Bayesian and Frequentist Methods and Analyses of Genome-Wide Association Studies

Damjan Vukcevic
St Hugh's College

Supervisor:
Professor Peter Donnelly

Trinity Term, 2009

Department of Statistics
University of Oxford

A dissertation submitted in partial fulfilment of the
requirements for the degree of Doctor of Philosophy

Bayesian and Frequentist Methods and Analyses of Genome-Wide Association Studies

Damjan Vukcevic

St Hugh's College, Department of Statistics, University of Oxford

Trinity Term, 2009

D.Phil. Thesis

Abstract

Recent technological advances and remarkable successes have led to genome-wide association studies (GWAS) becoming a tool of choice for investigating the genetic basis of common complex human diseases. These studies typically involve samples from thousands of individuals, scanning their DNA at up to a million loci along the genome to discover genetic variants that affect disease risk. Hundreds of such variants are now known for common diseases, nearly all discovered by GWAS over the last three years. As a result, many new studies are planned for the future or are already underway. In this thesis, I present analysis results from actual studies and some developments in theory and methodology.

The Wellcome Trust Case Control Consortium (WTCCC) published one of the first large-scale GWAS in 2007. I describe my contribution to this study and present the results from some of my follow-up analyses. I also present results from a GWAS of a bipolar disorder sub-phenotype, and a recent and on-going fine mapping experiment.

Building on methods developed as part of the WTCCC, I describe a Bayesian approach to GWAS analysis and compare it to widely used frequentist approaches. I do so both theoretically, by interpreting each approach from the perspective of the other, and empirically, by comparing their performance in the context of replicated GWAS findings. I discuss the implications of these comparisons on the interpretation and analysis of GWAS generally, highlighting the advantages of the Bayesian approach.

Finally, I examine the effect of linkage disequilibrium on the detection and estimation of various types of genetic effects, particularly non-additive effects. I derive a theoretical result showing how the power to detect a departure from an additive model at a marker locus decays faster than the power to detect an association.

Acknowledgements

I am very grateful to my supervisor, Peter Donnelly, and my college tutors, Mary Lunn and James Martin. Thank you for your superb guidance, encouragement, advice and enthusiasm. I am also grateful to my viva examiners, Geoff Nicholls and Shaun Purcell, as well as Chris Holmes and Andrew Morris who examined me in the earlier stages of my degree. Thank you for taking the time to consider my work and providing many helpful comments.

I have been fortunate to participate in a number of collaborative projects during my studies. The Wellcome Trust Case Control Consortium (WTCCC) has served as an umbrella under which much of my work has been done. A wealth of data has come from the various members of the consortium, which I use throughout this thesis. I particularly thank the Craddock group for providing the data for the bipolar disorder sub-phenotype I use in Chapter 3, Jeff Barrett for assistance with obtaining the Crohn's disease data for Chapter 6, and Julian Maller for providing the fine mapping data I use in Chapter 7.

The main WTCCC study was a large undertaking, involving many groups and individuals. I wish to emphasise that it was a big team effort and that attributing any part of it to particular individuals is often difficult. To the extent possible, in Chapter 2 I describe some of my contributions and extensions to the project. In doing so, I wish to acknowledge everyone who contributed computer code that I collated to form `tada`, most notably Jonathan Marchini, Dan Davison, Bryan Howie and Ingileif Hallgrimsdóttir. Special thanks also to Bryan who produced the combined-cases signal plots (Figures 2.9–2.11).

The work in Section 7.1 on disease models was done as a joint collaboration with Chris Spencer, Eliana Hechter and Peter Donnelly. My major contribution, described in the aforementioned section, is the theoretical derivation of observed disease models at marker loci. This has benefited from substantial helpful discussions, particularly with Eliana and also with Simon Myers.

My studies were made possible by a scholarship from the Commonwealth Scholarship & Fellowship Plan (CSFP), and I am also grateful for travel funding from the Department of Statistics, the CSFP and the Barbinder Watson Trust Fund from St Hugh's College.

I have greatly enjoyed my time at Oxford. The working atmosphere in our group has been wonderful and I am grateful for the support from my colleagues, especially Niall Cardin, Dan Davison, Teresa Ferreira, Joanne Gale, Eliana Hechter, Chris Holmes, Bryan Howie, Stephen Leslie, Julian Maller, Jonathan Marchini, Chris Spencer and Zhan Su. A particular highlight was living with Niall, Bryan and Alex—thank you for all the stimulating conversation, entertaining games and delicious food. I am also very grateful to Maja Starcevic and Kresimir Petrinec for their warm friendship and company—thank you for all the photos, stories and dinners, and for looking after me when my house flooded.

Finally, I am indebted to my friends, family and Joan, who have been so wonderfully supportive. Your encouragement was felt even from the other side of the world.

Contents Summary

Contents

Chapter 1

Introduction & Background

Contents

Medical research has progressed remarkably over the last century. A wide variety of human diseases are now well understood and can be successfully treated. Some types of diseases have been particularly amenable for study using currently (and historically) available tools, and many are now even able to be prevented. For example, vaccinations for many infectious diseases are routinely given in many parts of the world.

Despite these successes, the so-called *complex* diseases still pose a challenge to modern medicine. Broadly, these are disease with more varied and complex underlying causes. With

the improvements in living standards and life expectancy over the past century, many are now becoming more prevalent and are referred to as *common* diseases. Some examples include arthritis, cancer, diabetes and hypertension.

The causes of most complex diseases are generally not well understood. They usually involve a malfunctioning of some of our usual biological mechanisms, which may be due to either genetic or environmental factors, or a combination of the two. It is known that some of these diseases are more prevalent within certain families than in the wider population, suggesting they have a genetic component. Such diseases are termed *heritable*. With the discovery of DNA, genetic factors for such diseases could begin to be studied more precisely.

Genetic technology has advanced rapidly in the last few decades, and so has the study of the genetic basis of diseases. More than a thousand genes for rare, highly heritable diseases have now been discovered (HAPMAP 2005, JIMENEZ-SANCHEZ ET AL. 2001). Unfortunately, the techniques traditionally used (mainly linkage analysis) have limitations when studying complex diseases. Motivated to overcome these limitations, much research effort in the last decade has gone into developing the tools required to conduct genome-wide association studies (GWAS). With these in place, a large number of such studies have been published over the last two years. They have dramatically increased the number of known genetic factors for complex diseases. While there is still much to learn about these factors, and about the genetics of complex diseases in general, the success of GWAS heralds the beginning of a new era in genetic epidemiology.

This thesis presents my contribution to this field, both in the development of statistical methodology and the analysis of data from actual GWAS. In this chapter I provide a brief background of recent approaches to studying the genetics of diseases. In Chapter 2, I describe my contributions to the analysis of the landmark Wellcome Trust Case Control Consortium (WTCCC) study, a large GWAS of seven diseases and one of the first to be published. In Chapter 3, I describe and present the results of a GWAS that I carried out on a small sample of individuals with a particular form of bipolar disorder. In Chapter 4, I discuss some aspects of frequentist methods and derive some useful approximations. In Chapters 5 and 6, I outline a Bayesian approach to the analysis of GWAS and present comparisons, both theoretical and empirical, with frequentist approaches. Finally, in Chapter 7, I examine the effect of linkage disequilibrium (LD) on the detection and estimation of various genetic effects and show some analysis results from a recent and on-going fine mapping experiment.

Given the rapid advancement of the field, I present the chapters roughly in the chronological order corresponding to the work they describe. This entails that some results in the earlier chapters will not necessarily be indicative of latest knowledge (although I endeavour to comment on those that have been 'superseded'). It also means that some material will seem to be out of order—in particular, Bayesian methods are applied to data in Chapters 2 and 3, but a full introduction to them is deferred until Chapter 5 where I investigate them in greater depth.

1.1 Historical overview

I provide a brief overview of recent developments in genetic epidemiology and some of the main challenges that we still face. Further details are available in various books and reviews (THOMAS 2004, RISCH 2000, BALDING 2006, MCCARTHY ET AL. 2008, MANOLIO ET AL. 2008, HIRSCHHORN & DALY 2005, WANG ET AL. 2005, LAIRD & LANGE 2006, SMITH & O'BRIEN 2005).

1.1.1 Mendelian diseases & linkage analysis

Many diseases are caused by a mutation to a single gene. Such diseases are called monogenic or *Mendelian*. They are generally highly heritable, highly penetrant and (relatively) easy to understand. Thankfully, they are also quite rare, usually due to negative selection. Well-known examples of such diseases include haemophilia and cystic fibrosis.

Before the 1980s, the only way to identify genetic factors was by directly analysing the few genes that were known at the time, usually via case-control association studies (RISCH 2000) (see below). This was very limiting since for any given disease it was very unlikely that the genes important to that disease would be one of the few that were known. Some associations were discovered in this way, for example between the ABO blood-group system and traits involving the gastrointestinal tract (VOGEL & MOTULSKY 1982). However, the discoveries showed only weak association and the case-control studies carried out often suffered from high false-positive rates. This is generally attributed to confounding due to population stratification and the low prior probability that the few genetic variants studied are actually causal for the diseases studied (RISCH 2000).

In the 1980s, *linkage analysis* became practical, a new method that allowed the scanning of the whole genome for genetic factors. The breakthrough was the identification of many genetic markers along the genome that could be used to localise the chromosomal location of causal genetic factors. These included restriction-fragment length polymorphisms (RFLPs) (BOTSTEIN ET AL. 1980) and, about a decade later, polymorphic microsatellite loci (LITT & LUTY 1989, WEBER & MAY 1989).

Linkage analysis for a disease involves identifying families with affected individuals, collecting genetic material from as many family members as possible, and then comparing the genetic inheritance pattern between the affected and unaffected individuals. Due to recombination, family members will share different regions of the genome (although such regions will generally be quite large), and the sharing of regions can be analysed using the genetic markers. The aim is to look for regions that segregate with disease more often than expected by chance, with the rationale that causal variants are more likely to be located in such regions.

Linkage analysis has been remarkably successful for Mendelian traits. Some well known diseases whose genetic factors have been located by this method include cystic fibrosis (ROMMENS ET AL. 1989, RIORDAN ET AL. 1989, KEREM ET AL. 1989), Huntington's disease (GUSELLA ET AL. 1983, MACDONALD ET AL. 1993) and some forms of breast cancer (HALL ET AL. 1990). On the order of 1000 disease genes are now known (HAPMAP 2005, JIMENEZ-SANCHEZ ET AL. 2001).

A particular advantage of linkage analysis, as compared to some other methods we describe later, is that it is robust to allelic heterogeneity, the existence of multiple causal variants at the same locus. This follows from the fact that all families will show linkage to the same region even if they have different causal variants. A disease that is caused by many rare variants at one locus will be easy to study using linkage analysis, but not necessarily with the other methods.

Despite its successes, linkage analysis has a few disadvantages. One of these is poor performance in the presence of non-allelic heterogeneity, the existence of causal variants at *multiple* loci. This is made worse if any particular variant accounts for only a small proportion of disease cases, which would then require a very large number of families in order to detect them. Another disadvantage is poor localisation: because only meioses within families are

observed, causal variants can only be localised to large chromosomal regions, typically on the order of 10 centimorgans (approx. 10 Mb). This can be alleviated somewhat by combining data across families. An alternative is provided by population-based approaches, which harness all the meioses in a population sample to increase the resolution with which we observe the genome. Such methods (discussed below) require that we abandon the linkage approach.

1.1.2 Complex diseases

More common than the Mendelian are the so-called *complex* diseases. These are believed to be affected by variants in many regions of the genome, all with small to modest effects, combined with environmental factors. Further complexity is added by the possibility of gene-gene and gene-environment interactions. This fairly unrestrictive definition is expected to cover most of the common diseases whose causes are still not understood—for example, arthritis, cancer, diabetes and hypertension. Genetic factors are believed to exist for such diseases since they have been observed to be heritable.[1]

Genetic factors discovered by linkage analysis mostly share the properties of having low allele frequency and large effect size. In other words, they show Mendelian (or near-Mendelian) inheritance. While some common causative alleles have been found by linkage, notably the role of the HLA in type 1 diabetes (CONCANNON ET AL. 1998) and ApoE in late-onset Alzheimer's disease (CORDER ET AL. 1993), they are not expected to be the norm. In general, linkage analyses are not expected to be able to detect the sort of causal variants expected in common diseases—to achieve satisfactory power, an impractically large number of families are required. While they have been been successful for Mendelian diseases (often termed the 'low-hanging fruit' of genetic studies), complex diseases will most likely require a different approach. Indeed, linkage screens of complex diseases have generally led to results that have not replicated (RISCH 2000).

Due to their complex nature, the aetiology of most common diseases is not well understood. As such, we do not necessarily expect studies to uncover direct causes like they have for Mendelian diseases. However, by uncovering loci that affect susceptibility we hope to gain

[1] Simply observing that a disease is more prevalent in families than in the wider population might point to shared environmental factors as well as genetic factors. Care must be taken to minimise such confounding when studying the heritability of a trait.

further insight into the biological mechanisms involved. See Section 1.2 for a more detailed discussion of the goals of studying common diseases.

1.1.3 Association studies of candidate genes

In an association study, genetic variants are directly typed at a given locus in affected and unaffected individuals and are tested for correlation with the disease. If a significant difference from zero is observed, the locus is said to be *associated* with the disease. An association indicates either that the locus is causative, or is correlated with the causative locus through LD; although, spurious associations can be caused by population stratification and experimental artefacts (see Section 1.1.5).

A seminal paper by RISCH & MERIKANGAS (1996) showed that, for loci with a moderate effect, a test of association at the disease locus has much higher power than a linkage analysis approach. While only a particular form of linkage analysis was analysed, it is true that the linkage approach is generally not well-powered for complex diseases. As a result, much of the effort in the last decade has been directed at making association studies practical.

To perform an association study, the ability to type variants at many loci is required. Initially, the high cost of large-scale typing limited the number of loci able to be studied, making the choice of loci very important. The usual approach was to select a variety of genes that are plausibly related to the disease, based on prior knowledge of their functions. While this has led to some successes (HUGOT ET AL. 2001, OGURA ET AL. 2001, LOHMUELLER ET AL. 2003), it is not an adequate approach in general. It relies on accurate knowledge about gene function, a good prediction of which genes are involved and the assumption that causal variants are in genes (they may not be, we still have fairly limited knowledge of the function of most our genome). Even for a disease whose aetiology is at least partially understood, this approach is likely to only identify a fraction of the genetic risk factors, and is clearly inadequate for other diseases (HIRSCHHORN & DALY 2005).

There are further complications to do with the choice of genes. Our lack of knowledge about the function of many genes is likely to lead to them being excluded from these studies. Even where we do have some understanding of gene function, we rarely have a full understanding of how the relevant genes interact, which hampers our ability to predict which might be

the most important ones to study for a given disease. Finally, even if it was possible to test *all* genes, this might still be inadequate since genetic factors might lie in non-coding regions (for example, transcription factors). Given these uncertainties, any study that focuses only on a few regions runs a great risk of missing many important and interesting findings.

1.1.4 Genome-wide association studies

While both linkage analysis and candidate gene association studies have had some notable successes for common diseases, progress has been limited due to their inherent limitations (HAPMAP 2005). The most promising alternative are genome-wide association studies (GWA/GWAS[2]), which exploit the power of association studies but without the guesswork of the candidate gene approach (HIRSCHHORN & DALY 2005). In the GWA framework, individuals are typed at markers spread over the whole genome, with the hope of covering most of the genetic variation that exists. In a sense, this is just the candidate gene approach taken to the extreme—every locus is a candidate, whether genic or non-genic (WANG ET AL. 2005). While it may be just a difference in scale, it does remove the arbitrariness of selecting candidate genes.

The GWA approach is appealing in principle, but practical limitations prevented it being applied until recently. In particular, knowledge of the distribution of polymorphic loci had not been extensive enough to carry out a genome-wide search, and typing variants at many loci had been both laborious and expensive. Both of these have been overcome in recent years.

The variant of choice for GWAS have been single nucleotide polymorphisms (SNPs). These are very abundant across the genome, and we now have an extensive collection of known SNPs via the dbSNP database (SHERRY ET AL. 2001). We also know the distribution of variants at most of these SNPs in four population samples via the International HapMap Project (HAPMAP 2003, 2005). The latter resource is particularly useful since it documents genome-wide variation and LD in those populations. This enables the selection of a subset of SNPs, known as *tag* SNPs, that capture most of the variation in a population (JOHNSON ET AL. 2001). Out of the approximately 11 million SNPs with minor allele frequency (MAF)

[2]I will use the acronym GWA to refer to 'genome-wide association', and GWAS to refer to 'genome-wide association study/studies'.

greater than 1% (KRUGLYAK & NICKERSON 2001), a few hundred thousand are enough to tag most of the variation[3] (HAPMAP 2005). The use of tag SNPs minimises the genotyping effort required when performing a genome-wide scan.

Large-scale genotyping technology has recently become economically viable (SYVÄNEN 2005). Companies like Affymetrix and Illumina have developed dense genotyping chips which can genotype hundreds of thousands of SNPs in many individuals, at a cost that is not prohibitive. Designing and mass-producing chips with standard sets of tag SNPs, they deliver a product ready-made for GWAS.

More recently, interest has been shown in using copy number variants (CNVs) as a complement to SNPs in association studies (MCCARROLL & ALTSHULER 2007). Recent surveys have characterised more than 1,000 CNVs across the genome REDON ET AL. (e.g. 2006), and companies are developing products that aim to measure such variation economically in thousands of individuals.

In contrast to linkage analysis, association studies have generally been performed on unrelated individuals sampled from the population instead of related individuals sampled from selected families (but family-based association studies are also possible, see Section 1.1.6). A popular paradigm is the case-control study, which is common in standard epidemiological settings. Using unrelated individuals offers some advantages. Firstly, it is usually easier to collect large samples, especially for late-onset disease where collecting data from parents or other family members is difficult. Secondly, such samples allow the localisation of association signals to smaller regions of the genome, since they will carry the information from a larger number of recombination events. Thirdly, control samples that are collected can be re-used in different studies. As an example, the WTCCC (2007) study used a shared set of controls for studying multiple case cohorts.

Despite the advantages of going from linkage analysis to association studies, and from family-based data to population-based data, it must not be forgotten that large sample sizes are nevertheless still required to detect the moderate effects that are expected in common diseases (see Chapter 2 for power estimates for the WTCCC study and Section 4.5 for power calculations under a range of scenarios). Thus, the collection of large samples should be seen as a fundamental step in conducting association studies for common diseases. In this

[3]The exact number varies by population.

respect, the collection by the respective disease groups of the samples used by the WTCCC was an important advance that made this GWAS possible.

