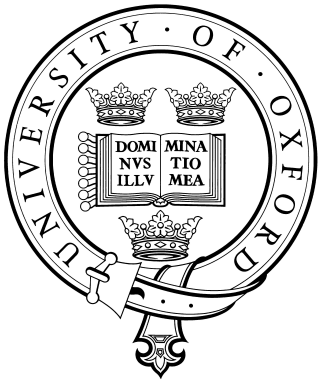
Design and analysis of genome-wide association studies



Jeffrey Charles Barrett

Department of Statistics

Brasenose College

Thesis submitted for the degree of Doctor of Philosophy

Hillary Term, 2008

Design and analysis of genome-wide association studies

Jeffrey Charles Barrett
Brasenose College
D. Phil Thesis, Hillary Term 2008

Abstract

Despite many years of effort, linkage and candidate gene association studies have yielded disappointingly few risk loci for common human diseases such as diabetes, auto-immune disorders and cancers. Large sample sizes, increased understanding of the patterns of correlation in genetic variation, and plunging genotyping costs have enabled genome-wide association studies, which have good power to detect common risk alleles of modest effect. I present an evaluation of SNP choice in study design and show that overall, despite substantial differences in genotyping technologies, marker selection strategies and number of markers assayed, the first generation platforms all offer good levels of genome coverage ($\sim 70\%$). I next describe the largest such project undertaken to date, the Wellcome Trust Case Control Consortium, which consisted of 2000 cases from each of seven common diseases and 3000 shared controls. It identified nearly two dozen new associations. I demonstrate the importance of careful data quality control, including both standard and unorthodox analyses. I next focus on the association results therein for Crohn's disease. I present a replication experiment in over 1000 additional Crohn's patients which unambiguously confirmed six previously published loci and four new loci. Next I describe, in a general context, several issues impeding the combination of genome-wide scans, including data annotation, population structure and differences in genotyping platform. Each of these problems is shown to be tractable with available methods, provided that these methods are applied prudently. I present the results of a meta-analysis of three genome-wide scans for Crohn's disease. The data showed a striking excess of significant associations, and a replication experiment involving over 4000 independent Crohn's patients verified twenty new risk loci. Finally, I discuss the early success of genome-wide association and its consequences for further understanding the biology of human disease.

Acknowledgements

I would like to acknowledge primarily my advisor, Professor Lon Cardon; if I could start my degree again there's noone I would rather have worked for. This work was funded in his lab by the Wellcome Trust.

I am indebted to all of my colleagues on the WTCCC Analysis Group, including Andrew Morris, Jonathan Marchini and Chris Spencer, and especially Peter Donnelly and David Clayton for providing inspiration and mentorship. I would also like to thank all of my collaborators in the UK IBD Genetics Group, especially Miles Parkes, Chris Mathew and Jack Satsangi as well as the members of the International Crohn's Meta-analysis Group. It was a great privilege to have been involved in all of these projects.

It would have been a far more stressful and less interesting three years without the support of many friends and colleagues at the WTCHG, including Dave Evans, Rob Lawrence, Fred Pettersson, Mahim Jain, Krina Zondervan, Nic Timpson, Blanca Herrera and Cecilia Lindgren. I would especially like to acknowledge my desk-mate and partner in IBD crime, Carl Anderson.

I thank my mother and father for teaching me the difference between intelligence and wisdom, and Jenny for helping me see this through to the end. Finally, I would like to thank Mark Daly for encouraging me to become a scientist for my whole life, providing me with my first opportunity to work in genetics, urging me to move to Oxford, and giving me two invaluable pieces of advice: "Don't bother with linkage, and if you're looking for a good phenotype, try Crohn's."

Attributions

I have had the good fortune to participate in a number of collaborative projects during my DPhil, sometimes on a very large scale. I am indebted to all of my collaborators for the incredibly collegiate atmosphere in which I've had the pleasure to work. The following individuals directly did work which is discussed in this thesis:

In Chapter 3, normalisation of the intensity data was done by Hin-Tak Leung. CHIAMO was written by YY Teo, Chris Spencer and Jonathan Marchini. Within-UK population structure evaluation was carried out by David Clayton and Dan Davison.

Samples used in Chapter 4 were collected in the labs of Miles Parkes, Jack Satsangi, John Mansfield, Derek Jewell and Chris Mathew as part of the UK IBD Genetics Consortium. Genotyping and sequencing was done by Natalie Prescott, Roland Roberts, Richard Bagnall and Denis Burke.

Replication samples in Chapter 6 came from the UK IBD Genetics Consortium, the NIDDK IBD Genetics Consortium, and the Belgian-French IBD Genetics Consortium. Genotyping was carried out at the Sanger Institute in the UK, at Sequenom Genetics Services in the US and at the Centre National de Génotypage in France. Todd Green and Sarah Hansoul generated the imputed data for the NIDDK and Belgian-French groups, respectively. The eQTL analysis was performed by members of the Georges group in Liège.

Associated Publications

Barrett, J.C. and Cardon L.R. (2006). Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**:659–62.

The Wellcome Trust Case Control Consortium[1] (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**:66–78.

Parkes, M.[2], **Barrett, J. C.**[2], Prescott, N. J.[2], Tremelling, M., Anderson, C. A., Fisher, S. A., Roberts, R. G., Nimmo, E. R., Cummings, F. R., Soars, D., Drummond, H., Lees, C. W., Khawaja, S. A., Bagnall, R., Burke, D. A., Todhunter, C. E., Ahmad, T., Onnie, C. M., McArdle, W., Strachan, D., Bethel, G., Bryan, C., Lewis, C. M., Deloukas, P., Forbes, A., Sanderson, J., Jewell, D. P., Satsangi, J., Mansfield, J. C., Cardon, L. and Mathew, C. G. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**:830–2.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton A., Dassopoulos, T., Wu Datta, L., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R., the NIDDK IBD Genetics Consortium, Libioulle, C., Sandor,

[1]Full author list in Appendix A
[2]These authors contributed equally to this work

C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.P., de Vos, M., Vermeire, S., Louis, E., the Wellcome Trust Case Control Consortium, Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghori, J., Bumpstead, S., Gwillam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathrew, C.G., Parkes, M., Georges, M., Daly, M.J. (2008). Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**:955–62.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human disease genetics

For most common human diseases it is impossible to predict with certainty that any particular person will become sick, or even retrospectively to explain the causes of illness. Epidemiological studies have identified many *environmental* risk factors for disease, but even the largest of these are not truly predictive: heavy smoking increases risk for lung cancer by twenty fold (Khuder, 2001), nonetheless, most smokers do not get cancer and thousands of non-smokers do. It is not surprising that disease susceptibility is difficult to dissect, because the underlying biological processes which drive human disease (and indeed all phenotypic differences) are incredibly complex and subtle. Each person's 'portfolio' of disease risk is affected not only by his environment, but also by his unique *genetic* make-up, as well as interactions between and among these factors. This, then, is the challenge of human disease genetics: to translate genetic experiments into an understanding of fundamental biology.

Because it is this detailed biological knowledge which is sought, rather than the identity of any particular genetic risk factor, it is possible for genetic studies to be useful even for diseases which are not heritable in the traditional sense. If, for instance, a particular genotype increases risk only in the presence of a

relatively rare environmental factor, no familial aggregation would be observed, despite the importance of that gene in disease aetiology. Still, it is obviously more straightforward to discover genetic effects for traits which are, in fact, heritable to some extent. Traditional genetic epidemiology findings which indicate heritability include a high sibling-relative-risk (the ratio of the risk of an individual with an affected sibling to the general population prevalence, denoted λ_S) and increased trait concordance between monozygotic twins compared with dizygotic twins.

Human disease genetics has been historically focused on conditions at the most tractable end of the heritability spectrum, those following a simple inheritance pattern implicating a single, controlling gene. These basic patterns of inheritance include *dominant*, where a single copy of the predisposing genotype causes disease, and *recessive* where two copies are required for disease, and indviduals with one copy are known as *carriers*. These patterns were first elucidated in plants by Austrian monk Gregor Mendel (1865) and then described in humans by Archibald Garrod (1902) when he observed a recessive pattern of inheritance of alkaptonuria (a rare disorder related to tyrosine metabolism). Of course, simply identifying the pattern of inheritance itself is not very helpful in understanding the pathology of disease; ideally the gene involved could be identified (e.g. *HGD* in alkaptonuria) and offer clues toward biological understanding.

The idea of finding disease genes by tracing the inheritance patterns in families of chromosomal segments which are 'linked' to the disease, that is, segments which segregate with the gene, was proposed by Fisher (1935) but was severely limited in practice by the paucity of heritable markers, such as red cell antigen and serum protein levels, which could be biologically measured at the time. Theoretical advances in linkage methodology in the mid 20th century (for instance, the formulation of the LOD score by Morton (1955) and the pedigree likelihood approach of Elston and Stewart (1971)) set the stage for systematic gene mapping for these 'Mendelian' diseases, but there was an unbalanced state

between sophistication of the statistics and the laboratory practices of the time. For the first of many occasions the potential of disease genetics was enabled by advances in one discipline but temporarily held back by another, underscoring the cross-disciplinary nature of the field.

The discovery of DNA as the genetic molecule by Watson and Crick (1953) was the first step towards making gene mapping routine, but the problem of finding biological markers of inheritance remained. A critical breakthrough was the identification of restriction fragment length polymorphisms (RFLPs), where polymorphic stretches of DNA could be measured by the relative sizes of restriction fragments yielded by endonuclease cutting. RFLPs are randomly distributed throughout the genome, relatively (compared to measuring protein markers) easy to genotype by agarose gel electrophoresis and could be discovered and queried without any prior knowledge of their genomic location. This biological discovery allowed Botstein *et al.* (1980) to propose the construction of a genome-wide linkage map of RFLPs. This scientific condition of both feasible biological experiments to test for linkage and statistical methods to analyse those data quickly yielded disease genes, beginning with the linkage of Huntington's Disease to chromosome 4p using only 12 RFLPs by Gusella *et al.* (1983).

The rate and efficiency with which such Mendelian linkages could be pinpointed (known as 'positional cloning') exploded over the next 20 years as a result of parallel advances in genotyping (notably the development of the polymerase chain reaction (PCR) by Mullis *et al.* (1986) and the discovery of abundant short tandem repeats or 'microsatellites', which are easily genotyped by PCR, by Weber and May (1989)) and methods for statistical analysis (Kruglyak *et al.*, 1996; Abecasis *et al.*, 2002). As of 2007 over 2400 Mendelian traits have been mapped (National Institutes of Health, 2007), a success rate which encouraged the hope that the same approach could be applied to more complex traits with a larger societal health burden such as diabetes, psychiatric illness and cancer. Unfortunately, numerous large scale linkage scans for complex traits have yielded disappointingly few genes (see discussion in McCarthy (2002), for

instance).

This lack of success is perhaps unsurprising in light of the differences between monogenic and complex diseases in terms of their aetiologies. The former are controlled by a single, rare genetic factor which explains all (or nearly all) of the variance for the trait — precisely the type of effect for which linkage studies are designed. Complex diseases, on the other hand, are expected to have many risk factors of modest effect found at relatively high frequency in the general population (the 'common disease, common variant' hypothesis). Tracing the segregation of these effects in families by linkage is difficult because the genetic risk is distributed across many disparate genomic locations rather than concentrated in one chromosomal region.

Whereas linkage approaches work via the observation of large chromosomal segments segregating in pedigrees, the alternative approach of 'association mapping' seeks to directly test for different allele frequencies between disease cases and healthy controls at a population level. The interplay between these two approaches was discussed by Risch and Merikangas (1996), who described a straightforward thought experiment about the relative number of families required to find a modest genetic risk factor under several plausible scenarios by linkage or association. While linkage studies are underpowered to find such effects without enormous numbers of families, association studies are well suited to the task with much more tractable sample sizes — provided the critical assumption that one could choose the right variant to test.

1.2 Patterns of linkage disequilibrium and the HapMap

The most simplistic association study design would be to test every polymorphic marker in the genome. Early attempts at this *direct* association approach in candidate genes or candidate regions were crippled by incomplete catalogues of variation or relied on expensive and unpredictable resequencing. If only a

handful of variants in a region are tested there is no way of distinguishing a true absence of association from a lack of information about the full spectrum of variation at the locus. The first critical ingredient to expanding the utility of the association approach was thus a massive effort at discovering and cataloguing genomic variation.

Creating this catalogue required a draft of the human genome sequence (Lander *et al.*, 2001; Venter *et al.*, 2001), which provided a template against which to compare other sequences and, importantly, created an academic infrastructure for generating large amounts of DNA sequence. Several of the biggest contributors to the Human Genome Project turned their sequencing capacity towards sequencing new individuals, discovering genetic variants throughout the genome.

The sparse map of microsatellite repeats which was so important to the success of linkage studies did not lend itself well to association mapping for several reasons. First, microsatellites are not abundant enough in the genome to provide the requisite density of markers. Second, they are highly mutable, making them imperfect records of human population history. Finally, their multiallelic nature makes them difficult to genotype in a high-throughput fashion. Single nucleotide polymorphisms (SNPs), on the other hand, suffer from none of these drawbacks: they are the most common kind of genetic variation, have a very low recurrent mutation rate and are highly amenable to high-throughput genotyping. Thus the International SNP Map Working Group, which consisted of several large sequencing centres, generated a genome-wide catalogue of variation consisting of nearly 1.5 million SNPs (Sachidanandam *et al.*, 2001).

Further resequencing efforts have since discovered over 10 million SNPs, all deposited in the public domain (Figure 1.1). In very short order the first problem confronting researchers seeking to perform association studies — a lack of markers — was replaced by a superabundance. The SNP map is now so dense, in fact, that genotyping all known polymorphisms in a gene (taking the direct approach to its logical conclusion) has become a prohibitively expensive and

time consuming task.

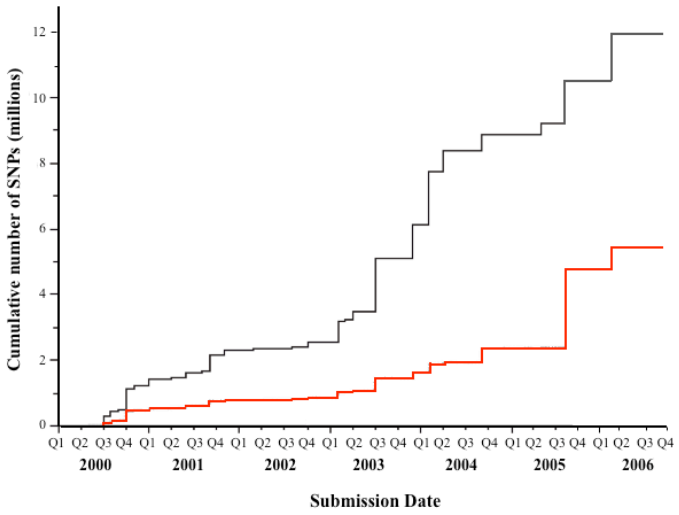


Figure 1.1: The cumulative number of non-redundant SNPs in dbSNP is shown in black. Number of SNPs validated by genotyping is shown in red. Originally published as part of the International HapMap Consortium (2005), updated in personal communication from Robert Lawrence.

Nearby variants in the human genome are not independent, however and the next step towards realising the vision of efficient, comprehensive association studies was understanding this pattern of correlation among the new library of millions of variants. Theoretical examinations of these statistical correlations (called linkage disequilibrium, or LD) based on population genetics models predicted strong correlation only in nearby markers with a rapid drop off to independence (Kruglyak, 1999). These models were contradicted, however by an intriguing pattern discovered by Daly *et al.* (2001) in a region of chromosome 5q31, where long stretches of SNPs were highly correlated, with only a handful of patterns of SNP alleles (called haplotypes) explaining nearly all the observed variation. These 'blocks' of LD were separated by highly punctate breakpoints across which there existed very little correlation, even over short distances.

This observation sparked both an effort to discover whether this phenomenon

was common in the genome and a debate about the causes of these patterns. The first relatively large scale study (Gabriel *et al.*, 2002) aimed at resolving these questions examined 51 regions from across the genome and observed the same pattern of stretches of high LD (and low haplotype diversity) divided by narrow breakpoints. As these patterns were seen in still more experimental studies (Reich *et al.*, 2002), an explanation was posited by unrelated work which examined single recombination events observed in sperm meioses (Jeffreys *et al.*, 2001). Nearly all detected recombinations were confined to narrow 'hotspots' as opposed to being homogenously distributed across the sequence. A picture emerged of long stretches of the genome characterised by low levels of historical recombination broken by intense recombination hotspots.

The hotspot explanation did not, however, win immediate widespread acceptance, nor was the 'block view' of LD considered to fully capture the subtlety of the underlying pattern of variation. In order to resolve these issues and to generate a public resource of human variation to complement the public genome sequence the International HapMap Consortium (2005) was launched. The project set out to discover and genotype a common (minor allele frequency (MAF) $> 5\%$) SNP every five kilobases in samples from four human populations: 90 individuals (30 parent-offspring trios) from the Yoruba in Ibadan, Nigeria (abbreviation YRI); 90 individuals (30 trios) in Utah, USA (CEU); 45 Han Chinese in Beijing, China (CHB); 45 Japanese in Tokyo, Japan (JPT). This dense genotype set, along with initial analyses was published in 2005 and followed up (International HapMap Consortium, 2007) with over two million additional SNPs and more in-depth analysis. Results from the HapMap yielded insight in a wide variety of areas of human genetics research, but the two most relevant to the development of association studies were the concept of choosing a subset of markers which capture as much information as possible (called *tagging*) and the technological advances which brought the price of genotyping down by several orders of magnitude.

1.2.1 Tagging

The idea of trying to exploit the correlation among nearby variants to select an efficient subset which captures as much variation as possible is straightforward, and a simple method proposed by Carlson *et al.* (2004b) and further developed by deBakker *et al.* (2005) took advantage of this correlation (LD) on a genome-wide scale. The basic concept consists of using the r^2 measure of LD to thin out redundant SNPs. Carlson *et al.* (2004b) proposed a simple greedy algorithm whereby one first picks the SNP which is most highly correlated with the largest number of other SNPs and chooses it as a tag. All the SNPs which have r^2 greater than a given threshold are considered captured by that tag and are removed from further consideration. The procedure is now repeated on the remaining SNPs until all SNPs are captured.

The first large scale evaluation of this approach (deBakker *et al.*, 2005) found that choosing tag SNPs could maintain 100% of the power of testing all known SNPs while reducing the actual number of genotyped markers by a factor of 2-3 depending on the population sample. The authors next tried using multimarker proxies to further increase their efficiency. In essence, they looked for two- or three-marker haplotype combinations of tag SNPs which in turn captured additional SNPs. This approach generally yielded an additional 25-30% efficiency over the simpler, pairwise approach. A trade-off exists between power and efficiency: even after removing the totally redundant SNPs, one could continue to remove largely redundant SNPs with only a marginal loss in power. This concept would prove important in designing association studies, since researchers could maximise their power for a given SNP genotyping budget.

1.2.2 Genotyping

The progress of disease genetics has often been mirrored by discoveries about the ways in which the genome sequence varies and the accompanying technologies for querying those variations. Just as linkage mapping was first enabled by the

discovery of RFLPs and blossomed with microsatellite maps, now a catalogue of SNPs made association studies feasible. Existing techniques for genotyping, however, were both slow (requiring a separate reaction for each SNP) and expensive. This hurdle was lowered over the course of the next few years as the speed and price of genotyping SNPs plunged. Driven in part by the HapMap, technologies which allow a handpicked set of thousands of SNPs to be assayed simultaneously and cheaply have become commonplace.

1.3 Genome-wide association studies

All of the developments culminating in the completion of the HapMap have application to association studies of any size. Indeed, until recently only studies within a candidate gene or region were technologically and economically feasible. Studies were designed to investigate biologically plausible candidates or regions implicated by other evidence such as modest linkage. These approaches yielded a few positive results (examples include *PPAR*γ for type 2 diabetes (Deeb *et al.*, 1998) as a biological candidate and *NOD2* for inflammatory bowel disease (Hugot *et al.*, 2001) as a linkage candidate), but are generally regarded as having limited success. Reasons for this include the aforementioned historical lack of a good SNP map but also inadequate samples for detecting realistic effect sizes. Diligent collection of large, carefully phenotyped samples have recently made progress on this front.

Finally, candidacy-driven associations have been limited by the ability to actually choose good biological candidates. This difficulty highlights one clear advantage linkage studies have traditionally had over association studies: they require no prior hypothesis about the genomic location or function of risk variants. This was principally an economic distinction, because only a few hundred polymorphisms (an easily affordable number) are needed for genome-wide linkage coverage, compared to hundreds of thousands for association. As has been described, the types of risk loci which were well suited to discovery by linkage

(rare, highly penetrant mutations) could affordably and routinely be positionally cloned. The vast expense of genome-wide association, by comparison, prohibited comprehensive searches for common, modest effects expected to be involved in complex traits.

This impediment to mapping genes via association was overcome by the completion of the HapMap and the plunging cost of genotyping. These advances have made it possible to test up to a million markers across the genome, which, in conjunction with large, carefully phenotyped disease collections have made it possible to undertake genome-wide association studies (GWAS). Experimental and methodological science once again combined to enable a new generation of disease genetics.

1.4 Outline

The aim of this thesis is both to describe and develop approaches for carrying out GWAS and to describe and interpret results from several early studies. While many of the analytical methodologies for GWAS have been in use for years, the scale of these studies presents a number of new challenges. These issues will be addressed both in a general context and in light of specific datasets.

The first chapter evaluates the genomic coverage of GWAS tag sets. The choice of SNPs to genotype will fundamentally affect the outcome of a study, but it is not technologically practical to construct different panels of SNPs to suit each GWAS. I therefore compare not only theoretical considerations related to selecting tag SNPs on a genome-wide scale, but also the specific properties of commercially available products.

Chapter 3 describes the Wellcome Trust Case Control Consortium, the largest of the first generation GWAS. This study was not only successful at identifying new disease loci, but was also informative as a template for the types of analyses required for GWAS. I describe my work on several aspects of the project, including genotype calling, data quality control, gross population

structure and association analysis.

Chapter 4 describes my focus in particular in GWAS for Crohn's disease, a common form of inflammatory bowel disease. I discuss the nature of the phenotype, with special attention given to its history of genetic studies, from linkage scans to several recent GWAS. I then describe a replication experiment undertaken as a follow-up to Chapter 3 where several potential hits are examined in independent samples.

After the early successes of GWAS it became apparent that additional power to detect associations could be obtained by combining individual studies. However, this approach is complicated by a number of issues, including data annotation, sample curation, population structure and genotype imputation. These are addressed in Chapter 5, which presents examples of these potential problems as well as solutions for them.

Chapter 6 describes a worked example as part of an international meta-analysis of three Crohn's disease GWAS. I describe the process of merging these data, selecting regions for follow-up, and the results of a replication experiment in independent samples from all three collaborating consortia. The results of this study entail the most comprehensive catalogue to date of genetic risk for a common human disease.

Finally, I summarise these results and discuss their place in the early period of GWAS. I examine likely directions in the near future as additional studies are published, and consider their impact on the fundamental understanding of human disease biology.

Chapter 2

Evaluating coverage of genome-wide association studies

2.1 Genome-wide marker sets

As described in Chapter 1, falling genotype costs and the completion of the HapMap (International HapMap Consortium, 2005; Hinds *et al.*, 2005) have made genome-wide association studies of complex diseases possible (Wang *et al.*, 2005; Hirschhorn and Daly, 2005; Palmer and Cardon, 2005). Such studies have the potential to assay 100,000 - 1,000,000 genetic markers from the $>$ 5 million validated genetic variants now available. While genotyping most or all of the genetic variants would be desirable in many settings, present economic and experimental conditions render it practically necessary to reduce the complete set of genetic variants down to a tractable, but maximally informative subset.

There are a number of potential approaches to this problem, motivated both by individual investigators' interests and broader questions such as the degree to which importance should be focused on full coverage of the genome or on po-

tentially functional variants (Botstein and Risch, 2003; Neale and Sham, 2004). Most of the ongoing or planned GWAS aim to evaluate most of the common genetic variants in the human genome, irrespective of their genic location (Wang *et al.*, 2005; Hirschhorn and Daly, 2005). For such designs, an obvious marker selection approach for any particular study is to pick a theoretically ideal set of SNPs for the study and genotype them in large samples. This method is appropriate for studies of small regions or candidate genes, but it is impractical for GWAS as the cost of ordering a *de novo* SNP set for each new genome scan is prohibitive. Instead, genome-wide studies must choose from several commercially available alternatives. These pragmatic concerns of what is currently available in a high-throughput capacity will be at least as important as theory-driven marker selection for the first generation of scans that are now underway or being planned.

The practical necessity of having a fixed set of GWAS markers has obvious advantages, such as the potential to combine datasets across disease laboratories and the ability to design statistical methods for commonly-used panels, as has been done for linkage studies over the past decade. This broad usage makes it important to appreciate the properties of different marker selection strategies in terms of genomic coverage, allele frequency representation and population diversity. Here I evaluate the different strategies (described in Section 2.2) employed in several commercially available GWAS panels, including sets of exclusively nonsynonymous SNPs (nsSNPs) (Clayton *et al.*, 2005), LD-based tagging panels (deBakker *et al.*, 2005), and random SNP collections across the genome (Dong *et al.*, 2001). In order to provide as comprehensive an evaluation as possible, I use the HapMap Phase II data (International HapMap Consortium, 2007) (1 SNP / 1250 bp across the entire genome) to provide a framework for testing and comparison of common variation.

2.2 SNP selection methods

Tag selection methods referenced in this chapter are described below:

1. **Tag** An LD based set of tag SNPs carefully chosen to maximize the amount of variation captured per SNP. I first examine the theoretically most efficient set of tag SNPs and then consider illustrative examples from commercially available products. There are at least two tag sets presently in widespread use: the Illumina HumanHap300 set of 317K markers and the HumanHap550 set of ∼550K SNPs (http://www.illumina.com/).

2. **Random** A set of SNPs distributed approximately randomly across the genome, that ignores LD patterns. I evaluate two LD agnostic marker sets currently available from Affymetrix with 111K and 500K markers (http://www.affymetrix.com/).

3. **Combined** A combination of these two methods consisting of a set of "random" SNPs augmented by a carefully chosen fill-in set. Due to cost efficiencies and speed of genotyping, several disease investigations have combined the random sets of Affymetrix with bespoke tag sets that aim to fill-in the gaps of random coverage.

4. **Functional** Either a panel of all known, polymorphic nsSNPs (the MegAllele system marketed by Affymetrix involves 12,000 nsSNPs, while Illumina's NS-12 has over 14,000) or a larger set focused more broadly on genes, such as Illumina's Human-1 BeadChip. I do not evaluate nsSNP-specific sets for genome-wide coverage since they are designed to directly test the functional variants as opposed to the other products which instead indirectly test many variants genome-wide without any prior hypothesis.

2.3 Evaluating coverage

I examine coverage as measured by simple pairwise correlation (r^2) between a member of the tag set and a potentially captured SNP (Carlson *et al.*, 2004b; Ke

et al., 2005). This approach is attractive in that it makes few assumptions about downstream analysis and has a simple relationship to sample size for disease association studies under certain conditions (International HapMap Consortium, 2005; Pritchard and Przeworski, 2001), but it is not always the most efficient (deBakker *et al.*, 2005). There also exist approaches to make use of multiple-marker tests derived from the HapMap when the initial panel is fixed (Pe'er *et al.*, 2006).

Coverage calculations may be confounded by two important factors: First, there is an upward bias in coverage when including the tag SNPs themselves as part of the coverage calculation. Second, coverage estimates will be biased upwards if all or part of the reference set used for the estimate was used to select tag SNPs. In order to equate coverage of a reference set to coverage of the genome, the reference set must be considered representative of all common SNPs. If, on the other hand, tag SNPs have been chosen specifically to capture all or part of the reference set then they are overfitted to the reference set compared to the set of all common SNPs in the genome. Detailed corrections for these biases are discussed in the Methods.

Figure 2.1 shows cumulative coverage of Phase II HapMap of a maximally efficient set of tag SNPs (Section 2.2, strategy 1) in three HapMap panels for three r^2 thresholds. The data for the CEU and JPT+CHB panels indicate that nearly all common variation in the genome can be captured with $r^2 \geq 0.8$ with approximately 500,000 carefully selected SNPs. As expected, the YRI panel requires more than twice as many SNPs to capture the Phase II HapMap.

