. A. McVean, S. R. Myers, S. Hunt, et al. The fine-scale structure of recombination rate variation in the human genome. Science, 304(5670):581–584, Apr 2004.

R. Mead. A generalised logit-normal distribution. Biometrics, 21(3):721–732, 1965.

G. Mendel. Versuche über pflanzenhybriden. Verh. Naturforsch., 4, 1866.

I. L. Mero, A. R. Lorentzen, M. Ban, et al. A rare variant of the TYK2 gene is confirmed to be associated with multiple sclerosis. Eur. J. Hum. Genet., 18(4): 502–504, Apr 2010.

Y. Miki, J. Swensen, D. Shattuck-Eidens, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science, 266(5182):66–71, Oct 1994.

R. E. Mills, W. S. Pittard, J. M. Mullaney, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res., 21(6): 830–839, Jun 2011a.

R. E. Mills, K. Walter, C. Stewart, et al. Mapping copy number variation by population-scale genome sequencing. Nature, 470(7332):59–65, Feb 2011b.

Y. Minegishi, M. Saito, S. Tsuchiya, et al. Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome. Nature, 448(7157): 1058–1062, Aug 2007.

A. Molven and P. R. Njølstad. Role of molecular genetics in transforming diagnosis of diabetes mellitus. Expert Rev. Mol. Diagn., 11(3):313–320, Apr 2011.

S.P. Morgan and J.D. Teachman. Logistic regression: Description, examples, and comparisons. Journal of Marriage and Family, 50(4):929–936, 1988.

T. M. Morgan, H. M. Krumholz, R. P. Lifton, and J. A. Spertus. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. JAMA, 297(14):1551–1561, Apr 2007.

J. A. Morris, J. C. Randall, J. B. Maller, and J. C. Barrett. Evoker: a visualization tool for genotype intensity data. Bioinformatics, 26(14):1786–1787, Jul 2010.

N. E. Morton. Sequential tests for the detection of linkage. Am. J. Hum. Genet., 7(3):277–318, Sep 1955.

J. Mudter, B. Weigmann, B. Bartsch, et al. Activation pattern of signal transducers and activators of transcription (STAT) factors in inflammatory bowel diseases. Am. J. Gastroenterol., 100(1):64–72, Jan 2005.

K. Mullis, F. Faloona, S. Scharf, et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb. Symp. Quant. Biol., 51 Pt 1:263–273, 1986.

R. M. Myers, J. Stamatoyannopoulos, M. Snyder, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol., 9(4):e1001046, Apr 2011.

R. P. Nair, K. C. Duffin, C. Helms, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet., 41(2):199–204, Feb 2009.

M. A. Nalls, V. Plagnol, D. G. Hernandez, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet, 377:641–649, 2011.

B. M. Neale, Y. Kou, L. Liu, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature, 485(7397):242–245, May 2012.

B. M. Neale, M. A. Rivas, B. F. Voight, et al. Testing for an unusual distribution of rare variants. PLoS Genet., 7(3):e1001322, Mar 2011.

M.C. Neale and L.R. Cardon. Methodology for genetic studies of twins and families. Number 67 in Series D: Behavioural and Social Sciences. Springer, 1992.

A. C. Need, D. K. Attix, J. M. McEvoy, et al. Failure to replicate effect of Kibra on human memory in two large cohorts of European origin. Am. J. Med. Genet. B Neuropsychiatr. Genet., 147B(5):667–668, Jul 2008.

S. Nejentsev, J. M. Howson, N. M. Walker, et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature, 450(7171): 887–892, Dec 2007.

S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science, 324(5925):387–389, Apr 2009.

J. A. Nelder and R. Mead. A simplex algorithm for function minimization. Computer Journal, 7:308–313, 1965.

P. C. Ng, S. S. Murray, S. Levy, and J. C. Venter. An agenda for personalized medicine. Nature, 461(7265):724–726, Oct 2009.

S. B. Ng, K. J. Buckingham, C. Lee, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet., 42(1):30–35, Jan 2010.

NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server, 2012. URL `http://evs.gs.washington.edu/EVS/`.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet., 12(6):443–451, Jun 2011.

Y. Nievergelt. How (not) to solve quadratic equations. The College Mathematics Journal, 34(2):90–104, 2003.

L. Nistico, R. Buzzetti, L. E. Pritchard, et al. The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. Hum. Mol. Genet., 5(7):1075–1080, Jul 1996.

L. D. Notarangelo, A. Fischer, R. S. Geha, et al. Primary immunodeficiencies: 2009 update. J. Allergy Clin. Immunol., 124(6):1161–1178, Dec 2009.