Having overcome the practical hurdles, GWAS of common diseases finally started appearing about two years ago and have been successful at identifying new causal loci, with many of the findings replicated in subsequent studies (MANOLIO ET AL. 2008). The National Human Genome Research Institute (NHGRI) maintains a running catalogue of GWAS results (HINDORFF ET AL. 2009).

1.1.5 Challenges for GWAS

Before the appearance of the first GWAS, possible problems and limitations of this method had been highlighted. These include technical issues to do with large-scale studies and statistical issues to with the ability of the method to detect the effects of interest. I will now briefly outline some of the more important of these.

Population stratification. The most widely discussed possible source of bias in GWAS is that due to population stratification, also known as *population structure*. This occurs when cases and controls are sampled from different subgroups of the population being studied, and where allele differences in these subgroups give rise to spurious association signals. Numerous strategies for avoiding such bias have been proposed. These range from careful sampling of controls to ensure they are well-matched to the cases (ARDLIE ET AL. 2002), to employing methods that take advantage of the fact that population structure effects should be present across the whole genome to detect and correct for them (DEVLIN & ROEDER 1999, PRITCHARD ET AL. 2000b). In well-matched studies even mild stratification is likely to exist (FREEDMAN ET AL. 2004), but the results from theoretical simulations (WACHOLDER ET AL. 2000) indicate that, at least within certain populations, this will lead to only a small bias. In the WTCCC study, the degree of population stratification was observed to be minimal over most of the genome (WTCCC 2007).

Technical artefacts. GWAS necessitate collecting DNA from thousands of individuals, genetic typing on sophisticated machines possibly in multiple labs, and collating and managing large data sets. There are therefore many possible sources of artefactual data, whether it

may be due to errors in the lab or even just to poor data handling. Artefacts that affect cases and controls differentially are of particular concern. Over the large data sets that would be employed in a typical GWAS, even a very low error rate can lead to spurious associations at a large number of loci. Since the focus will be on loci with the strongest association signals, artefacts that are not properly dealt with can lead to very misleading conclusions.[4] Minimising and removing such artefacts is therefore a vital component of the analysis of GWAS data. I emphasise this further in Chapter 2 where I discuss the analysis of the WTCCC study.

Tag SNPs and rare alleles. Due to ascertainment bias, rare alleles are less well represented in SNP databases than common alleles. In addition, tag SNPs are usually chosen to tag common SNPs (HIRSCHHORN & DALY 2005). For these reasons, even rare alleles with strong effects might not be detected in GWAS. One obvious solution is simply to use more SNPs to cover more of the genetic variation, combined with expanding SNP databases to capture more variation in the population. Such an approach can only go as far as available funding allows, and leads to gradually decreasing cost-efficiency (WANG ET AL. 2005). Nonetheless, SNP genotyping technologies are still improving, and the latest products can type up to a million SNPs at a time. A complementary approach is to use statistical methods that combine information across nearby loci in order to detect association signals at untyped loci (LI ET AL. 2006, SCHEET & STEPHENS 2006, MARCHINI ET AL. 2007, BROWNING & BROWNING 2007, 2009).

Allelic spectra of common diseases. The allelic spectra, or 'genetic architecture', of common human diseases has an important bearing on the success of GWAS. The common disease/common variant (CDCV) hypothesis states that most of the genetic variants that underlie common diseases are themselves common in the population. At the other extreme is the multiple rare-variant hypothesis, also known as the genetic heterogeneity hypothesis, which postulates a large number of rare variants. These should be seen as two extremes within which the true allelic spectra of common diseases will lie.

The CDCV hypothesis is suggested by the fact that most linkage scans for common diseases

[4] As an example, we received some correspondence in response to the WTCCC study where the correspondees highlighted a substantial number of loci as showing strong association that were not published in the original study. From these, they hypothesised particular genetic disease models for some of the diseases. However, all of the loci were artefactual

failed to find any reproducible results. That is, since such scans are well-powered to detect rare, highly penetrant loci, their lack of findings entails that any causal loci are either common or weak, although an alternative explanation is that the sample sizes used were too small (ALTMÜLLER ET AL. 2001). Some other arguments that have been put forward in favour of the hypothesis include:

- Alleles which are now risk factors for a disease may have been advantageous in the past, hence may be common due to positive selection. An example of this is the *thrifty gene* hypothesis (NEEL 1962), which posits that genes that allow us to process and store fat more efficiently were positively selected in the past when food was more scarce, but now predispose us to metabolic disease like diabetes.

- Late-onset diseases and variants conferring a small risk are both expected to be only under weak selective pressure, so will not necessarily be rare.

The CDCV hypothesis is the best case scenario for GWAS, since they have the greatest power for common variants. This has led some to question whether the popularity of this hypothesis is overly optimistic (PRITCHARD & COX 2002). Theoretical models of disease allelic spectra have led to differing conclusions. REICH & LANDER (2001) show how the CDCV hypothesis could arise as a consequence of recent population expansion, while the models of PRITCHARD (2001) lead to less optimistic scenarios. A more recent simulation study that builds on these favours the CDCV (PENG & KIMMEL 2007), while some empirical-based studies have concluded that rarer variants will be more enriched for deleterious effects (KRYUKOV ET AL. 2007, GORLOV ET AL. 2008).

Another perspective is to compare the allelic spectra of common diseases with that of the whole genome (WANG & PIKE 2004). The most neutral null hypothesis for any common disease would be that its allelic spectrum is the same as that of the genome. Particular diseases might then be expected to show a 'common shift', towards the CDCV model—for example, metabolic diseases with variants that were previously under positive selection—and other diseases might show a 'rare shift', towards the heterogeneity model (WANG ET AL. 2005). In the neutral case, even though most susceptibility variants would be rare (MAF less than 0.01), SNPs with MAF greater than 0.01 would account for more than 90% of the genetic variation between individuals (HAPMAP 2003).

We currently have too little data on common diseases to measure the extent of any common shift. However, the argument given above suggests that at least some common variants are likely to contribute to common diseases, and because they are common they will most likely have an important population-wide impact. Since association studies are able to detect common variants well, they are worth attempting. Even if they discover common variants that confer only a small increase in risk, if such discoveries elucidate causal mechanisms then they will be considered a success.

1.1.6 Other approaches

Family-based association approaches. While association studies are often carried out with population-based data, they can also be done with family-based samples. The simplest and most widely known design uses data from trios (an affected individual and his/her two parents) and the transmission disequilibrium test (TDT) to test for association (SPIELMAN ET AL. 1993), although more sophisticated and powerful methods are now available (LAIRD & LANGE 2006). Although differences in power compared to a population-based design are small, family-based approaches are more robust to population stratification (LAIRD & LANGE 2006). However, the collection of samples usually requires more time and money, more genotyping is usually required than an equivalent population-based design, and the studies are also more sensitive to genotyping errors (LAIRD & LANGE 2006).

Admixture mapping. Mapping by admixture linkage disequilibrium, also known as admixture mapping, uses case samples from a population formed by recent admixture of two or more populations with different disease prevalences. The aim is to look for excess sharing among cases of alleles more common in the high-risk ancestral population, which is an indication that this allele probably confers an increased risk to disease. Lying somewhere between linkage analysis and association studies in its approach, admixture mapping has some of the advantages and disadvantages of both (SMITH & O'BRIEN 2005). It involves typing fewer markers than GWAS (approximately 200–500 times fewer), while retaining similar statistical power. It also allows the use of case-only data. The main disadvantages are the need for different prevalences in different populations, and the existence of a recently admixed population from these. Luckily, European and African populations fit these criteria

for many diseases, with the African-American population being admixed from these and so a good source population for the method. Like GWAS, this method has had to wait for the creation of marker libraries and technological advances before being put to use, but admixture mapping studies of common diseases have recently started to be published (FREEDMAN ET AL. 2006, REICH ET AL. 2005, ZHU ET AL. 2005).

1.2 Goals for studies of complex diseases

Given the ample resources and strong focus currently placed on the study of complex diseases, and GWAS in particular, it is worth reflecting on the goals and possible outcomes of such studies.

We expect to find that many genetic and environmental factors play a role in any given complex disease. As such, the discovery of some of the genetic factors, which is about as much as we can hope to get from GWAS initially, will probably be of limited use in predicting the risk of disease for an individual or of fully accounting for the prevalence of the disease within the population. Instead, the primary goal will be to use new discoveries to guide studies which aim to determine their function. By narrowing down the focus to specific biological pathways, we hope to gain a better understanding of the disease aetiology. For this purpose, the effect size observed at any particular locus may not be relevant, since even loci with weak effects may point us to the appropriate biological mechanisms (MCCARTHY ET AL. 2008). An increased understanding of the disease will hopefully then lead to the development of new treatments and preventative measures.

A secondary goal is the development of diagnostic tests to identify individuals who are at higher risk of disease, to help better target any preventative measures. As outlined above, our initial studies are unlikely to achieve this goal. Indeed, for some complex diseases the already known environmental factors can have a greater effect on risk than any of the newly discovered genetic factors. For example, a newly discovered risk factor for myocardial infarction (heart attack) on chromosome 9 has a population attributable risk (PAR) of 21% (HELGADOTTIR ET AL. 2007), which is unusually high compared to other discovered complex disease loci. In contrast, it is known that nine potentially modifiable lifestyle factors,

such as smoking and diet, together have a PAR of at least 90%[5] (YUSUF ET AL. 2004). In addition, the genotype information does not increase the predictive performance above what is possible with traditionally measured risk factors (PAYNTER ET AL. 2009).

While this is a limited comparison, it illustrates the point that even if we take a promising genetic discovery, the predictive power of already known environmental factors is much greater. The development of diagnostic tests is therefore more of a long-term goal and will possibly result from discoveries due to follow-up studies rather than from GWAS directly. Having said this, the above example also shows that for some diseases there might already be something to be gained from new discoveries, and holds promise that diagnostic tests might be even more useful in the future.

A third goal is to better characterise the diseases themselves. Given our lack of knowledge of many complex diseases, what we call one disease may turn out to be a collection of similar, but different, phenotypes. Through these studies, we may be able to identify these different phenotypes and also determine genetic (or other biological) markers for them. If successful, it will make such diseases both easier to study and to diagnose. This is of particular interest for psychiatric diseases, which are difficult to diagnose and sometimes even have multiple definitions. Chapter 3 represents a first step down this road, where I study a sub-phenotype of bipolar disorder which seems to have a distinctive genetic signature.

1.3 Models & methods

A GWAS aims to find genetic loci where the genetic variation is associated with the phenotype of interest. For computational reasons, GWAS have mostly been analysed using single-locus approaches. My motivation is to explore and develop the methods that will be applied most often and most routinely by researchers in the field. For this reason, I will primarily focus on the most common GWAS scenario: single-locus analysis of binary phenotypes (case-control samples) at biallelic SNPs.

While there is certainly scope for developing more sophisticated methods or more complex designs, and they will no doubt play an important role in future studies, there are com-

[5]Note that PAR is not additive when considering multiple factors, so the fact that the quoted PARs sum to more than 100% is not of concern.

pelling reasons to also focus on this simpler setting:

- This design is ubiquitous for disease studies, and is likely to remain so in the immediate future.
- GWAS are generally all analysed with at least a single-locus approach, and sometimes this will be the only analysis. Even when other approaches are attempted, the single-SNP analyses will invariably still be included when publishing the results of the study, and are likely to form the core of the results.
- Since they are simple and quick, single-locus analyses are generally the very first to be done for any GWAS. They might then be used to target more sophisticated analyses to the most promising genomic regions, or to direct follow-up studies (especially replication studies). In that case, they play an important role in prioritising research effort and funding. It would then also be true that the majority of SNPs in the study would only have been analysed by a single-locus approach.
- Researchers are most familiar with single-locus analyses. Even where other approaches highlight regions of interest, they will want to go back to check what the simpler methods have to say in those regions, both as a diagnostic tool and also as a mental reference point to understand the other approaches.

Although I focus on binary phenotypes, the methods I describe naturally extend to continuous phenotypes. Where appropriate, I will highlight any of the results that are easily generalisable to a more complex setting, and the conclusions that are more widely applicable.

I first introduce notation that I use throughout the thesis and then I describe the commonly used single-SNP methods. All of these are frequentist methods. In subsequent chapters, I use them to analyse data alongside Bayesian methods, which are described more fully in Chapter 5, and I also compare the two approaches in Chapters 5 and 6.

1.3.1 Notation

Consider a single, biallelic SNP. Denote the two alleles by A and B. The three possible genotypes are AA, AB, and BB; label these by 0, 1, and 2, respectively, the label representing

Table 1.1: **Genotype counts at a SNP in a case-control study.**

	G			
	0	1	2	Total
Cases	s_0	s_1	s_2	S
Controls	r_0	r_1	r_2	R
Total	n_0	n_1	n_2	N

the number of B alleles. Let the genotype for the ith individual be G_i and the case-control status be Y_i, where $Y_i = 1$ denotes a case and $Y_i = 0$ denotes a control. I will also use Y and G to refer to these quantities without reference to any specific individual.

When considering a case-control sample, let the number of cases and controls be R and S respectively, and the total number of individuals be N. Let the proportion of cases in the sample be $\phi = S/N$. At a given SNP of interest, Table 1.1 shows the notation for the resulting genotype counts. Some individuals in the sample might have missing genotypes at the SNP, in which case they will usually be excluded from the sample for that SNP.

We are often interested in the allele frequency of a SNP. I consider four versions: the frequency in the population, in a case sample, in a control sample, or in the whole case-control sample. I will refer to these as f, f_1, f_0 and $\bar{f}$ respectively. While I will use consistent notation throughout, depending on the context these may refer to a specific allele of interest or the minor allele. For example, the sample frequency of allele B is,

$$\bar{f} = \frac{n_1 + 2n_2}{2N},$$

whereas the sample MAF is,

$$\bar{f} = \min\left(\frac{2n_0 + n_1}{2N}, \frac{n_1 + 2n_2}{2N}\right).$$

A simplification that is sometimes made is to assume haploid data. Each individual is treated as actually being two individuals, respectively having one of the two alleles of the original individual. For example, the number of 'cases' with allele A is $2s_0 + s_1$ (and note that the sample size doubles to $2N$). The data is then of simpler structure, a 2×2 contingency table. In this situation, I will refer to the allele for each 'individual' as H (as in 'haplotype'), taking value 0 for allele A and 1 for allele B.

1.3.2 Models

The commonly-used methods for association analysis (described in the next section) can all be derived from logistic regression models. That is, they naturally fit within such a framework, even if they were not all initially derived from such models. Viewing them using this framework thus allows a unified description and also makes it easier to understand the connections between them.

Let $p = \Pr(Y = 1 \mid G)$, the probability of an individual being a case (i.e. having the disease) given its genotype at the SNP of interest. A logistic regression model writes the log-odds of disease,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

in terms of the parameters of interest. A number of models can be considered. Firstly, the *null* model postulates that the genotype has no effect on disease risk; i.e. the risk is the same for all genotypes,

$$\text{logit}(p) = \mu\,.$$

For modelling a disease risk, the most commonly used is the so-called *additive* model, where the log-odds of disease increase (or decrease) by β with each copy of the B allele,

$$\text{logit}(p) = \mu + \beta G\,.$$

I will refer to β as the *additive* parameter. If it is positive, then B is the risk allele; if it is negative, B is protective; if it is zero, the model is equivalent to the null model. I will refer to μ as the *baseline* parameter. In a case-control study it primarily reflects the proportion of cases to controls, so is not of primary interest. While non-additive models are also of interest, and I consider other models below and in later chapters, the additive model will be of primary interest in this thesis. As well as being ubiquitous for modelling disease associations, the results from Chapter 7 suggest that additive effects are likely to be the norm for SNPs in GWAS.

With up to three possible genotypes possible at the SNP, there is scope for more complex models. The most general model will have three parameters and would allow a different disease risk for each genotype. Such a model is said to be *saturated*, being able to fit exactly

to any observed risk distribution. Various parameterisations are possible, I use the following which is based on the additive model (and is similar to that of BALDING (2006)),

$$\text{logit}\,(p) = \mu + \beta G + \gamma \mathbf{1}_{G=1}\,, \tag{1.1}$$

where $\mathbf{1}_{G=1}$ is an indicator function that takes value 1 for heterozygotes and 0 for homozygotes. I will refer to this as the *general* model. The extra parameter, γ, models the deviation from an additive model at the heterozygote, and I will refer to it as the *dominance* parameter. Other commonly used models are special cases of this model and can be recovered by setting the dominance parameter to specific values: $\gamma = 0$ gives an additive model, $\gamma = |\beta|$ a dominant model and $\gamma = -|\beta|$ a recessive model. Note that my parameterisation is different from that used by the WTCCC (2007),

$$\text{logit}\,(p) = \mu + \beta \mathbf{1}_{G=1} + \phi\,(2\beta \mathbf{1}_{G=2})\,, \tag{1.2}$$

where the extra parameter, ϕ, models the deviation from an additive model in a multiplicative way by scaling the log-odds at the $G = 2$ homozygote. I believe the parameterisation given by equation (1.1) is easier to interpret and also found it had better numerical properties (see Section 5.4.3).

Recessive and dominant models, familiar for Mendelian diseases, can be fitted directly by setting up the appropriate parameterisations. For example,

$$\text{logit}\,(p) = \mu + \alpha \mathbf{1}_{G=2}\,,$$

gives a recessive model. If treating the data as haploid, only the additive model makes sense since it is a saturated model,

$$\text{logit}\,(p) = \mu + \beta H\,.$$

Under the assumption of Hardy-Weinberg equilibrium (HWE), this is equivalent to the diploid additive model (SASIENI 1997).

These models are well-defined for biallelic SNPs on the autosomes. While I primarily focus on autosomal SNPs in this thesis, it is generally the case that these and related methods can be suitably modified for use with X chromosome data, so I briefly comment on the

relevant issues. Due to differences between males and females, SNPs on the X chromosomes need special treatment. Various biologically plausible choices are possible. One option is to analyse males and females separately, with the idea that genetic effects could act quite differently in the two sexes. In that case, a haploid model would be used for males and any of the above (diploid) models for the females. Another option is to jointly analyse the two sexes, hoping to pick up genetic effects that act similarly in both. In that case, there are two natural ways to treat the genotypes in males to make them comparable to females. The first is to have sex-specific μ parameters and a common β, capturing effects proportional to the number of alleles. A second and perhaps more natural way is to treat the males as if they were homozygous females (i.e. code the male genotypes as $G = 0$ and $G = 2$) and fit the same models as before. This is motivated by the fact that most loci on the X chromosome are subject to X chromosome inactivation. Some more discussion is provided in WTCCC (2007).