In all cases, the coverage versus increasing number of tags exhibits a striking pattern of diminishing returns. For example, while approximately 500,000 SNPs are required to completely capture common variation in CEU, a set of 250,000 already captures 85%. This pattern has been previously observed when selecting large pools of tag SNPs (deBakker *et al.*, 2005) and follows from the fact that tagging algorithms initially select the SNP with the largest number of proxies and then include the next most useful and so on (deBakker *et al.*, 2005). One

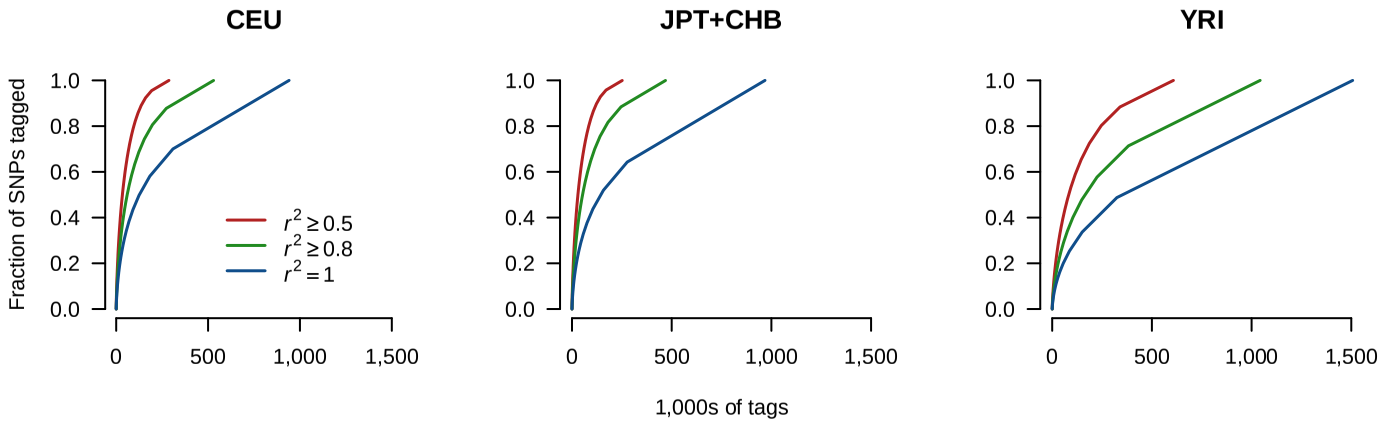


Figure 2.1: Genomic coverage by maximally efficient tag sets for three HapMap panels and three r^2 cutoffs. Evaluation of common SNPs is performed against the Phase II HapMap data, which provides a near-complete catalogue of common variation (MAF ≥ 0.05), including 5 million SNPs in 270 individuals from populations in North America (CEU), Africa (YRI) and Asia (CHB+JPT) (International HapMap Consortium, 2005). The finished Phase II HapMap contains a common SNP every 1250 bp in the CEU population and is estimated to capture 94% of common variation in CEU and CHB+JPT and 81% in YRI (International HapMap Consortium, 2005).

consequence of this effect is that a large fraction of any genome-wide tag set is devoted to capturing 'singleton' SNPs that are not in strong LD with any other SNPs. The magnitude of this effect varies strongly depending on the r^2 threshold and population, from about 1/3 of all tags when tagging CEU or JPT+CHB at $r^2 \geq 0.5$ to nearly 1.1M singletons in YRI at $r^2 = 1.0$, representing almost 80% of the total tag set.

Several pragmatic factors influence the coverage of emerging tag SNP panels when compared to the theoretically most efficient set. Table 2.1 (evaluated at $r^2 \geq 0.8$) and Figure 2.2 show coverage for several strategies, all of which are lower than the theoretically most efficient values in Figure 2.1. Most of these differences are explained because the tag SNPs in sets now becoming available were necessarily selected from a less complete reference set. While 250,000 SNPs selected from the Phase II HapMap capture 85% of common variation in CEU, this efficiency is reduced to only 72% when selecting tags from the Phase I HapMap dataset. The effect of incomplete reference data is exacerbated by the differences in underlying allele frequency spectra of the HapMap phases. Figure 2.3 shows that while both phases of the HapMap are enriched for common SNPs, the skewing is less pronounced in Phase II, which has a larger proportion of rare (MAF < 0.05) and moderately rare ($0.05 <$ MAF < 0.10) variants. Since rare alleles are more likely to be singletons, this difference reduces efficiency.

Coverage values for marker panels which do not incorporate LD patterns (Section 2.2, strategy 2) are also shown in Table 2.1. Two of the products in this category, the older Affymetrix 111K and widely used 500K panels, are estimated to capture 31% and 65% of common variation in CEU at $r^2 \geq 0.8$, respectively. As expected, these products have lower overall coverage and lower efficiency (coverage per SNP genotyped) than sets of carefully selected tag SNPs. However, their randomness offers a protective redundancy against the failure of any given SNP. Figure 2.4 shows the relative coverage of the Affymetrix Genechip 500K and Illumina HumanHap300 products for random failure rates between 0–20%. This relationship between efficiency and redundancy could be

		CEU		JPT+CHB		YRI	
	Type	Coverage %	Mean r^2	Coverage %	Mean r^2	Coverage %	Mean r^2
Illumina HumanHap300	tag	75	0.961	63	0.964	28	0.961
Affymetrix Genechip 500K	random	65	0.975	66	0.974	41	0.971
Affymetrix 111K	random	31	0.960	31	0.957	15	0.957
Affymetrix Genechip 500k + 175K tag	combination	86	0.975	79	0.978	49	0.973
Illumina Human-1	gene	26*	0.957	28*	0.955	12*	0.956

Table 2.1: Genomic coverage of commercial GWAS products for common SNPs at $r^2 \geq 0.8$ evaluated in the Phase II HapMap. Despite the r^2 cutoff of 0.8, the mean r^2 for tagged SNPs is very high; also, untagged SNPs are covered with intermediate values of r^2, providing modest power to detect such alleles (Figure 2.2). Coverage estimates for the Human-1 product (marked with a *) are underestimates because some of its SNPs were not genotyped in the HapMap project. Since these SNPs are largely rare genic SNPs it is not expected that they would substantially raise coverage of common variation.

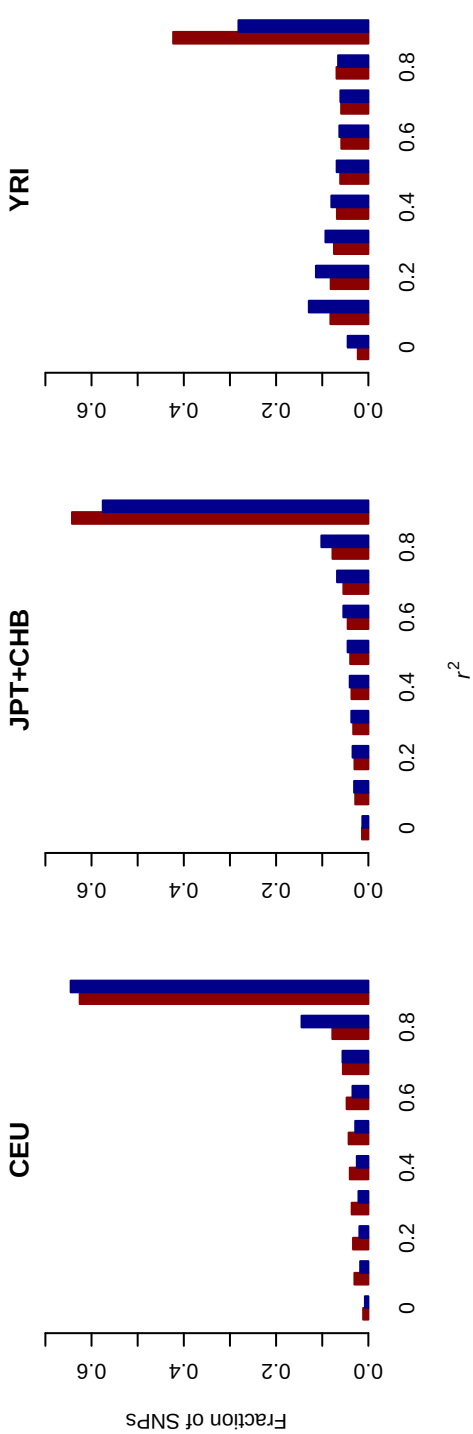


Figure 2.2: Fraction of SNPs in HapMap Phase II captured by r^2 of at least a given threshold by Illumina 300K (blue) and Affymetrix 500K (red).

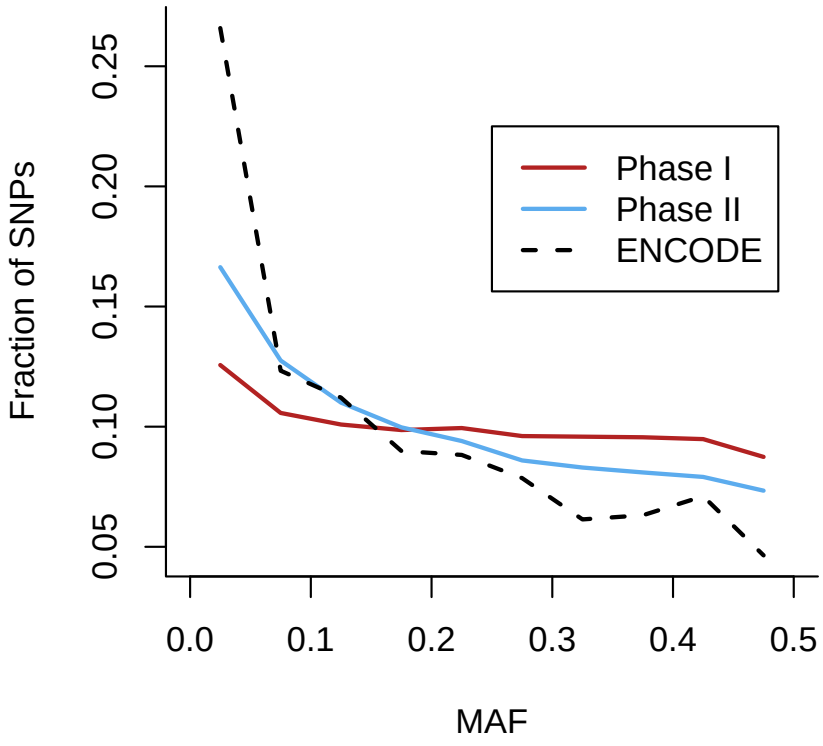


Figure 2.3: MAF in CEU for 900,000 polymorphic Phase I HapMap SNPs, 2M distinct polymorphic SNPs added during the second phase of HapMap and 10,000 polymorphic ENCODE SNPs (which approximate the underlying frequency distribution in the genome). Both phases of HapMap are intentionally biased toward common SNPs, but the bias for Phase II is less extreme. The pattern in the JPT+CHB and YRI analysis panels is not substantially different.

important, given the serious consequences of failure rate for disease association studies (Wang *et al.*, 2005; Clayton *et al.*, 2005) and initial indications of non-trivial failure rates for at least some genome-wide scans (Klein *et al.*, 2005).

Another factor which affects choice of a genome-wide SNP panel is the population being studied. Table 2.1 shows that the advantage of carefully selected tags is reduced when considering coverage in another of the HapMap analysis panels. While the Illumina HumanHap300 outperforms the Affymetrix Genechip 500K for the CEU samples (Illumina 75% — Affymetrix 65%), this advantage is no longer apparent with the JPT+CHB samples (Illumina 63% — Affymetrix 66%), and the advantage is reversed for the YRI samples (Illumina 28% — Affymetrix 41%). Figure 2.4 further shows that the most efficient design, selecting tags from the same population to be studied, is most susceptible to marker failure because nearly all redundancy has been intentionally eliminated.

While coverage of common variation by the larger versions of currently available products is promising, rare variants are much less well covered due to the preferential selection of common SNPs on these products. At present, there do not yet exist publicly available data in which large genomic regions have been resequenced in large sample sizes, though such studies of candidate genes are increasing (Rieder *et al.*, 2005). In order to gain some insight into correlation patterns among rare SNPs I used data from the ENCODE project (International HapMap Consortium, 2005), which suggest that fewer than 10% are well captured (pairwise $r^2 \geq 0.8$) by any of the genome-wide products in any of the HapMap analysis panels. Even these low coverage levels may be overestimates for several reasons. First, because only a modest number of individuals were resequenced, a large number of rare SNPs in these regions remain undiscovered. Second, estimates of r^2 for rare SNPs have a high variance, indicating that observations of "perfectly correlated" rare SNPs in the HapMap samples may be due to chance. Finally, many "private SNPs", that is, rare alleles observed only in one of the resequenced individuals, will appear to be perfectly correlated, thus inflating estimates of LD. The combination of the low estimates derived

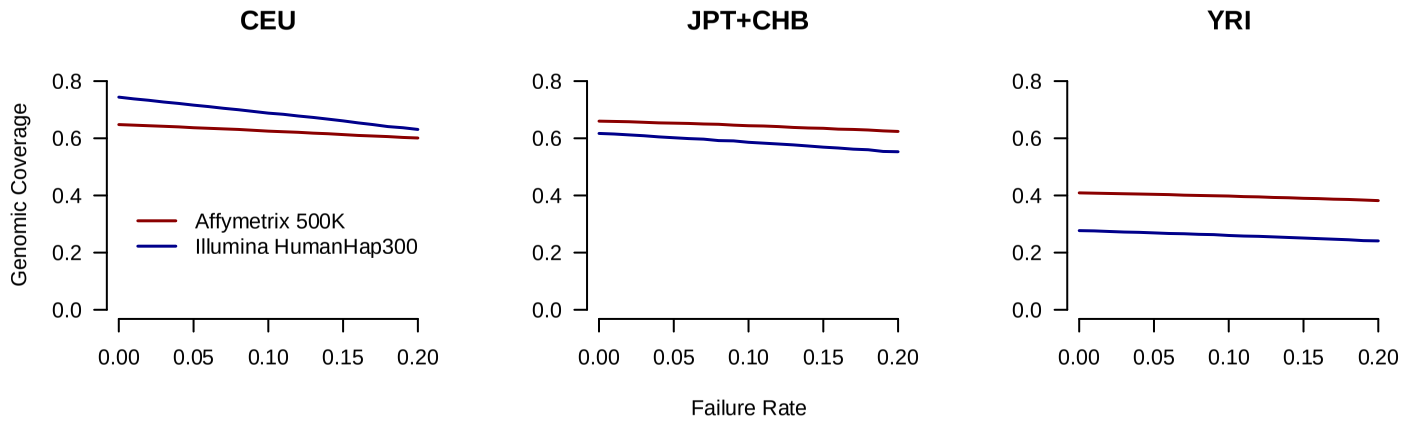


Figure 2.4: Coverage of common variation in the Phase II HapMap by the Affymetrix Genechip 500K and Illumina HumanHap300 products plotted as a function of random genotype failure rate. For each level of failure rate markers were excluded at random and coverage was calculated using only the remaining markers. Each point is the average of 1000 replicates. The larger number of SNPs and increased redundancy in the Affymetrix array provide it greater resistance to decreased coverage due to marker failure. The Illumina array performs worse in non-CEU populations because its SNPs are so carefully targeted toward CEU.

from ENCODE and the numerous factors which indicate that true coverage is lower underline the assumption of these GWAS products that common diseases are at least partially influenced by common genetic variants.

One possible compromise among the considerations of cost, efficiency and redundancy is to couple a set of 'random' SNPs with a smaller, personalized set that is highly selected to fill-in gaps (Section 2.2, strategy 3). In this case, the exact extent to which coverage improves depends on the goals for the selection of the fill-in set. Using 500K random SNPs as a baseline, approximately 360,000 additional SNPs would be required to capture all remaining common variation in CEU. Of course, the majority of these would be spent in the singleton tail (capturing only themselves). Instead, some groups have designed panels of non-singleton tags to bring coverage to 86% (Table 2.1). The remainder of the SNPs in the combined approach can then be used to add targeted redundancy to LD bins with many surrogates, or focus on known nonsynonymous variation.

While coverage of common variation in genes is generally comparable to the rest of genome, other focused classes of functional variants are captured poorly by SNP sets aimed at common variation. Figure 2.5 shows the CEU allele frequency distribution of the ∼15,500 nsSNPs that are polymorphic in the HapMap. Clearly there is an overwhelming preponderance of rare nsSNPs in this set. Capturing all 50,000 nsSNPs catalogued in dbSNP would be a significant challenge given that coverage of rare SNPs by genome-wide products is poor. The most direct way to ensure coverage of these SNPs is to genotype them directly (Section 2.2, strategy 4), either through a separate platform (Hardenbol *et al.*, 2005) or by adding them to the tag set. Both the HumanHap300 and combined approaches eschew some singleton tags to include all known polymorphic nsSNPs. The Illumina Human-1 BeadChip offers an intermediate approach by concentrating 109K SNPs in exons and conserved regions. This combines good coverage for a broad variety of functional SNPs and modest genome-wide coverage of common SNPs.

It is interesting that despite the many differences in panel focus, marker

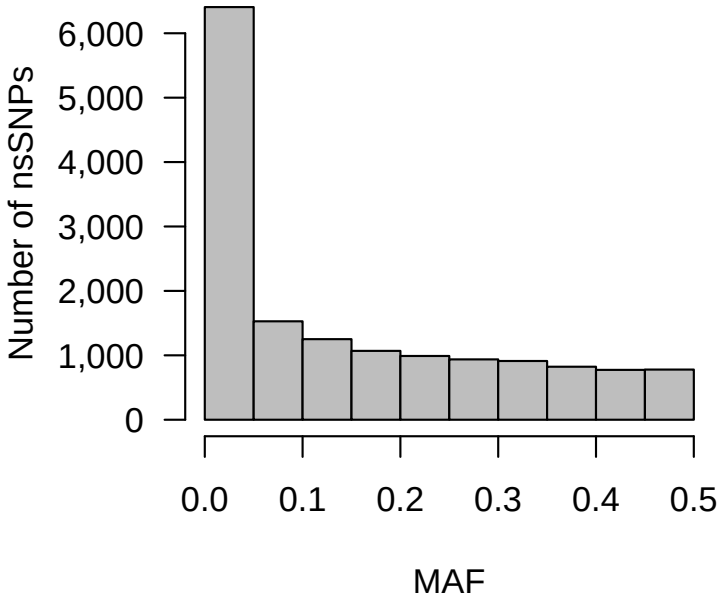


Figure 2.5: CEU minor allele frequencies for nsSNPs polymorphic in the Phase II HapMap.

selection strategies and technology platforms, the random and tag-SNP sets generally capture a similar fraction of common variants in the human genome. Choosing a GWAS SNP set requires careful evaluation of levels of coverage, population of interest and the balance between efficiency (via careful tag selection) and redundancy. Newer, larger GWAS products, together with functionally focused assays, promise an ever more complete picture of genetic variation. The challenge will then turn from how many of the variants are missed to how to extract meaningful information from those which are not.

2.4 Methods

2.4.1 Data

Genome-wide coverage evaluation was performed on Phase II HapMap (release 20) combined with Affymetrix genotypes on the HapMap samples for those markers contained on their product but not on HapMap. This dataset contained between 2.5 and 3 million polymorphic SNPs, depending on HapMap analysis panel (CEU, YRI, JPT+CHB), genotyped on 90 samples per panel. Evaluation of rare coverage was done in the HapMap ENCODE dataset (release 16c.1) which was created by genotyping all variable sites observed after resequencing 48 unrelated individuals in the full set of HapMap samples. The JPT and CHB samples were combined into one analysis panel for all analyses.

Complete SNP lists for all evaluated products are available at:

http://www.well.ox.ac.uk/~jcbarret/gwas/

2.4.2 Tag SNP selection and coverage evaluation

I selected tag SNPs using the program Haploview (Barrett *et al.*, 2005). A SNP was considered tagged by another if they had pairwise r^2 greater than a certain threshold. The maximum allowed physical distance between an allele and a tag was 200 kb. Common coverage of predetermined tag sets was evaluated by

using Haploview with the same parameters but force-including the SNPs in the tag set and force-excluding all other SNPs. The raw value for coverage is the fraction of all common (MAF ≥ 0.05) SNPs which are captured by the tag set.

Coverage calculations may be confounded by two important factors: First, there is an upward bias in coverage when including the tag SNPs themselves as part of the coverage calculation. Consider a reference set of SNPs R such as the Phase II HapMap. For a given tag set T, some SNPs are captured either because they are contained in T, or they are in LD with a SNP in T (call this set of SNPs L). Thus, the naive estimate of coverage of all SNPs in the genome, G, would be:

$$\frac{L+T}{R} \tag{2.1}$$

That is, count all SNPs either in the tag set or in LD with a tag and divide this by all the SNPs in the reference set. However, this overestimates the true fraction of captured SNPs which are actually tags, since $G > R$. To correct for this, the genome-wide coverage is defined as:

$$\frac{(\frac{L}{R-T})(G-T)+T}{G} \tag{2.2}$$

The $\frac{L}{R-T}$ part of the formula computes the fraction of the reference set captured because it is in LD with a tag but not a tag itself. Multiplying this fraction by $G-T$ yields a value for the number of SNPs captured by LD genome-wide. Adding this number to the number of tags gives an estimate of the total number of SNPs genome-wide which are captured by either LD or by being a member of the tag set. Dividing by G yields the coverage estimate. In the case where the reference set is the entire genome (i.e. $R = G$), Equation 2.2 sensibly reduces to Equation 2.1.

Coverage estimates will also be biased upwards if all or part of the reference set used for the estimate was used to select tag SNPs. In order to equate coverage of a given reference set to coverage of the genome, the reference set

must be considered representative of all common SNPs. If, on the other hand, tag SNPs have been chosen specifically to capture all or part of the reference set then they are "overfitted" to the reference set compared to the set of all common SNPs in the genome. In the case of tag SNPs chosen using the Phase I HapMap data, this bias is corrected by considering coverage of the Phase II-only subset of R, which is unbiased with respect to the tag set. In that case the correction factor in Equation 2.2 is split into two parts, representing the overfitted portion of the reference, R_1, and the remainder, R_2:

$$\frac{\frac{L_2}{R_2-T_2}(G-R_1-T_2)+T_2+L_1+T_1}{G} \tag{2.3}$$

where L_1, T_1, L_2, T_2 are the portion of SNPs captured by LD (L_1, L_2) and the tag SNPs (T_1, T_2) in the Phase I and Phase II data respectively. In this case the coverage of the markers exclusively in Phase II can be estimated: $\frac{L_2}{R_2-T_2}$. Multiplying this by the number of commons SNPs in the non-Phase I genome minus T_2 yields an estimate for the number of SNPs captured genome-wide via LD. Adding the number of Phase I and Phase II tags, plus the number of SNPs captured by LD from Phase I and dividing by G yields a corrected estimate of coverage. In the case where there is no overfitting (i.e. $R_1, T_1, L_1 = 0$), Equation 2.3 reduces to Equation 2.2. In the case where the two reference sets constitute the entire genome (i.e. $R_1 + R_2 = G$), it reduces to Equation 2.1.

2.4.3 Failure Rate

Effects of marker failure on coverage were evaluated by randomly selecting some fraction (between 0–20%) of markers and calculating genome-wide coverage based on the remainder of the "successful" markers. Each data point for a failure rate of a particular SNP set in a particular analysis panel is the average of 1000 replicates.

Chapter 3

The Wellcome Trust Case Control Consortium

3.1 Introduction

The Wellcome Trust Case Control Consortium (WTCCC) was a UK based collaboration of genetics researchers aimed at assessing the promise of GWAS to discover common DNA variants that influence complex disease. The largest component of the consortium was a genome-wide association study of 2000 British cases from each of seven complex diseases (bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D)). Each of these collections was compared to a total of 3000 shared control individuals from two sources: 1500 individuals selected from the 1958 British Birth Cohort (58C) and 1500 blood donors from the UK Blood Service recruited as part of this project (UKBS). All 17,000 samples were genotyped on the Affymetrix GeneChip 500K Mapping Array.

This study was the largest first generation GWAS and as such provided an important test of the validity of the approach. I described several potential

reasons for the limited success to date in finding genes for complex human traits in Chapter 1. The WTCCC design addressed three of these challenges: (i) the Affymetrix chip has high SNP density and good coverage of common variation (as described in Chapter 2), (ii) the SNPs are distributed without any *a priori* hypotheses about which parts of the genome are relevant to a given phenotype and (iii) the sample sizes were large enough to have good power to detect modest-sized effects (43% power for OR of 1.3 at $p = 5 \times 10^{-7}$, averaged across allele frequencies). The WTCCC was therefore not only aimed at finding associations, but also at fundamentally evaluating the approach. A failure to identify any loci would imply that very little of the disease heritability lies with common variants acting alone, pointing instead towards other explanations such as gene-gene or gene-environment interactions, or many rare variants.

In this Chapter I will present details of several analyses which I performed as part of the WTCCC Design and Analysis Group. First I will describe issues relating to genotype calling, interpretation of intensity cluster plots and quality control metrics for both samples and SNPs. These efforts collectively led to a clean dataset for which I describe results from association tests. Finally I comment on these results and the necessary future work required to utilise the data most effectively.

3.2 Genotype calling

The throughput (in terms of number of SNPs genotyped per unit time) of genotyping has been accelerating for the past decade, ranging from technicians transcribing by hand the result of a single genotyping reaction on an electrophoretic gel to assays which simultaneously measure genotypes at over one million SNPs across the genome. One of the most widely implemented modern methods for discriminating between alleles at a SNP is to tag DNA fragments with an allele-specific fluorescent molecule and to measure the relative intensities of fluorescence for the two alleles. This tagging can be done in a variety of ways, such

as allele-specific primer extension or probe hybridization (both the commonly used TaqMan single SNP assay and the Affymetrix chips use the hybridization approach).

Specifically, the Affymetrix GeneChip comprises millions of 25mer oligonucleotide probes synthesized in an ordered fashion on a solid surface by photolithography (Kim and Misra, 2007). Each SNP is measured by 24-40 probes, some of which contain the SNP in a perfectly matched probe sequence and some which are mismatches, which are useful in measuring the background level of hybridization. These probes are arranged in pairs or quartets on the array, and each has two versions corresponding to the two potential alleles. Digested, amplified genomic DNA from the sample is then hybridized to the chip, yielding fluorescence relative to the stability of the binding between the probes and these DNA targets. For each SNP and each sample, the challenge is thus to measure a few dozen fluorescent intensities and condense this information into a pair of normalised intensities corresponding to the two alleles. Members of the WTCCC developed a modified quantile-normalisation routine which used both the match and mismatch probe data. Interestingly, subsequent iterations of the Affymetrix chips have discarded the mismatch probes in favour of increased SNP density, which presents a normalisation problem distinct from that confronted by the WTCCC.

In order to keep up with the rapid pace of the biological and technical aspects of genotyping it has been critical to develop fast, accurate algorithms for translating these intensities to genotypes in an automated fashion. There has been a growing realisation in human genetics, however, that calling algorithms are subject to subtle biases which can lead to increased type I error (Clayton *et al.*, 2005). Furthermore, the scale of GWAS prohibits careful examination of cluster plots of allelic intensities (see, for example, Figure 3.1), which is the best way to assess the quality of genotype calls. Members of the WTCCC therefore implemented a program called CHIAMO, which improved on existing genotype calling algorithms and provided a measurement of the uncertainty of each call

(in the form of posterior probabilities from zero to one for each sample, for each possible genotype call). The vast majority of the SNPs on the Affymetrix platform provide high quality data similar to Figure 3.1, which shows cluster plots for all nine WTCCC collections. In cases such as these, genotype calling is relatively straightforward, and all currently available algorithms yield accurate, high confidence results.

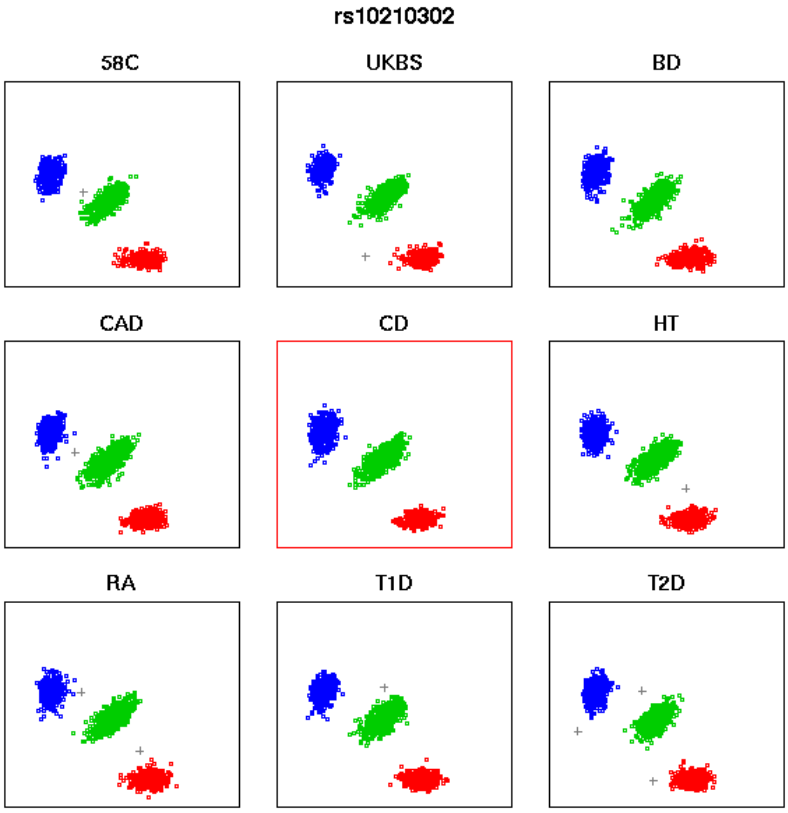


Figure 3.1: Cluster plots for the 9 WTCCC collections for a typical SNP, which happens to be associated with CD. Each point is a sample's normalized intensities for allele A on one axis and allele B on the other. Colours correspond to CHIAMO genotype calls, with homozygotes blue and red and heterozygotes green. Genotypes called as missing are shown as grey crosses.