M. C. O'Donovan, N. Craddock, N. Norton, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat. Genet., 40 (9):1053–1055, Sep 2008.

A. Papassotiropoulos, D. A. Stephan, M. J. Huentelman, et al. Common Kibra alleles are associated with human memory performance. Science, 314(5798): 475–478, Oct 2006.

M. Parkes, J. C. Barrett, N. J. Prescott, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet., 39(7):830–832, Jul 2007.

S. Y. Patel, R. Doffinger, G. Barcenas-Morales, and D. S. Kumararatne. Genetically determined susceptibility to mycobacterial infection. J. Clin. Pathol., 61 (9):1006–1012, Sep 2008.

N. Patil, A. J. Berno, D. A. Hinds, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294 (5547):1719–1723, Nov 2001.

L. Pauling, H. A. Itano, S.J. Singer, and I.C. Wells. Sickle cell anemia: a molecular disease. Science, 110(2865):543–548, Nov 1949.

P. D. Pharoah, A. Antoniou, M. Bobrow, et al. Polygenic susceptibility to breast cancer and implications for prevention. Nat. Genet., 31(1):33–36, May 2002.

J.K. Pickrell and J.K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. available from nature precedings http://hdl.handle.net/10101/npre.2012.6956.1. Nature proceedings, 2012.

M. Pop. Genome assembly reborn: recent computational challenges. Brief. Bioinformatics, 10(4):354–366, Jul 2009.

C. Power and J. Elliott. Cohort profile: 1958 British birth cohort (National Child Development Study). International Journal of Epidemiology, 35(1):34–41, 2006.

A. L. Price, M. E. Weale, N. Patterson, et al. Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet., 83(1):132–135, Jul 2008.

A.L. Price, N.J. Patterson, R.M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics, 38(8):904–909, 2006.

J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet., 69(1):124–137, Jul 2001.

The 1000 Genomes Project. A map of human genome variation from population-scale sequencing. Nature, 467(7319):1061–1073, Oct 2010.

The 1000 Genomes Project. An integrated map of genetic variation from 1,092 human genomes. Under revision at Nature, 2012.

R.C. Punnett. Mendelism in relation to disease. Proceedings of the Royal Society of Medicine, 1:135, 1908.

S. Purcell, B. Neale, K. Todd-Brown, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 81(3): 559–575, Sep 2007.

S.M. Purcell, N.R. Wray, J.L. Stone, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460(7256):748–752, 2009.

H. Q. Qu, S. F. Grant, J. P. Bradfield, et al. Association of RASGRP1 with type 1 diabetes is revealed by combined follow-up of two genome-wide studies. J. Med. Genet., 46(8):553–554, Aug 2009.

M. Quail, M. E. Smith, P. Coupland, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics, 13(1):341, Jul 2012.

M. Raabe. Hemophilia. Chelsea House Pub, 2008.

S. Raychaudhuri, R. M. Plenge, E. J. Rossin, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet., 5(6):e1000534, Jun 2009a.

S. Raychaudhuri, B. P. Thomson, E. F. Remmers, et al. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet., 41(12):1313–1318, Dec 2009b.

L.J. Reed and J. Berkson. The application of the logistic function to experimental data. The Journal of Physical Chemistry, 33(5):760–779, 1929.

T. Reich, JW James, and CA Morris. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits*. Annals of human genetics, 36(2):163–184, 1972.

J. D. Reveille, A. M. Sims, P. Danoy, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat. Genet., 42 (2):123–127, Feb 2010.

F.V. Rijsdijk and P.C. Sham. Analytic approaches to twin data using structural equation models. Briefings in bioinformatics, 3(2):119–133, 2002.

J. R. Riordan, J. M. Rommens, B. Kerem, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science, 245(4922): 1066–1073, Sep 1989.

J. Rios, E. Stein, J. Shendure, H.H. Hobbs, and J.C. Cohen. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum. Mol. Genet., 19(22):4313–4318, 2010.

N. Risch. Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet., 46(2):222–228, Feb 1990.

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. Science, 273(5281):1516–1517, Sep 1996.

M. A. Rivas, M. Beaudoin, A. Gardet, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet., 43(11):1066–1073, Nov 2011.

J. M. Rommens, M. C. Iannuzzi, B. Kerem, et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science, 245(4922):1059–1065, Sep 1989.

K. R. Rosenbloom, T. R. Dreszer, J. C. Long, et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res., 40(Database issue):D912–917, Jan 2012.

E. J. Rossin, K. Lage, S. Raychaudhuri, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet., 7(1):e1001273, 2011.

D. Rosskopf, A. Bornhorst, C. Rimmbach, et al. Comment on "A common genetic variant is associated with adult and childhood obesity". Science, 315(5809):187; author reply 187, Jan 2007.