The models above all assume *prospective* sampling of individuals, where the disease status is observed after the sample is taken so that it acts as the response variable. In other words, the models are of the form $\Pr(Y \mid G)$. In case-control studies, the individuals are sampled *retrospectively*, with ascertainment based on the disease status and genotype being the actual response variable. Thus, the appropriate models should be of the form $\Pr(G \mid Y)$. While it is possible to model the data retrospectively (e.g. EPSTEIN & SATTEN 2003), such models tend to be much harder to handle than prospective models. The latter also have the advantage that they are more interpretable and versatile, allowing easy inclusion of covariates and population stratification. There is theory to show that assuming prospective sampling does not make a difference when inference concerns odds ratios (e.g. β, but not μ), in both a frequentist (PRENTICE & PYKE 1979, MCCULLAGH & NELDER 1983) and Bayesian (SEAMAN & RICHARDSON 2004) setting, under quite general conditions. For these reasons, it is standard to model the data prospectively. Furthermore, a comparison of the prospective and retrospective approaches (SATTEN & EPSTEIN 2004) concluded that both perform similarly for detecting loci with additive effects. As noted above, we mainly expect to find additive effects in GWAS, so prospective methods should be adequate. Sections 4.1 and 4.6.2 feature some related discussion.

1.3.3 Association testing

The initial goal when analysing a GWAS is to detect SNPs that are associated with the disease. This is commonly done by carrying out a hypothesis test at each SNP and using the p-value as a measure of the strength of evidence. The most widely used is the Cochran-Armitage trend test statistic (ARMITAGE 1955),

$$T_{\text{add}} = \frac{N}{RS} \frac{\left(S\left(r_1 + 2r_2\right) - R\left(s_1 + 2s_2\right)\right)^2}{N\left(n_1 + 4n_2\right) - \left(n_1 + 2n_2\right)^2} \,. \tag{1.3}$$

This corresponds to a test of the additive model against the null model,

$$H_0 \colon \beta = 0 \quad \text{vs} \quad H_1 \colon \beta \neq 0 \,.$$

Formally, it is the score test (see below) of the null hypothesis under the additive model (SASIENI 1997), and is close to optimal, for large sample sizes, under this model (GART & TARONE 1983). I will also refer to it as the *additive* test. Under the null hypothesis of no association, the test statistic has a χ^2_1 distribution.

A score test is one that is based on the distribution of the score function (the first derivative of the log-likelihood) under the null. This is faster to compute than the more standard choice, the maximum likelihood ratio test (MLRT), where calculating the test statistic involves maximising the likelihood over both the null and the alternative models. The two tests are asymptotically equivalent for large sample sizes and the score test is also the most powerful test for small deviations from the null (COX & HINKLEY 1974). For GWAS, both of these are generally true (i.e. large sample sizes and small deviations), and we generally want to run tests on hundreds of thousands of loci, making computational speed paramount. This makes score tests the best choice for genome-wide scans (SCHAID ET AL. 2002).

We could also test the general model against the null,

$$H_0 \colon \beta = \gamma = 0 \quad \text{vs} \quad H_1 \colon \beta \neq 0, \gamma \neq 0 \,.$$

The score test statistic for this is,[6]

$$T_{\text{gen}} = \frac{N}{RSn_0n_1n_2}\left(n_0\left(r_1s_2 - r_2s_1\right)^2 + n_1\left(r_0s_2 - r_2s_0\right)^2 + n_2\left(r_0s_1 - r_1s_0\right)^2\right), \qquad (1.4)$$

and will have a χ^2_2 distribution under the null model. I will refer to this as the *general* test.

Many studies only use the additive test, but some also do a separate scan with the general test (e.g. WTCCC 2007). The latter is more sensitive to effects that are not additive, so can potentially detect loci that would be missed with an additive test. However, it does so at the expense of a degree of freedom, so will have reduced power for detecting effects that are additive. It is also more sensitive to genotyping errors (AHN ET AL. 2007). Studies will usually detect an effect not at the causal locus but at a surrogate SNP that is correlated through LD. It is known that this will act to dampen the size of the effect, and in Chapter 7 I show that it also acts to make it look closer to an additive model ($|\gamma|$ is reduced proportionally more than $|\beta|$).

The score test based on allele counts (i.e. the haploid model) is the familiar test of association for a 2×2 contingency table. However, it will only be valid if HWE holds, so it is better to use the additive test (SASIENI 1997). Tests based on the dominant and recessive models can also be formulated as 2×2 contingency tables—for each, two of the genotypes have equal disease risk, allowing us to collapse down to two risk classes.

SNPs that are highlighted from a genome-wide scan are often examined further. For example, we can test for deviation from an additive model. This is done by comparing the general and additive models,

$$H_0 \colon \gamma = 0 \quad \text{vs} \quad H_1 \colon \gamma \neq 0\,,$$

typically using a MLRT. I will refer to this as the *non-additivity* test. We can also test for interaction (epistasis) between the highlighted SNPs (e.g. MARCHINI ET AL. 2005, WTCCC 2007). For example, the following model specifies two SNPs acting additively and with an 'additive' interaction,

$$\text{logit}\left(p\right) = \mu + \beta_1 G_1 + \beta_2 G_2 + \phi G_1 G_2\,,$$

where the subscripts indicate the two SNPs and ϕ is the interaction parameter. The test for

[6] A partial derivation is shown in MARCHINI ET AL. (2007), and it is a special case of equation (4.1).

interaction will simply be a test on that parameter,

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi \neq 0 \,.$$

The methods described above assume complete data and accurate genotype calls. There are variations of these available that use the same underlying models but that take into account the extra uncertainty due to missing or uncertain genotypes (MARCHINI ET AL. 2007).

In regard to applying these methods and reporting results from GWAS, two different perspectives are possible. One perspective regards a particular GWAS as an end in itself and asks which SNPs, if any, show 'significant' departures from the null hypothesis. Another perspective is to regard the GWAS as the first part of a larger study: analysis of the GWAS will highlight a number of SNPs as potentially associated and these will then be examined further in follow-up or replication studies. In practice under the second perspective, SNPs will be ranked, and (roughly speaking) follow-up will be attempted from the top of this ranked list as far as available funds allow, with possible follow-up of SNPs lower down the list on grounds of biological candidacy. (For simplicity, I ignore the setting in which the GWAS itself involves a multi-stage design.) My view is that both perspectives have merit. The latter more naturally describes the actual overall role of GWAS, while the former serves to calibrate our intuition about the evidence of association at any given SNP of interest. The former also describes the typical style in which GWAS results are published, even where follow-up studies using the same data adopt the ranking approach.

A very low p-value threshold will typically be used for declaring 'significance', on the order of 10^{-8} to 10^{-6}. Various justifications are possible for such a threshold, including multiple testing correction or low prior expectation. I discuss these issues and the general interpretation of p-values in Section 5.1.

1.3.4 Effect size estimation

As well as quantifying the evidence of association, it is also of interest to estimate the genetic effects at each SNP. This is usually done with the maximum likelihood estimate (MLE) of the parameters in the relevant logistic regression model, typically the additive model, using standard optimisation techniques. Standard errors for these estimates are obtained from the observed information matrix.

According to likelihood theory, the MLE will be approximately normally distributed for large sample sizes, which can be used to construct confidence intervals. There is a direct connection between hypothesis testing and estimation in the frequentist setting. A test, with a specified significance level, will reject the null hypothesis exactly when the confidence interval, with the corresponding confidence level, does not contain the null value.

For saturated models, there are closed form expressions for the MLE. For the general model,

$$\begin{aligned} \hat{\mu} &= l_0 \,, \\ \hat{\beta} &= \frac{1}{2}\left(l_2 - l_0\right) \,, \\ \hat{\gamma} &= l_1 - \frac{1}{2}\left(l_2 - l_0\right) \,, \end{aligned}$$

where $l_i = \log\left(s_i / r_i\right)$ are the observed log-odds of disease for genotype i. For the haploid model,

$$\begin{aligned} \hat{\mu} &= l'_0 \,, \\ \hat{\beta} &= l'_1 - l'_0 \,, \end{aligned}$$

where l'_i are the observed log-odds for allele i, e.g. $l'_0 = \log\left(\left(2s_0 + s_1\right) / \left(2r_0 + r_1\right)\right)$. However, the MLE for the parameter of most interest, the additive parameter in the additive model, must be calculated numerically and will not in general be equal to $\hat{\beta}$ for either of the two models above (but will be close that under the haploid model when HWE holds approximately).

Effect sizes are often expressed in terms of the *odds ratio* (OR). That is, the ratio of the odds of disease under two different genetic variants. For the haploid model, it is the two alleles that are compared,

$$\begin{aligned} \mathrm{OR} &= \frac{\mathrm{odds}\left(Y = 1 \mid H = 1\right)}{\mathrm{odds}\left(Y = 1 \mid H = 0\right)} \\ &= \frac{\Pr\left(Y = 1 \mid H = 1\right)}{\Pr\left(Y = 1 \mid H = 0\right)} \frac{\Pr\left(Y = 0 \mid H = 0\right)}{\Pr\left(Y = 0 \mid H = 1\right)} \,. \end{aligned} \tag{1.5}$$

This is equivalent to e^{β} under the model above. The quantity e^{β} under the additive model is also an OR, where it compares the odds of disease between genotypes 0 & 1 or 1 & 2, both being the same under that model. Under either model, the parameter β is often called the

log odds ratio but I will usually refer to it as the *additive effect* for definiteness. The MLE of the OR is $e^{\hat{\beta}}$ and is the familiar cross product estimate under the haploid model (e.g. SASIENI 1997),

$$e^{\hat{\beta}} = \frac{(2r_0 + r_1)(s_1 + 2s_2)}{(2s_0 + s_1)(r_1 + 2r_2)}.$$

Under the general model, different ORs are possible between genotypes 0 & 1 and between 1 & 2.

Another commonly used quantity is the *relative risk* (RR), also called the *risk ratio*. It is similar to the OR, but compares the probability of disease ('risk') rather than the odds of disease,

$$\mathrm{RR} = \frac{\Pr(Y = 1 \mid H = 1)}{\Pr(Y = 1 \mid H = 0)}. \tag{1.6}$$

Similarly to the OR, we can posit a model where the RRs between genotypes 0 & 1 and 1 & 2 are the same,

$$\frac{\Pr(Y = 1 \mid G = 1)}{\Pr(Y = 1 \mid G = 0)} = \frac{\Pr(Y = 1 \mid G = 2)}{\Pr(Y = 1 \mid G = 1)},$$

or a more general one where they differ.

Odds ratios and relative risks are different quantities and are not interchangeable. A model that posits the same RR between genotypes 0 & 1 and 1 & 2, will not be equivalent to the additive model, which posits the same OR between the respective genotypes. Which quantity is more meaningful or useful is sometimes a cause for debate, with each having its own advantages. A relative risk is often considered easier to interpret, allowing statements of the form 'allele B increases your risk of disease by 10%'. The OR is usually more mathematically convenient since it is related to logistic regression and can be directly estimated from a case-control sample, whereas estimating the RR requires knowledge or assumptions about the disease prevalence. We usually have such knowledge, allowing us to estimate and report either quantity. Therefore, the choice of which to report should be based on the intended use and audience. When the disease is rare, RR $\approx$ OR, and can be seen in the above equations by letting $\Pr(Y = 0 \mid H) \to 1$. Thus, it is often adequate to assume they are equal.

GWA studies generally use a cohort sample in place of a true control sample. That is, they use individuals sampled randomly (or opportunistically) from the population, with some of those possibly having the disease being studied. This will reduce the power to detect

an association and also dampen the apparent genetic effect. Somewhat conveniently, it also has the effect that estimates of the OR are actually estimates of the RR; I show this in Section 4.6.2 and cite previous work. Thus, a straightforward analysis of this design is both mathematically convenient and gives the more interpretable RR, at the expense of a loss in power. Given this fact, in Chapters 2 and 3 where I present analysis results from real data, I will refer to effect size estimates as RRs. However, for theoretical results I will refer to them as ORs.

Effect size estimates taken from the most promising GWAS loci will generally have an upward bias due to the so-called *winner's curse*. This is caused by the fact that these SNPs were ascertained because they showed a strong effect. Methods exist that try to correct for this bias (ZÖLLNER & PRITCHARD 2007), and an unbiased estimate can also be obtained from a replication study. Replication is important for GWAS, not only because of the winner's curse but also due of the very lower prior expectation that any particular locus is associated with the disease. Combined with the poor track record of earlier association studies, it is now standard practice not to publish GWAS findings until they have been replicated (CHANOCK ET AL. 2007).

The above methods and discussion pertain to estimation of effect sizes at a given SNP. It should be remembered that SNPs typed in a study are likely to be surrogates for actual causal variants. Thus, observed disease effects will be imperfect replicas of true effects, depending on the amount of LD between them. It is known that LD acts to weaken the observed effect size (ZONDERVAN & CARDON 2004). In Chapter 7 I show that it also distorts the observed disease model, making it closer to being additive.

Chapter 2

The Wellcome Trust Case Control Consortium Study

Contents

Two years ago, the Wellcome Trust Case Control Consortium (WTCCC) published one of the first large genome-wide association studies (WTCCC 2007). The study investigated seven common human diseases and identified 24 loci showing strong evidence of association, roughly doubling the number of known associations at the time for those diseases. As

part of the analysis group of the consortium, I contributed to the analysis of the data for the study and the development of associated software. I also conducted some further analyses to look for weaker signals that were not explored in the initial study.

In this chapter, I first give a brief overview of the main study. I then describe some of my contributions, focusing on a few aspects not described in great detail in the main publication. Finally, I present the results from my further analyses. It should be remembered that these were completed more than two years ago; given the fast-paced nature of the field, some of the results and conclusions will now have been superseded by more recent studies.

After the main study, I conducted a GWAS of a sub-phenotype of bipolar disorder, one of the diseases in the study. I discuss this work in Chapter 3.

2.1 Overview of the study

To set the scene, I summarise the design of the WTCCC study and note connections to my work that is in later sections and chapters. Full details of the study are available in the main publication (WTCCC 2007).

The WTCCC study investigated the following seven 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). These are all common human diseases which have been shown to be heritable in previous family studies but for which few causative loci were known. Some basic information about these diseases is shown in Table 2.1; previous evidence of genetic effect is shown by the estimated *sibling recurrence risk ratio*,

$$\lambda_s = \frac{\Pr(\text{disease} \mid \text{sibling has disease})}{\Pr(\text{disease})},$$

which shows the increase in risk when a sibling is known to have the disease.

For each of the seven diseases, a case group of approximately 2,000 individuals was collected. Two control collections of about 1,500 individuals each were used, coming from the 1958 Birth Cohort (58C) and from the UK Blood Services (UKBS). All recruited individuals were living in Great Britain and self-identified as white Europeans.[1]

[1]Some of the individuals turned out to have non-Caucasian ancestry and were excluded from the analysis.

Table 2.1: **WTCCC diseases.** Sibling recurrence risk ratio estimates ($\hat{\lambda}_s$) and phenotype descriptions from the WTCCC study (WTCCC 2007). Disease prevalences are for the USA (ADVIWARE PTY LTD 2007).

Disease	$\hat{\lambda}_s$	**Prevalence**	**Phenotype description**
Bipolar disorder	7–10	1.2%	Manic depressive illness featuring an episodic recurrent pathological disturbance in mood and usually accompanied by disturbances in thinking and behaviour.
Coronary artery disease	2–7	4.85%	A chronic and generative condition in which lipid and fibrous matrix is deposited in the walls of the coronary arteries to form atheromatous plaques.
Crohn's disease	17–35	0.18%	A common form of chronic inflammatory bowel disease.
Hypertension	2.5–3.5	18.38%	A clinically significant increase in blood pressure.
Rheumatoid arthritis	5–10	0.92%	A chronic inflammatory disease characterised by destruction of the synovial joints resulting in severe disability.
Type 1 diabetes	$\sim$15	0.12%	A chronic autoimmune disorder with onset usually in childhood.
Type 2 diabetes	$\sim$3	5.88%	A chronic metabolic disorder typically first diagnosed in the middle to late adult years.

Table 2.2: **Strength of association descriptors.** Used for convenience to concisely describe the order of magnitude of p-values for association results. The first two correspond exactly to descriptors used in the WTCCC (2007).

Descriptor	p-value range
Strong	$p < 5 \times 10^{-7}$
Moderate	$5 \times 10^{-7} < p < 1 \times 10^{-5}$
Weak	$1 \times 10^{-5} < p < 1 \times 10^{-4}$

Each individual in the study was genotyped using the Affymetrix GeneChip 500K Mapping Array Set, which types 500,568 SNPs across the genome. The SNPs on this chip are not tag SNPs—they were chosen based on the cost and ease of genotyping rather than LD considerations. While tag SNPs would lead to greater coverage of the genetic variation, this more 'random' choice of SNPs offers protective redundancy against failure of particular SNPs (BARRETT & CARDON 2006). Using an LD threshold of $r^2 = 0.8$, the coverage of the chip is 65% in the HapMap Caucasian (CEU) sample (BARRETT & CARDON 2006). Genotype calling was done using the CHIAMO algorithm (`http://www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html`), developed as part of the study.

The full data set, for the study overall, consisted of 8.5 billion genotypes. Not all of these were used; some individuals and SNPs were excluded from the final analysis for quality control (QC) reasons:

- 809 individuals were excluded due to high missing data rates, inconsistent or corrupt data, non-Caucasian ancestry and relatedness;
- 31,011 SNPs were excluded due to high missing data rates, deviation from Hardy-Weinberg equilibrium and high differentiation between the two control groups.

The standard analysis consisted of a case-control comparison for each disease, applying the additive and general tests at each SNP. For convenience, we described the evidence of associations as either *strong* or *moderate* depending on the p-value, as shown in Table 2.2. In this thesis I add a further classification, *weak*, for p-values an order of magnitude greater than in the moderate range; this is also shown in the table.

Strong associations represent a very high level of evidence, and nearly all regions in the WTCCC study that showed a strong association have now been verified to be true positives (PARKES ET AL. 2007, TODD ET AL. 2007, ZEGGINI ET AL. 2007, HELGADOTTIR ET AL. 2007,

MCPHERSON ET AL. 2007). Moderate associations show less evidence, but enough to justify follow-up studies, especially given the low number of such regions (on the order of ten per disease). Many of the weak associations are likely to be false positives, but their relatively low number (no more than a few hundred per disease) might make them practical for certain types of follow-up as well.