3.2.1 Radial intensity shift

Some SNPs, however, were characterised by aberrant cluster plots which often fell into consistent patterns, such as poorly separated clusters very close to the origin. SNPs in one such pattern had cluster plots similar to those in Figure 3.2. Three collections (UKBS, CAD, RA) showed normal clusters, whereas the other six showed the three normal clusters plus three more shifted radially toward the origin of the graph. While CHIAMO did not have difficulty in correctly calling these genotypes, it was unclear whether this phenomenon represented a technical artefact or a potentially interesting biological process. Identifying such SNPs systematically was accomplished by converting the Cartesian intensity coordinates to polar coordinates:

$$r = \sqrt{x^2 + y^2}$$
$$\theta = \arctan\left(\frac{y}{x}\right)$$

Common SNPs showing the pattern in Figure 3.2 have a high ratio of radial variance between the 58C and UKBS in all three clusters (the behaviour of the cluster variance for rare SNPs was too erratic to be useful). Different SNPs were differentially affected, with some (such as Figure 3.2) showing six distinct clusters and twofold more radial variance in the 58C than the UKBS. Others showed a more subtle pattern with elongated, but continuous clusters, presumably ranging down to effects which were too small to observe compared to the within-cluster noise. Overall, approximately 7% of SNPs showed a noticeable pattern (at least a 1.25-fold increase in radial variance in all three clusters in the six shifted samples compared to the three unshifted).

Investigation of which specific samples appeared in the clusters closer to the origin revealed that these samples were always the same. Furthermore, these samples were non-randomly distributed across the 96-well DNA plates used in the project. As seen in Figure 3.3, all the shifted samples came from

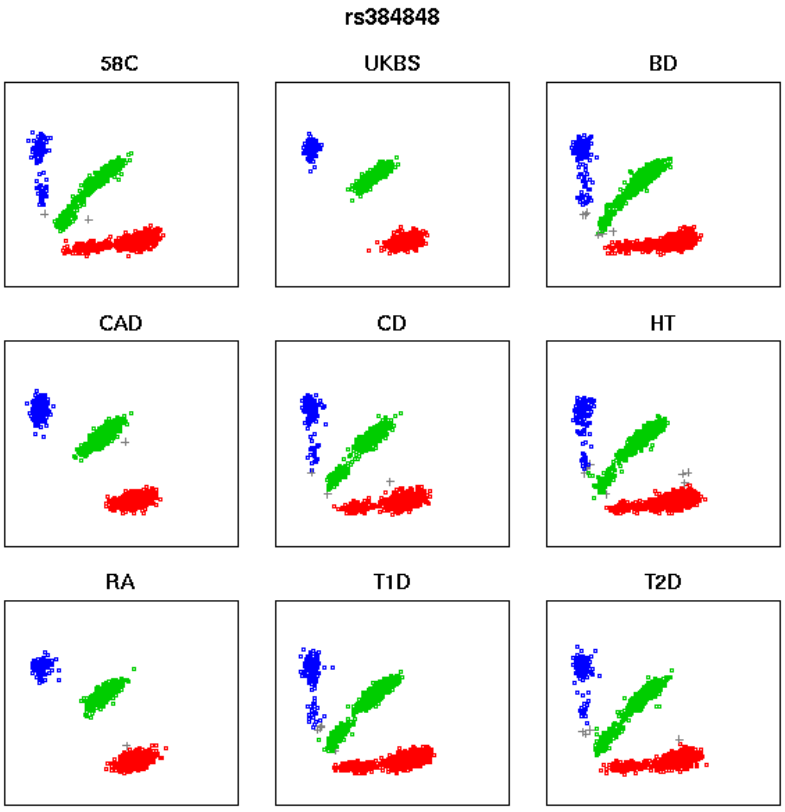


Figure 3.2: Approximately 7% of SNPs showed a pattern of six clusters (three in the expected locations and three shifted radially toward the origin) in six of the nine collections. This shift never occurred in the UKBS, CAD and RA collections.

plates processed early in the project, at the Affymetrix R&D facility. Clearly some difference in laboratory procedure or conditions affected the intensities for these SNPs. Intriguingly, this subset of SNPs did not show any consistent patterns in genomic location, primer sequence, proximity to various genomic features (including recombination hotspots and genes), or the distribution of probe quartet and pair locations on the chip. While this phenomenon could not be fully explained, it is worth noting for two reasons: first that specific laboratory conditions can significantly affect the raw data generated on this chip, and second that such experimental aberrations might be incorrectly followed up as possible copy number variation or other biological processes.

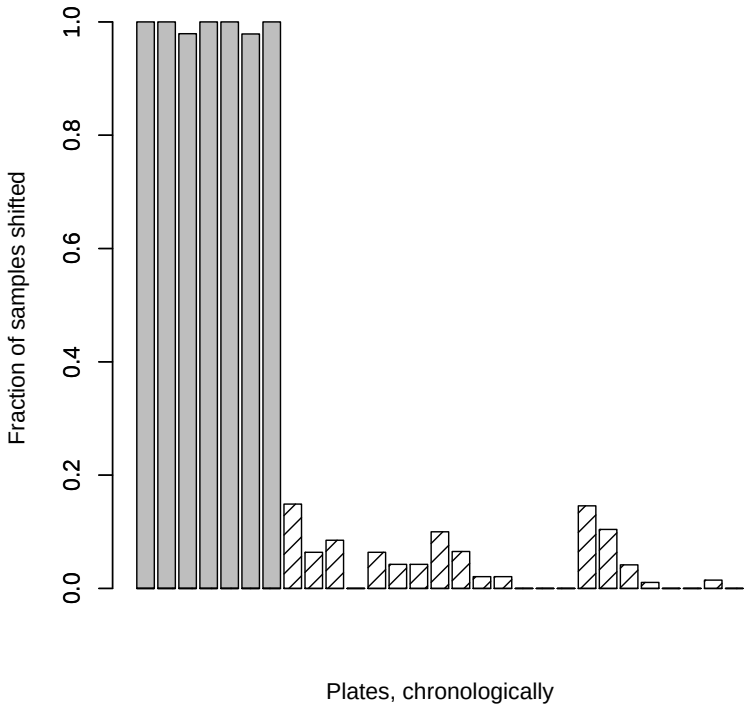


Figure 3.3: Plates containing 58C samples are arranged chronologically along the x-axis. Each bar shows the fraction of samples on that plate which are in the clusters closer to the origin. Grey bars show plates processed at the Affymetrix R&D facility, hashed bars show plates processed at the Affymetrix Services Lab. Since the shifted clusters are not completely distinct, the later samples which appear shifted are likely simply at the low end of the standard cluster.

3.3 Quality control

3.3.1 Sample filters

A variety of checks were employed on samples both to detect swaps or errors at any stage in the project (DNAs were sent from labs across the UK to a central location and then shipped to Affymetrix for genotyping) and to remove low quality samples from further analysis (see Table 3.1 for a summary per collection). For some collections, blood types or genotypes for a small proportion of SNPs were available, and 153 samples showing discrepancies were excluded. Similarly,

295 duplicate (>99% identity) and 86 related (86-98% identity) samples were excluded. A sample's missing data rate acts as an indicator of low DNA quality. Most samples had very low rates of missing data (study-wide average 0.00925, standard deviation 0.0187), and while any exclusion threshold is to some extent arbitrary, the empirical distribution of missingness (Figure 3.4) showed the bulk of the samples to have <3% missing data, so the 250 samples above this cutoff were excluded. Setting similar empirical thresholds on autosomal heterozygosity (excess heterozygosity in particular may indicate contamination) excluded a further six samples with >30% heterozygosity and three with <23% (Figure 3.4).

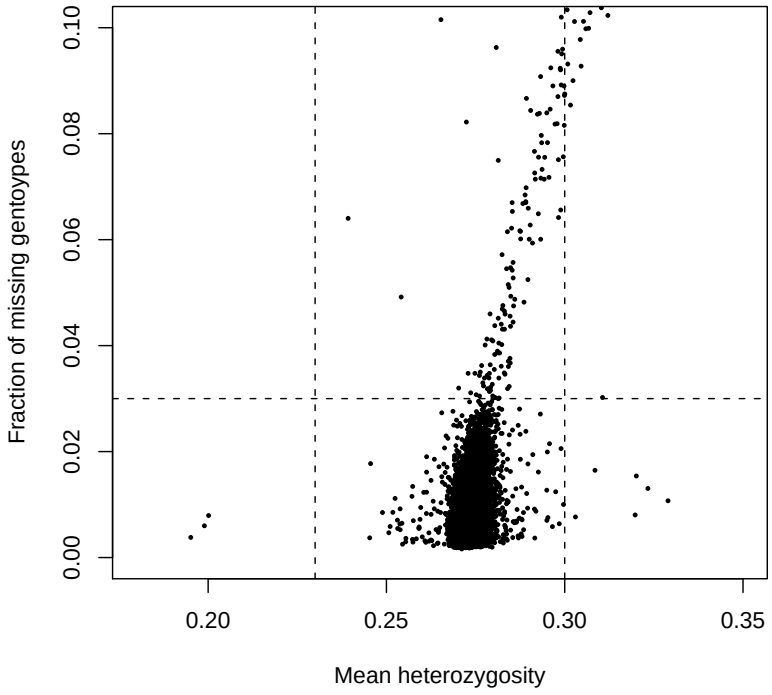


Figure 3.4: Autosomal heterozygosity vs. the proportion of missing data for each individual in the study. Dashed lines delimit the threshold used for exclusion of individuals from further analysis (boundaries of this figure chosen to show the bulk of the samples; extreme outliers are not shown).

Collection	Missingness	Heterozygosity	External discordance	Non-European ancestry	Duplicate	Relative	**Total**
58C	9	0	4	6	4	1	**24**
UKBS	8	0	5	14	0	15	**42**
BD	30	0	0	9	77	13	**129**
CAD	41	1	0	13	2	5	**62**
CD	43	4	6	54	131	18	**256**
HT	29	0	0	2	6	11	**48**
RA	47	1	0	26	53	9	**136**
T1D	7	2	1	18	6	3	**37**
T2D	36	1	0	11	16	11	**75**
Total	**250**	**9**	**16**	**153**	**295**	**86**	**809**

Table 3.1: WTCCC sample exclusion filters by collection.

3.3.2 Non-European ancestry

Samples were required to be self identified white Europeans in order to be included in the WTCCC. While analyses of within-Britain population structure had always been planned as part of the project, the data were assumed to be free of broader patterns of geographical genetic variation. In early analyses, however, several collections showed an unacceptable amount of over-dispersion of the distribution of association test statistics, a hallmark of population structure. CD in particular was affected, with a genomic control value (λ_{GC}) of 1.26.

The well known statistical concept of multi-dimensional scaling (MDS) can be used to provide a two-dimensional projection of a multi-dimensional dataset. The 270 HapMap samples (representing samples of Northern European, East Asian and Subsaharan African ancestry) were used to seed the MDS axes to represent intercontinental geographic genetic variation. In the interest of computational efficiency and to avoid confounding the MDS by extended linkage disequilibrium the data were thinned to a set of 71,458 SNPs within which no pair were correlated with $r^2 > 0.2$. For this set of nearly independent SNPs genome-

wide average identity by state was computed for all WTCCC and HapMap samples, defined between individuals j and k for a set of N SNPs as:

$$\mathrm{IBS}_{jk} = 1 - \left(\frac{\sum_{i=1}^{N} |g_{ij} - g_{ik}|}{2N} \right) \quad \text{where } g_{ij} \text{ and } g_{ik} \in \{0, 1, 2\}$$

The matrix of distances between individuals (1 − IBS) was used as input to the MDS. As can be seen in Figure 3.5, 153 samples were clearly separate from the main cluster of WTCCC individuals. Exclusion of these individuals resulted in a substantial reduction in estimates of over-dispersion in test statistic distributions (λ_{GC} for CD dropped to 1.1). Furthermore, careful investigation of sample records from some collections revealed that samples near the YRI cluster were, in fact, Afro-Caribbean, and samples part way along the Europe-Asia axis were of South Asian ancestry. This provides an excellent example of how a relatively simple statistical analysis can yield new insights when applied to the very large datasets available from GWAS.

3.3.3 SNP filters

A dataset of genotype calls can be generated by setting a threshold on the maximum posterior probability: any sample whose most likely call is less likely than this threshold is considered missing and all other samples are called their maximum likelihood genotype. The lower this threshold is set, the more complete the data will be, but the dataset will contain more incorrect genotypes. However, setting the threshold too strictly is counter-productive because of the possibility of differential missingness. Consider that samples between two clusters should sensibly be labelled as missing, since neither of two possible calls is sufficiently likely (this situation is exacerbated when the boundaries of two clusters actually overlap). Often, only two of the clusters are close together (the heterozygotes and one set of homozygotes), which leads to most of the missing individuals

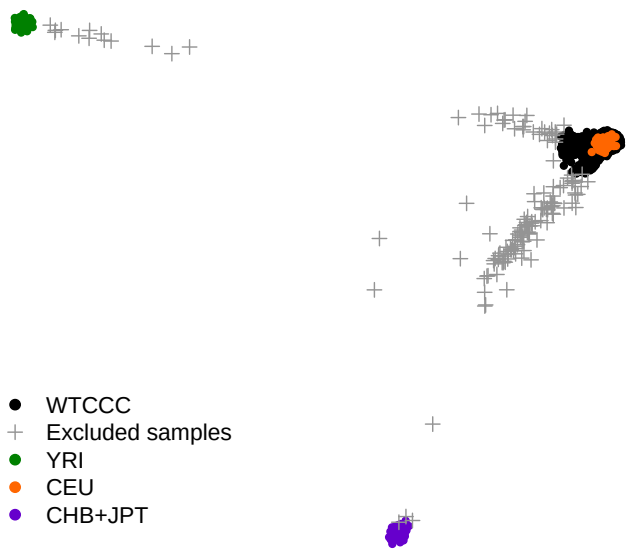


Figure 3.5: MDS analysis of WTCCC and HapMap samples. The substantial majority of samples cluster very tightly with the CEU samples (there are over 16,000 points in the dense black cloud). Outliers (grey crosses) near the YRI cluster were subsequently identified in sample records as Afro-Caribbean; the large cluster one-third of the way between CEU and CHB+JPT were subsequently identified as South Asian (India/Pakistan).

preferentially having that particular allele. If this happens differentially among collections (either due to plate- or DNA-specific effects), the biased missingness will generate false positive associations.

For these SNPs where clusters did not show sufficient inter-cluster angular distance, or, as mentioned earlier, radial distance from the origin, the challenge shifts from calling genotypes to recognition and exclusion, as they represent low quality data. A number of different statistics related to genotype clustering, including ratios of cluster mean and variance within and across sample collections were examined in order to automate the identification of problematic

SNPs. By far the most informative of these metrics (in terms of sensitivity to detecting bad clusters and specificity to avoid needlessly flagging high quality data) was the proportion of missing data in the CHIAMO calls. A threshold of 0.9 empirically yielded the best balance between high call rate and low differential missingness. CHIAMO consistently marked many individuals missing for SNPs with poorly defined or overlapping clusters, whereas it successfully called genotypes for nearly all individuals on high-quality SNPs.

Figure 3.6 shows a summary of missing data rates for all SNPs. 26,567 SNPs with study wide missing data rate >5% or >1% for SNPs with MAF <5% (both because rare SNPs are more susceptible to genotype calling problems and because relatively minor problems can lead to very significant apparent associations). Additionally, 4351 SNPs with Hardy-Weinberg exact $p < 5.7 \times 10^{-7}$ in the controls and 93 SNPs with $p < 5.7 \times 10^{-7}$ for either a one- or two-degree of freedom test of association between the two control groups (corresponding to a 1 d.f. χ^2 statistic of about 25) were excluded. As with sample filters, all of these cutoffs are to some degree arbitrary, but were motivated by examining the empirical distributions of each statistic and seeking values which demarcate a transition from the variance within the performance of good SNPs toward clear outliers.

No single criterion will be completely successful in excluding poor SNPs without removing genuine associations. These quality filters were lax compared to those used by some other studies (especially for Hardy-Weinberg equilibrium), but they were chosen to produce as clean a dataset as possible without excluding any true associations. Cluster plots for all putative associations were thus visually inspected to supplement QC, and a further 638 SNPs were excluded on this basis.

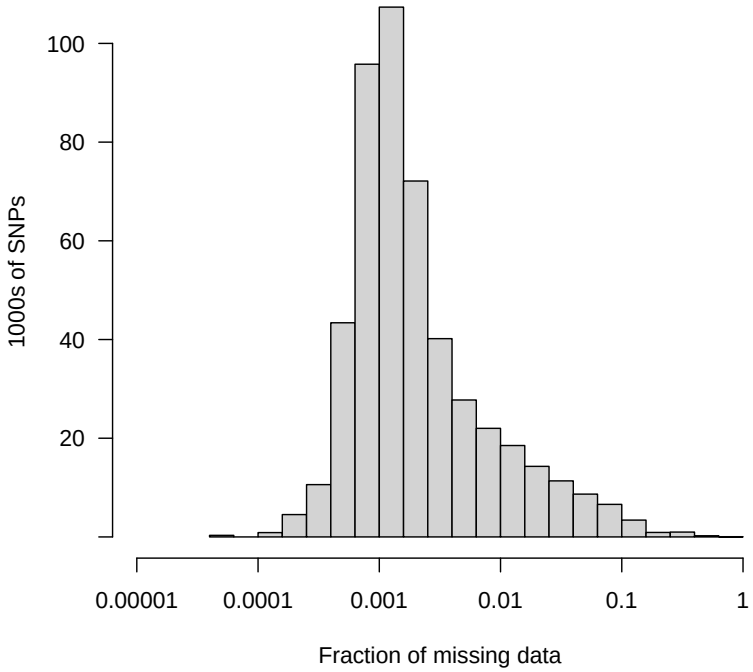


Figure 3.6: Histogram of proportion of individuals called missing for each SNP (i.e. with maximum posterior probability < 0.9). The large majority of SNPs had very high completeness.

3.4 Association analyses

3.4.1 Population structure

One cause of false positive findings in association studies is hidden population structure. Case and control samples may differ in the distribution of their ancestry, either due to ascertainment or to confounding when different ancestries carry higher disease risk and are, as a result, over-represented in cases. Even after exclusion of individuals with evidence of recent non-European ancestry, as described above, the British population is heterogeneous. Members of the WTCCC thus undertook two analyses to determine the extent of this problem. First, they looked across all 17,000 samples in an 11 d.f. test of association with

12 broad geographical regions within Britain. This analysis revealed 13 highly differentiated regions, including such expected loci as the lactase gene and the HLA region, but showed no evidence for genome-wide differences within Britain. Second, principal components analysis were used to look for within-Britain geographical axes of variation. Two principal components did correspond modestly to a NW to SE trend, but these were less significant than six principal components which were attributable to local linkage disequilibrium. Thus, the overall effect of population structure in the WTCCC dataset was small, confirmed by low estimates of over-dispersion of association statistics ($\lambda_{GC} < 1.1$).

3.4.2 Disease association results

Both trend tests (Balding *et al.*, 2003) (1 degree of freedom) and general genotype tests (Balding *et al.*, 2003) (2 d.f., referred to as genotypic) for association were carried out between each case collection and the pooled controls using the PLINK software package (Purcell *et al.*, 2007). The results of the trend test genome-wide for all seven diseases are shown in Figure 3.7. Of 15 variants for which there was strong prior evidence of association with one or more of the diseases studied (based on previous extensive replication), seven showed strong ($p < 5 \times 10^{-7}$) association and six showed modest association in the WTCCC (consistent with previously published effect sizes). The other two comprised *APOE* in CAD, which is poorly tagged on the chip, and *INS* in T1D, for which the relevant SNP marginally failed study-wide QC. In total, 21 regions exceeded a p value threshold of 5×10^{-7}, as shown in Table 3.2.

The findings in Table 3.2 are unlikely to be due to either technical artefacts or external confounders. As was described earlier, population structure within these samples was minimal, and the absence of gross systematic inflation in any case group implies that the excess of test statistics in the tail of the distribution was not due to structure or other forms of bias. Two lines of evidence excluded genotyping error as an explanation for these associations: (i) cluster plots were visually inspected, and (ii) 20 of the 21 regions contained many nearby, corre-

lated SNPs all showing a consistent pattern of association. The gold standard for verifying these associations, however, is replication of the effect in independent samples. A number of replication projects for these loci were carried out separately from the WTCCC; I will describe the CD replication effort in detail in Chapter 4.

Table 3.2 also illustrates one of the shortcomings of the GWAS approach. The region boundaries for the 19 non-MHC loci were defined by the flanking recombination hotspots which were strong enough to both extinguish the signal of association and to nearly completely break down LD in the HapMap samples. The sum of these bounded loci is nearly 7 megabases of sequence, highlighting the fact that while the WTCCC has demonstrated the potential of GWAS to identify associated regions, there is a great deal of effort required to translate this list of novel risk loci into deeper understanding of the biology of disease. Some regions with one or only a few genes offer a potentially obvious path forward in the targeted resequencing of exons, but others are much thornier, containing dozens of genes. Finally, for one of the CD loci, there are zero annotated transcripts in the critical region, providing a challenge to conventional thinking about the types of variation which can be causal with respect to disease association (this particular case, 5p13, was separately found to show an intriguing connection to expression levels of a nearby gene (Libioulle *et al.*, 2007)).

In addition to the set of convincing associations described in Table 3.2, many of the diseases showed a statistical excess of p values showing modest association ($1 \times 10^{-5} > p > 5 \times 10^{-7}$). While the WTCCC discovered more *bona fide* associations than had been generated by all previous candidate gene or positional cloning efforts, the sum of these effects still represents a small fraction of the overall heritability of these diseases. Several possibilities exist to explain the remaining fraction of heritability, including rare variants and gene-gene and gene-environment interactions. Still, it seems likely that many of the modestly associated variants will be found to be true associations once they have been evaluated in larger sample sizes. Chapters 5 and 6 discuss this possibility in

greater detail.

3.5 Discussion

The WTCCC convincingly demonstrated, via the identification of a number of new risk loci for complex disease, successful execution of a GWAS. In addition to these immediately exciting results, there are a number of other useful lessons to be learned. For example, typical effect sizes are even smaller than estimates recently considered realistic (Carlson *et al.*, 2004a). It was thus critical that the WTCCC experiment was designed with this consideration in mind: with 2000 cases and 3000 controls the WTCCC had 64% power to detect a 20% variant with a multiplicative OR of 1.3 (at $p < 5 \times 10^{-7}$), whereas a design more typical of the first generation of GWAS, with 1000 cases and 1000 controls, would have had less than 6% power.

The study can serve as a guide for how generally to approach a GWAS. Concerns before the project began included the problem of multiple hypothesis testing, computational burden, population structure and genomic coverage of available chips. In the end, however, none of these issues posed a significant challenge. Developments from early GWAS, including the WTCCC, have indicated that true associations surpass even conservative thresholds (e.g. 5×10^{-8} (Dudbridge, 2007)), once a sufficiently large sample has been assembled. Efficient software packages including PLINK and SNPTEST have made the standard analyses feasible even on the largest datasets. Population structure within the UK did not lead to systematic increase in type I error, and while incomplete, the coverage of the Affymetrix chip was sufficient to detect nearly all previously documented associations.

More important than any of these expected challenges was careful quality control. In such large datasets, small systematic differences can readily produce effects which appear to be disease associations, or which are capable of obscuring true associations. The detailed analyses of genotype calling allowed

sensible thresholds for heuristic QC filters which removed as much suspect data as possible without discarding actual associations. It was in large part due to this rigorous attention to the details of quality control that the results of the WTCCC were relatively straightforward to interpret.

The WTCCC yielded many exciting, novel associations, but represented only the beginning of the process of identifying, verifying and understanding these loci. The WTCCC itself was only a hypothesis-generating exercise, and requires careful replication to satisfactorily demonstrate that these loci are associated with disease. None of the loci had effects comparable in size to large, longstanding associations, such as the HLA in T1D, and typical per-allele ORs were less than 1.4. This implies that still larger sample sizes will be required in order to identify putative loci with even smaller effects than these. The crucial next steps, therefore, as presented in the next three chapters, are replication of described loci and agglomeration of more samples to pursue additional loci.

Collection	Chr	Region (Mb)	SNP	Trend p	Genotypic p	Risk allele	Minor allele	Het odds ratio	Hom odds ratio	Control MAF	Case MAF
BD	16p12	23.3–23.62	rs420259	2.19×10^{-4}	6.29×10^{-8}	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-7}	3.58×10^{-8}	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-8}	3.15×10^{-7}	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-7}	1.75×10^{-6}	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-8}	5.82×10^{-8}	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-8}	2.03×10^{-7}	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617	3.44×10^{-76}	5.18×10^{-75}	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346	2.42×10^{-134}	5.47×10^{-134}	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-8}	4.92×10^{-7}	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-6}	3.34×10^{-7}	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-8}	1.91×10^{-7}	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453

Table 3.2: Regions of the genome with at least one SNP with a p value $< 5 \times 10^{-7}$ for our primary analyses. Region boundaries are defined by recombination hotspots and return of test statistics to background levels. Minor alleles are defined in controls. Positions are in NCBI build 35 coordinates.

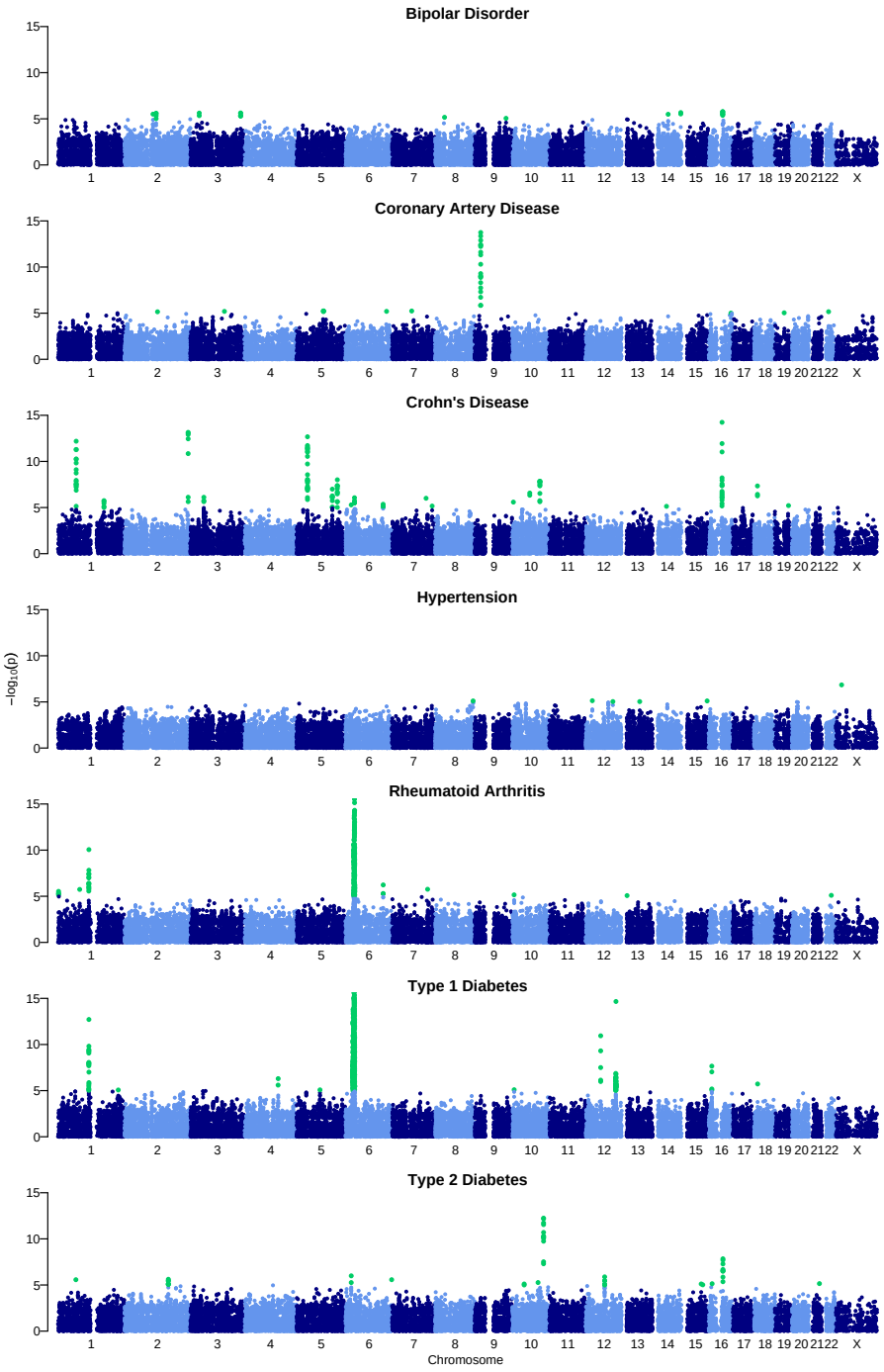


Figure 3.7: For each of seven diseases, $-\log_{10}$ of the trend test p value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with p values $< 1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-log_{10}(p) = 15$, although some markers (for example in the MHC in T1D and RA) exceed this threshold.

Chapter 4

Replication of WTCCC Crohn's disease associations

4.1 Background

4.1.1 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a set of related disorders characterised by chronic inflammation of the gastrointestinal tract. The two most common forms of the disease are ulcerative colitis (UC), first publicly described by Wilks and Moxon (1875) and Crohn's disease (CD), first described by the eponymous Crohn *et al.* (1932). Together, the discovery of these disorders represented a change in the understanding of bloody colitis, which was previously considered to be universally attributed to infection by dysenteric pathogens.