M. Ruffalo, T. LaFramboise, and M. Koyuturk. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics, 27(20):2790–2796, Oct 2011.

R.K. Russell and J. Satsangi. IBD: a family affair. Best Practice & Research Clinical Gastroenterology, 18(3):525–539, 2004.

R. Sachidanandam, D. Weissman, S. C. Schmidt, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409(6822):928–933, Feb 2001.

D.L. Sackett, J.J. Deeks, and D.G. Altman. Down with odds ratios! Evidence Based Medicine, 1(6):164–166, 1996.

N. J. Samani, J. Erdmann, A. S. Hall, et al. Genomewide association analysis of coronary artery disease. N. Engl. J. Med., 357(5):443–453, Aug 2007.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A., 74(12):5463–5467, Dec 1977.

J. Satsangi, D. P. Jewell, and J. I. Bell. The genetics of inflammatory bowel disease. Gut, 40(5):572–574, May 1997.

S. Sawcer. The genetic aspects of multiple sclerosis. Ann Indian Acad Neurol, 12 (4):206–214, Oct 2009.

S. Sawcer, M. Ban, J. Wason, and F. Dudbridge. What role for genetics in the prediction of multiple sclerosis? Ann. Neurol., 67(1):3–10, Jan 2010.

S. Sawcer, G. Hellenthal, M. Pirinen, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature, 476(7359):214–219, Aug 2011.

S. Sawcer and J. Wason. Risk in complex genetics: "All models are wrong but some are useful". Ann Neurol, Apr 2012.

E. E. Schadt, C. Molony, E. Chudin, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol., 6(5):e107, May 2008.

K. Schaper, H. Kolsch, J. Popp, M. Wagner, and F. Jessen. KIBRA gene variants are associated with episodic memory in healthy elderly. Neurobiol. Aging, 29 (7):1123–1125, Jul 2008.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet., 78(4):629–644, Apr 2006.

N.R. Schneider, WR Williams, RSK Chaganti, et al. Genetic epidemiology of familial aggregation of cancer. Advances in cancer research, 47:1–36, 1986.

H. Schunkert, I. R. Konig, S. Kathiresan, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet., 43: 333–338, 2011.

L.J. Scott, P. Muglia, X.Q. Kong, et al. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc. Natl. Acad. Sci. U.S.A., 106(18):7501–7506, 2009.

R. A. Scott, V. Lagou, R. P. Welch, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat. Genet., Aug 2012.

J.M. Seddon, R. Reynolds, J. Maller, et al. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. Invest. Ophth. Vis. Sci., 50(5):2044–2053, 2009.

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet., 3(7):e114, Jul 2007.

E. Sezgin, J. M. Lind, S. Shrestha, et al. Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. Hum. Genet., 125(3):281–294, Apr 2009.

T. S. Shah, J. Z. Liu, J. A. Floyd, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics, 28(12): 1598–1603, Jun 2012.

M. H. Shaw, N. Kamada, N. Warner, Y. G. Kim, and G. Nunez. The ever-expanding function of NOD2: autophagy, viral recognition, and T cell activation. Trends Immunol., 32(2):73–79, Feb 2011.

P. Sherlock, B. M. Bell, H. Steinberg, and T. P. Almy. Familial occurrence of regional enteritis and ulcerative colitis. Gastroenterology, 45:413–420, Sep 1963.

S. T. Sherry, M. Ward, and K. Sirotkin. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res., 9 (8):677–679, Aug 1999.

S. Shivananda, J. Lennard-Jones, R. Logan, et al. 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(5):690–697, Nov 1996.

S.I. Shyn, J. Shi, J.B. Kraft, et al. Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. Mol. Psychiatr., 16(2):202–215, 2009.

A. Siepel, G. Bejerano, J. S. Pedersen, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15(8):1034–1050, Aug 2005.

M. S. Silverberg, J. H. Cho, J. D. Rioux, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat. Genet., 41(2):216–220, Feb 2009.

L. M. Smith, J. Z. Sanders, R. J. Kaiser, et al. Fluorescence detection in automated DNA sequence analysis. Nature, 321(6071):674–679, 1986.

D. J. Smyth, V. Plagnol, N. M. Walker, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med., 359(26):2767–2777, Dec 2008.

H. Song, S. J. Ramus, J. Tyrer, et al. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. Nat. Genet., 41:996–1000, 2009.

E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol., 98(3):503–517, Nov 1975.