To examine different types of effects, and hopefully boost power, a number of other analyses were carried out in addition to the standard one. These include:

- **Expanded reference group.** An analysis using samples combined into larger control groups to increase statistical power. For each disease, the control collections are supplemented with case collections from diseases that are expected to have little phenotypical overlap. This larger set is referred to as the *expanded reference group* for that disease. For BD and T2D, all the other diseases were used. For CAD and HT, both being cardiovascular disorders, all the other diseases except each other were used. Likewise for the autoimmune disorders, CD, RA and T1D. I make use of the expanded reference group for some of the analyses I present later.

- **Combined cases.** Tests using samples combined into larger case groups to study loci involved in multiple diseases and to increase power for such loci. More details are provided in Section 2.5, where I explore further results from this analysis.

- **Sex-differentiated.** Tests sensitive to effects that differ in males and females. More details are provided in Section 2.6, where I explore further results from this analysis.

- **Bayesian.** An analysis using Bayes factors (BFs) as an alternative to p-values. I discuss this in detail, and compare to the frequentist approaches, in Chapters 5 and 6.

- **Multi-locus.** An analysis involving imputing genotypes at unobserved SNPs and those with missing data. Imputation was done using the IMPUTE software package (MARCHINI ET AL. 2007) and the additive and general tests were applied to the then denser set of SNPs. I do a similar analysis for my study of a BD sub-phenotype in Chapter 3.

For each analysis for each disease, we carried out the following procedure, which I will refer to as a *genome-scan*:

1. Assemble the case and control groups (after exclusions of individuals) as appropriate for the given analysis.

2. Carry out the relevant tests on every SNP in the data (after exclusions of SNPs).

3. Take all SNPs which pass a pre-determined p-value threshold. We used different p-value thresholds for the various analyses (see below).

4. Perform extra data-quality checks on these SNPs to identify spurious associations due to obvious artefacts. The main method we used, apart from adjusting pre-testing exclusion criteria, was cluster plot inspection. I describe this further in Section 2.3.

5. The remaining SNPs typically cluster into small regions along the genome, due to LD inducing similar association signals at nearby SNPs. For each region, the SNP with the strongest association is usually chosen to represent it when reporting results.

For the standard analysis in the WTCCC study, we reported all regions with strong or moderate associations. In addition, in the Supplementary Information we reported all weak associations that were within 200 kb of a SNP with a p-value less than 1×10^{-3}. For the multi-locus analysis, we reported strong and moderate associations, and for the other analyses we only reported strong associations. In Section 2.5 I describe and extend the combined cases analysis to report moderate associations, and do the same for the sex-differentiated analysis in Section 2.6.

The X chromosome needs to be treated differently from autosomes because of differences in males and females. It was analysed separately in the WTCCC and the extra analyses described above were only carried out on the autosomes. Similarly, the further analyses I describe later were also only performed on the autosomes.

We address some other analytical questions in the study, including the effect of population structure. After excluding individuals inferred to have non-Caucasian ancestry, a principal components analysis showed a small amount of genome-wide variation attributable to geographical differences. Inclusion of principal components as covariates in association tests changed results only slightly, leading to the conclusion that confounding due to population

structure is at most a small effect. Thus, we did not correct for structure in our published results. Similarly, I report uncorrected results for the analyses I present later.

On a more local scale, 13 short genomic regions (generally much shorter than 1 Mb) showed strong evidence of geographic differentiation. A few of these were known to either show within-UK differentiation or contain genes where selection has been implicated, such as *LCT* (lactase), but most were new findings. Due to this strong differentiation, any putative disease associations in these regions should be treated with caution. None arose in the WTCCC study. However, some of these appear in the analyses I present later.

The use of large sample sizes is necessary to ensure the study had adequate power. Based on simulations (WTCCC 2007), the estimated power of a design that compares 2,000 cases to 3,000 controls and a p-value threshold of 5×10^{-7} was estimated to be 43% for a true RR of 1.3, rising to 80% for a true RR of 1.5. Thus, even a study of this size is expected not to detect some susceptibility loci. A striking demonstration of the necessity of large samples was a subsampling experiment carried out on the 16 loci that showed the strongest associations. Studies of smaller sample size were emulated by repeatedly sampling subsets of the full data and testing for association. For 1,000 cases and 1,000 controls, the expected number of loci (of the 16) detected is about 6, and rises to about 9 for 1,500 cases and 1,500 controls (WTCCC 2007). Some care is needed in interpreting this result since it is based on an arbitrary p-value threshold, but it at least illustrates that using smaller samples entails a substantial loss of power.

2.2 Overview of my role

The analysis of the WTCCC data was done collaboratively, involving multiple researchers across many institutions. As a member of the team I was involved in many parts of the analysis, all the way from making sense of the raw data through to preparation of text and figures for the final publication.

As with any large data analysis project, there are many steps along the way to a final set of results. These range from statistical tasks (e.g. exploring the data, trying out different analyses, calibrating algorithms, determining quality control criteria, interpreting analysis results) through to tasks of a more logistical nature (e.g. writing software, managing a team,

timely meetings, communicating progress). For large collaborative projects in particular, good logistics and quality control will be quite important. Despite this, such procedures only get a brief mention, if at all, in publications.

Rather than describe every task I performed in the study, I focus on two aspects of the analysis that highlight my contributions and which are not described in detail, or at all, in the final publication. This also serves to give some insight 'behind the scenes', illuminating key steps in the study. I summarise these two aspects here, and provide more detailed discussion later in the chapter.

Cluster plot inspection. An important issue that arose was the presence of genotype calling errors and how best to deal with them. While a lot of effort was invested in producing an accurate genotype calling algorithm, calling errors could not be completely eliminated because of the large number of SNPs that were being studied and the imperfect state of the data obtained from the genotyping chips. For this reason, checking for such errors became an important part of the study and this was done via inspection of cluster plots. This was quite a labour intensive process and because of the human element it was necessary to invest some extra time and effort to ensure this was performed adequately. I discuss this in more detail in Section 2.3.

Software for collaborative analysis. Processing large amounts of data requires good planning, ample IT resources and the right software. Developing the latter was an important part of the analysis. Generally, different pieces of software were written for different, well-defined tasks, like genotype calling and carrying out statistical tests. Connecting all these components together was a set of scripts that I helped develop and compile into a cohesive unit. The script also defined a *de facto* standard for storing and reporting results within the group, which facilitated easier collaboration. I discuss this in more detail in Section 2.4.

Following these, in Sections 2.5 and 2.6 I describe my further analyses of the WTCCC data, as examples of my work that is more statistical in nature. These look at the combined cases and sex-differentiated analyses respectively.

2.3 Cluster plot inspection

Genotype calling errors can easily lead to spurious associations. These can be attributed either to errors made by the calling algorithm or to poor data quality. The latter can make genotype calling very difficult, or in some cases impossible.

The calling algorithm used in the WTCCC study, CHIAMO, correctly calls the vast majority of SNPs. However, the large number of SNPs in the study means that even a small error rate or a small amount of poor quality data can lead to non-trivial numbers of false-positive associations. Thus, all SNPs that were reported as showing an association (whether strong, moderate or weak) were subject to cluster plot inspection to verify their genotype calls.

A cluster plot is a graphical representation of the genotyping of a SNP, showing both the output of the genotyping process done in the laboratory and the genotype calling done computationally. It is a scatter plot of normalised summary probe intensities from the genotyping chip, with each point representing one individual. Each point is then coloured to indicate how the genotype calling algorithm decided to classify that individual: either as a homozygote for one of the two alleles (red and blue), a heterozygote (green), or a 'null' (missing) call (cyan).

The aim of examining a cluster plot is twofold: to determine whether the given SNP has good quality intensity data and whether the calling algorithm has called the clusters correctly. In particular, we look to see whether there are clear, distinct clusters on the plot that would correspond to the three genotypes, and whether these are coloured correctly. If both of these are true, we can be confident that the genotype counts are accurate. If these are not true, it is likely that any associations we observe are caused by the resulting incorrect genotype counts. Figure 2.1 shows an example of a good cluster plot.

For the WTCCC study, a genome-scan that looked for at least a moderate association in all the autosomes typically resulted in 100–200 putatively associated SNPs. Across all the analyses on the autosomes, a total of 5,091 cluster plots were inspected, of which 1,525 were rejected and subsequently excluded from the analysis. Many SNPs were flagged up on more than one genome-scan (whether it be on more than one type of analysis or for more than one disease) but were inspected separately for each since errors on one genome-scan may not be problematic for a different one (e.g. an error in only one of the case collections). The rejected

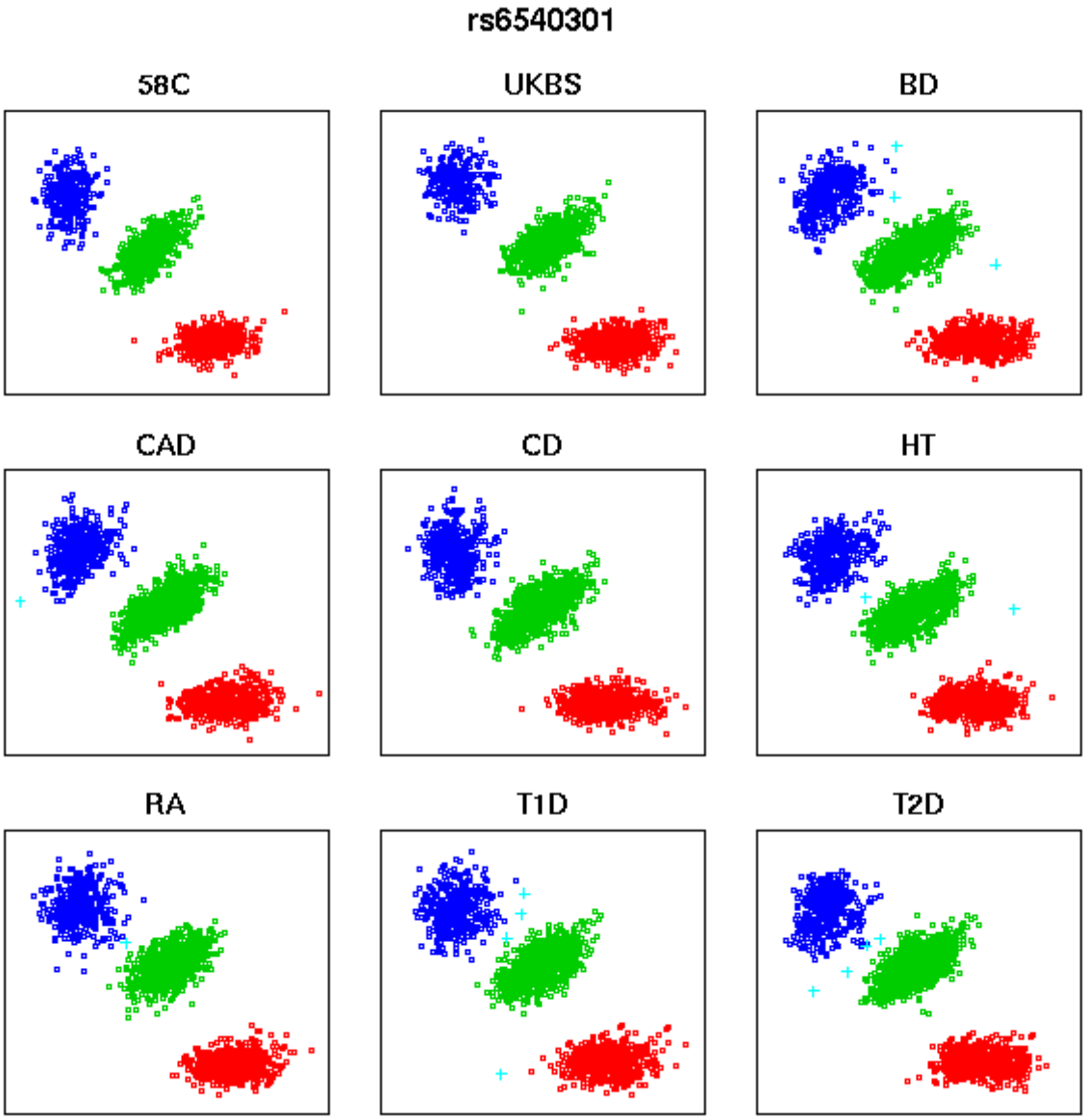

Figure 2.1: **A good cluster plot.** Each panel shows intensity and genotype calling data for one collection, all at a single SNP. The first two panels show the control collections (58C and UKBS), and the other seven the case collections. See the main text for a full description and guide to interpretation.

cluster plots corresponded to 579 different SNPs. Depending on the kind of error and any other evidence of association, some method of error recovery may be attempted. For example, re-running CHIAMO, re-genotyping using a different platform, manual calling or imputation of genotypes. In our study, only the last of these was attempted systematically. Most of the regions identified to have strong associations did so for many SNPs in the region, for which the loss of a few does not pose a problem. It is more a problem for SNPs near recombination hotspots, for which imputation will generally be less successful as well. However, no other method of recovery was attempted, since it would involve a large time investment and was deemed unlikely to result in convincing new findings—the existence of other SNPs in a region showing an association signal, even if weakened, plays an important role in confirming the association, and this would be lacking for such SNPs. This latter point motivates the use of signal plots (see Section 2.4).

With seven diseases and a few different types of analyses to conduct, the WTCCC study involved multiple genome-scans. Thus, it became necessary to inspect a few thousand cluster plots, a task we split amongst a number of people. The labour intensive nature of this task combined with its inherent subjectivity made it one of the more error prone aspects of the analysis. Since cluster plot inspection was inevitable, we invested some extra time and effort to make it quicker and easier, to ensure it was performed adequately, and to minimise the amount of inspection required. The aim was to control the 'human element' and make the process as repeatable as possible.

The first improvement was optimising CHIAMO and pre-filtering SNPs which were likely to have errors using criteria such as missing data rates and deviations from Hardy-Weinberg equilibrium. This reduced the rate of error such that only 100–200 cluster plots needed to be checked per genome-scan (this figure was well over 1,000 prior to these improvements). We subsequently focused on the inspection process itself. The creation of whole batches of cluster plots was automated, which made it possible to later inspect them at the natural pace of the operator rather then being limited by how fast the computer can create them. The creation process itself was sped up by changing to a new file format, which was also smaller and allowed the data to be dispersed over multiple servers, reducing the computational burden when conducting multiple genome-scans. The net effect of these changes was that the time it took to complete an inspection for a genome-scan dropped to about an hour for an experienced operator. (Section 2.4 describes the software used for creating cluster plots.)

Some cluster plots are clearly fine and some show clear problems, while others can be hard to judge and subjectivity becomes an important issue. To make sure we were all evaluating them in a similar way, I wrote a guide to cluster plot inspection that explained what they were, how to evaluate them, and gave many examples of typical problems to look out for (a brief version of the latter is reproduced below; the full guide is available from me by request). I also prepared a meeting for all the researchers involved in the inspection to agree on standards and to train each other. The outcomes included the adoption of a more streamlined and flexible verdict rating system, and a two-tiered inspection 'pipeline' to re-inspect all plots deemed either good or borderline. I become the *de facto* coordinator of the inspection team and our analysis results. To facilitate communication of verdicts, I instituted a consistent method of tracking and reporting them, which also doubled as an archive and proved useful for later retrieval.

2.3.1 Types of cluster plot errors

Errors observed during cluster plot inspection can be broadly classified into four types:

1. Differentially called monomorphic SNPs
2. High and differential missingness
3. Mis-called clusters
4. Unusual clustering patterns

The first two were observed more frequently than the second two. This is unlikely to be representative for the genome as a whole since we only inspected cluster plots for highly associated SNPs. Some types of errors might not cause spurious association, or at least not very often, and such errors will be observed rarely in our inspections.

Figure 2.2 shows an example of a cluster plot for a monomorphic SNP that has been called differentially in cases and controls. A number of individuals, particularly in 58C, have been called as heterozygotes. The imbalance of heterozygotes gives the impression that the rare allele is protective and gives rise to a false association.

Figure 2.3 is an example of a cluster plot showing differential missingness. The proximity of two clusters causes a large number of missing genotypes on the adjacent cluster boundaries.

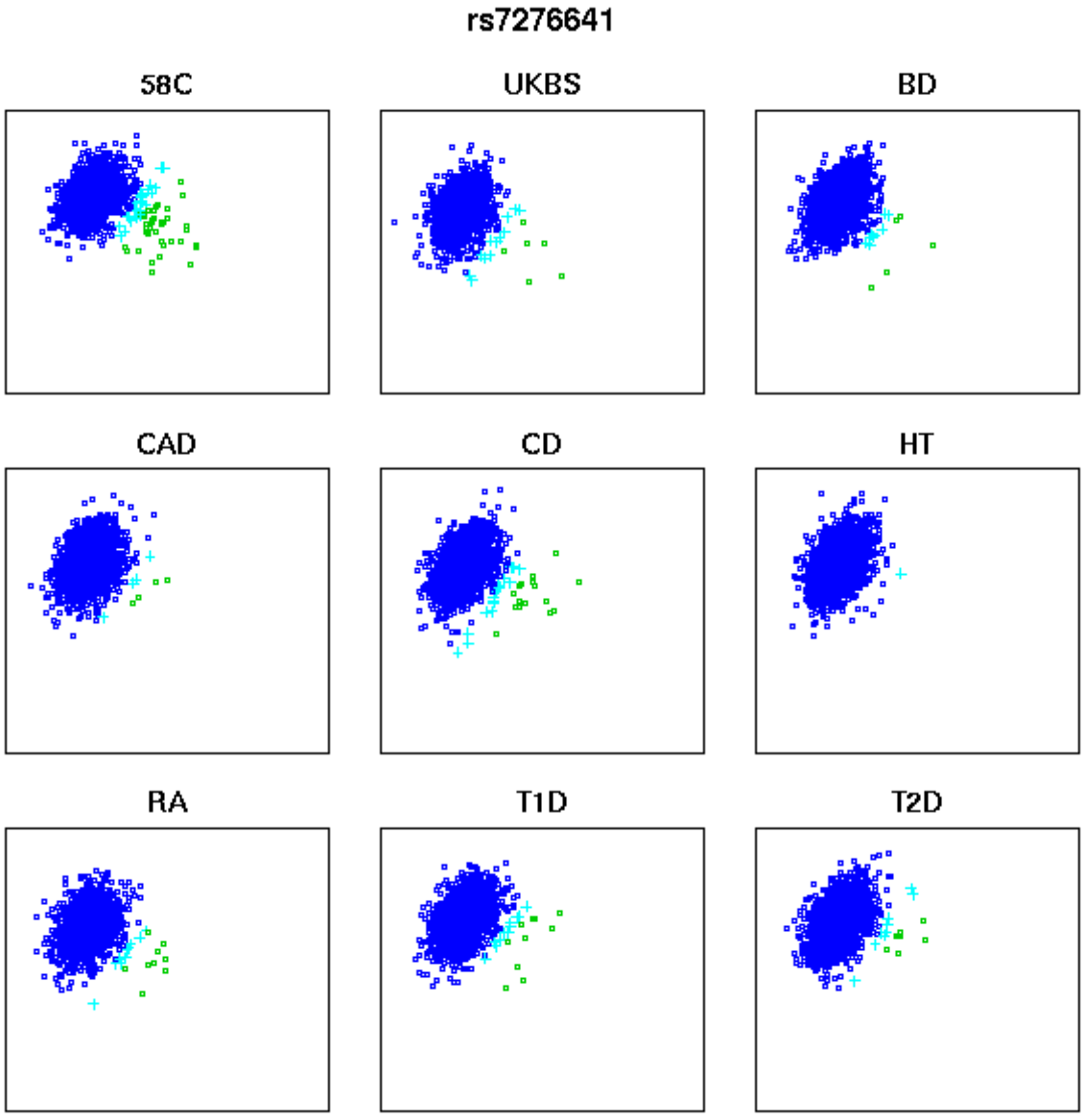

Figure 2.2: **A cluster plot showing a differentially called monomorphic SNP.** See Figure 2.1 for a general description of cluster plots.