The twentieth century has seen great progress in empirically defining the clinical, pathological and radiological features of these diseases (Xavier and Podolsky, 2007), but their aetiology is still almost entirely unresolved. Both subtypes involve similar clinical features, including fever, abdominal pain, diarrhoea, rectal bleeding and weight loss, but they are endoscopically quite distinct. CD is characterised by discontinuous or segmented inflammation principally in the

ileum and colon, but with possible involvement of any part of the gastrointestinal tract. Complications including stricture, fistulae, and perianal involvement are relatively common. UC, on the other hand, involves continuous inflammation exclusive to the colon without any of the aforementioned complications. The two conditions are thus considered to be distinct, if not discrete, entities, with a misdiagnosis rate of approximately 10% (Podolsky, 2002).

Treatment options vary according to the severity of disease, ranging from longstanding use of aminosalicylates and cortical steroids to infusions of infliximab, an anti-tumour necrosis factor agent, to surgery in the most extreme cases. These treatments are imperfect, however, both due to inconsistent response and drug resistance (20% of treated patients are steroid resistant, and 30% become steroid-dependent (Munkholm *et al.*, 1994)). Some of the more severe treatments (e.g. infliximab) are associated with significant side effects and potential toxicity, including neutropenia, sepsis and lymphoma. The annual cost of treatment of inflammatory bowel disease in the UK was recently estimated to be £376 million (Bassi *et al.*, 2004), marking it as a substantial public health burden. A British population cohort of subjects born in 1970 reported a CD prevalence of 375 cases per 100,000 people (Mathew, 2008). Further estimates of prevalance in Western countries vary, but are widely agreed to be increasing (Loftus, 2004), even among populations with a historically low rate of disease.

4.1.2 Early genetics of CD

There are a number of reasons to believe that CD has a large heritable component. Monozygotic twin concordance is estimated to be 45%, compared with 10% for dizygotic twins (Tysk *et al.*, 1988), and λ_S is consistently estimated to be > 20. This is more than four times greater than a disease like type 2 diabetes, and is even larger than highly heritable auto-immune disorders like multiple sclerosis and type 1 diabetes (Merikangas and Risch, 2003). Unlike these latter two, for which a large fraction of the heritability derives from effects in the HLA region, CD does not have a single overwhelming risk locus.

All of this evidence has fuelled two decades of research into the genetics of CD. A number of international groups carried out family based linkage scans (Satsangi *et al.*, 1996a; Hugot *et al.*, 1996; Rioux *et al.*, 2000) which had been successful in monogenic disorders. A region on 16q (designated IBD1) was the first reported (Hugot *et al.*, 1996) and most widely replicated linkage finding, including among pooled data from 12 centres (Cavanaugh, 2001). Positional cloning efforts by Hugot *et al.* (2001) led to the discovery of a strong association to the *NOD2* gene within the IBD1 locus. Three relatively rare variants (less than 10% in control populations) were found to account for the association signal, and are presumed to be causal: two missense mutations, Arg702Trp and Gly908Arg, along with a frameshift mutation (Hugot *et al.*, 2001; Ogura *et al.*, 2001; Hampe *et al.*, 2001). The population genetics of *NOD2* are also interesting, as the frequency of the risk mutations is much lower in Northern Europe than in Southern Europe (Gaya *et al.*, 2006), and they are essentially absent from East Asian populations (Yamazaki *et al.*, 2002). Thus, the fraction of genetic risk contributed from a given locus can vary widely among populations.

NOD2 is a rare success story among efforts to map genes for complex traits via linkage, but upon close consideration it is clear that the nature of the effect is perfectly suited to the linkage approach. First, heterozygote ORs for these mutations have been estimated to be roughly three (Gaya *et al.*, 2006), marking *NOD2* as a substantial outlier (in terms of effect size) from other confirmed associations to complex disease. Second, the risk alleles are relatively rare among healthy individuals (Gaya *et al.*, 2006), making it easier to trace the segregation of chromosomes carrying the risk allele within families. Finally the presence of multiple risk alleles plays to the strength of linkage methodology, which does not suffer loss of power in regions of allelic heterogeneity.

Follow-up of a weaker (but still significant) linkage signal on chromosome 5q (IBD5) led to an association in a long stretch of several hundred kilobases of strong LD (Rioux *et al.*, 2001). While this association has been replicated numerous times, the susceptibility genes within IBD5 have not been conclusively

identified (Peltekova *et al.*, 2004; Vermeire *et al.*, 2005; Waller *et al.*, 2006). This effect is much smaller than *NOD2*, with an OR = 1.35 (the risk allele is quite common, existing at 40% among controls). Interestingly, this effect by itself is not enough to explain the linkage signal in the region, implying that, unlike *NOD2*, the story is not simply the positional cloning of a locus discovered by linkage.

Other interesting results from linkage scans include chromosome 6p (IBD3), which spans the MHC. This effect has been widely replicated (Satsangi *et al.*, 1996b), but is neither as large, nor as well understood, as it is for other autoimmune diseases. Furthermore, the effect appears to be substantially larger for UC than CD (Gaya *et al.*, 2006). Stoll *et al.* (2004) identified an association to the gene *DLG5* during a positional cloning effort on a linkage peak on chromosome 10q. This association (estimated in the original study to be of similar effect size as IBD5) has not been as consistently replicated (Vermeire *et al.*, 2005; Noble *et al.*, 2005) as either IBD5 or *NOD2*, and thus requires additional data before it can be convincingly labelled a CD risk gene, or ruled out as a false positive. Neither the IBD2 (chromosome 12) nor the IBD4 (chromosome 14) loci have yet had fine-mapping or gene identification reported, despite meeting established criteria for significant linkage. Nor have any of the many additional candidate gene studies in CD yielded consistently replicable associations.

Some insight into the pathophysiology of IBD has come from these early discoveries, coupled with additional evidence from mouse, histological, and biological studies (Xavier and Podolsky, 2007). It seems clear, for instance, that a perturbation of the dynamic balance between commensal flora in the gut and the host is critical to IBD development. This is reinforced, for example, by the observation that some colitis-susceptible strains of mice show no symptoms when maintained in a gnotobiotic state, but develop disease once bacteria are introduced to their environment (Xavier and Podolsky, 2007). IBD seems to develop as a result of bacterial penetration of the epithelial barrier, resulting in the symptomatic inflammatory response from the adaptive immune system.

While incompletely understood, the role of *NOD2* in CD seems to be related to impaired anti-bacterial immunity in the innate immune system, as it is expressed in dendritic cells, Paneth cells and the intestinal epithelium (Xavier and Podolsky, 2007). Despite this progress it is clear that additional insight about the genetic risks will help to understand the full pathogenesis of IBD.

4.1.3 Genome-wide association in CD

The longstanding evidence for a large genetic component for CD risk, coupled with its relatively high public health burden, makes it an ideal phenotype for GWAS. The effect at IBD5 demonstrates the plausibility of common loci of modest risk for CD (for which GWAS is most useful), and the disparity between total heritability and the few known loci implies that many remain to be found. Early GWAS in CD span the spectrum of designs in terms of both SNPs and samples, and provide a useful overview of the general prospects and pitfalls involved, in addition to discovering new genes.

Yamazaki *et al.* (2005) described the first large scale association study in CD, which genotyped about 73,000 gene-based SNPs in 94 Japanese CD patients and 752 healthy controls. Nearly 2000 SNPs with $p < 0.01$ were subsequently genotyped in the second stage of the experiment in an additional 390 CD patients. One association, at *TNFSF15*, remained extremely significant ($p < 10^{-13}$), and was replicated in a UK population, consisting of both families and unrelated cases and controls. The success of this study is somewhat surprising because the relatively small SNP set was chosen without respect to LD and thus did not provide anything close to the coverage of GWAS chips described in Chapter 2. Furthermore, the sample size in the first stage was small, and even the combined sample size from stages 1 and 2 was modest by current standards, and was only well-powered to find very large effects. Indeed, the OR estimate in the Japanese sample for this effect was roughly two, but was much smaller (~ 1.3) in the UK samples. It remains to be seen whether there is truly population-differential risk at this locus. Another possible explanation would be systematic inflation

of test statistics in the Japanese samples (as evidenced by a three-fold excess of $p < 0.01$ in their screening set).

A scan of about 7000 common nonsynonymous SNPs (Hampe *et al.*, 2007) in 735 German CD patients and 368 controls detected a number of strong signals, which were followed up in additional German and UK samples (including both unrelated cases and controls and families). Interestingly, their four strongest initial signals ($10^{-7} > p > 10^{-14}$) showed absolutely no evidence for replication, implying they may have been technical artefacts left behind by insufficiently rigorous QC (as described in Chapter 3). Three signals replicated: the known associations at *NOD2* and IBD5, plus a new finding in the gene *ATG16L1*.

The first true CD GWAS was conducted using approximately 1000 cases and 1000 controls from North America, genotyped on the Illumina HumanHap300 (Duerr *et al.*, 2006; Rioux *et al.*, 2007). This scan identified two new loci with extremely convincing replication ($p_{\text{rep}} < 10^{-6}$): *IL23R* and an intergenic region on chromosome 10q21. An additional three regions (of 19 attempted) which showed modest evidence for association in the GWAS showed nominal replication, and require further validation. Of the previously identified loci, only *NOD2* and *ATG16L1* showed strong signals. IBD5, *TNFSF15*, and *DLG5* showed modest association at best.

The second CD GWAS (and final important study prior to the WTCCC) genotyped 547 Belgian CD patients and 928 healthy controls from Belgium and France, also on the the Illumina HumanHap300. They detected (and subsequently replicated) a signal in a large gene desert on chromosome 5p13, as well as the previously reported hits in *IL23R* and *NOD2*. No significant association was seen in the other previously reported hits, although the *ATG16L1* was confirmed upon genotyping the specific coding SNP reported by Hampe *et al.* (2007). The authors performed dense genotyping in the newly discovered region at 5p13, and observed that some SNPs in the region associated with CD were also associated with expression of the nearby gene *PTGER4*. Curiously, however, the strongest associations with gene expression did not correspond to the

strongest associations with CD, suggesting that the causal relationship to CD has not been fully elucidated.

Knowledge of CD genetics had thus been substantially expanded in a very short space of time. In addition to explaining a larger fraction of the genetic risk of CD, these loci identified disease pathways which were previously unsuspected. *ATG16L1*, for example, highlighted the process of autophagy, the degradation of both a cell's own components and intracellular bacteria. Indeed, cells with small interfering RNA induced knockdown of *ATG16L1* show a significant reduction of autophagic activity when infected with *Salmonella typhimurium* (Rioux *et al.*, 2007).

These successes are slightly tempered, however, by questions raised when considering the full collection of historical results: how could the relatively modest overlap among newly identified regions, and inconsistency with respect to longstanding associations be explained? Were results from GWAS following the same pattern of non-replication as candidate gene studies described in Chapter 1? These issues, along with the validation of hits described in Chapter 3, are addressed in the next section.

4.2 Replication of hits from the WTCCC

As was described in Chapter 3 and highlighted above, the ultimate test of the validity of a genetic association is replication in an independent set of samples. Thus, a CD replication experiment separate from the main WTCCC project was undertaken to follow-up the most significant results. Nine genomic regions showed strong association to CD ($p < 5 \times 10^{-7}$), including four which had been convincingly replicated, as described above: *NOD2*, *IL23R*, 5p13 and *ATG16L1*. The other five loci showing very strong association, plus an additional 25 loci with $p < 10^{-5}$ on initial analysis of the WTCCC dataset were taken forward to the follow-up study. The aim of this project was to focus on the regions which showed the greatest statistical evidence for association in the WTCCC,

irrespective of assumptions about candidacy. Some of the SNPs described below have $p > 10^{-5}$ because support for these markers diminished in the final WTCCC analysis after extensive data filtering.

As expected, nearly all associations show strong signal at many nearby SNPs in LD. As described in Chapter 3, any set of associated SNPs flanked by recombination hotspots was considered to be a single risk locus. Occasionally, however, multiple subsets of variants for which there was strong LD within a subset, but almost no correlation between subsets, were observed within a single locus. The interpretation of such observations is open to debate: are there, in fact, two independent causal variants in the same region, each correlated with a distinct set of proxies? Or is the underlying pattern of LD complex enough that one causal mutation is strongly correlated to both subsets without them being correlated to each other? Two markers were selected from each of the six loci exhibiting this pattern to evaluate this question further.

The process of replicating a genetic association is twofold: confirming that the original estimate of the population frequency of the allele (in controls) is accurate, and confirming that disease cases show a consistent difference from that frequency. The additional disease samples in the WTCCC provide a unique resource to assess how close the control frequency is to the true population frequency, provided that any CD variants are not involved in the other diseases. Allele frequencies for our 37 SNPs were therefore examined in 5746 non-autoimmune WTCCC cases (comprising the BD, CAD, and HT collections). To verify a consistent frequency difference in CD cases the SNPs were then genotyped in a new panel of 1182 Caucasian CD cases using TaqMan assays (Methods and Table 4.3). Concordance with Affymetrix data was 99.7%, based on genotyping 96 WTCCC cases on both platforms. Twelve SNPs showed a difference in allele frequency between these two groups (Table 4.1), while for 25 other markers the allele frequencies converged (Table B.1, Appendix B). To exclude possible bias introduced by this approach, the 12 markers implicated in this preliminary comparison were then genotyped in 2024 independent population controls from the

1958 British Birth Cohort to test formally for association. Results are presented in Table 4.1, with odds ratios estimated from replication samples only.

Among new loci achieving genome-wide significance in the WTCCC scan, the strongest replication for a region containing a known gene was for SNPs rs13361189 and rs4958847 ($p_{\text{rep}} = 6.6 \times 10^{-4}$ and 3.1×10^{-4} respectively) immediately flanking the *IRGM* gene [MIM 608212] on chromosome 5q33.1. Association in the combined panels of 2930 CD cases and 4962 controls was highly significant ($p_{\text{comb}} = 2.1 \times 10^{-10}$ and 3.8×10^{-9} respectively). Figure 4.1 shows the associated region, which also contains a zinc-finger, *ZNF300* and part of the transcript of *NID67*. Two lines of evidence point to *IRGM* as the relevant gene in the interval. First, the SNPs with the strongest signal cluster directly around *IRGM*. While this does not rule out the possibility that different, ungenotyped SNPs in the broader LD window are causal, it increases the prior probability that the causal variant is in this narrow interval. Second, functional knowledge of *IRGM* indicates that it is involved in the autophagy pathway, which was implicated in CD with the discovery of *ATG16L1*, described above. *IRGM* belongs to the p47 immunity-related GTPase family. *Irgm*, its mouse homologue, critically controls intracellular pathogens by autophagy, and $Irgm^{-/-}$ mice showed markedly increased susceptibility to *Toxoplasma gondii* and *Listeria monocytogenes* (Collazo *et al.*, 2001). Consistent with this, IRGM induces autophagy and thereby control of intracellular *Mycobacteria tuberculosis* in human macrophages (Singh *et al.*, 2006).

The long coding exon of the human *IRGM* gene encodes a 20 kD protein of 181 amino acid residues (Bekpen *et al.*, 2005; Singh *et al.*, 2006). Since none of the associated SNPs were known to be functional, this coding exon of *IRGM* and the 4 small putative downstream exons were re-sequenced in 48 CD cases homozygous or heterozygous for risk alleles (Methods). Two novel non-synonymous sequence variants, c.51G→C (p.Glu17Asp) and c.281C→A (p.Thr94Lys) and an exonic synonymous SNP, c.313T→C (rs10065172, p.L105) were detected. These SNPs were then genotyped in 769 unselected CD cases

SNP	Chr	Pos (Mb)	Gene	RAF	WTCCC p	Rep p	Combined p	OR (95% CI)
rs12035082	1q24	169.53 – 169.67		0.389	2.15×10^{-6}	0.01692	2.07×10^{-7}	1.14 (1.02 – 1.27)
rs10801047	1q31	188.17 – 188.32		0.067	8.15×10^{-5}	4.78×10^{-5}	2.83×10^{-8}	1.47 (1.22 – 1.76)
rs9858542	3p21	49.3 – 49.87	Many	0.282	7.71×10^{-7}	0.01058	4.91×10^{-8}	1.17 (1.04 – 1.31)
rs17234657	5p13	40.32 – 40.66		0.125	2.13×10^{-13}	0.05305	3.80×10^{-12}	1.16 (1.00 – 1.35)
rs9292777	5p13	40.32 – 40.66		0.606	1.89×10^{-12}	2.88×10^{-7}	3.20×10^{-18}	1.34 (1.20 – 1.50)
rs10077785	5q23	131.40 – 131.90	*IBD5*	0.766	1.81×10^{-6}	0.008086	7.53×10^{-8}	1.19 (1.05 – 1.36)
rs13361189	5q33	150.15 – 150.31	*IRGM*	0.067	4.38×10^{-8}	0.000664	2.09×10^{-10}	1.38 (1.15 – 1.66)
rs4958847	5q33	150.15 – 150.31	*IRGM*	0.113	2.41×10^{-6}	0.00031	3.77×10^{-9}	1.36 (1.17 – 1.59)
rs6887695	5q33	158.61 – 158.76	*IL12B*	0.318	0.009871	8.42×10^{-5}	9.21×10^{-6}	1.26 (1.12 – 1.41)
rs10883365	10q24	101.26 – 101.32	*NKX2-3*	0.477	1.41×10^{-8}	0.00373	3.71×10^{-10}	1.18 (1.05 – 1.32)
rs2542151	18p11	12.76 – 12.91	*PTPN2*	0.163	4.56×10^{-8}	0.0479	3.16×10^{-8}	1.15 (1.00 – 1.32)
rs2836754	21q22	39.21 – 39.24		0.353	1.07×10^{-5}	0.01224	5.48×10^{-7}	1.15 (1.03 – 1.28)

Table 4.1: Association results for 12 SNPs genotyped in 1182 cases from the replication CD panel. Risk allele frequency (RAF) estimated from the WTCCC controls. Some SNPs with $p > 10^{-5}$ showed stronger evidence in preliminary data from which we fast-tracked them for follow-up. Odds ratios are estimated from the replication samples only.

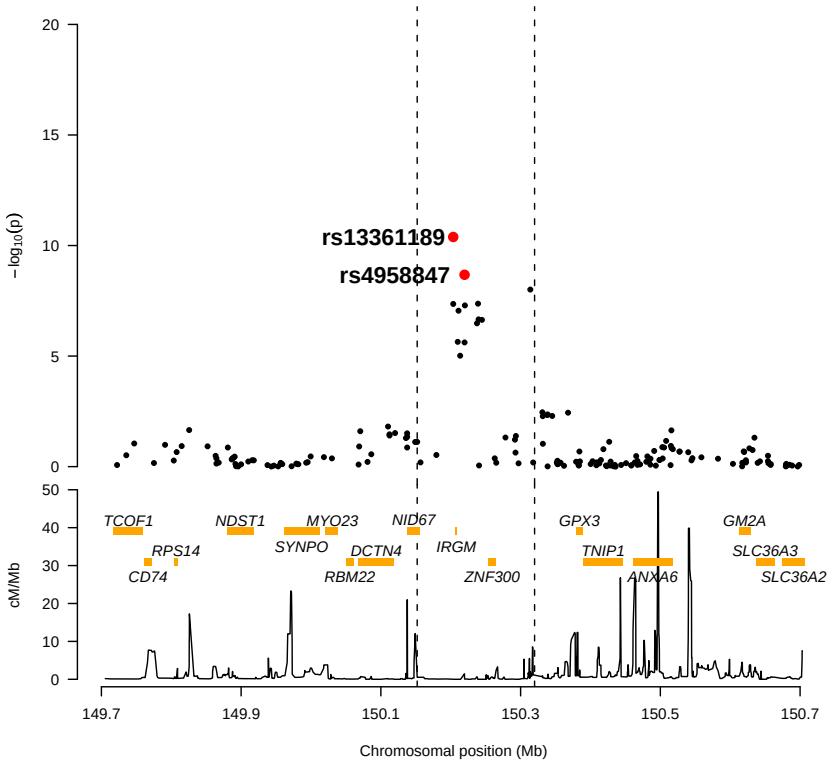


Figure 4.1: Associated region at *IRGM*. The upper panel shows strength of association of Affymetrix 500K SNPs in the region, with combined p values shown in red for replicated SNPs. The lower panel is a recombination map showing how the signals are demarcated by recombination hotspots. Nearby genes are shown in yellow.

from our study and 705 controls. Only the silent 313T→C variant was associated with CD ($p = 0.008$), and was in near perfect LD with SNP rs13361189 (r^2= 0.91). Sequencing of the IRGM coding region in a further 100 unselected CD cases and 100 controls did not detect additional variants. These results strongly suggest that the causal variants do not change the amino acid sequence of the IRGM protein. They may lie in regulatory sequences, in LD with the associated SNPs, and affect IRGM expression; alternatively, the exonic (c.313T→C) SNP itself might affect transcript splicing or the rate of translation of the protein.

Three other associations with genome-wide significance in the WTCCC scan produced strong evidence of replication ($p < 0.01$). The strongest was SNP

rs9292777 ($p_{\text{rep}}= 2.9 \times 10^{-7}$; $p_{\text{comb}}= 3.2 \times 10^{-18}$) which maps to the previously described 1.2 Mb gene desert on chromosome 5p13 (Libioulle *et al.*, 2007) (follow-up was already underway when this association was published). The second most significant previously unreported association was SNP rs10883365 ($p_{\text{rep}}= 0.0037$, $p_{\text{comb}}= 3.7 \times 10^{-10}$), which maps within the *NKX2-3* gene (NK2 transcription factor related, locus 3) on chromosome 10q24.2. *Nkx2.3*-deficient mice develop splenic and gut-associated lymphoid tissue abnormalities with disordered segregation of T and B cells (Pabst *et al.*, 2000). The novel locus at rs9858542 ($p_{\text{rep}}= 0.010$, $p_{\text{comb}}= 4.9 \times 10^{-8}$) on chromosome 3p21 is a 1 Mb region of high LD that contains over 20 genes, including *MST1* (macrophage stimulating 1), encoding a protein which induces phagocytosis by resident peritoneal macrophages. Interestingly, a detailed analysis of this region was subsequently published as part of a follow-up to previous linkage on chromosome 3p, with some evidence that coding changes in *MST1* account for the signal (Goyette *et al.*, 2008).

The modest evidence of replication for SNP rs2542151 ($p = 0.048$) at the *PTPN2* locus (protein tyrosine phosphatase, non-receptor type 2) on chromosome 18p11 ($p_{\text{comb}}= 3.2 \times 10^{-8}$) is of interest since *PTPN2* encodes a T cell protein tyrosine phosphatase, a key negative regulator of inflammation and is also associated with type 1 diabetes (Todd *et al.*, 2007). Puzzlingly, rs10761659 on chromosome 10q21, a region previously reported and replicated by the North American scan (Rioux *et al.*, 2007) and strongly associated in the WTCCC scan, did not show evidence for replication in our data (Table B.1).

Allele frequencies for most of the markers from the 25 other loci studied that did not achieve genome-wide significance in the WTCCC scan but had an initial $p < 10^{-5}$ converged with control frequencies in the CD replication panel (Table B.1). Five of these loci, however, provided at least nominal evidence of replication, indicating that a systematic follow-up in larger samples of all CD hits in the 10^{-5} to 10^{-7} range is likely to yield a number of additional CD susceptibility loci. Four of these loci are novel putative CD risk loci; intrigu-

ingly they include two gene deserts, both on chromosome 1q. SNP rs10801047 ($p_{\text{rep}}= 4.8 \times 10^{-5}$, $p_{\text{comb}}= 2.8 \times 10^{-8}$) is located in a gene desert of 1.7Mb on chromosome 1q31.2, and rs12035082 ($p_{\text{rep}}= 0.017$, $p_{\text{comb}}= 2.1 \times 10^{-7}$) maps to chromosome 1q24 (the association signal does not extend to the nearest gene, *TNFSF18*). Although previous studies did not implicate *IL12B* (interleukin 12B) (Hampe *et al.*, 2007; Rioux *et al.*, 2007), modest WTCCC association at rs6887695 (in the gene) replicated here ($p_{\text{rep}}= 8.4 \times 10^{-5}$, $p_{\text{comb}}= 9.2 \times 10^{-6}$) and warrants follow-up, particularly given its association with psoriasis (Cargill *et al.*, 2007) and the common association of *IL23R* with psoriasis and CD (Duerr *et al.*, 2006; Cargill *et al.*, 2007). SNP rs10077785 ($p_{\text{rep}}= 0.008$) tags the previously described IBD5 locus. Finally, rs2836754 ($p_{\text{rep}}= 0.012$, $p_{\text{comb}}= 5.5 \times 10^{-7}$) maps to an intron in *FLJ45139*, a gene of unknown function on chromosome 21 that is expressed in bone marrow and colon.

Within-cases interaction analysis was performed between known CD susceptibility variants and replicating SNPs. For *NOD2*, rs2066844, rs2066845 and rs2066847 were tested independently or in combination classifying individuals as homozygous wild type, heterozygous or compound heterozygote/rare-allele homozygote. No significant interactions were seen. Testing CD associated variants rs11805303 in *IL23R* and rs10210302 in *ATG16L1* also showed no interactions. These risk loci in CD appear to act independently.

Six sub-phenotypes of CD and two modifiers were examined for differential effects within cases at markers studied. Cases were classified as affected or unaffected for disease site ('pure ileal', 'pure colorectal', 'perianal involvement') and behaviour ('inflammatory', 'stenosing' and 'penetrating'); and by age of onset and smoking status (Table 4.2). Inflammatory behaviour appeared less strongly associated with rs10210302 (in *ATG16L1*, $p = 0.0003$, Bonferroni $p = 0.010$) than stenosing or penetrating behaviour. No further significant differential associations were seen, although 'pure ileal' tends to associate more strongly with most replicating SNPs than 'pure colorectal'.

The understanding of CD pathogenesis has been further refined, as the

genetic evidence regarding *IRGM*, *ATG16L1* (converging on autophagy pathways), *NOD2* and *IL23R* strongly implicates defects in the early immune response, particularly innate immune pathways and the handling of intracellular bacteria. Whether the latter is restricted to a single pathogen or, more likely, a class effect of sub-pathogenic bacteria relying on defective host innate immunity for survival, awaits further investigation.

This study confirmed ($p_{\text{comb}} < 5 \times 10^{-8}$) all four previously unpublished CD loci which achieved genome-wide significance in the WTCCC scan (*IRGM*, *NKX2-3*, *PTPN2*, and 3p21) as well as less conclusively implicating four additional new risk loci. Along with the WTCCC scan, these data also confirm previously published associations at *NOD2*, IBD5, *IL23R*, *ATG16L1* and 5p13. Supporting evidence was also found for both *TNFSF15* and 10q21: the former barely missed the cutoff for follow-up from the WTCCC ($p = 7.9 \times 10^{-5}$) and the latter showed very strong evidence in the WTCCC, but replication was equivocal. The cumulative evidence to this point had thus pinpointed 11 CD risk loci, with varying degrees of evidence for an additional handful.

Upon initial inspection, the WTCCC does not seem to have completely resolved the question posed at the end of Section 4.1.3: why have these studies found different subsets of these loci? Upon more careful consideration, however, the pattern of associations in these studies makes a great deal of sense. Every study found consistent evidence for the two largest effects, *NOD2* and *IL23R* (provided relevant SNPs were genotyped). The remaining complement of nine verified associations are all small enough that each of these studies had imperfect power to detect them. Even for the largest of these effects, *ATG16L1* and 5p13, the pre-WTCCC scans had as little as 10–20% power to detect them at the significance thresholds required to be taken forward into their respective replication experiments. The WTCCC, by contrast, had 85% power to detect these effects, which is borne out by the fact that the WTCCC identified half a dozen modest effects, compared to one or two by each of the smaller scans. Clearly, then, there is great potential in assessing even larger sample sizes to

continue identifying variants of even smaller effect. I shall address this issue in detail in Chapters 5 and 6.

4.2.1 Methods

Participants

1182 European Caucasian CD patients distinct from the WTCCC panel gave written consent and blood for DNA extraction following Research Ethics Committee approval. Standard endoscopic, radiological and histological diagnostic criteria were applied and phenotypic details obtained by case notes review (Lennard-Jones, 1989). The Montreal classification (Silverberg *et al.*, 2005) (age of diagnosis, disease location and behaviour) was used for CD sub-phenotype (Table 4.3). Only one member of multiply affected families was included. The 2024 ethnically matched controls came from the 1958 British Birth Cohort which includes all subjects born in one week of March 1958 in England, Scotland and Wales (Power and Elliott, 2006). There was no overlap with the 1500 samples genotyped as controls in the WTCCC.

SNP genotyping and sequencing

37 SNPs were genotyped using validated Taqman assays (rs1931047 was custom designed, ABI). The human IRGM gene contains one large exon encoding all 181 amino acids of the major isoform; 4 small exons extending 50 kb 3′ to this have been proposed to encode additional isoforms with short extensions of the C-terminal end of the protein (Bekpen *et al.*, 2005). The large coding exon was screened by direct re-sequencing of two overlapping amplimers using primer pairs:

IRGM1_EX1.1F 5′ — GCCAGGATGGTCTCAATCTC

IRGM1_EX1.1R 5′ — GCAGATGCAACCATGATGAA

IRGM1_EX1.2F 5′ — CCCTTCGAAACACAGGACAT

IRGM1_EX1.2R 5′ — CCTTCTGCACCAGGTGTGTT

with ABI BigDye v3.1 dye-terminator chemistry and analysed on an ABI 3730XL Sequencer. Primers for the 4 small putative 3′ exons were:

IRGM1_EX2F 5′ — CGGGTGACCTAAATAACCACT

IRGM1_EX2R 5′ — GCCTCACAAAGTTAGGAAGCA

IRGM1_EX3F 5′ — TTAGTCTGGCCAGGCACTGT

IRGM1_EX3R 5′ — TGAGCTGTGACCCACACTATG

IRGM1_EX4F 5′ — TGCCTCAGTCACTGTTTTGC

IRGM1_EX4R 5′ — TGATGAGACCCCACTTTTTCA

IRGM1_EX5F 5′ — CCCATGTCTGCTACAGAACTTTT

IRGM1_EX5R 5′ — TGGATTTCCTCTGTATTCCACA

Identified IRGM variants were genotyped by direct re-sequencing as above.