P. H. St George-Hyslop, R. E. Tanzi, R. J. Polinsky, et al. The genetic defect causing familial Alzheimer's disease maps on chromosome 21. Science, 235(4791): 885–890, Feb 1987.

E. A. Stahl, S. Raychaudhuri, E. F. Remmers, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet., 42(6):508–514, Jun 2010.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68(4):978–989, Apr 2001.

A. Sterner-Kock, I. S. Thorey, K. Koli, et al. Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer. Genes Dev., 16(17):2264–2273, Sep 2002.

R. Stiratelli, N. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. Biometrics, 40(4):961–971, Dec 1984.

A. Strange, F. Capon, C. C. Spencer, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat. Genet., 42(11):985–990, Nov 2010.

P. E. Stuart, R. P. Nair, E. Ellinghaus, et al. Genome-wide association analysis identifies three psoriasis susceptibility loci. Nat. Genet., 42(11):1000–1004, Nov 2010.

Y. J. Sung, L. Wang, T. Rankinen, C. Bouchard, and D. C. Rao. Performance of genotype imputations using data from the 1000 Genomes Project. Hum. Hered., 73(1):18–25, 2012.

I. Surolia, S. P. Pirnie, V. Chellappa, et al. Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. Nature, 466(7303):243–247, Jul 2010.

A. D. Swafford, J. M. Howson, L. J. Davison, et al. An allele of IKZF1 (Ikaros) conferring susceptibility to childhood acute lymphoblastic leukemia protects against type 1 diabetes. Diabetes, 60(3):1041–1044, Mar 2011.

A. C. Syvanen. Toward genome-wide SNP genotyping. Nat. Genet., 37 Suppl: 5–10, Jun 2005.

Y. Y. Teo, A. E. Fry, K. Bhattacharya, et al. Genome-wide comparisons of variation in linkage disequilibrium. Genome Res., 19(10):1849–1860, Oct 2009.

Y. Y. Teo, M. Inouye, K. S. Small, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics, 23(20):2741–2746, Oct 2007.

Y. Y. Teo, K. S. Small, and D. P. Kwiatkowski. Methodological challenges of genome-wide association analysis in Africa. Nat. Rev. Genet., 11(2):149–160, Feb 2010.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 489:57–74, Sep 2012.

The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. . Cell, 72(6):971–983, Mar 1993.

The MalariaGEN Consortium. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat. Genet., 41(6):657–665, Jun 2009.

W. Thomson, A. Barton, X. Ke, et al. Rheumatoid arthritis association at 6q23. Nature genetics, 39(12):1431–1433, 2007.

T. Thye, E. Owusu-Dabo, F. O. Vannberg, et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat. Genet., 44(3):257–259, Mar 2012.

J. A. Todd, J. I. Bell, and H. O. McDevitt. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature, 329 (6140):599–604, 1987.

J. A. Todd, N. M. Walker, J. D. Cooper, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet., 39(7):857–864, Jul 2007.

S. P. Tsai, R. J. Hardy, and C. P. Wen. The standardized mortality ratio and life expectancy. Am. J. Epidemiol., 135(7):824–831, Apr 1992.

L. C. Tsui, M. Buchwald, D. Barker, et al. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science, 230(4729):1054–1057, Nov 1985.

C. Turnbull, S. Ahmed, J. Morrison, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat. Genet., 42(6):504–517, 2010.

D. A. van Heel, S. A. Fisher, A. Kirby, et al. Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. Hum. Mol. Genet., 13(7):763–770, Apr 2004.

B. Vanhaesebroeck, J. Guillermet-Guibert, M. Graupera, and B. Bilanges. The emerging mechanisms of isoform-specific PI3K signalling. Nat. Rev. Mol. Cell Biol., 11(5):329–341, May 2010.

M.H. Vatn. Environmental factors in the epidemiology of inflammatory bowel disease. Crohn's Disease and Ulcerative Colitis: From Epidemiology and Immunobiology to a Rational Diagnostic and Therapeutic Approach, page 17, 2011.

J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. Science, 291(5507):1304–1351, Feb 2001.

P.F. Verhulst. Notice sur la loi que la population suit dans son accroissement. Correspondance Mathématique et Physique, publiée par A. Quetelet, 10:113–121, 1838.

B. F. Voight, G. M. Peloso, M. Orho-Melander, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet, 380(9841):572–580, Aug 2012.

B. F. Voight, L. J. Scott, V. Steinthorsdottir, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet., 42:579–589, 2010.

S. Wacholder, P. Hartge, R. Prentice, et al. Performance of common genetic variants in breast-cancer risk models. New Engl. J. Med., 362(11):986–993, 2010.