Calling those genotypes is difficult because it is not clear to which cluster they belong, and for many the uncertainty is high enough that they are not called (we required a posterior probability of at least 0.9 from CHIAMO for a call to be made). Most SNPs will have a few null calls, but adjacent clusters will generate many more. A high amount of missingness generally leads to biased genotype counts. This is especially visible when a homozygote and a heterozygote cluster abut—the null genotypes will preferentially consist of one allele and the resulting allele counts will be biased. If this bias occurs differently in cases and controls, as it inevitably will for many SNPs, a spurious association results.

Figure 2.4 shows a situation where a whole cluster has been called incorrectly. This is a gross error which gives rise to a massive spurious association and is thankfully rare. Such errors are also rarely reproducible—re-running CHIAMO generally gives a much better result.

Some SNPs fail to display a three-cluster pattern and hence cannot be called properly. Figure 2.5 shows an example with apparently five clusters. Such SNPs may suffer from data quality issues (e.g. poor DNA quality, plate effects), or inherent properties of the SNP that cause it to be hard to genotype (e.g. copy number variation, tri-allelic SNPs, indels).

2.3.2 Challenges for genotype calling

The inspection of thousands of cluster plots gave us a good sense of the most challenging problems facing genotype calling algorithms. These relate to the two main types of cluster plot errors we observed: monomorphic SNPs and differential missingness.

SNPs with rare genotypes (especially those which are monomorphic in our sample) present a set of data that is troubling for many calling algorithms, which try to force a three-cluster model where only one or two are appropriate. The lack of individuals with certain genotypes means there is nothing to 'anchor' those clusters in the model, which then might try to 'steal' some individuals from nearby clusters. In the WTCCC study this problem was alleviated to some extent because near-monomorphic SNPs were filtered out before reaching the cluster plot inspection stage, and the use of nine separate collections (two controls, seven cases) helped in this regard. CHIAMO incorporates a prior on allele frequency based on data from the HapMap, which also helped (we generally observed more problems at SNPs not in Hapmap). To further deal with this problem, a future version of CHIAMO might try to incorporate a model that can vary the number of clusters it fits depending on the data.

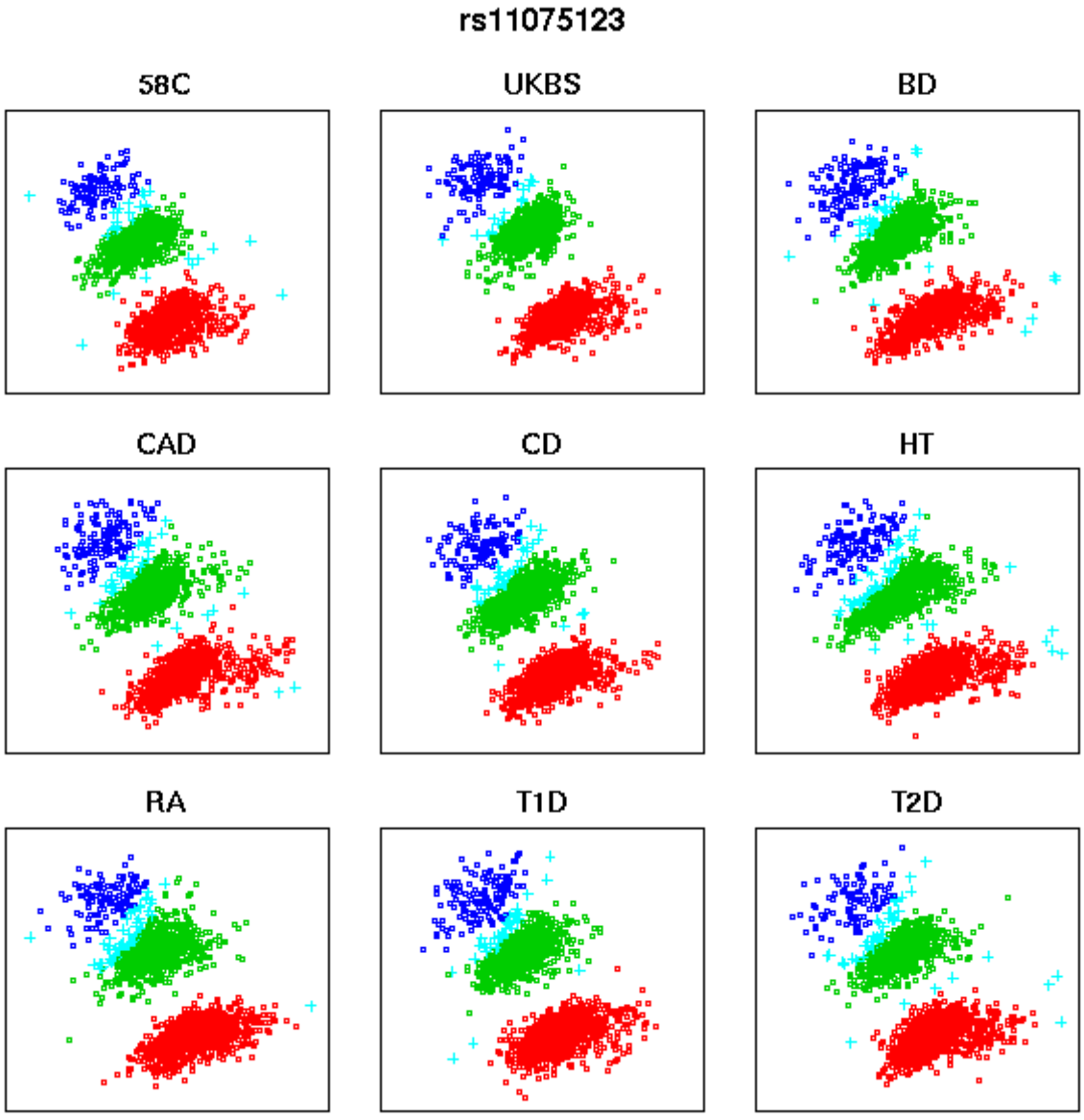

Figure 2.3: **A cluster plot showing differential missingness.** See Figure 2.1 for a general description of cluster plots.

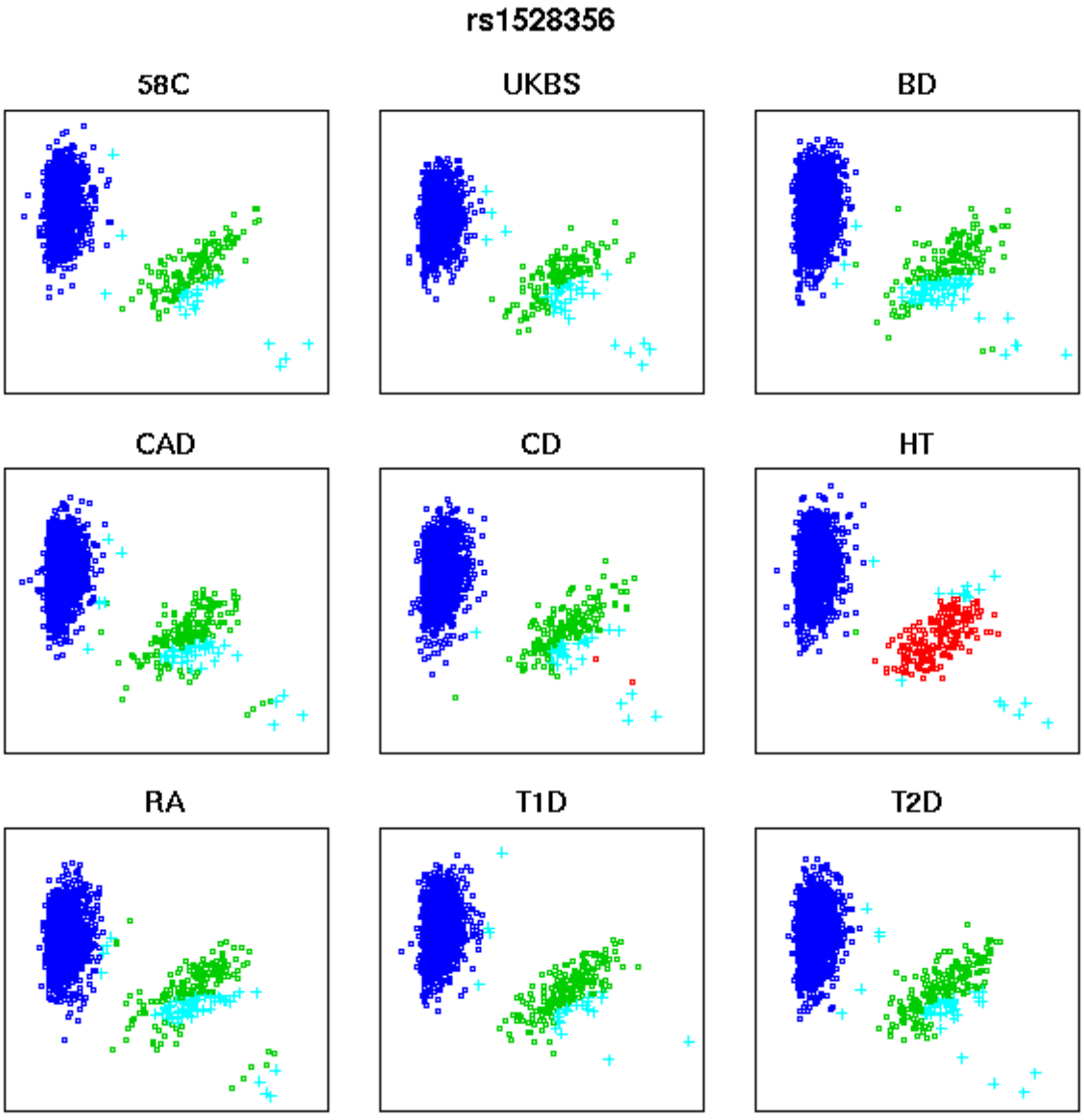

Figure 2.4: **A cluster plot showing a mis-called cluster.** See Figure 2.1 for a general description of cluster plots.

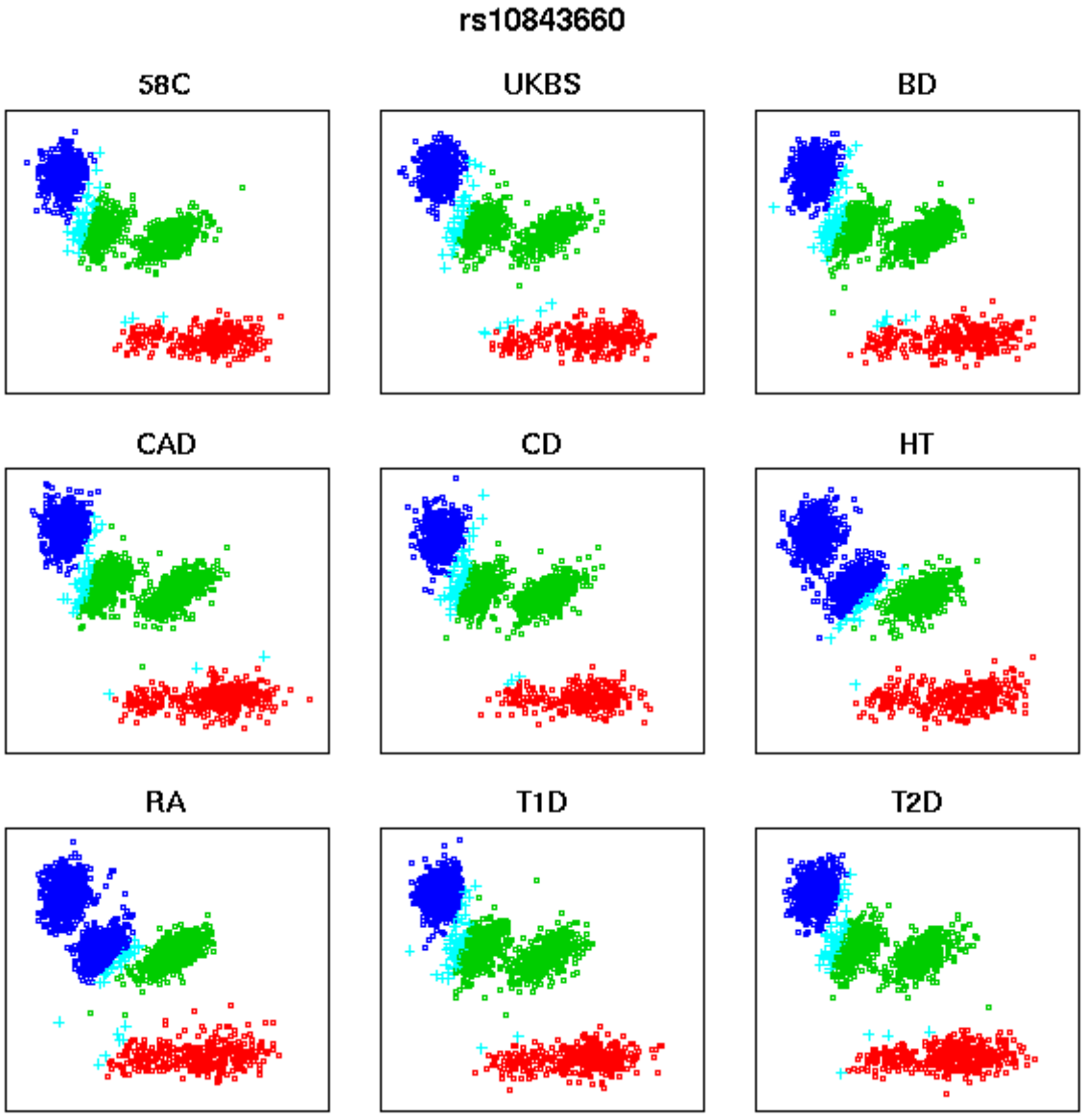

Figure 2.5: **A cluster plot showing unusual clustering patterns.** See Figure 2.1 for a general description of cluster plots.

Clusters that are adjacent or overlapping are problematic for calling algorithms because the cluster boundaries become hard to discern. The best an algorithm can do is call most of each cluster correctly and call the genotypes on the overlapping cluster boundaries as null. However, we have seen that this gives rise to biased genotype counts. This is an example of a bias that occurs even when the calling algorithm performs *correctly*. In other words, the problem here lies with the data itself and cannot be fixed by a better calling algorithm. Re-genotyping is one option for dealing with this problem, but would be expensive to do for all such SNPs. An alternative is imputation, which uses the nearby SNPs and our knowledge of the underlying recombination rates to impute the null genotypes (MARCHINI ET AL. 2007).

CHIAMO outputs a posterior distribution on the genotype of each individual at each SNP. To actually call the genotype, we used a posterior probability threshold of 0.9, treating the genotype as missing when the most probable call fell below this threshold. The rate of missingness with this choice of threshold was low—the missing data rate over all non-excluded SNPs in the study was 0.3%, compared to 0.7% for BRLMM (WTCCC 2007). BRLMM is the standard genotype calling algorithm used by Affymetrix (AFFYMETRIX 2006). Counter-intuitively, we found that increasing the threshold in an attempt to improve data quality was counter-productive, leading to increased false positives because of differential missingness (more null calls on the boundaries between adjacent clusters).

We observed that calling errors generally did not occur together but were scattered throughout the genome. Thus, a spurious association caused by a calling error will be inconsistent with surrounding SNPs, some of which will be in high LD and would therefore be expected to show an elevated association signal. A future version of CHIAMO might be able to use data from nearby SNPs to improve its calls, or at least flag which associations it believes might be spurious. This latter idea can be attempted independently of any calling algorithm and could help to reduce the amount of cluster plot inspection.

An effect that was observed but that did not cause a problem for CHIAMO was a systematic difference between samples typed in different laboratories. Some individuals in the WTCCC study were typed in an early phase of the study at a different laboratory. For some SNPs, these individuals show a radial shift towards the origin on the cluster plot. Such a SNP is shown in Figure 2.6. With data from both laboratories combined, the cluster plot shows either three elongated clusters or three cluster pairs for six of the nine collections (three

collections did not have individuals typed at the other laboratory). CHIAMO was able to call most of these SNPs correctly despite the model mis-specification and the differences between the collections.

2.4 Software

The WTCCC study involved analysing a very large amount of data, which necessitated both substantial computing power and ample storage space. Our group was quite well-equipped, having access to a computing cluster with 160 processing units, each with 1 GB of RAM, and a total of approx. 2 TB storage space. We also had the use of a few smaller servers and each member of the group had a personal workstation which was quite powerful on its own. The hardware resources were not a limiting factor. Rather, our attention was focused on using those resources efficiently and on developing software specific to our needs.

Software development occurred both prior to and during the actual study. New software was necessary since this was one of the first studies of this type and we needed software that would run quickly and was customised to our analyses. Much of the development occurred in tandem with the study, as the results of the analyses and experiences from other members of the group fed back to those developing the software. The core pieces of software, mostly developed by Jonathan Marchini, were CHIAMO, IMPUTE and SNPTEST (`http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html`). Respectively, they do genotype calling, genotype imputation and calculation of test statistics. They were designed to be mutually compatible, with CHIAMO and IMPUTE both producing output in the format required by SNPTEST.