Statistical analysis and quality control

Using data from a preliminary analysis of the WTCCC GWA scan, 37 SNPs showing allelic association at $p < 10^{-5}$ using the Cochran-Armitage trend test were selected for follow-up. Table B.2 in Appendix B shows missing data rates and HWE p values for the 37 SNPs. Twelve SNPs showed consistent case/control allele frequencies in additional genotyped cases and three non-autoimmune samples from the WTCCC. These 12 SNPs were evaluated using the trend test as implemented in PLINK (Purcell *et al.*, 2007), both alone and combined with the WTCCC dataset. For a multiplicative effect, with a perfect proxy to the locus, the power to detect the effect with the replication sample set of 1182 cases and 2024 controls at $p < 0.01$ was as follows:

OR	MAF	Power (%)
1.2	0.1	38
	0.2	66
1.3	0.1	78
	0.2	96

Case-only interaction tests were performed for replicating SNPs against *CARD15*, *IL23R* and *ATG16L1* using PLINK. We tested for association at these loci within-cases in the original WTCCC sample for six sub-phenotypes and two modifying effects. Significance would indicate the variant preferentially affects one subphenotype more than the others.

SNP	Ileal	Colonic	Perianal	Inflammatory	Stenosing	Penetrating	Smoking	Age at Onset
rs11805303	0.970	0.457	0.071	0.079	0.147	0.586	0.240	0.626
rs12035082	0.363	0.175	0.400	0.121	0.087	0.498	0.911	0.555
rs10801047	0.719	0.158	0.125	0.247	0.929	0.204	0.194	0.642
rs10210302	0.458	0.024	0.667	**0.0003**	0.005	0.125	0.374	0.019
rs17234657	0.104	0.058	0.863	0.077	0.882	0.092	0.768	0.904
rs9292777	0.049	0.143	0.252	0.370	0.026	0.175	0.580	0.794
rs10077785	0.108	0.990	0.230	0.698	0.634	0.606	0.683	0.770
rs13361189	0.876	0.448	0.436	0.512	0.847	0.260	0.505	0.952
rs4958847	0.466	0.162	0.200	0.416	0.994	0.116	0.561	0.570
rs6887695	0.463	0.425	0.326	0.468	0.277	0.024	0.656	0.504
rs2542151	0.329	0.290	0.763	0.050	0.080	0.307	0.277	0.387
rs2836754	0.961	0.810	0.946	0.285	0.765	0.093	0.187	0.059

Table 4.2: Within-cases test of association for 6 CD sub-phenotypes (ileal, colonic and perianal specify disease location; inflammatory, stenosing and penetrating specify disease behaviour) and 2 modifiers (smoking and age at onset). The one association which remained significant after correcting for multiple tests is shown in bold.

	WTCCC (n = 1748)	**Replication** (n = 1182)
Median age (yrs)	45.7	43.9
Median age at diagnosis (yrs)	26.1	25.5
Gender (female: male)	1.55	1.48
Jewish ancestry	1.75%	1.03%
Ileal location	33.1%	38.1%
Colonic location	31.2%	29.6%
Ileo-colonic location	35.7%	32.3%
Perianal involvement	27.8%	23.8%
Inflammatory behaviour	45.6%	40.2%
Stenosing behaviour	37.0%	38.7%
Penetrating behaviour	17.5%	21.2%

Table 4.3: Demographic and subphenotype details for CD samples.

Chapter 5

Integrating publicly-available genome wide association studies

5.1 Introduction

Chapters 3 and 4 have demonstrated the success of genome-wide association at identifying novel genetic loci associated with multifactorial human diseases including diabetes, cardiovascular disease, breast, prostate and colon cancer and several auto-immune disorders (Duerr *et al.*, 2006; Wellcome Trust Case Control Consortium, 2007a; Easton *et al.*, 2007; Zeggini *et al.*, 2007; Parkes *et al.*, 2007; Gudmundsson *et al.*, 2007; Todd *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007b). Because of careful attention to statistical rigour and independent replication, many of these associations have achieved levels of statistical support which frequently eluded earlier candidate gene studies (Chanock *et al.*, 2007). Most of the GWAS conducted thus far have employed case-control designs with 500-2000 cases and similar numbers of controls, and in general the associated variants have OR of about 1.25, with very few loci at

or above OR = 1.5 (Cardon, 2006).

Despite the success of the first generation of genome-wide scans, there are likely many more susceptibility loci awaiting identification. Although some of these will comprise rare mutations (Cohen *et al.*, 2004), structural variants (Sebat *et al.*, 2007) and population-specific loci (Negoro *et al.*, 2003), there are also additional loci to be found using the existing marker sets and study designs, but with larger sample sizes. In Type 2 diabetes, for example, seven new regions of replicated association were identified by combining three genome scans (total of 4549 cases/5579 controls). None of these loci met the study-specific screening thresholds in all three studies, yet in combination the loci yielded significance levels ranging from 10^{-7} to 10^{-16} and were subsequently validated in independent samples (Wellcome Trust Case Control Consortium, 2007a; Scott *et al.*, 2007; Saxena *et al.*, 2007). Similar studies have recently been completed for lipid concentrations (Kathiresan *et al.*, 2008; Kooner *et al.*, 2008; Willer *et al.*, 2008)(yielding 7 new loci), and Crohn's disease, described in Chapter 6 (increasing the number of validated associations from 11 to over 30). These outcomes illustrate the theoretical expectation that larger sample sizes lead to increased power to detect modest effects (Altshuler *et al.*, 2000; Clarke *et al.*, 2006).

Many of the recent and ongoing GWAS have made the genotype and phenotype data available to the scientific community. Indeed, there are already more than 50,000 samples available in public repositories, with additions occurring regularly (see, e.g., the WTCCC (http://www.wtccc.org.uk/) and dbGAP (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap)). In principle, these data have the potential to yield direct increases in statistical power to detect smaller genetic effects via increased sample sizes, which is one of the main reasons that funding agencies have strongly encouraged or mandated public data release (http://grants.nih.gov/grants/gwas/index.htm). In practice, however, it is unclear how to make use of public data to increase power while avoiding an increase in type I error rate.

The potential applications of public genotype and phenotype data to association studies can broadly be divided into two categories: (i) performing a meta-analysis of multiple scans for the same phenotype, each consisting of a distinct set of cases and controls; (ii) directly combining genotype data from multiple studies with the intent of treating them as a single dataset for subsequent analyses. The meta-analysis approach is attractive because it provides insulation from some sources of bias and preserves whatever case-control matching is done within each study. The most obvious scenario for use (ii) is to increase power by including additional controls, as shown to be successful in a recent nsSNP scan of four diseases (Wellcome Trust Case Control Consortium, 2007b). Either approach, but especially the second, requires that the public data be combined in such a way that the power gain is not negated by the inevitable effects of genetic heterogeneity, variability in genotype quality, differences in the composition of marker panels and other genetic confounders. The aim of the present study is to elucidate the sources and magnitude of genetic variability in presently available public data and assess how they can be integrated for practical applications in association mapping.

5.2 Results

In order to gain insight into the issues related to integrating GWAS, four publicly-available datasets were selected which vary for several interesting properties, including nationality, genotyping technology, calling algorithms used, and means of data distribution (Table 5.1).

5.2.1 Data annotation and distribution

The first step in combining these datasets was to correctly annotate SNPs relative to some genomic reference. Genomic build and strand information was obtained from the WTCCC manuscript for the UKBS and 58C Affymetrix datasets, from the download site for NINDS, and was not obviously available for

Dataset	N Samples	QC+ SNPs	Platform	Nationality
UKBS	1433	459,450	Affymetrix 500K	UK
NINDS	817	308,235	Illumina HumanHap300	USA
NIMH	1229	440,439	Affymetrix 500K	USA
58C	1449	459,450	Affymetrix 500K	UK
58C	1424	537,862	Illumina HumanHap550	UK

Table 5.1: Samples used for evaluating GWAS integration. UKBS and 58C are WTCCC controls described earlier, NINDS samples are from the National Institute of Neurological Disorders and Stroke(Schymick *et al.*, 2007) , NIMH samples are controls from the National Institute of Mental Health Center for Collaborative Genetic Studies (http://www.nimhgenetics.org/).

the NIMH data. Official annotations from genotyping companies such as Illumina and Affymetrix are frequently updated and modified, and it is difficult to map these to the versions used for any particular project unless annotations are distributed with the datasets. While unknown or conflicting genome builds lead to difficulty in precisely placing particular SNPs, incorrect strand orientation can render data unusable or yield spurious results. For example, the 58C and NIMH Affymetrix datasets share 394,044 polymorphic SNPs which pass QC in both samples. Of these, 330,136 have non-complement alleles (i.e. A/C, A/G, C/T, G/T) for which strand flips are easily spotted. Almost exactly half of these (165,003) show strand discrepancy between the two datasets. This leaves, however, 63,908 SNPs with complement alleles (i.e. A/T, C/G) which cannot be reliably matched without annotation files.

One possible means of correctly determining alleles in such cases is to match minor alleles from two samples. This approach fails, however, for markers with allele frequency close to 50%. In essence, such an approach would have no power to detect true associations at these nearly equifrequent markers, because cases and controls would have different minor alleles and would be incorrectly flipped to match. This is especially problematic because power to detect an effect increases with allele frequency: a study with 1000 cases and 1000 controls would have 23% power to detect ($p < 10^{-6}$) a MAF 0.45 locus with an OR of 1.3, compared to just 8% for the same OR but 0.20 frequency. This is borne out by the observation that 4 of 24 associations found in the WTCCC have opposite

minor alleles in cases vs. controls. Figure 5.1 shows that adding additional data without reliable strand information substantially increases power for most frequencies, but is indeed problematic close to 50% MAF.

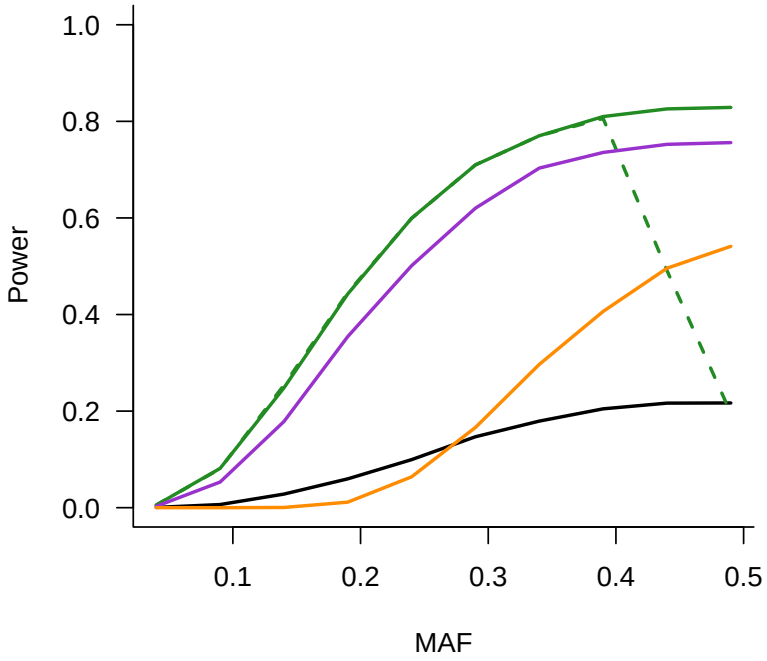


Figure 5.1: Power to detect an effect with OR=1.3 vs. frequency under a variety of circumstances in combined analyses: black line shows power for 1000 cases and 1000 controls; solid green line shows 2000 cases and 2000 controls from one study; dashed green line shows two studies of 1000/1000 each, with alleles matched on frequency, demonstrating reduced power at high frequency; purple line shows 1000/1000 from US and 1000/1000 from UK, corrected by genomic control for residual structure (after MDS), showing less power than a homogeneous set of the same size, but much more than either alone; orange line shows a SNP with 1000/1000 samples genotyped, and 1000/1000 badly imputed (see text), with the most severe attenuation at low frequencies.

Most public distributions of GWAS data consist of the genotype calls, but there are potential problems with releasing only these processed data, instead of the raw intensity values upon which the genotype calls are based. It has been demonstrated, for instance, that mistakes in genotype calling on many platforms

(even after strict QC filtering) lead to spurious associations (Clayton *et al.*, 2005). From the perspective of merging datasets, different calling algorithms can also introduce false positives. For instance, when BRLMM genotypes were compared to CHIAMO genotypes on the identical set of intensities for the 58C (which should ideally lead to identical genotypes), nearly 700 SNPs show allele frequency differences with $p < 0.01$. The best way to address these issues is to use the same calling algorithm whenever possible, and to visualise cluster plots of the intensity data (Figure 5.2) for all putative associations — both of which require distribution of the raw intensity data.

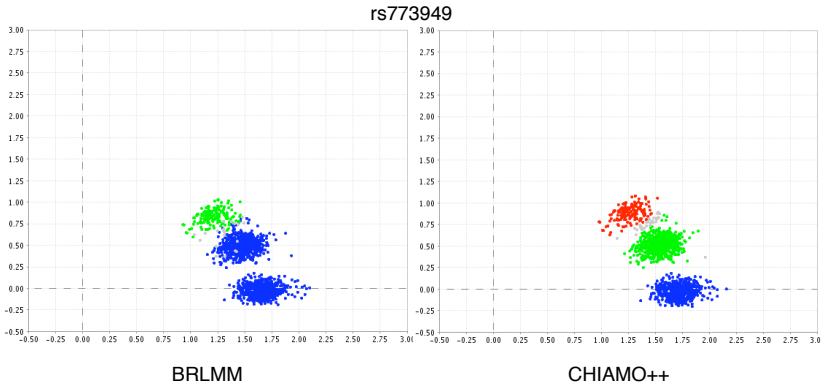


Figure 5.2: Extreme difference in genotype calling between two algorithms, BRLMM (L) and CHIAMO++ (R). While BRLMM has clearly made an egregious calling error, both sets of calls pass QC because they have very little missing data, and coincidentally both fall into Hardy-Weinberg equilibrium.

Annotation of DNA samples is also important, as both plate- and lab-specific effects on genotyping have been previously documented (Clayton *et al.*, 2005; Wellcome Trust Case Control Consortium, 2007a) and are likely to be exacerbated by the sheer size of GWAS datasets. It is therefore recommended that such information is provided, whenever possible, in much the same way as raw genotyping data. A useful example of using SNP data to annotate sample data is the comparison of specified gender in sample manifests with empirically determined genetic gender. Figure 5.3 simultaneously shows both that manifest genders are often incorrect, and that attributes of platform and calling algorithm

can have unforeseen effects.

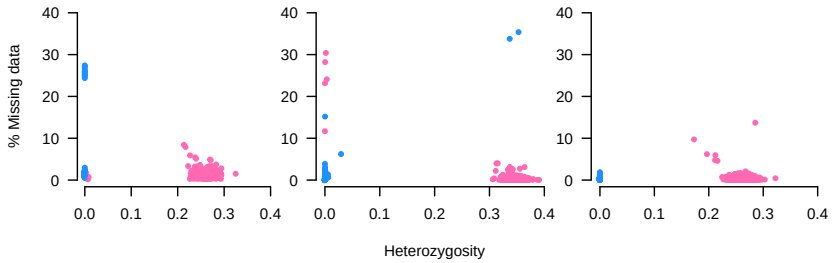


Figure 5.3: X chromosome heterozygosity and missingness for NIMH(Affymetrix-BRLMM), NINDS(Illumina) and WTCCC(Affymetrix-CHIAMO++) data (left to right) with males coloured blue and females pink. BRLMM disallows male heterozygotes and marks such genotypes as missing. This non-randomly inflates missingness in males (which should look more like the samples incorrectly labeled as females near the origin of the left panel), and makes females mislabeled as males appear to be completely homozygous, but with a missing rate approximately equal to the expected heterozygosity rate in females.

5.2.2 Population structure

Another potential impediment to combining genome-wide datasets is population stratification between samples, which can both reduce power and create false positive signals (Marchini *et al.*, 2004). A number of methods have been detailed for addressing this issue, including genomic control (Bacanu *et al.*, 2000) and structured association (Pritchard *et al.*, 2000). Genomic control applies a uniform adjustment to all markers, which has the drawback of heavily penalizing markers with little difference in frequency while under-correcting for markers with extremely divergent frequencies in different populations. Structured association approaches are potentially more sensitive to this problem, but are currently computationally unfeasible for genome-wide data. For these reasons the data were analysed via MDS, as implemented in the PLINK software package (Purcell *et al.*, 2007).

MDS was employed for two distinct reasons: (a) to identify and remove individuals with non-European ancestry and (b) to construct axes of variation to remove the more subtle, within-population structure. Given that all of these

samples are meant to be of European ancestry (likely true for most exercises in combining GWAS) samples with a large proportion of non-European ancestry should be excluded. As was described in Chapter 3, the MDS was first seeded with the three HapMap panels (representing European, East Asian and Sub-Saharan African ancestry) to identify outlier samples (Figure 5.4).

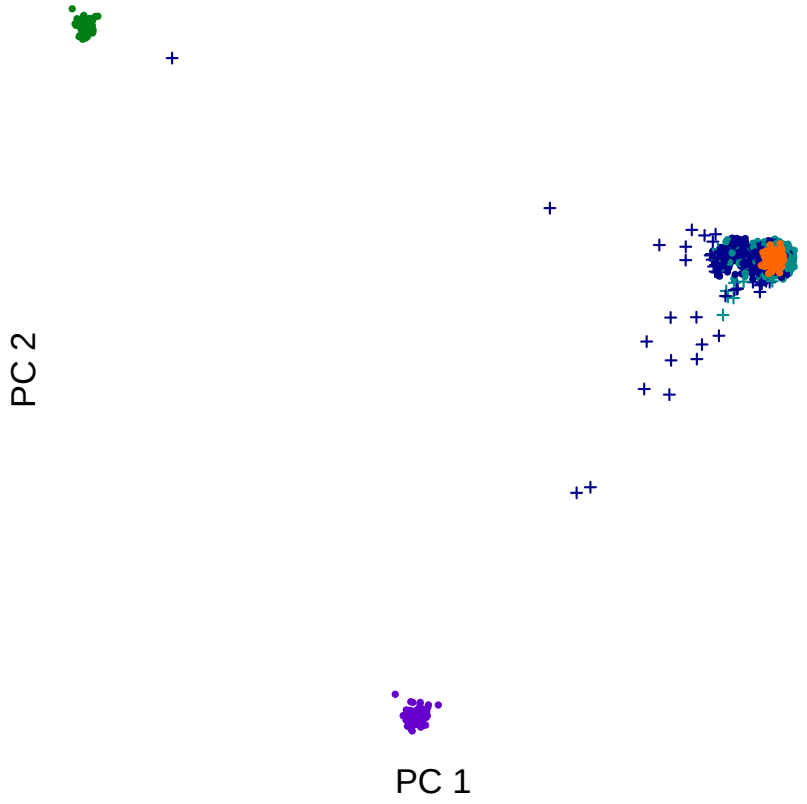


Figure 5.4: MDS of all samples, plus three HapMap panels to identify outliers (crosses). Note that even on these axes, which are driven to correspond to continental differences by the presence of the HapMap samples, it is possible to see more dispersion in US samples (dark blue) than UK samples (cyan).

The second goal was to tease apart modest levels of within-Europe structure, which is much more sensitive to other data issues. While refining the within-Europe MDS (using just the GWAS samples after excluding outliers), for example, axes were found which corresponded to duplicate samples, undocumented relatives, and a small number of SNPs which were grossly mis-called

(see above) in one dataset. Even after filtering out these artefacts, only the first axis robustly represents population structure, corresponding to the previously reported North-South cline in Europe (Tian *et al.*, 2008). The second axis represents a large, documented inversion on chromosome 8p23.1 (Barber *et al.*, 1998), the third and fifth correspond to long patterns of LD encompassing the MHC region (it is not yet clear whether these are also inversions, or result from some other process, such as selection) and the fourth to very modest geographic variation that has no effect on the inflation of test statistics (see below).

As expected, the US samples have a more widely spread distribution of the first, demographically-relevant, component (Kolmogorov-Smirnov comparison between the UK and US distributions, $p < 10^{-16}$), which represents the greater variability of Northern vs. Southern European ancestry, and more samples at the Southern tail, which correspond well to a small number of US samples with self-reported grandparental ancestry in these regions (data not shown). The sample composition of the MDS analysis strongly influences the results, as evidenced by the fact that this sample, which is over half British, yields this North-South cline.

Population structure can both decrease power and generate false positive associations. The latter was evaluated by considering the genomic control metric, λ_{GC}, a measure of the overall inflation of test statistics. Table 5.2 shows values of λ_{GC} between several pairs of datasets both before and after correcting for the previously discussed geographic axis of variation.

Because λ is directly proportional to sample size, calculations were performed on size-matched sets from each dataset (N = 792, the number of NINDS samples which pass outlier-QC). Thus, the values in each column are directly comparable. As expected, there is little evidence for structure between two UK samples. Comparing UK with US samples, on the other hand, suffers from an unacceptably high level of stratification. Correcting for the most significant MDS axis, however, eliminates a large fraction of this inflation. Comparing

N	Sample 1	Sample 2	$\lambda_{\text{uncorrected}}$	$\lambda_{\text{corrected}}$
792	UKBS	NIMH	1.25	1.08
	UKBS	NINDS	1.29	1.07
	UKBS	58C	1.03	1.03
	58C	NIMH	1.21	1.07
	58C	NINDS	1.26	1.08
	NIMH	NINDS	1.10	1.09
1584	UK	US	1.43	1.15
	58C + NIMH	UKBS + NINDS	1.10	1.10

Table 5.2: Genomic control lambda comparing various datasets before and after correcting for the demographically-relevant MDS axis.

the two US groups reveals more structure than within the UK, but less than between the US and UK.

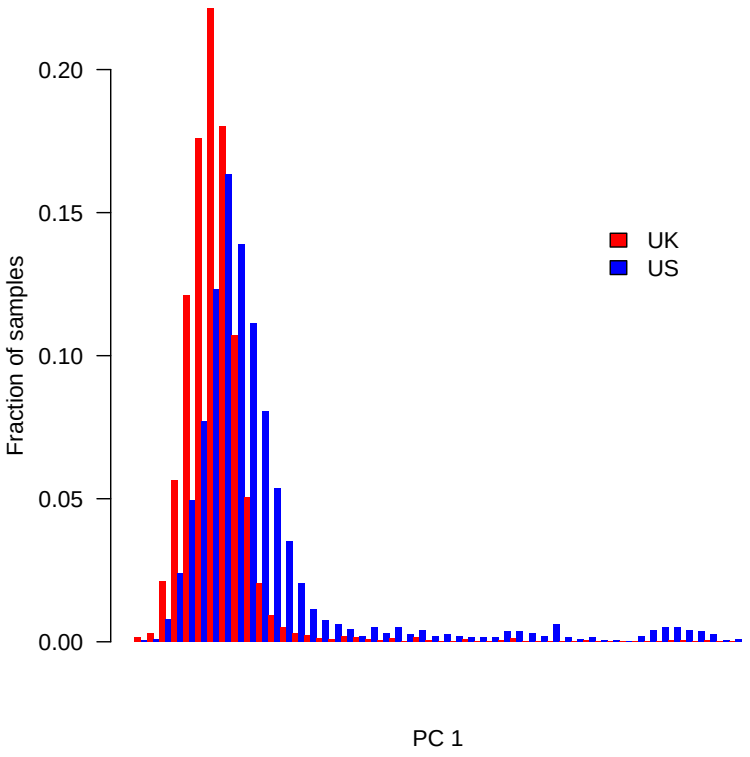


Figure 5.5: Histogram of MDS axis corresponding to NE-SW variation in Europe for UK samples (red) and US samples (blue). US samples are shifted overall toward the right (farther South) and have more outlying samples.

Two analyses were performed on combined pairs of the data. First, com-

paring all UK with all US samples shows the expected increase in λ due to increased sample size, but most of this is removed by conditioning on the PC. This is in essence the worst possible scenario (controls from one locale, cases from another), so a comparison was also made of 58C plus NIMH to UKBS plus NINDS (combining cases and controls from both locales), which substantially decreases the level of inflation.

5.2.3 Imputation

Recently developed imputation methods enable the prediction of ungenotyped SNPs by combining information from genotype data with reference patterns of variation observed in the HapMap. To evaluate the effects of using imputed genotypes when combining GWAS, the 58C samples were used because they have been genotyped on both the Affymetrix 500K and Illumina 550K chips. The samples were divided into two equal groups, with one half considered to be cases (genotyped on Affymetrix) and the other half controls (genotyped on Illumina). In such a circumstance a fraction of the SNPs overlaps on the two platforms for which there will be actual genotypes for all samples, and some fraction for which there will be genotypes in cases and imputed data in controls, or vice versa (the possibility of testing SNPs in the HapMap but on neither chip was not considered). Figure 5.6 shows that the 71,994 overlapping QC+ SNPs behave as expected under the null, with neither systematic inflation ($\lambda_{GC} = 1.04$) nor any significant results in the tail.

There were 735,025 SNPs which passed QC in genotyped samples and for which imputed data was available in the other samples. The IMPUTE program (Marchini *et al.*, 2007) provides a measure of certainty in imputed genotypes, the average maximum posterior probability (AMPP), which is defined for each SNP as the sum over all individuals of the posterior probability of the most likely genotype, divided by the number of individuals. Individual SNPs will obviously have poor predictability if they are in areas of low LD, or if the relevant SNP chip has poor nearby density. For this reason a stringent threshold was applied

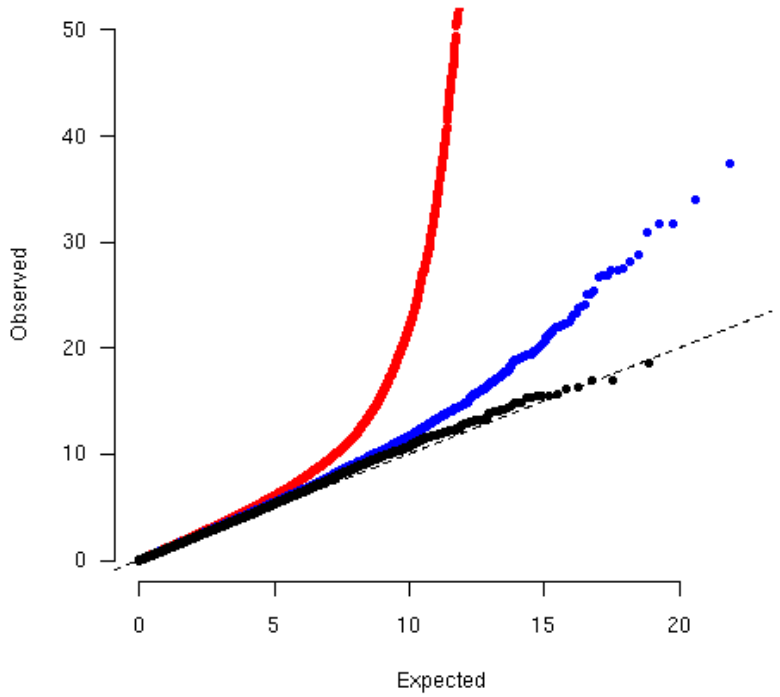


Figure 5.6: Quantile-quantile plot of observed vs expected χ^2 statistics between 688 samples genotyped on the Illumina 550 chip and 688 samples genotyped on the Affymetrix 500K chip. SNPs present on both chips are shown in black. All SNPs on one chip and imputed from the other with AMPP > 0.98 are shown in red. A subset of these SNPs passing further QC (see text) are shown in blue.

of AMPP > 0.98, which left 475,423 SNPs with high-confidence imputation, shown in light blue in Figure 5.6. There is a substantial enrichment in significant associations, with 2756 (0.6%) $p < 10^{-3}$ and 658 (0.1%) $p < 10^{-6}$.

Two observations help to address some of these issues. First, imputation methods tend to make more confident mistakes (i.e. errors for which the AMPP is high) in rare SNPs. SNPs with MAF less than 0.1 are more than ten times as likely to have $p < 10^{-6}$ as common SNPs. Since most association studies have little power to detect the expected modest effect sizes at these rare SNPs, little is lost by excluding them. Second, the use of imputed data adds an assumption that the HapMap data is of uniformly high quality. Because of its small sample

size, and the desire of the project to have as complete a dataset as possible, the filtering was not very strict. It allowed, for example, occasional Mendelian inheritance errors and relatively large departures from Hardy-Weinberg equilibrium. To address this, SNPs which fail any of the following criteria were filtered out:

- Newly documented strand flips between chip annotation and HapMap.
- SNPs removed from latest chip annotation files.
- HapMap SNPs with $> 5\%$ missing data.
- SNPs with multiple discrepancies between official HapMap data and subsequent genotyping performed by Affymetrix or Illumina on HapMap samples.
- SNPs with substantial differences between BRLMM and CHIAMO.