B. J. Wainwright, P. J. Scambler, J. Schmidtke, et al. Localization of cystic fibrosis locus to human chromosome 7cen-q22. Nature, 318(6044):384–385, 1985.

C. Wallace, D. J. Smyth, M. Maisuria-Armer, et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat. Genet., 42(1):68–71, Jan 2010.

D. G. Wang, J. B. Fan, C. J. Siao, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science, 280(5366):1077–1082, May 1998.

K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38(16):e164, Sep 2010.

Y. H. Wang and Y. J. Liu. OX40-OX40L interactions: a promising therapeutic target for allergic diseases? J. Clin. Invest., 117(12):3655–3657, Dec 2007.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171(4356):737–738, Apr 1953.

J. L. Weber and P. E. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet., 44(3): 388–396, Mar 1989.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447 (7145):661–678, Jun 2007.

K. A. Wetterstrand. DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. Retrieved 16/08/2012, Available at: www. genome. gov/sequencingcosts. Accessed [2011-10-25], 2012.

C. C. Whitacre. Sex differences in autoimmune disease. Nat. Immunol., 2(9): 777–780, Sep 2001.

R. White, S. Woodward, M. Leppert, et al. A closely linked genetic marker for cystic fibrosis. Nature, 318(6044):382–384, 1985.

C. N. Williams, K. Kocher, E. S. Lander, M. J. Daly, and J. D. Rioux. Using a genome-wide scan and meta-analysis to identify a novel IBD locus and confirm previously identified IBD loci. Inflamm. Bowel Dis., 8(6):375–381, Nov 2002.

L. H. Wise, J. S. Lanchbury, and C. M. Lewis. Meta-analysis of genome searches. Ann. Hum. Genet., 63(Pt 3):263–272, May 1999.

M. G. Wolfs, M. H. Hofker, C. Wijmenga, and T. W. van Haeften. Type 2 Diabetes Mellitus: New Genetic Insights will Lead to New Therapeutics. Curr. Genomics, 10(2):110–118, Apr 2009.

K. Wong, T. M. Keane, J. Stalker, and D. J. Adams. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol., 11(12):R128, 2010.

E. A. Worthey, A. N. Mayer, G. D. Syverson, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med., 13(3):255–262, Mar 2011a.

E.A. Worthey, A.N. Mayer, G.D. Syverson, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med., 13(3):255–262, 2011b.

N. R. Wray and M. E. Goddard. Multi-locus models of genetic risk of disease. Genome Med, 2(2):10, 2010.

N. R. Wray, M. E. Goddard, and P. M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res., 17(10): 1520–1528, Oct 2007.

N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet., 6(2): e1000864, Feb 2010.

C. F. Wright and S. Gregory-Jones. Size of the direct-to-consumer genomic testing market. Genet. Med., 12(9):594, Sep 2010.

J. Yang, P. M. Visscher, and N. R. Wray. Sporadic cases are the norm for complex disease. Eur. J. Hum. Genet., 18(9):1039–1043, Sep 2010.

S. Yazdanyar, M. Weischer, and B. G. Nordestgaard. Genotyping for NOD2 genetic variants and crohn disease: a metaanalysis. Clin. Chem., 55:1950–1957, 2009.

S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res., 19(9):1586–1592, Sep 2009.

A. P. Yu, L. A. Cabanilla, E. Q. Wu, P. M. Mulani, and J. Chao. The costs of Crohn's disease in the United States and other Western countries: a systematic review. Curr Med Res Opin, 24(2):319–328, Feb 2008.

E. Zeggini, L. J. Scott, R. Saxena, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet., 40(5):638–645, May 2008.

E. Zeggini, M. N. Weedon, C. M. Lindgren, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science, 316(5829):1336–1341, Jun 2007.

T. Zeller, P. Wild, S. Szymczak, et al. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. PLoS ONE, 5(5):e10693, 2010.

D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18(5):821–829, May 2008.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 4:Article17, 2005.

F. Zhang, H. Liu, S. Chen, et al. Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. Nature genetics, 43(12):1247–51, 2011.

H. Zhang, D. Massey, M. Tremelling, and M. Parkes. Genetics of inflammatory bowel disease: clues to pathogenesis. Br. Med. Bull., 87:17–30, 2008.

J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res., 115(3-4): 205–214, 2006.

X. J. Zhang, W. Huang, S. Yang, et al. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. Nat. Genet., 41(2):205–210, Feb 2009.

J. Zhu, M. C. Wiener, C. Zhang, et al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. PLoS Comput. Biol., 3(4):e69, Apr 2007.

O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U.S.A., 109(4):1193–1198, Jan 2012.