The core software was developed and run centrally, producing large amounts of output spread across numerous files. Post-processing of these files was required to actually complete each analysis. This included tasks that were easy to automate, like scanning the results for low p-values, and also others that were very interactive, like cluster plot inspection. Initially, each member of the group wrote their own scripts to perform these tasks, using the R statistical computing package (R DEVELOPMENT CORE TEAM 2007). While some code was shared, a lot of redundant programming occurred as each person tackled the same basic tasks in their own way. Once it became clear that the post-processing stage involved a

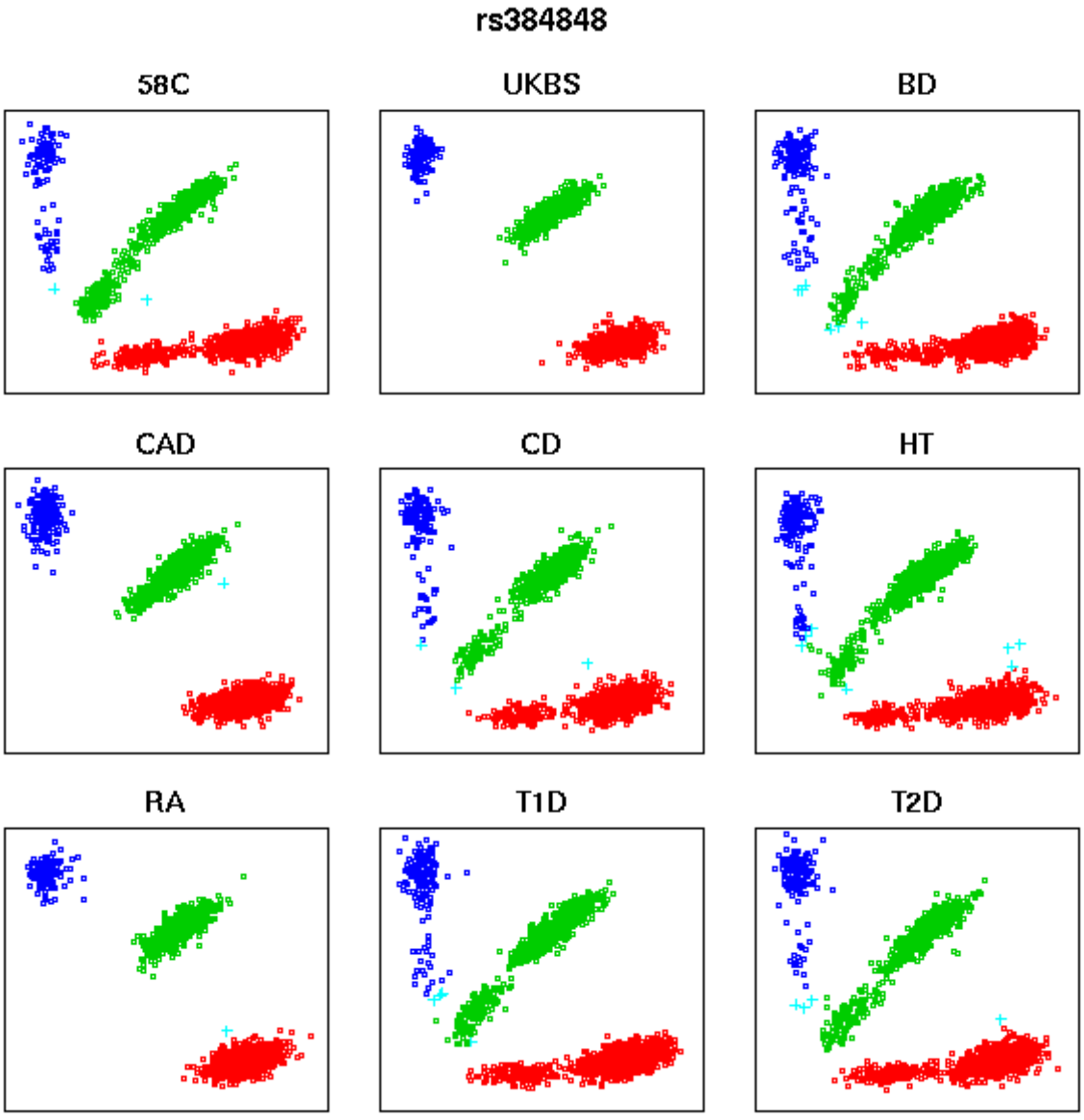

Figure 2.6: **A cluster plot showing a radial shift.** The individuals showing a shift were typed in an early phase of the study; UKBS, CAD and RA did not have any such individuals. See Figure 2.1 for a general description of cluster plots.

non-trivial amount of work, and that this would need to be done by multiple people, we decided to pool and standardise our code. With some help from Dan Davison, I amalgamated code from various members of the group, re-wrote much of it and wrote extra functions to make it all work as a cohesive unit. The resulting set of scripts, named `tada`, was a large and comprehensive collection and played an important role in our analysis effort. It was used to complete all the non-imputation analyses and served as base for more complicated tools (for example, the code that creates the information-rich signal plots shown in WTCCC (2007)). It also provided a convenient set of tools for accessing the output from CHIAMO and SNPTEST.

2.4.1 `tada`

The source code for `tada` consists of 2,300 lines of R code (including comments) which is too long to attach as an appendix. It is available from me by request.

The `tada`[2] scripts serve a number of functions. The main one is to automate the post-processing tasks for a given analysis, which include scanning the genome to identify all SNPs that meet the required association threshold, create cluster plots for them, and then collating all the results after verdicts for them have been given. Cluster plots can also be created on demand for given SNPs, a convenient feature since creating such plots from scratch is non-trivial. Once a genome-scan is complete, `tada` can create signal plots for regions with an association signal and also QQ plots for the whole genome-scan. In creating these, `tada` incorporates the cluster plot verdicts given earlier, using them to exclude certain points in a plot or colour them differently.

A functionality of `tada` that eventually became redundant was the ability, when doing a genome-scan, to filter out SNPs based on allele frequency, missingness, deviation from Hardy-Weinberg equilibrium or simply a given exclusion list. This was eventually done centrally, with unwanted SNPs being removed from the SNPTEST output completely, but the functionality was retained in `tada` in case it was needed in the future.

The design of `tada` was deliberately modular, splitting tasks into small, interacting functions. Some of these turned out to be convenient in their own right. In particular, a function

[2] *Tada* means 'then' in Serbian, referring to the fact that these scripts are intended to be used on the output of the SNPTEST software package.

that loads the correct SNPTEST file (out of the hundreds of choices, all with similar filenames) was very useful for most exploratory analyses.

A convenient consequence of writing `tada` was the creation of a *de facto* standard for reporting cluster plot verdicts and genome-scan results, which eased jobs further down the pipeline.

I now describe the main uses of `tada` in more detail from the user's perspective, making reference to particular functions and giving some examples of typical output.

Configuration and directory structure

Before using `tada`, you need to specify the directory where the data is stored and the directory where you wish output to be written. This is done by editing `data.dir` and `work.dir` respectively, which are the first two functions in the file. The following data is assumed to exist in `data.dir`:

- Genotype posterior probabilities from CHIAMO in gzipped binary format with filenames of the form: `58C_01_chiamo++_bin.gz` for SNPs from chromosome 1 for the individuals in the 58C collection.
- Normalised probe intensities from the genotyping in gzipped binary format with filenames of the form: `RA_11.txt_bin.gz` for SNPs from chromosome 11 for the individuals in the RA collection.
- SNPTEST output in gzipped ASCII format stored in a `snptest_output` subdirectory with filenames of the form:
  - `single_BD_02_snptest.gz` for SNPs on chromosome 2 for the analysis comparing BD with the two control collections.
  - `combined_HT_16_snptest.gz` for SNPs on chromosome 16 for the analysis comparing HT with the expanded reference group.
  - `combined_cases_T1D_RA_21_snptest.gz` for SNPs on chromosome 21 for the combined cases analysis comparing RA+T1D with the two control collections.

- Estimates of r^2 between SNPs on the Affymetrix 500K chip based on the HapMap CEU samples in an `r2_tables` subdirectory with filenames of the form: `Affy500k_r_squared_chr03.txt` for chromosome 3.
- Recombination rate estimates based on the HapMap CEU samples in a `rates` subdirectory with filenames of the form: `genetic_map_chr19.txt` for chromosome 19.
- The list of excluded individuals in a `sample_files` subdirectory in the file `exclusion-list-05-02-2007.txt`.
- The covariates for each individual in a `sample_files` subdirectory with filenames of the form: `wtccc_sample_T2D.txt` for the T2D collection.

Genome-scans

A genome-scan is performed with the `analyse.all` function, or with `analyse` if just a particular chromosome is of interest. You need to specify what type of analysis to run and for which disease(s). In particular, you can choose which test statistics to use, what threshold to specify for them and whether to use the expanded reference group. Which SNPTEST files need to be read is determined automatically. Each genome-scan will result in many files being written to an appropriately named subdirectory in `work.dir`. For example, `single_HT` for a genome-scan of HT using the normal controls or `combined_RA` for a genome-scan of RA using the expanded reference group. Within that subdirectory, the following files will be written:

- The file `summary`, with details on how many SNPs on each chromosome were filtered out and how many passed the threshold.
- A list of SNPs passing the threshold with a few summary statistics and space to fill in a cluster plot verdict, with filenames of the form `single_BD_01.hitSNPs.summary` for chromosome 1.
- The same lists but with the complete SNPTEST output for each SNP, with filenames of the form: `single_BD_01.hitSNPs.alldata` for chromosome 1.
- Cluster plots, one per SNP, will be written to the `cluster_plots` subdirectory.

A complete run of `analyse.all` takes about 20–30 minutes using an Intel Pentium 4 CPU with 3 GHz. Once it is complete, the cluster plots should be inspected and the verdicts recorded in the `hitSNPs.summary` files by changing the '?' for each SNP to either 'y', 'n' or 'm'. Once this is complete, running `collate.hits` will collect all the SNPs with a 'y' or 'm' verdict into a single file with a name of the form `single_BD.hits.summary.csv`. This can be used as a list for a second-pass cluster plot inspection, after which it can act as a summary of the results.

Cluster plots

The `make.cplots` function creates cluster plots for a given set of SNPs. An example of a typical cluster plot is shown in Figure 2.1. Usually, a cluster plot shows individuals from all nine collections excluding those on the exclusion list, but both of these options can be varied. It is also possible to change the calling threshold, use renormalised posteriors (excludes the null posterior from the CHIAMO output and renormalises) and create cluster plots for the X chromosome that distinguish between males and females.

Signal plots

At the conclusion of a genome-scan, running `create.signal.plots` will create a signal plot for each SNP with a 'y' verdict for a given test statistic. Figure 2.7 shows an example. The plots show the test statistics for a 500 kb region around the 'focal' SNP, colouring SNPs by their r^2 value with the this SNP, excluding SNPs with a rejected cluster plot, and also plot the estimated recombination rate for that region (McVean et al. 2004). The fine-scale recombination rate estimates are taken from Phase II HapMap (HapMap 2005), and also shown is the cumulative genetic distance from the focal SNP based on these estimates. Note that these signal plots are simpler that those shown in WTCCC (2007), which also include imputed SNPs, genes and information on conservation across 17 different species.

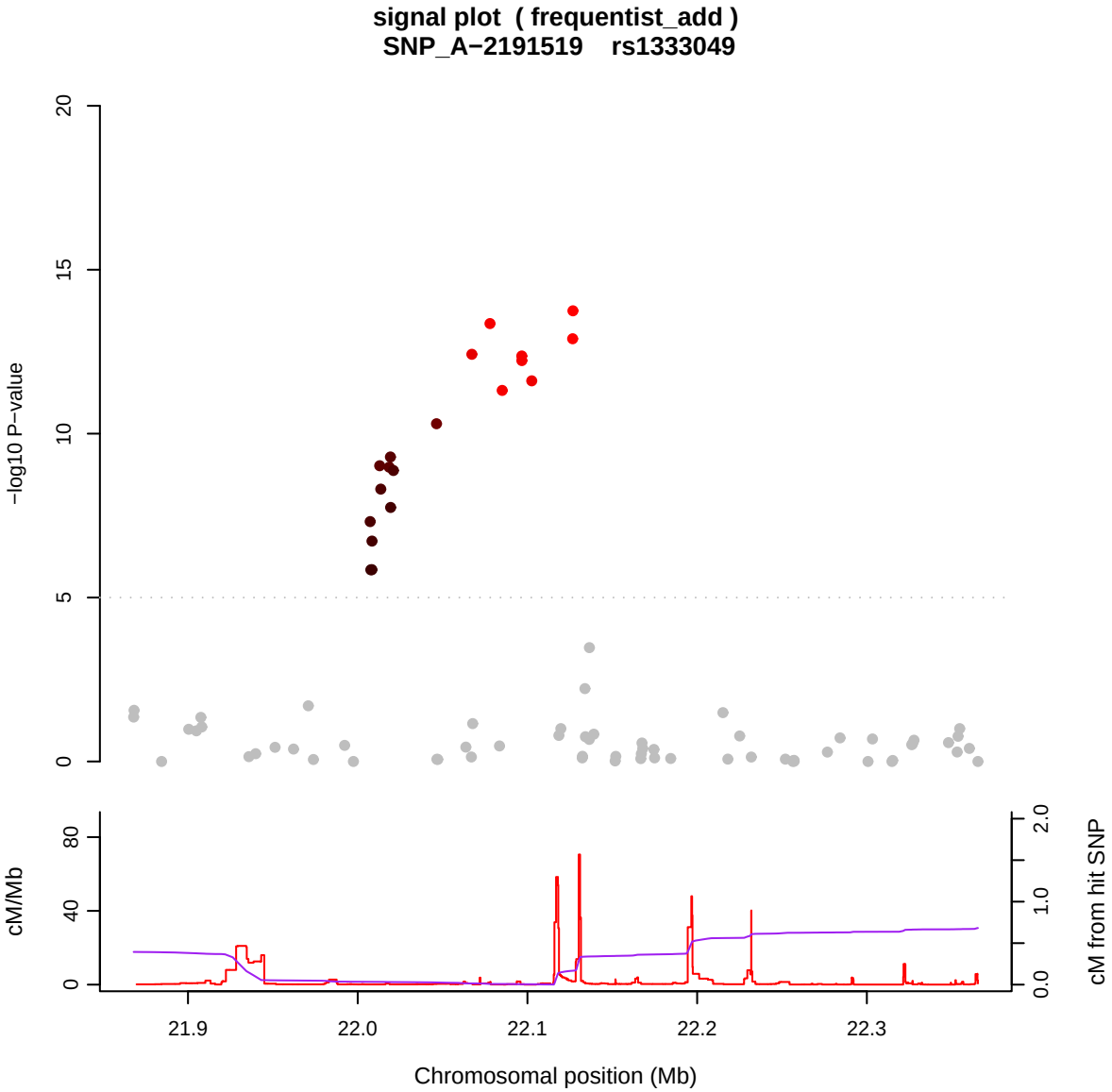

Figure 2.7: **An example signal plot produced by `tada`.** A visual display of the association test results in a region and known LD information. The plot is with reference to a chosen 'focal' SNP, usually the one with the smallest p-value, and extends for 250 kb in each direction. The top part shows p-values (on a log scale) for all SNPs in a region, after any exclusions. Points are coloured by their r^2 value with the focal SNP, using estimates from HapMap CEU, ranging from red ($r^2 = 1$) to black ($r^2 = 0.1$, the minimum value reported in HapMap), and grey ($r^2 < 0.1$). The bottom part shows fine-scale recombination rate estimates (red) and cumulative genetic distance (purple) from the focal SNP.

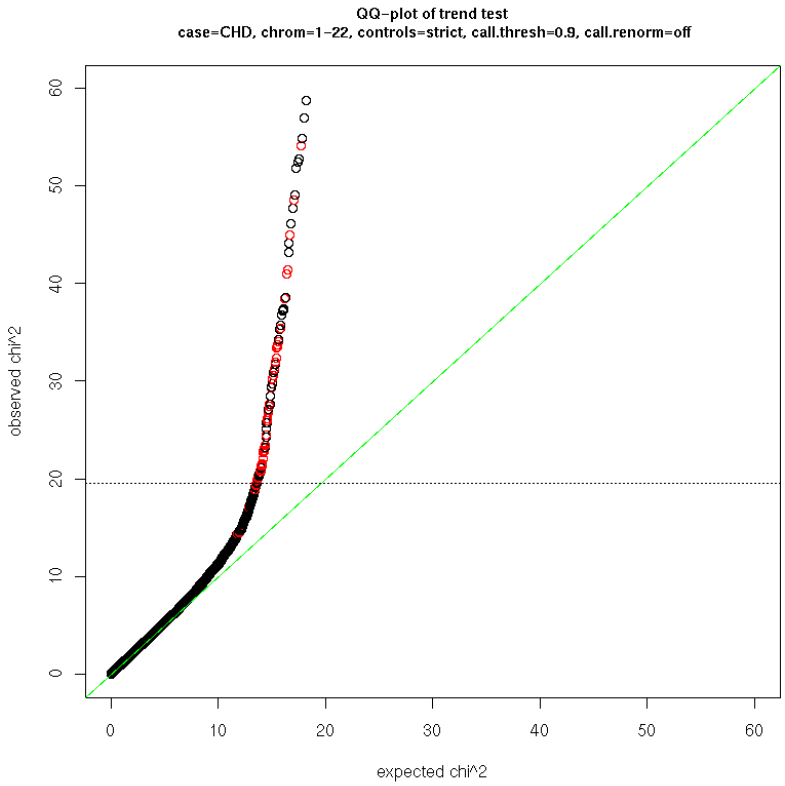

Figure 2.8: **An example QQ plot produced by `tada`.** Shows the distribution of test statistics for a given genome-scan. SNPs with rejected cluster plots are coloured red.

QQ plots

After a genome-scan, a QQ plot can be created for any of the frequentist test statistics using the function `qqchisq`. There are options to exclude SNPs in the major histocompatibility complex (MHC), to apply genomic control and to use data from particular chromosomes only. SNPs with rejected cluster plots are coloured red. Figure 2.8 shows an example.

Reading SNPTEST output

With the SNPTEST data dispersed over multiple files, the `read.snptest.output` function provides a convenient way of accessing the right file without having to remember the exact filename. For example,

```
read.snptest.output(chrom=18, case="BD")
```

will find the file corresponding to the analysis of BD on chromosome 18, and return the contents as a data frame. The same can be done for the combined cases analyses and those

using the expanded reference group, as follows:

```
read.snptest.output(chrom=18, case=c("T1D","RA"))
read.snptest.output(chrom=18, case="BD", extra.controls=TRUE)
```

If only a subset of SNPs is required, they can be loaded using:

```
read.snptest.output(13, "BD", affy=affy.ids)
```

where `affy.ids` is a vector of Affy IDs for the SNPs of interest.