The 347,560 SNPs which remained after the frequency and HapMap filters are shown in dark blue in Figure 5.6. Nearly all of this distribution behaves as expected under the null ($\lambda_{GC} = 1.04$), concurring with previous reports (Marchini *et al.*, 2007) that imputation is accurate. The inflation in the tail was substantially attenuated, with 777 (0.2%) $p < 10^{-3}$ and just 17 (0.005%) $p < 10^{-6}$. It is not clear what accounts for the errors in these last few SNPs, which are problematic even in different samples or using different imputation methods (data not shown).

It is worthwhile to consider the effect of using imputed data in a variety of the possible circumstances in merging GWAS datasets, illustrated by Table 5.3. In each row a high-confidence, but badly imputed SNP (from the tail of the dark blue distribution in Figure 5.6) was simulated both under the null and with a true association (OR = 1.5). The fractions of such simulations which achieve $p < 10^{-6}$ was then computed, representing the false positive rate in the former case and power in the latter. In the fully-genotyped scenario, a false association was never (in 1000 simulations) observed and the power to detect

the true association was 23%. In the scenario where two studies with the same ratio of cases to controls had genotype data on different platforms, and used imputation to predict SNPs on the other platform, statistical noise has been added in the form of bad imputation, which reduces power. Finally, in the case where the imputation noise is correlated with case status, both the Type I and II error rates are affected.

Cases	**Controls**	**Null**	**OR = 1.5**
Genotypes	Genotypes	0	0.23
Mix	Mix	0	0.13
Genotypes	Imputation	0.17	0/0.99

Table 5.3: Fraction of simulations reaching $p < 10^{-6}$ for a badly imputed SNP under the null (false positive rate) or a simulated effect (power) for scenarios with all samples genotyped; a mixture of imputation and genotypes in cases and controls; cases genotyped and controls imputed.

5.3 Discussion

While GWAS have already had great success in elucidating the pathology of complex human disease, it is obvious that increased sample size, and the concomitant increase in power (Figure 5.1) will help to fulfill their promise. This enormous quantity of data (the WTCCC alone generated 8 billion genotypes), combined with rapidly changing technological and statistical tools related to its generation has raised new analytical issues and revealed previously unconsidered aspects to existing problems, including the management of voluminous annotation data, population substructure and the use of imputation data. Publicly-available datasets were used to evaluate empirically these issues and find that, on the whole, they are tractable provided they are addressed carefully.

The most basic requirement for publicly available data to be useful is that they can be interpreted by researchers with no connection to the original project. It is imperative that detailed information relating to build, strand and genotyping technology version are released in a complete and consistent manner. Efforts like dbGAP are already underway to encourage this process, and other projects

should adopt a standardized set of basic analyses and policy of full data release, including raw intensities.

GWAS provide such a wealth of data that many straightforward analyses provide useful insight. MDS, for instance, can be employed to detect both gross and subtle levels of population structure. While simply removing outliers is the most sensible approach in the first case, conditioning on demographically-relevant axes of variation was better for small effects. Given that the useful axes will depend on the geographic composition of the samples, it might be sensible to use recently published genome-wide data from the Human Genome Diversity Panel (Jakobsson *et al.*, 2008) to select loadings for particular SNPs which would be useful in distinguishing particular populations. When performing such an analysis on a mixture of UK and US samples, a broader distribution was found in the US samples, which is unsurprising, given the greater heterogeneity of ancestry in the US. A less obvious observation, however, is that distinct US datasets do not show as much systematic inflation when compared with each other rather than to UK datasets. This makes sense upon consideration that the mere presence of heterogeneity is not enough to cause inflation—there also must be different mixture proportions of the sub-classes (Marchini *et al.*, 2004). These mixture proportions will be highly dependent on sample ascertainment, but these particular datasets (which were ascertained in markedly different ways) seemed to be only modestly distinct.

Imputation methods offer great promise in merging datasets with different SNP sets, but again must be treated with caution. For example, the QC standards for including SNPs in the HapMap, while relatively stringent at the time (International HapMap Consortium, 2005), are quite lax by current GWAS standards (Wellcome Trust Case Control Consortium, 2007a) (allowing, for instance, up to 20% missing data per SNP). While many of these SNPs could be excluded via filters on the current HapMap dataset, there remained a few SNPs which are prone to errors. This and other problems with the HapMap data (such as the relatively small sample size) percolate through the imputation

process to create problems in the downstream association analysis. A project is currently underway to generate a 'Phase III HapMap' consisting of the roughly 1.7 million SNPs on either the Illumina 1M or Affymetrix 6.0 chips, genotyped on twice as many samples as the current HapMap. Such a resource, subjected to rigorous quality control and phased into haplotypes, would be an excellent starting point for future projects using imputation.

To put problems with imputation in perspective, one must consider how common each possible scenario is. Combining imputed data with genotype data in equal proportions in cases and controls is problematic only in the extremely rare case of a badly imputed SNP being associated (the product of two low probabilities). If genotypes and imputation are mixed unequally between cases and controls, however, badly imputed SNPs under the null, which encompasses nearly the whole genome, will falsely appear to be associated. While this number is still small, the rejection of each is a potentially expensive genotyping experiment. While this challenge has been illustrated specifically for imputation, it follows that any source of experimental noise, including genotype calling and population structure, is most problematic when correlated with disease status.

Generous data release policies have afforded the field of human disease genetics the chance to mine GWAS datasets more deeply by combining these datasets. This opportunity could be impeded by issues of data annotation or derailed by artefacts introduced by genotype calling, population structure or other sources of bias. However, feasible methods exist for dealing with all of these problems in most circumstances, and careful attention to these details can unlock the full potential of these studies.

Chapter 6

Meta-analysis of three Crohn's disease GWAS

6.1 Introduction

As described in Chapters 3 and 4, the first GWAS have identified many common variants associated with complex diseases, and have rapidly expanded our knowledge of the genetic architecture of these traits. Progress in CD has been especially striking, with the previously described GWAS publications increasing the number of confirmed associated loci from two to eleven. The results have identified new pathogenic mechanisms of IBD and promise to advance fundamentally our understanding of CD biology. These recent discoveries highlight, for instance, the key importance of autophagy and innate immunity (Hampe *et al.*, 2007; Parkes *et al.*, 2007; Rioux *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007a) as determinants of the dysregulated host-bacterial interactions implicated in disease pathogenesis. Furthermore, genetic associations have been shown to be shared between CD and other auto-inflammatory conditions — for example, *IL23R* variants (Duerr *et al.*, 2006) are also associated with psoriasis (Cargill *et al.*, 2007) and ankylosing spondylitis (Wellcome Trust

Case Control Consortium, 2007a), and *PTPN2* variants with type 1 diabetes (Wellcome Trust Case Control Consortium, 2007a). As in other complex diseases, restricted sample sizes have resulted in early CD studies focusing on only the strongest effects, which turn out to explain only a fraction of the heritability of disease.

The most efficient way to increase the sample size, and achieve the increase in power necessary to detect loci of modest effect (Altshuler and Daly, 2007), is to combine information from multiple datasets, especially when multiple complete scans exist for the same phenotype. Chapter 5 described some potential pitfalls involved in such a meta-analysis, but also noted that the presence of both cases and controls in each constituent study (and thus the absence of strong correlation between disease status and inter-study variation) reduces the magnitude of some of these problems. Therefore, a meta-analysis of the three largest CD GWAS in CD in European-derived populations (Rioux *et al.*, 2007; Libioulle *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007a) was undertaken.

6.2 Results

The combined GWAS study samples (Table 6.1) consisted of 3230 cases and 4829 controls, all of European descent. While the individual scans did identify new risk factors, they were not well-powered to discover common alleles with effects below ORs of 1.3 (Figure 6.1). By contrast, the power of the combined sample is excellent at an OR of 1.2, allowing evaluation of the role of alleles with smaller effect sizes for the first time. As two different genotyping technologies were used in the constituent scans, recently developed imputation(Marchini *et al.*, 2007) methods were used to assess association across all three studies at 635,547 SNPs contained on one or both platforms. A quantile-quantile (QQ) plot of the primary meta-statistic (single SNP Z-scores, Figure 6.2) shows a striking excess of significant associations, well beyond what would be attributable to the modest overall distributional inflation ($\lambda_{GC} = 1.10$). Despite the large sample

size, the overall inflation is modest because (1) each group had separately tested for evidence of population stratification, and the meta-analysis used a test that combined the results from each study (rather than mixing the raw data and compromising the case-control matching of each study), and (2) imputation was done on all samples ignoring case status and thus would not introduce artefactual differences between cases and controls(Balding *et al.*, 2003).

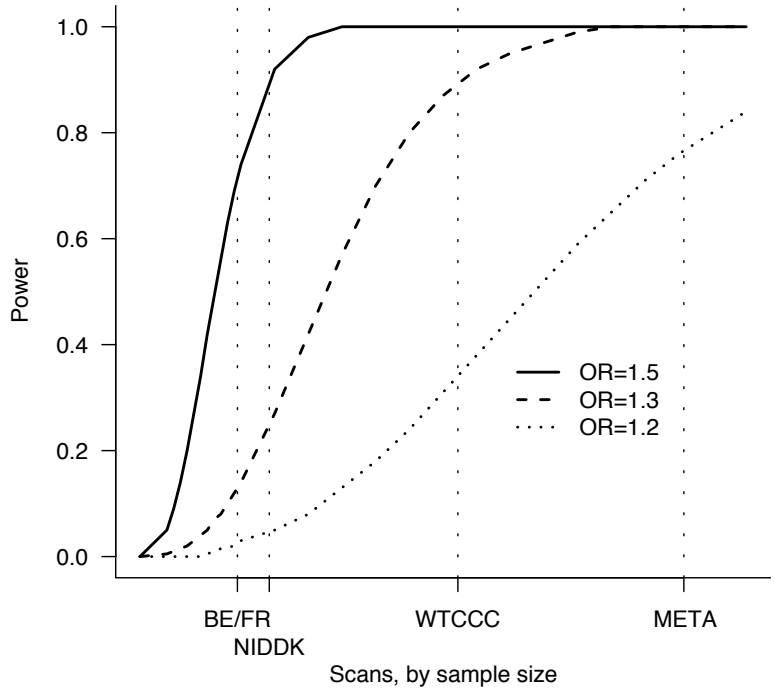


Figure 6.1: Power to detect a 20% allele with three different effect sizes (OR 1.2, 1.3, 1.5) under a multiplicative model versus study sample size. Power is reported here as the probability of $p < 5 \times 10^{-5}$ in a scan — the value used to define regions for attempting replication in a larger sample set. Vertical dotted lines show the sample sizes for the three constituent scans and the meta-analysis. Relatively large effects are likely to be detected by any of these scans, whereas only the combined analysis is well powered to detect more modest effects.

This study focused exclusively on the significant excess of highly associated SNPs in the tail of the distribution. Specifically, 526 SNPs from 74 distinct genomic loci were associated with $p < 5 \times 10^{-5}$ — more than seven times the

	NIDDK	BEL/FR	UKIBDGC	Total
Scan cases	946	536	1748	3230
Scan controls	977	914	2938	4829
Repl. cases	0	1170	1415	2585
Repl. controls	0	789	1054	1843
Repl. Trios	720	619	0	1346
Nationality	USA/Canadian	Belgian/French	British	
Scan Platform	Illumina HumanHap300	Illumina HumanHap300	Affymetrix 500	
Repl. Platform	Sequenom	Illumina GoldenGate	Sequenom	

Table 6.1: Samples used in this study

number of SNPs expected by chance even after correction for the modest overall inflation detected. While this threshold is arbitrary, it is important to remember that it is not meant to be representative of genome-wide significance. Rather, it is chosen with respect to the amount of statistical excess above this value (compared to the low overall inflation) and bearing in mind the available budget for the replication experiment. The GWAS data are used to generate hypotheses, it is the replication that will determine which loci are true associations.

All eleven associations previously replicated and established at genome-wide significance levels (Methods, Table 6.2), and described in Chapter 4, including both historical associations at *NOD2* and IBD5 as well as recent replicated findings from individual GWAS such as *IL23R*, *ATG16L1*, *IRGM*, *TNFSF15* and *PTPN22*, were among the 74 regions represented in this tail of the QQ distribution. Even after removing all SNPs in LD with these eleven loci, however, there continued to be a substantial excess of associated alleles genome-wide beyond that which would be expected by chance (Figure 6.2).

As for the other putative loci identified before this study (many of which were described in Chapter 4), the picture is unclear. One of those loci on chromosome 1q31 showed no evidence in the additional samples; upon further inspection of the replication genotype data from Chapter 4, it was discovered that the clusters had much higher than average variance, possibly giving rise to a false replication. None of the marginally replicated hits from Rioux *et al.*

(2007) showed signal here, nor did the postulated association at *DLG5*.

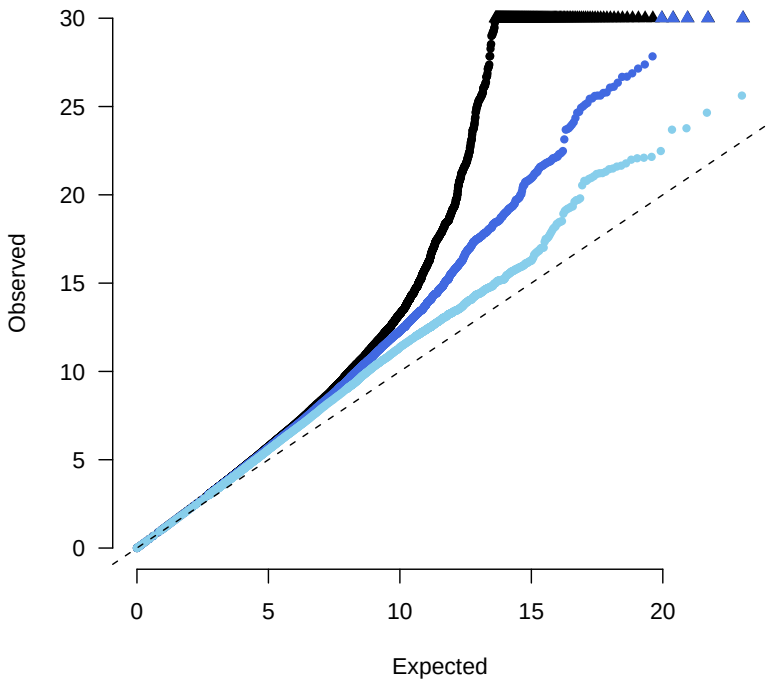


Figure 6.2: A quantile-quantile plot of observed χ^2 test statistics versus the expectation under the null. Black points represent the complete meta-analysis, with a substantial departure from the null at the tail (values > 30 are represented along the top of the plot as triangles). Dark blue points show the distribution after removing 11 previously published loci, demonstrating a still notable excess. Light blue points show the distribution after removing all 43 loci which replicate at least nominally. In all the cases the overall distribution is only marginally inflated ($\lambda_{GC} = 1.10$).

To identify the true risk factors from this set of 74 regions, a replication study was undertaken involving a total of 2585 additional Crohn's disease cases and 1843 controls alongside an independent family-based dataset of 1345 parent-parent-affected offspring trios. As these 74 regions included the 11 already reported as independently replicated and meeting genome-wide significance thresholds, this replication experiment effectively explored 63 putative associations in novel regions with 11 positive controls (Table B.3).

SNP	Chr	Critical region	p values			Num. genes	Gene of interest	RAF	Risk allele	Odds ratios	
			Scan	Replication	Combined					Case Ctrl	TDT
rs11465804	1p31	67.4*	4.56×10^{-35}	8.42×10^{-30}	8.64×10^{-63}	NA	*IL23R*	0.932	T	2.52	2.77
rs3828309	2q37	230.9*	8.58×10^{-21}	1.98×10^{-8}	2.94×10^{-27}	NA	*ATG16L1*	0.53	G	1.25	1.37
rs9858542	3p21	48.73 - 49.87	1.91×10^{-7}	0.01**	NA	35	*MST1*	0.567	A	1.17	NA
rs4613763	5p13	40.32 - 40.48	1.05×10^{-22}	8.37×10^{-9}	4.22×10^{-28}	0	*PTGER4****	0.125	C	1.35	1.28
rs2188962	5q23	131.44 - 131.90	8.86×10^{-9}	2.97×10^{-11}	2.76×10^{-17}	7		0.412	T	1.36	1.36
rs11747270	5q31	150.15 - 150.32	2.52×10^{-10}	1.27×10^{-7}	4.9×10^{-16}	3	*IRGM*	0.091	G	1.35	1.31
rs4263839	9q32	114.61 - 114.78	4.12×10^{-7}	0.000176	7.92×10^{-10}	2	*TNFSF15*	0.676	G	1.2	1.07
rs10995271	10q21	64.05 - 64.12	1.72×10^{-11}	3.03×10^{-8}	1.09×10^{-17}	1	*ZNF365*	0.393	C	1.26	1.53
rs11190140	10q24	101.26 - 101.32	1.19×10^{-10}	1.18×10^{-7}	1.51×10^{-16}	1	*NKX2-3*	0.477	T	1.2	1.28
rs2066847	16q12	49.3*	NA	3.72×10^{-24}	7.44×10^{-24}	NA	*NOD2*	0.018	C	4.01	2.57
rs2542151	18p11	12.73 - 12.88	2.96×10^{-11}	1.66×10^{-6}	9.14×10^{-16}	1	*PTPN2*	0.152	G	1.32	1.14
rs2476601	1p13	113.79 - 114.17	6.4×10^{-5}	5.69×10^{-5}	2.78×10^{-8}	7	*PTPN22*	0.899	G	1.33	1.17
rs9286879	1q24	169.54 - 169.67	3.84×10^{-7}	0.000504	2.52×10^{-9}	0		0.244	G	1.18	1.08
rs11584383	1q32	197.60 - 197.77	5.16×10^{-7}	0.00014	5.8×10^{-10}	3		0.674	T	1.2	1.21
rs10045431	5q33	158.69 - 158.76	1.81×10^{-8}	4.62×10^{-6}	7.98×10^{-13}	1	*IL12B*	0.705	C	1.11	1.36
rs6908425	6p22	20.63 - 20.84	2.4×10^{-7}	0.000209	6.24×10^{-10}	1	*CDKAL1*	0.78	C	1.21	1.09
rs7746082	6q21	106.52 - 106.62	1.33×10^{-5}	4.81×10^{-6}	5.5×10^{-10}	0		0.288	C	1.18	1.19
rs2301436	6q27	167.32 - 167.52	3.3×10^{-7}	3.24×10^{-8}	2.38×10^{-13}	3	*CCR6*	0.448	T	1.42	1.16
rs1456893	7p12	50.03 - 50.11	6.8×10^{-5}	2.99×10^{-5}	1.62×10^{-8}	0		0.679	A	1.18	1.14
rs921720	8q24	126.60 - 126.62	4.86×10^{-6}	2.17×10^{-5}	1.19×10^{-9}	0		0.606	G	1.17	1.29
rs7849191	9p24	4.94 - 5.26	2.48×10^{-5}	0.000714	1.43×10^{-7}	3	*JAK2*	0.602	C	1.12	1.12
rs17582416	10p11	35.30 - 35.60	8.34×10^{-6}	5.33×10^{-6}	4.32×10^{-10}	3		0.345	G	1.18	1.26
rs7927894	11q13	75.80 - 76.02	2.86×10^{-7}	0.000567	1.36×10^{-9}	1	*C11orf30*	0.376	T	1.23	NA
rs11175593	12q12	38.61 - 39.31	1.67×10^{-7}	6.52×10^{-5}	1.33×10^{-10}	3	*LRRK2, MUC19*	0.017	T	1.63	1.44
rs3764147	13q14	43.13 - 43.54	1.32×10^{-7}	1.93×10^{-7}	2.46×10^{-13}	3		0.221	G	1.25	1.19
rs2872507	17q21	34.63 - 35.34	3.14×10^{-6}	0.000221	5.44×10^{-9}	17	*ORMDL3*	0.473	A	1.13	1.24
rs744166	17q21	37.74 - 37.95	5.08×10^{-6}	1.1×10^{-7}	6.78×10^{-12}	4	*STAT3*	0.564	A	1.18	1.25
rs4807569	19p13	1.05 - 1.15	1.43×10^{-8}	0.000171	2.12×10^{-11}	2		0.209	C	1.16	1.31
rs1736135	21q21	15.73 - 15.76	3.08×10^{-5}	0.000149	3.48×10^{-8}	0		0.567	T	1.16	1.1
rs762421	21q22	44.43 - 44.48	1.28×10^{-5}	0.000382	3.56×10^{-8}	1	*ICOSLG*	0.388	G	1.09	1.21

Table 6.2: Convincingly replicated CD loci, with previously replicated associations above the horizontal line. RAF is risk allele frequency in control samples. Critical region is in NCBI B35 coordinates, definition is described in Methods. Risk alleles are defined relative to the + strand of the reference. * regions where causal variants have been convincingly mapped, rendering the LD window uninformative. ** genotyping for this SNP was problematic, replication data from Parkes *et al.* (2007) is shown, and additional genotyping is underway. ****PTGER4* is outside the critical region, but was implicated via eQTL analysis.

Results (significance levels and odds ratios) for strongly replicating loci, including all positive controls, are presented in Table 6.2. The distribution of Z-scores from the 63 novel regions shows a dramatic departure from the null distribution (Figure 6.3) with 19 novel regions showing significant replication ($p < 0.0008$ — Bonferroni corrected for tests of 63 potentially new associations). SNPs in the MHC (replication $p = 0.0025$, combined $p = 3.0 \times 10^{-9}$ - suspected but not previously conclusively established in Crohn's disease) did not reach this conservative threshold, but so convincingly satisfy proposed thresholds for genome-wide significance ($p < 5{\times}10^{-8}$) that it is proposed as the 20th additional Crohn's associated locus defined here. A further 12 of the 43 remaining loci showed nominal replication (Table 6.3).

It is possible that extreme population substructure in the replication sample could give rise to such a striking excess of hits. While unlikely, this was directly evaluated by the large family-based component of the replication study. Odds ratio estimates from the TDT analysis of the North American, French and Belgian families alone are consistent with those from the UK and Belgian case/control samples (Tables 6.2 and 6.3), with all 20 newly defined loci showing odds ratios in the same direction of association with the original scan in the family-based component (and half showing greater OR than in the case-control arm). Importantly, none of the significantly or nominally replicating loci show significant evidence for heterogeneity (across studies or between family-based and population-based arms) when corrected for the number of tests performed. This independent family based evidence confirms these alleles constitute true Crohn's disease loci.

For this newly expanded set of 31 unequivocally associated loci, a test was performed for evidence of significant pairwise interactions which could add further to the overall variance in liability explained by this set of loci. A case-only analysis of the 3931 cases in the replication study was performed and no interactions were observed that withstood a correction for the number of tests performed (Table B.4).

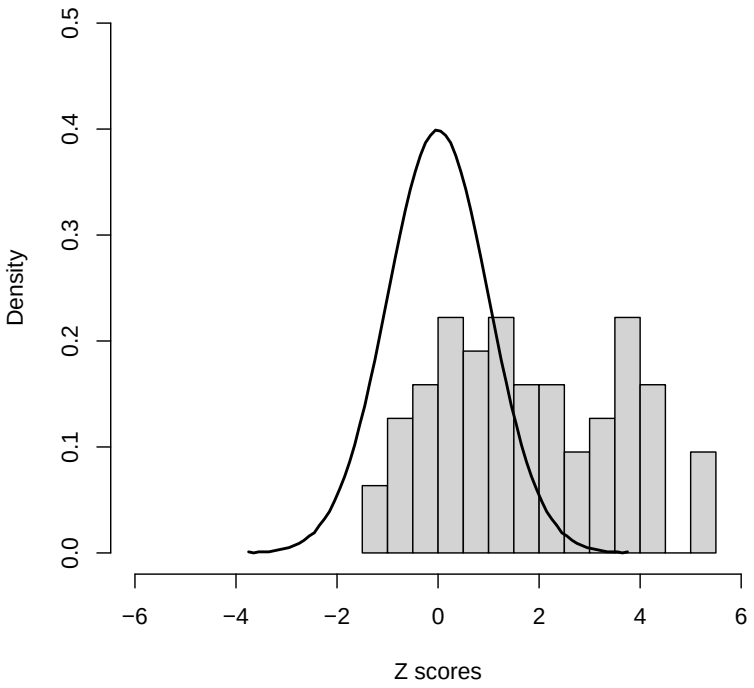


Figure 6.3: Distribution of observed Z scores from the 63 novel regions explored, along with the expected distribution under the null (a standard normal with mean 0 and variance 1). Even setting aside the 20 of the 63 that strongly replicate, the distribution is highly skewed — 7 more results exceed a Z of 2 (1 would be expected by chance under the null) whilst none showed a Z of less than -2 (same expectation under the null) suggesting that even more of the regions investigated here are likely to constitute true positive associations when additional data becomes available.

The contributions of the 31 loci to disease risk were computed using a standard liability threshold model(Lynch and Walsh, 1998) and are displayed as a histogram of individual variances (Figure 6.4). The observations from this variance analysis that many loci were detected for which the current study had low power, and that only a minority of the variance in risk is explained by these 31 loci, suggest that many additional loci are yet to be identified. This is reinforced by the additional 12 nominal replications (Table 6.3) where only 2 or 3 would be expected by chance, and by the continued excess in the Q-Q plot when these

43 total regions are removed (Figure 6.2).

While recognizing that fine-mapping is required to identify specific causal variants, several analyses were performed to gain some general insight into the CD associations. First the HapMap was queried to discover any instances where a non-synonymous SNP (nsSNP) was correlated ($r^2 > 0.5$) to the most associated variant discovered in this study. Accepting that HapMap is not a complete catalogue of nsSNPs, but including four loci where fine-mapping has identified coding variants, just 8 of the 31 genome-wide significant associations could be linked to a known nsSNP (Table 6.5). To explore whether any of the associations reflect a cis-acting regulatory effect on a nearby gene, genotype-expression correlation was evaluated using the panel of 400 lymphoblastoid cell lines described by Dixon *et al.* (2007). From all genes within 250 kb of the LD-based intervals defined in Tables 6.2 and 6.3, six correlations between gene expression and a nearby associated variant were identified (LOD > 2, Table 6.4). This was far in excess of chance ($p \sim 0.001$) and suggests that regulatory variation also contributes to the genetic architecture identified.

6.3 Discussion

Genome-wide association studies provide a systematic assessment of the contribution of common variation to disease pathogenesis. A limiting factor is often the size of the case-control dataset, and hence the power to detect any but the most strongly associated loci. Meta-analysis of existing data provides an obvious potential solution. As Figure 6.1 demonstrates, the expectation was that the additional power of the combined dataset would result in the identification of a substantially larger number of readily replicating associations than were derived from any of the smaller, constituent datasets. However, the paradigm of exploring common genetic variation with similar effects across studies (in this case all of European descent) needs testing before its results can be accepted as valid.

SNP	Chr	Critical region	← p values		→	Num. genes	Gene of interest	RAF	Risk allele	Odds ratios	
			Scan	Replication	Combined					Case Ctrl	TDT
rs3763313	6p21	29 – 32*	2.38×10^{-8}	0.00342	2.98×10^{-9}	MHC		0.187	C	1.21	1.01
rs13003464	2p16	61.09 – 61.14	1.31×10^{-5}	0.00255	2.24×10^{-7}	1	*CCDC139*	0.374	G	1.19	1.08
rs6128541	20q13	57.31 – 57.37	2.86×10^{-6}	0.0156	4x10-7	1	*EDN3*	0.735	T	1.15	1.08
rs991804	17q12	29.57 – 29.70	5.76×10^{-6}	0.0103	9.5×10^{-7}	4	*CCL2, CCL7*	0.727	C	1.11	1.08
rs11265519	1q23	157.65 – 157.72	1.5×10^{-6}	0.0221	9.64×10^{-7}	2	*ITLN1*	0.665	A	1.11	1.04
rs6743984	2q37	230.91 – 231.01	1.62×10^{-5}	0.011	1.11×10^{-6}	2	*SP110, SP140*	0.194	C	1.16	NA
rs12529198	6p25	5.04 – 5.11	2.62×10^{-6}	0.021	1.54×10^{-6}	1	*LYRM4*	0.061	G	1.12	1.19
rs780094	2p23	27.30 – 27.77	5.26×10^{-5}	0.00482	2.3×10^{-6}	22	*GCKR*	0.399	T	1.08	1.13
rs17309827	6p25	3.36 – 3.42	3.2×10^{-6}	0.048	5.08×10^{-6}	1	*SLC22A23*	0.639	T	1.09	1.02
rs7758080	6q25	149.54 – 149.65	8.88×10^{-6}	0.0299	5.7×10^{-6}	0		0.276	G	1.13	0.99
rs8098673	18q11	17.74 – 17.93	2.92×10^{-5}	0.0278	1.28×10^{-5}	0		0.329	C	1.06	1.09
rs11758386	6p22	21.51 – 21.59	2.08×10^{-5}	0.0439	1.36×10^{-5}	0		0.667	C	1.07	1.09
rs917997	2q11	102.31 – 102.64	2.76×10^{-5}	0.0445	2.3×10^{-5}	5	*IL18RAP*	0.222	T	1.05	1.11

Table 6.3: Nominally replicated CD risk loci. RAF is risk allele frequency in control samples. Critical region is in NCBI B35 coordinates, definition is described in Methods. Risk alleles are defined relative to the + strand of the reference. *this signal maps to the large stretch of strong LD in the human MHC.