2.5 Combined cases analysis

Studying multiple diseases simultaneously gives us the ability to look for joint susceptibility loci. Examples of such loci already exist amongst our seven diseases, raising the prospect of discovering more. Autoimmune diseases, in particular, seem to share many loci, and the currently known examples amongst our disease all relate to autoimmune disorders. Both RA and T1D had well-known associations in the major histocompatibility complex (MHC) on chromosome 6, and at rs2476601 in *PTPN22* on chromosome 1 (BEGOVICH ET AL. 2004, BOTTINI ET AL. 2004, SMYTH ET AL. 2004). Loci in *IL2RA/CD25* on chromosome 10, previously reported to be associated with T1D (VELLA ET AL. 2005), were recently shown to be associated with Grave's disease (BRAND ET AL. 2007) and evidence for association with RA is presented in the WTCCC study (and also below).

The approach we took to studying shared loci is to combine case collections for diseases that have putatively similar aetiology and treat them as a single case group. This enabled us to use the same statistical methods as for the standard analysis, and the increased sample size also meant that we had greater power to detect shared loci. We explored the three disease groupings shown in Table 2.3.

Table 2.3: **Disease groupings for the combined cases analysis.** Samples sizes are after exclusion of individuals as described in Section 2.1.

Disease grouping	Sample size	Reason for grouping
RA + T1D	3,823	Already known to share common loci
CD + RA + T1D	5,571	Autoimmune diseases
CAD + HT + T2D	5,802	Metabolic and cardiovascular diseases

The combined cases analysis in the WTCCC study reported regions with strong associations only. Four such regions were found. A strong association on this analysis is not necessarily indicative of a joint effect, however, since it may just be driven by a very strong effect in only one of the diseases. Further analysis of each region is therefore warranted to determine the nature of the association. Nevertheless, the small number of loci that this analysis identifies makes it easy to recommend that all loci be earmarked for further study in each disease, simply due to putatively similar aetiology.

Following the WTCCC study, I extended the combined cases analysis to find moderate associations, and further examined each highlighted region to determine which diseases contribute to the association signal. I now report the results of this analysis.

2.5.1 Results

Conducting a genome-scan for each of the combined case groups, following the procedures already describe, leads to the results shown in Table 2.4. The SNP with the lowest p-value was chosen to represent each region. Note that some regions showed an association for both the RA+T1D and CD+RA+T1D case groups. For each region, the strength of association in each disease analysed separately is shown, by describing it as strong, moderate or weak. I excluded the MHC region from the analysis. Note that the region on chromosome 20 was identified in the WTCCC study as showing high geographic variation, indicating possible bias due to population structure.

Ten regions showed at least a moderate association amongst the autoimmune diseases, in either or both of the RA+T1D or CD+RA+T1D case groups. Of those, four were strong associations and were reported in the WTCCC study. The CAD+HT+T2D metabolic disorders case group gave seven regions, all showing moderate association. Example signal plots for some of these regions are shown in Figures 2.9–2.11, one for each combined case group.

Some of associations might be due to genuine shared susceptibility in multiple diseases, while others will simply result from a very strong association in one disease. To try to distinguish between these possible scenarios, I look to see whether the genetic signal has the same character and strength in each disease separately. This is shown in Table 2.4.

Table 2.4: **Combined cases analysis results.** Regions with at least one SNP showing a strong or moderate association on the additive or general tests using the combined case groups. A representative SNP is shown for each region. Regions in bold show a strong association and were reported in the WTCCC study (WTCCC 2007). For each region it is also shown whether it has at least a weak association with any of the diseases that comprise that combined case group (such associations were reported in the WTCCC study). The MHC region was excluded. *Region shows high geographic variation, association result should be interpreted with extreme caution.

				p-value		Individual case associations		
Case group	**Chromosome**	**Position (Mb)**	**SNP**	**Additive**	**General**	**Strong**	**Moderate**	**Weak**
RA + T1D	**1**	**114.0**	**rs6679677**	2.54×10^{-34}	3.90×10^{-33}	RA, T1D		
	4	123.4	rs10015924	1.95×10^{-6}	1.20×10^{-5}	T1D		
	6	138.0	rs2327832	7.65×10^{-6}	3.27×10^{-5}		RA	
	9	27.7	rs10812655	1.20×10^{-3}	9.64×10^{-6}			
	10	**6.1**	**rs2104286**	5.92×10^{-8}	2.53×10^{-7}		RA, T1D	
	12	**110.9**	**rs17696736**	5.72×10^{-11}	3.29×10^{-10}	T1D		
	16	61.2	rs4265819	8.10×10^{-4}	7.84×10^{-6}			
	22	35.9	rs743777	1.25×10^{-6}	1.98×10^{-6}		RA	T1D
CD + RA + T1D	**1**	**114.0**	**rs6679677**	2.96×10^{-16}	2.55×10^{-15}	RA, T1D		
	2	141.6	rs16845023	5.64×10^{-5}	9.82×10^{-6}			
	4	123.7	rs17388568	4.78×10^{-6}	2.59×10^{-5}	T1D		
	9	27.7	rs10812655	2.39×10^{-3}	5.59×10^{-6}			
	12	**110.9**	**rs17696736**	9.33×10^{-10}	2.62×10^{-9}	T1D		
	16	61.2	rs4265819	3.25×10^{-4}	7.03×10^{-6}			
	18	**12.8**	**rs2542151**	9.29×10^{-8}	6.17×10^{-7}	CD	T1D	
CAD + HT + T2D	1	149.5	rs16834421	5.02×10^{-3}	7.96×10^{-6}			
	2	2.9	rs1470614	3.26×10^{-3}	2.91×10^{-6}			
	2	226.9	rs2943634	7.49×10^{-6}	3.36×10^{-5}			CAD
	9	22.0	rs10965219	7.00×10^{-7}	3.77×10^{-6}	CAD		
	12	18.5	rs12581163	1.09×10^{-2}	9.02×10^{-6}			T2D
	14	54.3	rs709939	3.41×10^{-6}	1.39×10^{-5}			HT
	20	38.5	rs16989035*	4.44×10^{-2}	2.99×10^{-6}			

Additive effects

For regions that showed an additive effect (the additive p-value is at least moderate strength), the relative risks in the combined and separate case groups are shown in Table 2.5. The region on chromosome 2 represented by rs16845023 has an additive test p-value that is not quite moderate strength but was added to this list because it had an effect that was somewhere between additive and recessive (by visual inspection). Together with Table 2.4, we can get a sense of which disease in each regions might be sharing susceptibility loci. I now comment on each locus in turn; some of these were already discussed in (WTCCC 2007), but I comment on all of them here for completeness.

Autoimmune diseases

rs6679677 (chromosome 1) The strongest joint association observed in the analysis. This is the known *PTPN22* shared locus for RA and T1D. A strong association is also observed in both T1D and RA individually. The association in CD is much weaker and in a different direction, so the effect of this locus in CD, if any, is likely to be through a different mechanism than in RA and T1D. The association with CD has now been replicated in other studies (BARRETT ET AL. 2008).

rs16845023 (chromosome 2) A moderate association for the autoimmune combined case group. It lies in the *LRP1B* gene, which is preferentially inactivated in non-small cell lung cancer, but which doesn't seem to have any known connection to autoimmune diseases. The relative risk is approximately 1.16 in all three diseases, a very small effect that only becomes statistically significant when the three disease groups are combined. Figure 2.9 shows a signal plot for this region.

rs10015924 & rs17388568 (chromosome 4) This region, which shows a moderate association for the autoimmune combined case group, was highlighted in the WTCCC study for showing a strong association for T1D in the multi-locus analysis. This region contains a few genes, including *IL-2* and *IL-21*, which have shown to be involved in autoimmune diseases

Table 2.5: **Estimated relative risks for regions with an additive effect.** Regions from Table 2.4 with an additive test p-value of at least moderate strength. For each, the estimated relative risk and 95% confidence interval is shown for the combined case groups and for each case group individually. For each SNP the allele coding that gives a relative risk greater than 1 is chosen. The first table shows the regions for the autoimmune diseases and the second for the metabolic/cardiovascular diseases. Relative risks that correspond to strong associations are in bold. The two SNPs on chromosome 4 are in moderate LD ($r^2 = 0.3$).

			Relative risk				
Chromosome	**Position (Mb)**	**SNP**	**CD+RA+T1D**	**RA+T1D**	**CD**	**RA**	**T1D**
1	114.0	rs6679677	**1.53 (1.38–1.69)**	**1.91 (1.72–2.12)**	0.77 (0.66–0.90)	**1.90 (1.68–2.15)**	**1.92 (1.70–2.17)**
2	141.6	rs16845023	1.16 (1.08–1.24)	1.15 (1.07–1.25)	1.16 (1.05–1.27)	1.16 (1.06–1.27)	1.15 (1.05–1.26)
4	123.4	rs10015924	1.15 (1.08–1.23)	1.19 (1.11–1.28)	1.08 (0.99–1.18)	1.15 (1.06–1.26)	**1.22 (1.12–1.33)**
4	123.7	rs17388568	1.18 (1.10–1.27)	1.20 (1.11–1.30)	1.14 (1.04–1.25)	1.14 (1.04–1.25)	1.26 (1.15–1.38)
6	138.0	rs2327832	1.15 (1.07–1.24)	1.20 (1.11–1.30)	1.05 (0.95–1.16)	1.23 (1.12–1.36)	1.17 (1.06–1.28)
10	6.1	rs2104286	1.16 (1.08–1.24)	**1.24 (1.15–1.34)**	1.00 (0.91–1.10)	1.24 (1.13–1.36)	1.23 (1.13–1.35)
12	110.9	rs17696736	**1.22 (1.15–1.30)**	**1.26 (1.17–1.35)**	1.15 (1.05–1.25)	1.13 (1.04–1.23)	**1.39 (1.28–1.50)**
18	12.8	rs2542151	**1.26 (1.16–1.37)**	1.21 (1.11–1.33)	1.35 (1.21–1.50)	1.14 (1.02–1.27)	1.29 (1.16–1.43)
22	35.9	rs743777	1.15 (1.07–1.23)	1.20 (1.11–1.29)	1.04 (0.94–1.14)	1.22 (1.12–1.34)	1.18 (1.08–1.28)

			Relative risk			
Chromosome	**Position (Mb)**	**SNP**	**CAD+HT+T2D**	**CAD**	**HT**	**T2D**
2	226.9	rs2943634	1.17 (1.09–1.25)	1.22 (1.12–1.33)	1.13 (1.03–1.23)	1.16 (1.06–1.26)
9	22.0	rs10965219	1.17 (1.10–1.25)	1.32 (1.21–1.43)	1.07 (0.99–1.16)	1.14 (1.05–1.24)
14	54.3	rs709939	1.16 (1.09–1.24)	1.13 (1.04–1.22)	1.20 (1.10–1.30)	1.17 (1.07–1.26)

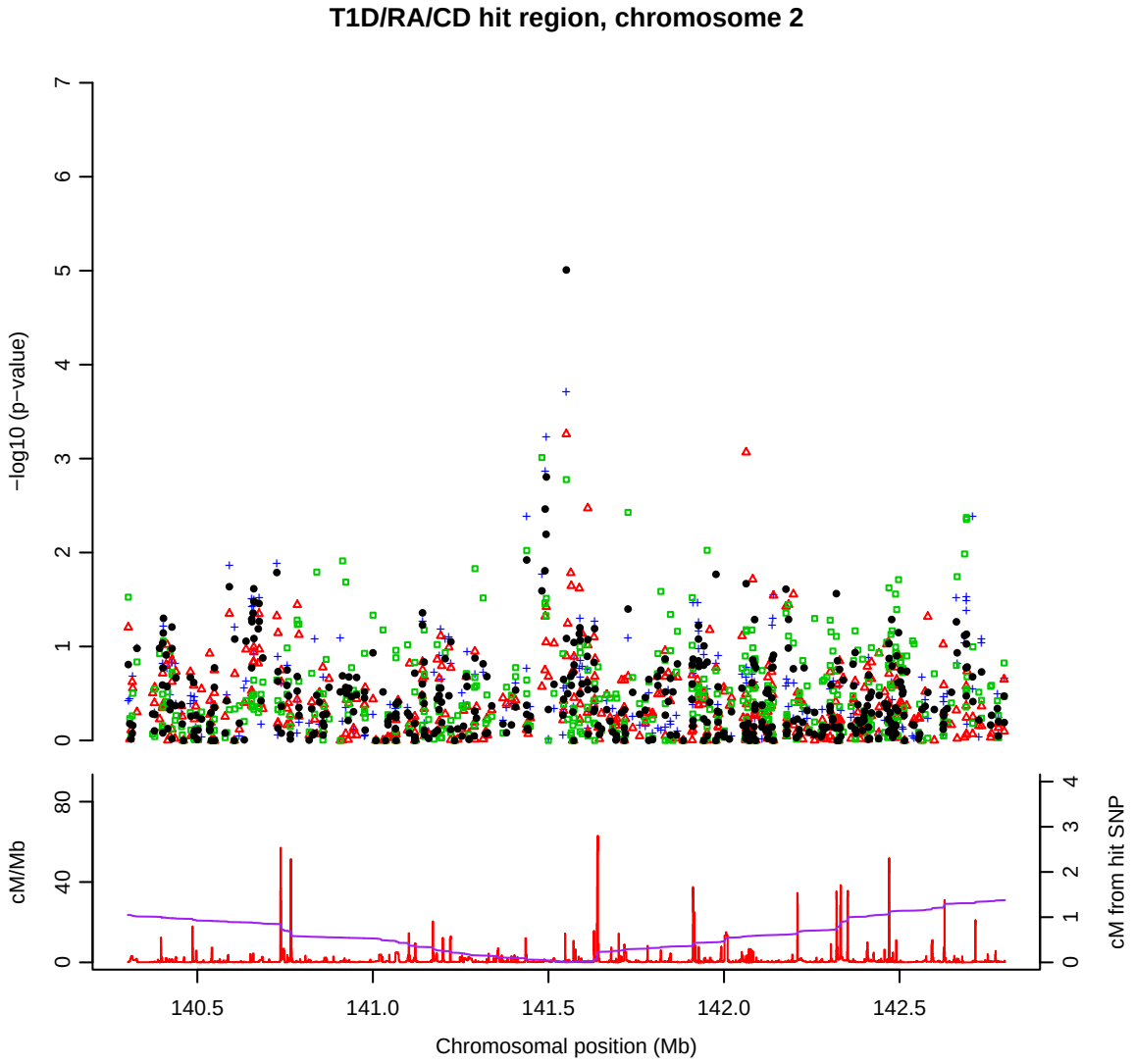

Figure 2.9: **Signal plot for rs16845023 (chromosome 2) in CD+RA+T1D using the general test.** Each point represents one SNP. Black filled circles are for the combined case group, green squares for CD, red triangles for RA and blue crosses for T1D. Estimated fine-scale recombination rate and cumulative genetic distance from rs16834421 are shown in the bottom panel. See Section 2.4.1 for more details.

in mice (YAMANOUCHI ET AL. 2007). The effect in CD and RA seems to be smaller, too small to show an association in those diseases individually, but it is at least in the same direction.

rs2327832 (chromosome 6) A region that shows moderate association for RA, and also RA and T1D combined. The effect in T1D is smaller than RA, and that in CD is very weak, but they are all in the same direction.

rs2104286 (chromosome 10) This region is in the *IL2RA/CD25* gene, which encodes the alpha chain of the IL2 receptor and is a known susceptibility locus for T1D and Grave's disease (BRAND ET AL. 2007, VELLA ET AL. 2005). Here it shows a moderate association with both T1D and RA, both with the same relative risk of about 1.24, which is encouraging evidence of a joint effect. There seems to be no effect at all in CD. A signal plot for this region is shown in Figure 2.10.

rs17696736 (chromosome 12) This region shows a very strong association with T1D, and weak association with both RA and CD. The T1D effect dominates in the joint analysis, but the effects in RA and CD, albeit weak, are in the same direction and about the same size, so are at least suggestive of a joint association. The region includes the *PTPN11* gene, which has not been shown to have a role in these diseases before, but is an attractive candidate because it has a major role in insulin and immune signalling (MUSTELIN ET AL. 2005) and is in the same family as *PTPN22* which is a known risk locus for both T1D and RA (see above).

rs2542151 (chromosome 18) This region is near that *PTPN2* gene, and shows a strong association in CD and a moderate association in T1D. The effect in RA is weaker, but it in the same direction as both of the other diseases. *PTPN2* is in the same family as *PTPN11* and *PTPN22*, which is highly suggestive that it plays a role in many autoimmune disorders.

rs743777 (chromosome 22) This region is near *IL2RB*, the beta chain of the IL2 receptor, which is thought to play an important role in preventing autoimmunity. The alpha chain *IL2RB*, on chromosome 10, is a known susceptibility locus for T1D and also shows a significant effect in RA (see above), which suggest this region is also likely to play a role. A

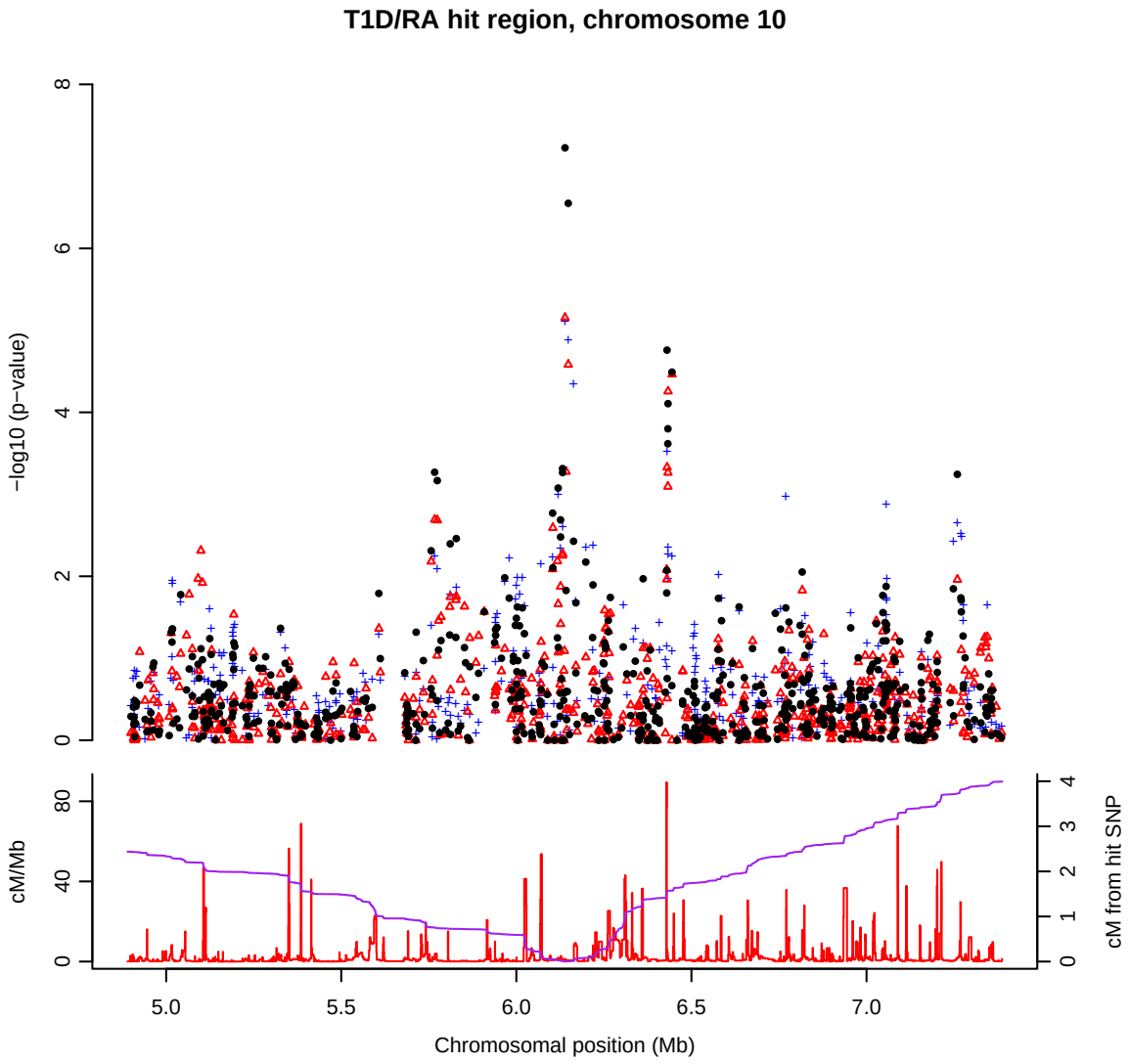

Figure 2.10: **Signal plot for rs2104286 (chromosome 10) in RA+T1D using the additive test.** Each point represents one SNP. Black filled circles are for the combined case group, red triangles for RA and blue crosses for T1D. Estimated fine-scale recombination rate and cumulative genetic distance from rs16834421 are shown in the bottom panel. See Section 2.4.1 for more details.

moderate association is observed in RA, and a weak association in T1D, both with relative risks of around 1.2. There seems to be very little or no effect in CD.