SNP	Chr	Gene	LOD score	Distance SNP-gene
rs6743984	2	*SP140*	6.1	0
		SP110	2.7	24,641
rs2188962	5	*SLC22A5*	2.9	39,502
rs2858331	6	*HLA-DRB4*	4.2	0
rs17582416	10	*CREM*	4.3	128,151
rs3764147	13	*C13orf31*	4.2	0
rs2872507	17	*GSDML*	2.6	122,801
		ORMDL3	20.3	139,249

Table 6.4: cis-eQTL effects for CD associated SNPs

Chromosome	Position (Mb)	Gene(s)
1	67.4	*IL23R*
1	113.79 - 114.17	*PTPN22, DCLRE1B*
2	230.9	*ATG16L1*
3	48.73 - 49.87	*MST1*
5	131.44 - 131.90	*SLC22A4*
12	38.61 - 39.31	*MUC19*
13	43.13 - 43.54	*C13orf31*
16	49.3	*NOD2*
17	34.63 - 35.34	*GSDML*

Table 6.5: Crohn's risk loci containing a highly correlated nsSNP

On the validity of the method the results were substantially reassuring. All 11 previously confirmed CD susceptibility loci were strongly replicated both in the meta-analysis and follow-up experiment. These include the two widely replicated findings from studies published in 2001 (Hugot *et al.*, 2001; Ogura *et al.*, 2001; Rioux *et al.*, 2001) as well as all of the compelling findings from individual GWAS (Table 6.2, upper section). Significantly, 20 new CD susceptibility loci have also been identified and replicated. Using a conservative threshold for significance (only 1 such region would be expected by chance in 20 such experiments), the loci with clear evidence for association in the replication panel include a very high proportion of those showing strongest signals in the meta-analysis (Table B.3) — 9 of 9 previously unreported regions with $p < 5 \times 10^{-7}$ in the combined scan were replicated convincingly — emphasising the validity of the meta-analysis results. Further emphasising the robustness of these results, 19 of these 20 loci exceed a conservative genome-wide level of significance ($p < 5 \times 10^{-8}$) — most by a significant margin (15 have $p < 5 \times 10^{-9}$)

- and equivalent strength of association was observed in the family-based subset of the replication sample.

In keeping with other regions recently identified as associated with CD, the 20 new loci do not conform to any obvious pattern in terms of gene content. Thus, as shown in Table 6.2, some loci (defined by HapMap recombination hotspots flanking the set of correlated, associated variants) contain just a single gene, some contain many genes and others none. Clearly the first category provides the most immediate clues regarding pathogenic mechanisms. Included among these are compelling functional candidates such as *STAT3*, *JAK2* and *IL12B* while others, such as *CDKAL1* and *PTPN22*, highlight potentially intriguing overlaps between the genetic architecture of Crohn's disease and other complex disorders. It is noteworthy — and consistent with previous findings from CD and other complex diseases — that no strong evidence of deviation from the model of multiplicative (random) effects was found when testing for gene-gene interactions among the 31 confirmed associations.

For loci containing multiple genes or no genes the picture is less well defined. The identified paucity of correlation between associated SNPs and coding variation suggests that these loci may, in particular, benefit from expression quantitative trait locus (eQTL) analysis. This seeks correlation between genotype and expression patterns — bearing in mind that such functional relationships need not respect the specific boundaries of LD around the association. An eQTL effect incriminating *PTGER4* at the 5p13 locus was previously reported by Libioulle *et al.* (2007). A striking outcome from the present analysis was at the established IBD5 locus 16, where CD-associated SNPs were associated with decreased *SLC22A5* mRNA expression levels. While Peltekova *et al.* (2004) had previously proposed a SNP as regulating *SLC22A5* transcriptional activity, this data suggests for the first time that the most associated variation in the IBD5 region, including a coding variant in neighboring *SLC22A4*, are the same variants that are most significantly associated to *SLC22A5* expression. Equally striking, the most significant Crohn's associated eQTL reported here affects

ORMDL3 (LOD = 20) on chromosome 17 and SNPs in precisely the same region were recently shown to be strongly associated with childhood asthma (Moffatt *et al.*, 2007). This suggests that the same polymorphisms might underlie susceptibility to both CD and asthma, possibly by perturbing *ORMDL3* expression.

The new loci that have been identified are of modest effect size, which is unsurprising given all loci with larger impact on disease risk were — as would be expected — discovered in the original scans. The small sizes of these effects explains the lack of overlap between linkage results in CD and these newly discovered loci, with the possible exceptions of combined effects of multiple high ranking associations on chromosomes 5q and 6p. Indeed, the linkage evidence that led to the discovery of the 5q31 locus was very likely boosted by the presence of other nearby loci. As expected, the only gene conclusively discovered via linkage (*NOD2*) is one of two loci which stand well out from the remainder of the distribution of effect sizes (Figure 6.4). The other outlier, *IL23R*, illustrates an interesting characteristic of linkage — because (unlike *NOD2*) the most penetrant risk allele has very high frequency (93%), it is nearly invisible to linkage analysis despite the high OR; highly protective rare alleles are simply not present in multiplex affected families and thus do not influence allele sharing substantially.

Using a liability-threshold model, the 31 loci identified to date explain about 10% of the overall variance in disease risk, which may be as much as a fifth of the genetic risk, given previous estimates of CD heritability of approximately 50% (Tysk *et al.*, 1988). It should be noted, however, that the full impact of the new loci cannot be determined until causal variants have been identified by directed sequencing and fine-mapping experiments. Until then the proportion of the variance in Crohn's disease risk explained must be measured from the confirmed SNPs, where association is due to LD with causal variants. Since multiple causal variants might exist at each locus (ranging in frequency from rare to common) these estimates of variance explained provide only a lower

bound for the true contribution of each locus.

Since this study is well-powered to identify loci that explain $> 0.2\%$ of the overall variance, but the sum of such loci explains a relatively small fraction of the total, it seems likely that many loci with even more modest effect sizes remain undiscovered. Of particular note is the continued excess of strong associations outside of the confirmed regions, as well as the nominal replication of an additional 12 loci, notably in excess of chance. Overall, the distribution of Z scores in the replication experiment is clearly skewed towards replication — only 11 of the 63 Z-scores in this replication experiment generate $Z< 0$. If only the 20 strongly confirmed loci were genuinely associated, half of the 43 remaining should end up with $Z< 0$. Indeed, observing 12 of the 43 remaining tests with $Z> 1.5$ is itself a highly significant observation ($p < 0.0001$). Although modest in terms of effect size, identification of such loci is likely to still provide important insights into pathogenic mechanisms, as biological importance need not be proportional to the statistical evidence for genetic association.

As was seen in Chapter 2 the generation of GWAS arrays used in the scans here did not offer complete genome coverage of common variation (additional loci may reside in poorly covered intervals) and did not address copy number variation effectively. Thus in spite of the wealth of new susceptibility genes and loci identified by the current study, it seems implausible that there are not more to be found — albeit very large datasets are likely to be required to achieve robust statistical support for them. With respect to the present findings, there is much work to be done in resequencing and fine mapping to identify causal variants. While there is not yet a complete understanding of the genetic architecture of Crohn's disease, dramatic progress has now been made towards this goal — and with it the prospect of directed functional exploration of the pathways identified, insight into how risk alleles interact with environmental modifiers, and the hope of new avenues for treatment.

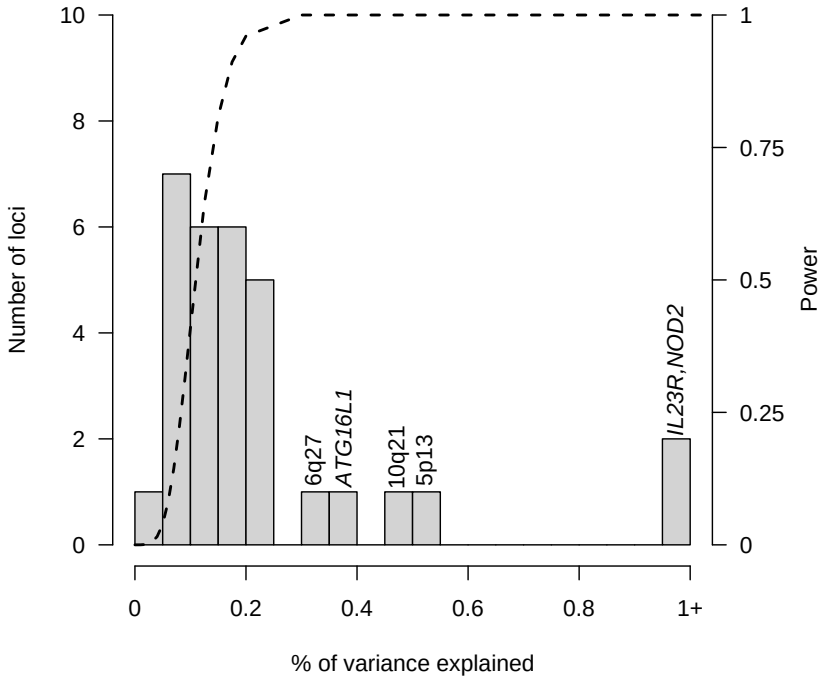


Figure 6.4: Histogram of percent variance explained by each of the 31 established CD risk loci. The distribution resembles the long postulated exponential distribution of effect sizes. Dashed line shows the joint power for our meta-analysis to detect ($p < 5 \times 10^{-5}$), and for our replication sample to replicate (at Bonferroni corrected p values), a 20% variant explaining a given fraction of variance. Note how quickly this curve moves from nearly zero power to detect tiny effects (less than one tenth of one percent variance explained) to nearly full power to detect larger effects (presuming they are well covered by the current generation of GWAS chips). Complete power near the origin would likely reveal a more complete exponential distribution, with many very small effects. *NOD2* and *IL23R* are distant outliers, each explaining 1-2% of total variance, partially because multiple causal variants have been discovered at these loci. Interestingly, two of the four next largest effects (5p13 and 10q21) are in gene deserts, highlighting the importance of non-coding variation in disease risk.

6.4 Methods

6.4.1 Crohn's disease patients and controls

The meta-analysis was based on data from the 3 genome-wide scans of the NIDDK (Rioux *et al.*, 2007), WTCCC (Wellcome Trust Case Control Consortium, 2007a) and Belgian/French (Libioulle *et al.*, 2007) studies. Details of the numbers of cases and controls genotyped in the respective scans and of the genotyping platforms used are shown in Table 6.1, as are case/control and family cohorts genotyped in the replication study of the meta-analysis. Details of the ascertainment and characterization of these cohorts were provided in the original scan and replication publications (Parkes *et al.*, 2007; Rioux *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007a; Duerr *et al.*, 2006; Libioulle *et al.*, 2007). Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants.

6.4.2 Imputation

Briefly, these methods rely on observed haplotype patterns in a set of reference data (the HapMap) and the actual genotype data from each project to make predictions (along with a measure of statistical certainty) at un-genotyped SNPs. The program MACH (G. Abecasis, personal communication) was used with the NIDDK and Belgian/French data, and IMPUTE (Marchini *et al.*, 2007) with the WTCCC data. Comparisons between the two algorithms yielded very similar results (data not shown). The superset of polymorphic markers which passed QC in the original scans was imputed. This set was comprised of SNPs on either the Affymetrix 500K only (n = 350,507), Illumina HumanHap300 version 1 only (n = 238,935), or both panels (n = 46,105) such that all association tests performed were partially based on observed genotype data.

6.4.3 Test for association and effect size estimation

Using the genotype probabilities (rather than best-guess genotypes) for imputed markers in the case and control tallies,the standard 1 d.f. allele-based test of association was summarised as a Z-score within each scan and combined scores across studies to produce a single meta-statistic for each SNP for all three datasets. Odds ratios were estimated separately in TDT samples and each case/control collection, and then combined and tested for heterogeneity (Kazeem and Farrall, 2005).

6.4.4 Critical regions

Given that most associations contain many correlated SNPs showing signal, independent loci were demarcated by first defining the set of HapMap SNPs with $r^2 > 0.5$ to the most significantly associated SNP. The critical region was then bound by the flanking HapMap recombination hotspots which contained this set. These windows very likely contain the causal polymorphisms explaining the associations.

6.4.5 Replication

Loci were defined to have been previously confirmed if an earlier study had both detected and replicated the association in independent samples and the association achieved $p < 5 \times 10^{-8}$ (recently proposed as an appropriate genome-wide significance level for GWAS). For replication genotyping, the most significantly associated SNP was selected from each region along with a second, correlated SNP with $p < 0.0001$ or a second assay on the opposite strand in order to have a technical backup should the first fail genotyping (Table B.3). Replication genotyping for all of the putatively associated loci was performed using primer extension chemistry and mass spectrometric analysis (iPLEX, Sequenom) using Sequenom Genetics Services (N. American panel) and Genome Research Limited, Wellcome Trust Sanger Institute (UK panel), and using a

custom-made Golden Gate assay on a Beadstation500 (Illumina), following the manufacturer's recommendations (Belgian/French panel). The more completely genotyped SNP of the two from each region was chosen to represent that regional association in analysis (if both were completely typed, the SNP that was more strongly associated in the scan was used).

6.4.6 Regional Annotation: eQTL analysis

Effects of SNPs in Table 2 on expression levels of neighbouring genes were studied using transcriptome data from the $\sim$ 400 lymphoblastoid cell lines described by Dixon *et al.* (2007). SNPs that were not genotyped on this panel (n=14) were replaced with a proxy with $r^2 > 0.95$ when possible (n=12). LOD scores > 2 for genes (probe average) located within 250 Kb of the corresponding LD windows were retrieved from http://www.sph.umich.edu/csg/liang/asthma/. To evaluate the significance of the findings with the CD associated SNPs, the observed (i) number of genes yielding LOD scores > 2, and (ii) sum of these LOD scores, were compared with the corresponding frequency distributions for 1000 randomly selected sets of 31 SNPs, matched for allele frequency (± 0.02) and gene context. Window sizes determined for associated SNPs were used as is for the matched simulated SNPs.

Chapter 7

Discussion

7.1 Summary

In this thesis I have presented results relating to several stages in the development of GWAS in the past three years, including marker selection and design, basic analysis of one study, replication of GWAS hits, and meta-analysis of multiple scans.

In the early stages of planning for the first generation of GWAS (all of which are now complete), researchers were faced with three major impediments to the discovery of genes influencing common human diseases: (i) small sample sizes with insufficient power, (ii) incomplete catalogues of genetic variation and (iii) failure to identify strong candidate genes. They hoped to overcome these obstacles by heeding past failures and carefully designing these studies with those lessons in mind. To this end large sample sizes numbering in the thousands, not hundreds, were collected for a number of common diseases and the SNP Consortium and the HapMap led SNP discovery efforts.

The final hurdle, then, was moving from candidate gene or region approaches to a hypothesis-free genome-wide approach. Advances in molecular biotechnology, in the form of genome-wide SNP chips enabled this possibility. The question remained, however: how much of the common variation in the genome would

these new products capture? I have shown that most of the commercial products available when these studies were being designed capture approximately two-thirds of the common variation in European populations. Subsequent technological development, and the introduction of million-SNP chips have increased these coverage levels in Europeans even further, and expanded to provide good coverage in East Asian populations and reasonable coverage in African populations (Li *et al.*, 2008). I also demonstrated that more targeted approaches, such as trying to capture all known nsSNPs, are not well served by typical genome-wide chips, but must instead be individually genotyped (an approach taken by a number of genotype products).

Once these issues of sample size and SNP density had been addressed, work on GWAS began in earnest. In Chapter 3 I presented my work on one of the largest of these studies, the WTCCC. One of the principal challenges confronted by this project was the laborious task of transforming an enormous set of raw intensities into a clean, usable dataset. This process led to a number of unusual observations, including lab-specific effects on genotyping and the discovery of a large number of non-European samples (contrary to the expectations of the project), in addition to investigating the standard measures of quality in genetic datasets.

Cleaning the WTCCC of these artefacts allowed the discovery of nearly two dozen *bona fide* associations. These results reinforced the idea that risk loci for common diseases have extremely modest ORs, and require concomitantly large sample sizes to detect. Additionally, several loci were identified which contributed to multiple diseases, including *PTPN2* for both CD and T1D. Such genetic connections between phenotypes have the potential to redefine our understanding of these diseases.

As the field began designing and executing these early GWAS it became very clear (Chanock *et al.*, 2007) that replication of GWAS signals in independent samples was critical to avoid the confusing state of affairs of the past decade, where many putative associations were heralded in the literature only to be

later determined to be false positives. Interestingly, the widespread concern regarding sensible thresholds for genome-wide significance, and corrections for multiple tests, did not turn out to be a large problem. True associations consistently exceed conservative thresholds, provided they are studied in a large enough sample, so the focus has shifted from defining the thresholds to increasing sample size and performing adequate replication experiments. I describe the CD replication experiment for the WTCCC in Chapter 4, where six previously published loci and four completely new loci were unambiguously confirmed. These results revealed new insight into the pathogenesis of CD, and hinted at the possibility that still more loci awaited discovery with increased sample sizes.

Chapter 5 addressed some of the problems facing researchers who wish to combine multiple GWAS datasets. I describe how the tedious process of data annotation is crucial to being able to sensibly merge disparate datasets. Distributing information about SNPs and samples, including raw intensity data is essential to enabling combined studies and meta-analyses. I next discussed the problem of population structure between UK and US samples and concluded that the problem is tractable so long as caution is used in both analysing and interpreting the data. I showed that imputation methods can be successfully used to combine datasets from different genotyping platforms, but again, that great care is needed to avoid spurious results. In all these cases the extent of the problem strongly depends on the particular way in which data are combined for analysis, the ratio of cases and controls in each dataset, and the degree to which any source of bias is correlated with disease status. This observation underscores the idea that there is unlikely to be any fixed recipe for carrying out such studies.

I describe one such meta-analysis, for CD, in Chapter 6. This project brought together data from three CD GWAS (UK, North America and Belgian/French), and attempted replication in a multi-national collection of both unrelated cases and controls, and families. The study confirmed the eleven loci previously demonstrated to be associated at a genome-wide significant level,

and discovered twenty further loci. Beyond these conclusively associated loci were an additional eleven showing nominal replication, of which at least some are expected to be validated once they have been studied in additional samples. In addition to these dividends, the study served as a proof of principle that a large number of risk loci can be found if enough samples are analysed. This is especially exciting, because there is no reason to believe that this conclusion does not also apply to other phenotypes.

7.2 Implications for the future

The first three years of GWAS have undoubtedly transformed the landscape of human disease genetics. Specialists in many diseases (not least Crohn's) have lurched from being confined to studying a handful of risk genes to being overwhelmed by the abundance of newly discovered loci. But this very fact that so many scientists have laboured for so very long *without* many verified genetic loci clearly illustrates just how much work is left to be done after the associations have been found. What, then, is next?

7.2.1 Identifying causal variants

One task resulting from these new discoveries, which is already underway, is to narrow down these broad regions of association and to identify causal variants. For some regions with only one or two transcripts, this might be relatively straightforwardly accomplished by resequencing exons. But this will not work in all cases, as was seen in Chapter 4, where no coding variation was found which could explain the association, despite deep exon resequencing in *IRGM*. Regions with many genes increase the cost and complexity of this problem.

Of course, regions without any known transcripts are even more challenging. These associations are in one sense the most interesting of all, because they challenge the fundamental ideas about what types of variation underlie disease risk. Obviously a whole spectrum of variation beyond the coding sequence is

affecting disease: by splicing, transcriptional regulation, or mechanisms not yet understood. Furthermore, it is becoming increasingly clear that copy-number variation plays an important role in human diversity, from widespread germline variation to *de novo* mutations affecting phenotypes such as autism (Weiss *et al.*, 2008).

Even if coding or regulatory variants are found which explain most of the signal discovered via GWAS, it does not preclude the existence of other functional variants. It seems likely that any gene with one or more common risk variants will also have an assortment of very rare mutations which also predispose carriers to disease, but are not detectable in GWAS. A complete understanding of this allelic spectrum will be necessary to estimate disease risk and to evaluate the predictive potential of genetics.

It is important, therefore, to begin broad resequencing and fine-mapping of these regions in order to try to pinpoint the most associated variants, even if their function is not yet known. Many such experiments are currently underway (enabled in part by new high-throughput sequencing technologies) and the early results will prove invaluable in pursuing causal variants in the dozens of novel associations being discovered.

Identifying causal variants will also enable a number of other genetic analyses. To date, almost no replicable examples of gene-gene or gene-environment interaction have been identified in common diseases. One possible explanation for this is that power to detect epistasis (whether with other genes or environmental factors) is substantially diminished when testing LD proxies compared with testing the causal variant. A full catalogue of functional variants might, therefore, unlock interactions within the set of known associations or with loci with undetectable marginal effects on their own.

7.2.2 Biological translation

Another critical undertaking will be the translation of these associations into biological observations, either *in vivo* or *in vitro*. Gene expression experiments,

animal models and small molecule screens of cells will help to understand the specific biological function of these SNPs. Gene expression arrays have only recently been applied to large numbers of samples (such as the Dixon *et al.* (2007) data described in Chapter 6), and are most often used with lymphoblastoid cell lines. This approach must be expanded to different tissue types and states of activation to understand transcriptional regulation more completely (Yamanouchi *et al.*, 2007).

The eventual hope is that identifying first the specific genomic features (genes, regulatory sequences, etc.), then the functional variants, and finally the biological mechanisms, will allow medical interventions through prevention, pharmaceuticals or better understanding of environmental risk factors. While it is unlikely that these modest risk loci will ever explain enough of the variance in disease risk to be explanatory, it is possible that a large panel of functional variants might be of diagnostic as a supplement to existing clinical practice. GWAS results have opened a door to a long and promising set of possibilities.

7.2.3 Mining GWAS

GWAS data has accumulated at an incredible pace, and coupled with widespread policies of public data release (see Chapter 5), the genetics community is blessed with un unprecedented volume of data available at no cost. One potential opportunity arising from these data is the joint analysis of multiple phenotypes. Chapter 6 already showed several promising overlaps among related (CD, T1D) and unrelated (CD, T2D) phenotypes. Some of these loci could be found via traditional meta-analysis, as the effects appear to be at the same variants, and in the same direction. Others, however, have inverse effects, or are associated with entirely distinct variants. This latter class presents an open question for methodological development to search for such regions across many phenotypes in a statistically robust way.

These data also offer opportunities in research areas unrelated to disease genetics. With the recent publication of genome-wide data for the samples from

the Human Genome Diversity Project by Jakobsson *et al.* (2008), the variants associated with many common diseases can be evaluated from the perspective of population genetics. Regional variation of allele frequencies at disease-predisposing variants could provide clues to selection and human population demography.

7.3 Conclusion

By the end of this decade medical research will have undergone a fundamental change. GWAS have transformed our knowledge of the underlying genetic pathology of many common human diseases. Many genetic variants found at high frequency in the general population have been indisputably shown to have small effects on disease risk. These findings have cracked open the black box of disease pathology, revealing a disassembled assortment of parts. New insights, including the possibilities described above, will be required to connect these pieces and fully comprehend biology's inner workings.

Appendix A

WTCCC authors

Management Committee: Paul R Burton[1], David G Clayton[2], Lon R Cardon[3], Nick Craddock[4], Panos Deloukas[5], Audrey Duncanson[6], Dominic P Kwiatkowski[3,5], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Nilesh J Samani[10], John A Todd[2], Peter Donnelly (Chair)[11]

Analysis Committee: Jeffrey C Barrett[3], Paul R Burton[1], Dan Davison[11], Peter Donnelly[11], Doug Easton[12], David Evans[3], Hin-Tak Leung[2], Jonathan L Marchini[11], Andrew P Morris[3], Chris CA Spencer[11], Martin D Tobin[1], Lon R Cardon (Co-chair)[3], David G Clayton (Co-chair)[2]

UK Blood Services & University of Cambridge Controls: Antony P Attwood[5,8], James P Boorman[8,9], Barbara Cant[8], Ursula Everson[13], Judith M Hussey[14], Jennifer D Jolley[8], Alexandra S Knight[8], Kerstin Koch[8], Elizabeth Meech[15], Sarah Nutland[2], Christopher V Prowse[16], Helen E Stevens[2], Niall C Taylor[8], Graham R Walters[17], Neil M Walker[2], Nicholas A Watkins[8,9], Thilo Winzer[8], John A Todd[2], Willem H Ouwehand[8,9]

1958 Birth Cohort Controls: Richard W Jones[18], Wendy L McArdle[18], Susan M Ring[18], David P Strachan[19], Marcus Pembrey[18,20]

Bipolar Disorder (Aberdeen): Gerome Breen[21], David St Clair[21]; **(Birmingham):** Sian Caesar[22], Katherine Gordon-Smith[22,23], Lisa Jones[22]; **(Cardiff):** Christine Fraser[23], Elaine K Green[23], Detelina Grozeva[23], Marian L Hamshere[23],

Peter A Holmans[23], Ian R Jones[23], George Kirov[23], Valentina Moskvina[23], Ivan Nikolov[23], Michael C O'Donovan[23], Michael J Owen[23], Nick Craddock[23]; **(London):** David A Collier[24], Amanda Elkin[24], Anne Farmer[24], Richard Williamson[24], Peter McGuffin[24]; **(Newcastle):** Allan H Young[25], I Nicol Ferrier[25]

Coronary Artery Disease (Leeds): Stephen G Ball[26], Anthony J Balmforth[26], Jennifer H Barrett[26], D Timothy Bishop[26], Mark M Iles[26], Azhar Maqbool[26], Nadira Yuldasheva[26], Alistair S Hall[26]; **(Leicester):** Peter S Braund[10], Paul R Burton[1], Richard J Dixon[10], Massimo Mangino[10], Suzanne Stevens[10], Martin D Tobin[1], John R Thompson[1], Nilesh J Samani[10]

Crohn Disease (Cambridge): Francesca Bredin[27], Mark Tremelling[27], Miles Parkes[27]; **(Edinburgh):** Hazel Drummond[28], Charles W Lees[28], Elaine R Nimmo[28], Jack Satsangi[28]; **(London):** Sheila A Fisher[29], Alastair Forbes[30], Cathryn M Lewis[29], Clive M Onnie[29], Natalie J Prescott[29], Jeremy Sanderson[31], Christopher G Mathew[29]; **(Newcastle):** Jamie Barbour[32], M Khalid Mohiuddin[32], Catherine E Todhunter[32], John C Mansfield[32]; **(Oxford):** Tariq Ahmad[33], Fraser R Cummings[33], Derek P Jewell[33]

Hypertension (Aberdeen): John Webster[34]; **(Cambridge):** Morris J Brown[35], David G Clayton[2]; **(Evry, France):** G Mark Lathrop[36]; **(Glasgow):** John Connell[37], Anna Dominiczak[37]; **(Leicester):** Nilesh J Samani[10]; **(London):** Carolina A Braga Marcano[38], Beverley Burke[38], Richard Dobson[38], Johannie Gungadoo[38], Kate L Lee[38], Patricia B Munroe[38], Stephen J Newhouse[38], Abiodun Onipinla[38], Chris Wallace[38], Mingzhan Xue[38], Mark Caulfield[38]; **(Oxford):** Martin Farrall[39]

Rheumatoid Arthritis: Anne Barton[40], The Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee*, Ian N Bruce[40], Hannah Donovan[40], Steve Eyre[40], Paul D Gilbert[40], Samantha L Hider[40], Anne M Hinks[40], Sally L John[40], Catherine Potter[40], Alan J Silman[40], Deborah PM Symmons[40], Wendy Thomson[40], Jane Worthington[40]

Type 1 Diabetes: David G Clayton[2], David B Dunger[2,41], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], Barry Widmer[2,41], John A Todd[2]

Type 2 Diabetes (Exeter): Timothy M Frayling[42,43], Rachel M Freathy[42,43], Hana Lango[42,43], John R B Perry[42,43], Beverley M Shields[43], Michael N Weedon[42,43], Andrew T Hattersley[42,43]; **(London):** Graham A Hitman[44]; **(Newcastle):** Mark Walker[45]; **(Oxford):** Kate S Elliott[3,7], Christopher J Groves[7], Cecilia M Lindgren[3,7], Nigel W Rayner[3,7], Nicholas J Timpson[3,46], Eleftheria Zeggini[3,7], Mark I McCarthy[3,7]

Tuberculosis (Gambia): Melanie Newport[47], Giorgio Sirugo[47]; **(Oxford):** Emily Lyons[3], Fredrik Vannberg[3], Adrian VS Hill[3]

Ankylosing Spondylitis: Linda A Bradbury[48], Claire Farrar[49], Jennifer J Pointon[48], Paul Wordsworth[49], Matthew A Brown[48,49]

AutoImmune Thyroid Disease: Jayne A Franklyn[50], Joanne M Heward[50], Matthew J Simmonds[50], Stephen CL Gough[50]

Breast Cancer: Sheila Seal[51], Breast Cancer Susceptibility Collaboration (UK), Michael R Stratton[51,52], Nazneen Rahman[51]

Multiple Sclerosis: Maria Ban[53], An Goris[53], Stephen J Sawcer[53], Alastair Compston[53] **Gambian Controls (Gambia):** David Conway[47], Muminatou Jallow[47], Melanie Newport[47], Giorgio Sirugo[47]; **(Oxford):** Kirk A Rockett[3], Dominic P Kwiatkowski[3,5]

DNA, Genotyping, Data QC and Informatics (Wellcome Trust Sanger Institute, Hinxton): Claire Bryan[5], Suzannah J Bumpstead[5], Amy Chaney[5], Kate Downes[2,5], Jilur Ghori[5], Rhian Gwilliam[5], Sarah E Hunt[5], Michael Inouye[5], Andrew Keniry[5], Emma King[5], Ralph McGinnis[5], Simon Potter[5], Rathi Ravindrarajah[5], Pamela Whittaker[5], David Withers[5], Panos Deloukas[5]; **(Cambridge):** Hin-Tak Leung[2], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], John A Todd[2]

Statistics (Cambridge): Doug Easton[12], David G Clayton[2]; **(Leicester):** Paul R Burton[1], Martin D Tobin[1]; **(Oxford):** Jeffrey C Barrett[3], David Evans[3], Andrew P Morris[3], Lon R Cardon[3]; **(Oxford):** Niall J Cardin[11], Dan Davison[11], Teresa Ferreira[11], Joanne Pereira-Gale[11], Ingeleif B Hallgrimsdottir[11], Bryan N Howie[11], Jonathan L Marchini[11], Chris CA Spencer[11], Zhan Su[11], Yik Ying Teo[3,11], Damjan Vukcevic[11], Peter Donnelly[11]

PIs: David Bentley[5,54], Matthew A Brown[48,49], Lon R Cardon[3], Mark Caulfield[38], David G Clayton[2], Alistair Compston[53], Nick Craddock[23], Panos Deloukas[5], Peter Donnelly[11], Martin Farrall[39], Stephen CL Gough[50], Alistair S Hall[26], Andrew T Hattersley[42,43], Adrian VS Hill[3], Dominic P Kwiatkowski[3,5], Christopher G Mathew[29], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Miles Parkes[27], Marcus Pembrey[18,20], Nazneen Rahman[51], Nilesh J Samani[10], Michael R Stratton[51,52], John A Todd[2], Jane Worthington[40]

Affiliations: [1]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester, LE1 7RH, UK; [2]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge, CB2 0XY, UK; [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK; [4]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK; [5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [6]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK; [7]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford, OX3 7LJ, UK; [8]Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 2PT, UK; [9]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge, CB2 2PT, UK; [10]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester, LE3 9QP, UK; [11]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK; [12]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK; [13]National Health Service Blood and Transplant, Sheffield Centre, Longley Lane, Sheffield S5 7JN, UK; [14]National Health Service Blood and Transplant, Brentwood Centre, Crescent Drive, Brentwood, CM15 8DP, UK; [15]The Welsh Blood Service, Ely Valley Road, Talbot Green, Pontyclun, CF72 9WB, UK; [16]The Scottish National Blood Transfusion Service, Ellen's Glen Road, Edinburgh, EH17 7QT, UK; [17]National Health Service Blood and Transplant, Southampton Centre, Coxford Road, Southampton, SO16 5AF, UK; [18]Avon Longitudinal Study of Par-

ents and Children, University of Bristol, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK; [19]Division of Community Health Services, St George's University of London, Cranmer Terrace, London SW17 0RE, UK; [20]Institute of Child Health, University College London, 30 Guilford St, London WC1N 1EH, UK; [21]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen, AB25 2ZD, UK; [22]Department of Psychiatry, Division of Neuroscience, Birmingham University, Birmingham, B15 2QZ, UK; [23]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK; [24]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park Denmark Hill London SE5 8AF, UK; [25]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK; [26]LIGHT and LIMM Research Institutes, Faculty of Medicine and Health, University of Leeds, Leeds, LS1 3EX, UK; [27]IBD Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge, CB2 2QQ, UK; [28]Gastrointestinal Unit, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU UK; [29]Department of Medical & Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London, SE1 9RT, UK; [30]Institute for Digestive Diseases, University College London Hospitals Trust, London, NW1 2BU, UK; [31]Department of Gastroenterology, Guy's and St Thomas' NHS Foundation Trust, London, SE1 7EH, UK; [32]Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne, NE1 4LP, UK; [33]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK; [34]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK; [35]Clinical Pharmacology Unit and the Diabetes and Inflammation Laboratory, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK; [36]Centre National de Genotypage, [2], Rue Gaston Cremieux, Evry, Paris 91057.; [37]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, G12 8TA, UK; [38]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London, Queen Mary's School of Medicine, Charterhouse Square, London EC1M 6BQ, UK; [39]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK; 40arc Epidemiology Research Unit, University of Manchester,

Stopford Building, Oxford Rd, Manchester, M13 9PT, UK; [41]Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK; [42]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter EX1 2LU UK; [43]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, Barrack Road, Exeter EX2 5DU UK; [44]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB UK; [45]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK; [46]The MRC Centre for Causal Analyses in Translational Epidemiology, Bristol University, Canynge Hall, Whiteladies Rd, Bristol BS2 8PR, UK; [47]MRC Laboratories, Fajara, The Gambia; [48]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Woolloongabba, Qld [4102], Australia; [49]Botnar Research Centre, University of Oxford, Headington, Oxford OX3 7BN, UK; [50]Department of Medicine, Division of Medical Sciences, Institute of Biomedical Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK; [51]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, SM2 5NG, UK; [52]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [53]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK; [54]PRESENT ADDRESS: Illumina Cambridge, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex, CB10 1XL, UK.

Appendix B

Supplementary SNP genotyping information

SNP	**Chr**	**Pos(Mb)**	**WTCCC p**	**WTCD**	**RCD**	**WTU1**	**WTU2**
rs2484676	1	50.6	0.0001882	0.512	0.494	0.472	0.485
rs11205760	1	50.9	1.67×10^{-5}	0.307	0.317	0.351	0.338
rs4506508	1	77	0.0001148	0.204	0.190	0.172	0.181
rs2814036	1	164	0.0001503	0.024	0.023	0.014	0.016
rs17419032	1	197.7	7.50×10^{-5}	0.268	0.273	0.306	0.297
rs363617	2	74.8	0.0002298	0.15	0.141	0.123	0.139
rs2322659	2	136.4	0.001851	0.221	0.222	0.194	0.191
rs9870678	3	57.5	3.67×10^{-5}	0.417	0.393	0.374	0.389
rs1462651	3	149.2	8.42×10^{-5}	0.136	0.194	0.167	0.154
rs9993022	4	59.8	0.0001554	0.029	0.024	0.017	0.02
rs1363670	5	158.7	3.19×10^{-5}	0.138	0.159	0.17	0.168
rs17309827	6	3.4	3.30×10^{-5}	0.333	0.344	0.375	0.366
rs12529198	6	5.1	1.65×10^{-5}	0.074	0.073	0.053	0.062
rs6908425	6	20.8	5.13×10^{-6}	0.19	0.202	0.23	0.218
rs6927210	6	138.1	9.05×10^{-6}	0.529	0.494	0.482	0.492
rs946227	6	138.1	1.31×10^{-5}	0.325	0.356	0.369	0.357
rs1558043	7	19.8	0.0003365	0.135	0.132	0.11	0.119
rs12704036	7	147.7	1.98×10^{-5}	0.264	0.301	0.305	0.288
rs4871612	8	126.6	2.58×10^{-5}	0.184	0.187	0.22	0.208
rs3936503	10	35.6	1.60×10^{-5}	0.344	0.356	0.301	0.319
rs10761659	10	64.1	2.68×10^{-7}	0.406	0.442	0.461	0.456
rs7081330	10	101.3	1.80×10^{-6}	0.339	0.361	0.388	0.384
rs1931047	13	92.1	0.009975	0.02	0.023	0.013	0.017
rs916977	15	26.2	0.001596	0.165	0.156	0.14	0.147
rs9895062	17	9.3	0.0002114	0.066	0.063	0.048	0.052

Table B.1: Allele frequencies for 25 SNPs genotyped in 1,182 cases from the replication CD (RCD) panel showing convergence with frequencies from the WTCCC Unaffected case samples (WTU2: CAD, HT, BD) when compared to the significant difference between the original WTCCC CD samples (WTCD) and the WTCCC Unaffected control samples (WTU1). Quoted p values are for the final, clean WTCCC dataset. Some SNPs with $p > 10^{-4}$ in this table showed stronger evidence in the preliminary data from which we fast-tracked them for follow-up.

SNP	**Missing data cases**	**Missing data controls**	$p_{\mathbf{HWE}}$
rs2484676	0.08968	NA	0.3004
rs11205760	0.1168	NA	0.8304
rs4506508	0.1083	NA	0.3011
rs2814036	0.06176	NA	0.1105
rs12035082	0.05161	0.08696	0.903
rs10801047	0.1252	0.02767	0.03348
rs17419032	0.1261	NA	1
rs363617	0.08799	NA	0.005125
rs2322659	0.1193	NA	0.5904
rs9858542	0.1328	0.07065	0.03211
rs9870678	0.1269	NA	0.6901
rs1462651	0.1100	NA	0.6761
rs9993022	0.2132	NA	0.4204
rs17234657	0.05668	0.08449	0.009521
rs9292777	0.07783	0.1052	0.2744
rs10077785	0.04653	0.126	0.116
rs13361189	0.1159	0.1196	0.5228
rs4958847	0.08545	0.02668	1
rs1363670	0.1125	NA	0.5607
rs6887695	0.08883	0.1428	0.455
rs17309827	0.1582	NA	0.8854
rs12529198	0.1176	NA	0.8169
rs6908425	0.08968	NA	0.4938
rs6927210	0.07783	NA	0.4677
rs946227	0.1151	NA	0.6353
rs1558043	0.08037	NA	0.1331
rs12704036	0.1041	NA	0.4183
rs4871612	0.05668	NA	0.09334
rs3936503	0.08037	NA	0.6359
rs10761659	0.0423	NA	0.07061
rs7081330	0.1015	NA	0.1086
rs10883365	0.1311	0.1259	0.3358
rs1931047	0.0643	NA	0.4456
rs916977	0.09052	NA	0.04415
rs9895062	0.1024	NA	0.1845
rs2542151	0.1049	0.1215	0.5582
rs2836754	0.0753	0.1146	0.7498

Table B.2: Fraction of replication cases and controls (where applicable) with null genotype calls, and HWE p values in replication cases for each of the 37 genotyped SNPs.

				p values			Replication data?			
Loc	**SNP**	**Chr**	**Position**	**Scan**	**Replication**	**Combined**	**N**	**U**	**B**	**F**
1	rs11465804	1	67414547	4.56×10^{-35}	8.42×10^{-30}	8.65×10^{-63}	Y	Y	Y	Y
1	rs11209026	1	67417979	1.23×10^{-33}	4.27×10^{-37}	3.16×10^{-68}	Y	Y	Y	Y
1	rs11805303	1	67387537	8.22×10^{-24}	6.59×10^{-17}	1.43×10^{-38}	Y	Y		
2	rs2066844	16	49303427	NA	NA	NA				
2	rs2066845	16	49314041	NA	3.69×10^{-8}	7.37×10^{-8}	Y	Y		
2	rs2066847	16	49321280	NA	3.72×10^{-24}	7.43×10^{-24}	Y	Y		
3	rs4613763	5	40428485	1.05×10^{-22}	8.37×10^{-9}	4.22×10^{-28}	Y	Y	Y	Y
3	rs17234657	5	40437266	2.31×10^{-22}	8.94×10^{-9}	7.84×10^{-28}	Y	Y	Y	Y
4	rs2241880	2	233965368	1.46×10^{-21}	3.66×10^{-5}	5.23×10^{-25}	Y			
4	rs3828309	2	233962410	8.58×10^{-21}	1.98×10^{-8}	2.94×10^{-27}	Y	Y		
5	rs10995271	10	64108492	1.72×10^{-11}	3.03×10^{-8}	1.09×10^{-17}			Y	Y
6	rs2542151	18	12769947	2.95×10^{-11}	1.66×10^{-6}	9.13×10^{-16}	Y	Y	Y	Y
6	rs1893217	18	12799340	8.34×10^{-11}	3.84×10^{-6}	6.35×10^{-15}	Y	Y	Y	Y
7	rs11190140	10	101281583	1.19×10^{-10}	1.18×10^{-7}	1.51×10^{-16}	Y	Y	Y	
7	rs10883371	10	101282445	1.27×10^{-10}	3.47×10^{-7}	4.87×10^{-16}	Y	Y	Y	
8	rs11747270	5	150239060	2.51×10^{-10}	1.27×10^{-7}	4.90×10^{-16}	Y	Y	Y	Y
8	rs13361189	5	150203580	3.23×10^{-10}	3.89×10^{-7}	2.17×10^{-15}	Y	Y	Y	Y
9	rs2188962	5	131798704	8.86×10^{-9}	2.97×10^{-11}	2.76×10^{-17}	Y	Y		
9	rs1050152	5	131704219	NA	2.75×10^{-7}	5.51×10^{-7}	Y	Y	Y	
10	rs2024092	19	1075031	1.23×10^{-8}	0.008743643	8.07×10^{-10}		Y		
10	rs4807569	19	1074378	1.43×10^{-8}	1.71478×10^{-4}	2.12×10^{-11}	Y	Y		
11	rs10045431	5	158747111	1.81×10^{-8}	4.62×10^{-6}	7.98×10^{-13}	Y		Y	Y
12	rs3763313	6	32484449	2.38×10^{-8}	0.003416345	2.99×10^{-9}	Y	Y	Y	Y
12	rs2858331	6	32789255	2.41×10^{-8}	0.00252653	1.84×10^{-9}	Y	Y	Y	Y
13	rs3924462	3	49499240	4.60×10^{-8}	0.1938421	6.82×10^{-8}	Y			
13	rs6784820	3	49425868	6.15×10^{-8}	0.1581088	5.02×10^{-7}	Y	Y		
14	rs3764147	13	43355925	1.32×10^{-7}	1.93×10^{-7}	2.45×10^{-13}	Y	Y	Y	Y
14	rs9562532	13	43497789	9.08×10^{-7}	0.007794938	1.72×10^{-7}	Y	Y	Y	Y
15	rs11175593	12	38888207	1.67×10^{-7}	6.52×10^{-5}	1.33×10^{-10}	Y	Y	Y	Y
15	rs11564144	12	39104262	1.88×10^{-7}	3.08×10^{-5}	5.89×10^{-11}	Y	Y	Y	Y
16	rs6908425	6	20836710	2.40×10^{-7}	2.09447×10^{-4}	6.24×10^{-10}	Y	Y	Y	Y
16	rs7748720	6	20797924	2.23×10^{-6}	0.02189073	8.23×10^{-7}		Y	Y	Y
17	rs7927894	11	75978964	2.85×10^{-7}	5.67305×10^{-4}	1.36×10^{-9}		Y		
18	rs2301436	6	167408399	3.29×10^{-7}	3.24×10^{-8}	2.37×10^{-13}	Y		Y	Y
18	rs7749278	6	167405736	3.78×10^{-7}	8.56443×10^{-4}	2.52×10^{-9}	Y	Y		
19	rs9286879	1	169593891	3.84×10^{-7}	5.03953×10^{-4}	2.52×10^{-9}	Y	Y	Y	Y

Continued...

Loc	SNP	Chr	Position	Scan	Replication	Combined	N	U	B	F
19	rs10489276	1	169594596	4.15×10^{-7}	0.02154895	1.52×10^{-7}	Y		Y	Y
20	rs4263839	9	114645994	4.12×10^{-7}	1.75689×10^{-4}	7.92×10^{-10}	Y	Y	Y	Y
20	rs6478109	9	114648320	4.62×10^{-7}	0.01374653	4.12×10^{-8}		Y		
21	rs2738758	20	61820069	4.16×10^{-7}	0.2957954	4.23×10^{-6}			Y	Y
21	rs2297441	20	61798026	2.12×10^{-6}	3.32708×10^{-4}	7.44×10^{-9}		Y		
22	rs11584383	1	197667523	5.17×10^{-7}	1.40273×10^{-4}	5.80×10^{-10}	Y	Y		
22	rs2297909	1	197691964	9.75×10^{-7}	1.35651×10^{-4}	1.06×10^{-9}	Y	Y		
23	rs9545740	13	80973593	6.86×10^{-7}	0.03438785	1.18×10^{-3}			Y	Y
23	rs6563216	13	80961793	1.09×10^{-6}	0.3459239	8.85×10^{-4}	Y	Y	Y	Y
24	rs2274910	1	157665119	8.02×10^{-7}	0.01650716	8.33×10^{-8}		Y		
24	rs11265519	1	157691986	1.50×10^{-6}	0.02212868	9.63×10^{-7}	Y	Y	Y	Y
25	rs2283790	22	20281207	1.13×10^{-6}	0.4387843	4.65×10^{-5}	Y	Y		
25	rs5754217	22	20264229	2.52×10^{-6}	0.3234629	1.19×10^{-5}		Y		
26	rs12529198	6	5096246	2.63×10^{-6}	0.0209643	1.54×10^{-6}	Y	Y	Y	Y
27	rs12604993	18	75866208	2.63×10^{-6}	0.2715172	4.62×10^{-4}	Y	Y		
27	rs12607007	18	75865061	4.17×10^{-6}	0.1130518	1.22×10^{-2}	Y	Y	Y	Y
28	rs6128541	20	57351084	2.85×10^{-6}	0.01560668	4.01×10^{-7}	Y	Y		
29	rs12985909	19	18300383	3.06×10^{-6}	0.0690986	1.01×10^{-5}	Y	Y	Y	Y
30	rs2872507	17	35294289	3.15×10^{-6}	2.21214×10^{-4}	5.45×10^{-9}		Y	Y	Y
30	rs2305480	17	35315722	6.33×10^{-6}	0.4447568	9.02×10^{-5}		Y		
31	rs17309827	6	3378317	3.19×10^{-6}	0.04803835	5.07×10^{-6}	Y	Y	Y	Y
31	rs4959830	6	3379241	3.39×10^{-6}	0.05170022	6.16×10^{-6}	Y	Y	Y	Y
32	rs5974	4	187576360	3.36×10^{-6}	0.0935971	1.92×10^{-5}	Y	Y	Y	Y
32	rs4253431	4	187585769	4.06×10^{-6}	0.1264694	3.99×10^{-5}	Y	Y	Y	Y
33	rs707472	1	7840274	3.62×10^{-6}	0.0748943	1.29×10^{-5}	Y	Y	Y	Y
33	rs707458	1	7766478	2.12×10^{-5}	0.228105	1.24×10^{-4}	Y		Y	Y
34	rs10512670	5	37949301	4.70×10^{-6}	0.2517905	5.07×10^{-5}	Y		Y	Y
34	rs11959616	5	37948752	4.52×10^{-5}	0.3620376	7.35×10^{-4}		Y	Y	Y
35	rs921720	8	126603853	4.86×10^{-6}	2.17×10^{-5}	1.19×10^{-9}	Y	Y		
35	rs1551398	8	126609233	5.28×10^{-6}	9.40×10^{-6}	6.71×10^{-10}	Y	Y		
36	rs744166	17	37767727	5.08×10^{-6}	1.10×10^{-7}	6.78×10^{-12}	Y	Y	Y	Y
36	rs12948909	17	37824128	1.52×10^{-5}	0.00637066	1.19×10^{-6}	Y	Y	Y	Y
37	rs10753415	1	222692358	5.13×10^{-6}	0.2792209	4.44×10^{-5}	Y	Y		
38	rs991804	17	29611838	5.75×10^{-6}	0.01027644	9.51×10^{-7}	Y	Y	Y	Y
38	rs8064381	17	29849794	3.45×10^{-5}	0.09200612	2.35×10^{-5}			Y	Y
39	rs158865	18	54054001	5.85×10^{-6}	0.1924176	1.98×10^{-5}	Y	Y		

Continued. . .

Loc	SNP	Chr	Position	Scan	Replication	Combined	N	U	B	F
39	rs158862	18	54054701	1.84×10^{-5}	0.1688187	1.22×10^{-2}	Y	Y	Y	Y
40	rs17582416	10	35327656	8.33×10^{-6}	5.33×10^{-6}	4.32×10^{-10}		Y	Y	Y
40	rs3936503	10	35589263	1.40×10^{-5}	2.62×10^{-5}	3.11×10^{-9}		Y	Y	Y
41	rs7759649	6	21578398	8.48×10^{-6}	0.05883719	7.28×10^{-6}	Y		Y	Y
41	rs11758386	6	21565929	2.08×10^{-5}	0.04386035	1.36×10^{-5}	Y	Y	Y	
42	rs7161377	14	75071147	8.59×10^{-6}	0.1931113	6.52×10^{-5}		Y	Y	Y
42	rs175718	14	75056332	2.69×10^{-5}	0.3468502	3.34×10^{-4}	Y		Y	Y
43	rs7758080	6	149618772	8.88×10^{-6}	0.02990224	5.70×10^{-6}	Y	Y	Y	Y
44	rs559928	11	63906946	8.95×10^{-6}	0.2446537	1.38×10^{-5}	Y			
44	rs596308	11	63967228	3.72×10^{-5}	0.07025349	5.30×10^{-5}	Y	Y	Y	Y
45	rs10863202	16	84545499	9.06×10^{-6}	0.1988238	2.19×10^{-5}			Y	Y
45	rs1915033	16	84542932	1.10×10^{-5}	0.2776923	1.59×10^{-4}	Y			
46	rs1200610	1	181883035	9.51×10^{-6}	0.4668432	6.37×10^{-4}		Y	Y	Y
47	rs4073149	15	72685472	1.10×10^{-5}	0.1182811	1.59×10^{-2}	Y	Y	Y	Y
47	rs7497036	15	72660732	1.36×10^{-5}	0.3404593	1.66×10^{-3}		Y	Y	Y
48	rs10758669	9	4971602	1.13×10^{-5}	2.41129×10^{-4}	2.08×10^{-8}	Y	Y	Y	
48	rs7849191	9	4978761	2.49×10^{-5}	7.13691×10^{-4}	1.43×10^{-7}	Y	Y	Y	Y
49	rs743479	21	44436378	1.17×10^{-5}	2.98098×10^{-4}	2.74×10^{-8}	Y	Y		
49	rs762421	21	44439989	1.28×10^{-5}	3.81605×10^{-4}	3.57×10^{-8}	Y		Y	Y
50	rs13003464	2	61098480	1.31×10^{-5}	0.002547838	2.23×10^{-7}	Y	Y		
50	rs10188217	2	61129193	2.35×10^{-5}	0.002828342	4.40×10^{-7}	Y	Y		
51	rs7746082	6	106541962	1.33×10^{-5}	4.81×10^{-6}	5.49×10^{-10}	Y	Y	Y	Y
52	rs6743984	2	230934834	1.62×10^{-5}	0.01100453	1.11×10^{-6}		Y	Y	
52	rs13397985	2	230916728	2.36×10^{-5}	0.07257604	1.76×10^{-5}	Y	Y		
53	rs2836757	21	39215894	1.69×10^{-5}	0.1417882	1.02×10^{-4}	Y	Y	Y	Y
54	rs11931489	4	7649390	1.80×10^{-5}	0.3490472	1.69×10^{-4}	Y	Y		
55	rs2688610	10	75324937	1.99×10^{-5}	0.08254245	4.24×10^{-5}	Y	Y	Y	Y
55	rs2675671	10	75302766	2.39×10^{-5}	0.02673483	4.82×10^{-6}	Y	Y		
56	rs12827915	12	13070503	2.05×10^{-5}	0.4219729	7.10×10^{-4}	Y	Y		
56	rs3825272	12	13046606	8.42×10^{-5}	0.3767292	2.19×10^{-3}	Y	Y		
57	rs1260326	2	27642591	2.10×10^{-5}	0.01159361	1.63×10^{-6}	Y	Y		
57	rs780094	2	27652888	5.27×10^{-5}	0.004820794	2.30×10^{-6}	Y	Y	Y	Y
58	rs6708413	2	102521887	2.15×10^{-5}	0.01472857	2.74×10^{-6}	Y		Y	Y
58	rs917997	2	102529086	2.77×10^{-5}	0.04447729	2.30×10^{-5}	Y	Y	Y	Y
59	rs7227145	18	59311578	2.26×10^{-5}	0.4255902	1.25×10^{-3}	Y	Y	Y	Y
60	rs1736020	21	15734423	2.46×10^{-5}	0.001014562	1.74×10^{-7}	Y	Y		

Continued...

Loc	SNP	Chr	Position	Scan	Replication	Combined	N	U	B	F
60	rs1736135	21	15727091	3.09×10^{-5}	1.49210×10^{-4}	3.48×10^{-8}	Y	Y	Y	Y
61	rs1040092	12	58052436	2.53×10^{-5}	0.3073626	6.26×10^{-3}	Y	Y	Y	Y
61	rs1996199	12	58059725	5.52×10^{-5}	0.1144238	5.37×10^{-5}			Y	Y
62	rs4543904	10	1453158	2.78×10^{-5}	0.3123357	1.59×10^{-3}	Y	Y		
63	rs9319943	18	55030807	2.86×10^{-5}	0.2763167	2.07×10^{-3}	Y	Y		
63	rs9961822	18	55028896	4.33×10^{-5}	0.2983416	7.90×10^{-3}	Y	Y	Y	Y
64	rs8098673	18	17927329	2.91×10^{-5}	0.0277945	1.28×10^{-5}	Y	Y	Y	Y
65	rs1456896	7	50081722	3.43×10^{-5}	2.05×10^{-5}	6.09×10^{-9}	Y	Y	Y	
65	rs1456893	7	50046933	6.80×10^{-5}	2.99×10^{-5}	1.62×10^{-8}	Y	Y	Y	Y
66	rs2944533	10	132842492	3.72×10^{-5}	0.331602	2.91×10^{-4}	Y	Y		
67	rs10010325	4	106463957	3.73×10^{-5}	0.234778	1.47×10^{-4}	Y	Y		
68	rs9829140	3	133674827	3.77×10^{-5}	0.1502965	1.93×10^{-4}	Y	Y	Y	Y
69	rs6679677	1	114015850	3.89×10^{-5}	4.88×10^{-5}	1.60×10^{-8}		Y	Y	Y
69	rs2476601	1	114089610	6.40×10^{-5}	5.69×10^{-5}	2.79×10^{-8}	Y	Y	Y	Y
70	rs8111071	19	50999246	4.09×10^{-5}	0.381589	4.12×10^{-3}	Y	Y	Y	
71	rs17814449	7	130385443	4.18×10^{-5}	0.4179221	1.92×10^{-3}	Y	Y	Y	Y
72	rs6987445	8	83235127	4.36×10^{-5}	NA	NA				
73	rs3011410	10	122495603	4.59×10^{-5}	0.2648274	2.20×10^{-4}	Y	Y		
74	rs1784493	8	107743073	5.00×10^{-5}	0.1962221	1.79×10^{-2}	Y	Y	Y	Y
74	rs1580428	8	107779719	6.85×10^{-5}	0.3332752	5.58×10^{-3}		Y	Y	Y

Table B.3: 129 SNPs from 74 distinct loci followed up in replication samples for CD meta-analysis. Final four columns show success of replication experiments: N - North American family samples, U - UK unrelated case/control samples, B - Belgian unrelated case/control samples, F - French family samples.

SNP 1	SNP 2	p
rs11265519	rs4613763	0.0496
rs11265519	rs2188962	0.001713
rs11265519	rs3764147	0.007312
rs9286879	rs13003464	0.0368
rs9286879	rs1456893	0.04274
rs9286879	rs7849191	0.006887

Continued...

SNP 1	**SNP 2**	p
rs11584383	rs6743984	0.01683
rs11584383	rs11747270	0.0243
rs780094	rs917997	0.03249
rs780094	rs6784820	0.02849
rs780094	rs7746082	0.04362
rs780094	rs1456893	0.00265
rs780094	rs7849191	0.04246
rs780094	rs10995271	0.02404
rs780094	rs744166	0.02902
rs780094	rs2542151	0.00169
rs780094	rs4807569	0.02133
rs13003464	rs2301436	0.02274
rs13003464	rs10995271	0.04577
rs13003464	rs2542151	0.01959
rs917997	rs6784820	0.02225
rs917997	rs4613763	0.03703
rs917997	rs7758080	0.02673
rs917997	rs17582416	0.02064
rs917997	rs8098673	0.01276
rs6743984	rs17582416	0.03731
rs3828309	rs2301436	0.03148
rs3828309	rs1736135	0.02527
rs6784820	rs7758080	0.0327
rs6784820	rs2301436	0.04931
rs6784820	rs10995271	0.03066
rs6784820	rs3764147	0.008501
rs4613763	rs7758080	0.0006182
rs2188962	rs7758080	0.02726

Continued...

SNP 1	**SNP 2**	p
rs2188962	rs2872507	8.17E-05
rs11747270	rs2301436	0.03407
rs11747270	rs2872507	0.03232
rs11747270	rs1736135	0.04903
rs17309827	rs6908425	0.02921
rs17309827	rs2542151	0.01143
rs17309827	rs1736135	0.04143
rs12529198	rs17582416	0.04465
rs12529198	rs2066847	0.02834
rs6908425	rs1456893	0.04442
rs6908425	rs762421	0.002305
rs11758386	rs1456893	0.03435
rs7746082	rs7849191	0.001362
rs7746082	rs4263839	0.02635
rs7746082	rs10995271	0.01366
rs7746082	rs8098673	0.03668
rs7758080	rs3764147	0.03878
rs7849191	rs17582416	0.0345
rs7849191	rs2542151	0.008799
rs17582416	rs11190140	0.025
rs10995271	rs6128541	0.02217
rs11190140	rs2542151	0.0007632
rs2066847	rs991804	0.04447
rs2066847	rs2872507	0.001565
rs991804	rs8098673	0.002871
rs2872507	rs6128541	0.04778
rs744166	rs2542151	0.02249
rs2542151	rs8098673	0.04003

Continued...

SNP 1	**SNP 2**	p
rs8098673	rs762421	0.01574

Table B.4: Nominally significant SNP-SNP interaction in CD meta-analysis