Metabolic/cardiovascular diseases

rs2943634 (chromosome 2) This region does not lie near any known genes, and shows a weak association in CAD. The effects in HT and T2D are even smaller but roughly equal and in the same direction.

rs10965219 (chromosome 9) This region shows a very strong association in CAD, and which was also recently reported by other studies (HELGADOTTIR ET AL. 2007, MCPHERSON ET AL. 2007). The effect in T2D is weaker, but has also been recently confirmed (ZEGGINI ET AL. 2007). The effect in HT is also weaker, but in the same direction.

rs709939 (chromosome 14) A region that shows weak association in HT, and similar effect sizes in all three metabolic/cardiovascular diseases. A few genes lie in the region, but none seem to be natural candidates for these diseases.

Non-additive effects

The remaining regions did not show an additive effect (the additive p-value is less than moderate strength). They were inspected to see what kind of genetic signal they had and whether they were consistent in all the individual case groups. They all showed a dominant/recessive signal, except for rs16989035 (chromosome 20) which displayed overdominance—but that SNP was identified as possibly affected by population structure. In all reported regions, the signal was consistent for all the diseases in their respective groups. The details of the signal types and corresponding relative risks are shown in Table 2.6.

All of these regions were not reported as showing any association signal in the WTCCC study, except for rs12581163 (chromosome 12) which has a weak association in T2D.

Of the two regions in the autoimmune diseases, neither lie near known genes and the first SNP, rs10812655 (chromosome 9), lies in a recombination hotspot. Both show a dominant

Table 2.6: **Genetic signal types & relative risks for regions with a non-additive effect.** Regions from Table 2.4 with an additive p-value of less than moderate strength. For each, the model that best suited each SNP was chosen by visually inspecting a plot of the log-odds for each genotype after fitting a general logistic regression model. The best fitting model and the corresponding relative risk estimates are shown, for both the combined case groups and each case group individually. The allele coding that gives a relative risk greater than 1 is chosen in each case. Most of these regions have low genotype counts for the 'smaller' genotype group (the protective homozygote in dominant models, the causal homozygote in recessive models), so the relative risks for the individual disease groups, particularly, are likely to be inaccurate. The first table shows the regions for the autoimmune diseases, and the second for the metabolic/cardiovascular diseases. *Relative risks are not shown for rs16989035, the only region not to show a dominant or recessive signal and also likely to be affected by population structure.

				Relative risk				
Chr.	**Pos. (Mb)**	**SNP**	**Model**	**CD+RA+T1D**	**RA+T1D**	**CD**	**RA**	**T1D**
9	27.7	rs10812655	dominant	1.38 (1.21–1.57)	1.41 (1.23–1.63)	1.31 (1.10–1.56)	1.41 (1.18–1.68)	1.42 (1.20–1.69)
16	61.2	rs4265819	dominant	2.65 (1.70–4.12)	3.25 (1.91–5.56)	1.88 (1.05–3.37)	2.50 (1.32–4.73)	4.54 (2.05–10.06)

				Relative risk			
Chr.	**Pos. (Mb)**	**SNP**	**Model**	**CAD+HT+T2D**	**CAD**	**HT**	**T2D**
1	149.5	rs16834421	recessive	1.77 (1.09–2.88)	1.60 (0.88–2.93)	1.80 (1.00–3.22)	1.90 (1.07–3.39)
2	2.9	rs1470614	dominant	6.24 (2.66–14.63)	7.26 (1.70–30.90)	3.67 (1.26–10.68)	14.50 (1.95–107.65)
12	18.5	rs12581163	recessive	2.00 (1.50–2.67)	2.00 (1.42–2.82)	1.92 (1.36–2.71)	2.08 (1.48–2.92)
20	38.5	rs16989035*	overdominant	——	——	——	——

signal, the first one being quite consistent among the diseases, while the second one is much stronger in T1D than the other two.

Two of the four regions for the metabolic/cardiovascular diseases are likely to be spurious: rs1470614 (chromosome 2) shows a dominant signal but the genotype counts for the protective homozygote are very low (22 in controls, 7 in the combined case group), while rs16989035 (chromosome 20) is likely to have population structure bias due to strong geographic variation. The remaining two regions show quite a consistent signal across the three diseases, both of them recessive. A signal plot for rs16834421 (chromosome 1) is shown in Figure 2.11.

2.5.2 Discussion

The combined cases analysis has confirmed previously known joint susceptibility loci and produced strong evidence for novel joint associations in autoimmune diseases. Evidence for joint associations in the metabolic/cardiovascular diseases was less strong. Most of the observed associations showed an additive genetic effect, while the other loci showed either a dominant or recessive effect. For the most part, moderate associations in the combined case groups did not show a strong or moderate association when each disease was analysed separately, showing that pooling the data across diseases was necessary to detect the weak effects present (assuming they are real).

The discovery of these joint associations is interesting and encouraging. It suggests that we can use the evidence from different but similar diseases to help boost power to detect associations for the diseases individually. As well as doing a combined analysis as we have here, a case can be made that any associations found for one disease should be checked for association in the other similar diseases as well. In the other diseases, evidence that might have looked *a priori* weak would now be looked at more favourably. Equivalently, a strong association at a locus in a similar disease is extra evidence that such an association might also exist in the disease of interest.

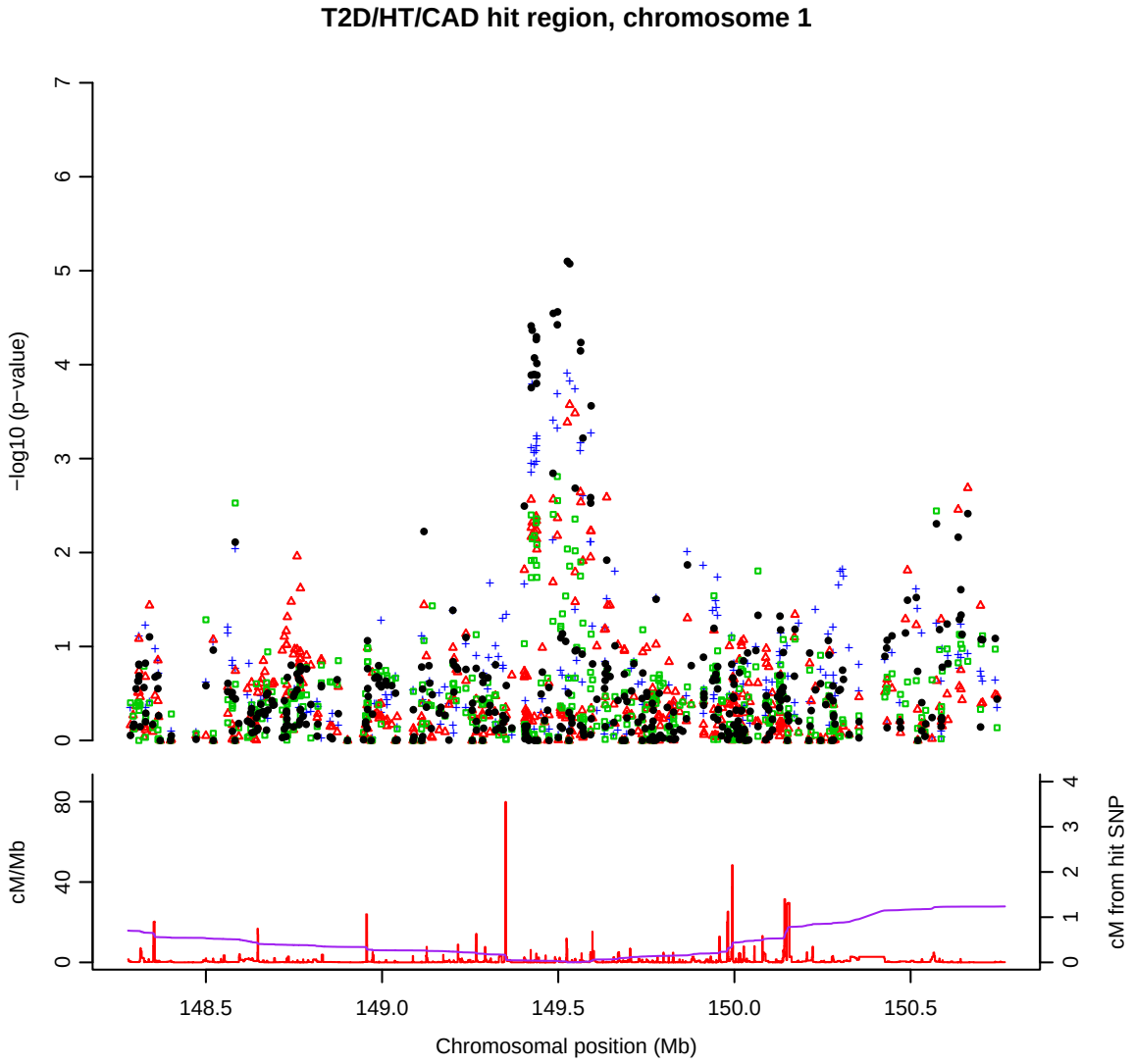

Figure 2.11: **Signal plot for rs16834421 (chromosome 1) in CAD+HT+T2D using the general test.** Each point represents one SNP. Black filled circles are for the combined case group, green squares for CAD, red triangles for HT and blue crosses for T2D. Estimated fine-scale recombination rate and cumulative genetic distance from rs16834421 are shown in the bottom panel. See Section 2.4.1 for more details.

Table 2.7: **Sample sizes & sex ratios of WTCCC samples.** These are after exclusion of individuals as described in Section 2.1.

	Collection	Sample size	Males (%)	Females (%)
Controls	58C	1,480	50	50
	UKBS	1,458	48	52
Cases	BD	1,868	37	63
	CAD	1,926	79	21
	CD	1,748	39	61
	HT	1,952	40	60
	RA	1,860	25	75
	T1D	1,963	51	49
	T2D	1,924	58	42

2.6 Sex-differentiated analysis

Genetic effects that act differently in males and females have been found in animal models (MACKAY & ANHOLT 2006), so we should also expect to find such effects in humans. No such effects have previously been found for a common human disease (except for disorders that are clearly sex-related like breast cancer), but studies have not yet been sufficiently powered to find such effects. The large sample sizes provided by the WTCCC and subsequent large GWAS now allow us to search for such effects with much greater power.

To find sex-differentiated effects in the WTCCC study, we performed a genome-scan using a test sensitive to effects that differ in the two sexes as well as effects that are the same (more details below). Effects of the latter type would already have been highlighted by the standard analysis (which is better powered to detect them anyway), so we focused only on newly identified SNPs.

The sex ratio for each WTCCC collection is shown in Table 2.7. In cases it varies from a roughly even split (e.g. T1D) to being dominated by one of the sexes (e.g. CAD is 79% males, RA is 75% females), due to the disease being more prevalent in that sex. In controls it is an even split.

The availability of X chromosome data enabled us to verify the reported sex of each individual in the study. We used the *heterozygosity* (the proportion of SNPs called as heterozygotes) of each individual on the X chromosome as an indicator of sex. Males only have one copy of the X chromosome, so should all be called as homozygotes and so have zero heterozygosity. In practice, the heterozygosity was generally observed to be non-zero due to genotyping

errors, but usually no more than a few percent. Females showed heterozygosity varying roughly between 20% and 30%. A number of individuals were observed to show heterozygosity not typical for their reported sex. These discrepancies were verified with the research groups who collected the DNA samples and amended where possible. The source of the discrepancy could not be discerned for approximately 80 individuals, but these were retained in the study on the grounds that it was unlikely that any mishandling would have introduced samples with a very different phenotype. Since the correct sex could be determined using the X chromosome, these were also retained for this analysis.

The sex-differentiated analysis in the WTCCC study reported regions with strong associations only. Two such regions were found, which I analysed further as part of the study to determine the nature of their effects. One of these (a female-only effect in RA) initially looked promising but unfortunately has failed to replicate in subsequent studies. The other one looks likely to be a false positive and has not been further pursued. I discuss both of these in more detail below.

Following the WTCCC study, I extended the sex-differentiated analysis to find moderate associations, developed a procedure to determine whether the detected associations show a genuine sex-differentiated effect, and further examined each highlighted region to determine the nature of the effect. I now describe and report the results of this analysis.

2.6.1 Methods

In the context of logistic regression models, there are a few choices of how to incorporate sex into the model. Considering the additive model, the baseline and additive effect parameters can each be allowed to vary based on sex, giving the following four possibilities:

$$\begin{aligned}
\mathrm{logit}(p) &= \mu \ + \beta G\,, && (2.1)\\
\mathrm{logit}(p) &= \mu_s + \beta G\,, && (2.2)\\
\mathrm{logit}(p) &= \mu \ + \beta_s G\,, && (2.3)\\
\mathrm{logit}(p) &= \mu_s + \beta_s G\,, && (2.4)
\end{aligned}$$

where $s \in \{\mathrm{male}, \mathrm{female}\}$ is a sex indicator. Equation (2.1) gives the standard additive model that ignores sex. Equation (2.2) allows different baselines for each sex but a common addi-

tive effect. This can be useful to prevent confounding due to sex, in the situation where we have different sex ratios in cases and controls. However, it is not of interest for this analysis because we seek loci where the genetic effect differs in sexes. Equation (2.3) allows for such different effects but retains a common baseline. Intuitively, this would appear to make sense if we had matching sex ratios in cases and controls, or equivalently, matching case-control ratios for each sex. However, this model assumes matching ratios specifically when $G = 0$, thus depending on the somewhat arbitrary genotype coding. Furthermore, the baseline does not correspond exactly to the overall case-control ratio and needs to be able to vary away from this nominal (null) value to accommodate any genetic effect. Thus, it would be more natural to allow the baseline to vary by sex as well (effectively, letting the model decide exactly where the ratios should match, if at all). This leads to the model in equation (2.4), which is the one we use. Comparing it with the sex-stratified null model, $\mathrm{logit}(p) = \mu_s$, gives the test,

$$H_0 : \beta_{\text{male}} = \beta_{\text{female}} = 0 \quad \text{vs} \quad H_1 : \beta_{\text{male}} \neq 0, \beta_{\text{female}} \neq 0\,.$$

We implement this as a score test and refer to it as the *sex-differentiated additive test*. The *sex-differentiated general test* is defined analogously. Under the null, the test statistics will have χ^2_2 and χ^2_4 distributions respectively.

Note that these models are equivalent to simply fitting the standard additive and general models in males and female separately. We exploit this fact in our implementation—we partition our data by sex, fit separate models in each, then add the test statistics together on the χ^2 scale. For example, the test statistics for each sex under the additive test will have independent χ^2_1 distributions.

I used the following procedure for my analysis:

1. Find all SNPs that show at least a moderate association on either the additive or general sex-differentiated tests.

2. Inspect cluster plots to eliminate all SNPs with genotyping calling errors.

3. Group SNPs into regions based on proximity, and select the SNP with the lowest p-value to act as a representative SNP. For each region, determine whether it has already been highlighted in any of the other analyses conducted in the main study, with either

a strong or moderate association. (Up to here, this is the same procedure used in the WTCCC study.)

4. We now wish to discard any regions that don't show a genuine differential sex effect, especially those that have already been highlighted in a different analysis and will already be earmarked for further study. Do this by fitting a full logistic regression with sex and genotype effects on the representative SNP and test for a significant interaction term (i.e. fit the general model with a sex effect, and test whether the genotype effect parameters are equal for both sexes). I used a p-value threshold of 0.05 for 'old' regions (already highlighted in another analysis), and a less stringent 0.10 for the 'new' regions.

5. The list now consists of all of regions that show at least a moderate association on the sex-differentiated tests and where this seems to be due to a differential effect in the two sexes. Since some false positives are expected in this list, I looked for two further pieces of evidence for each region:

   (a) The presence of SNPs in high to moderate LD with the representative SNP that show the same type of qualitative disease effect. I visually inspected all SNPs passing a threshold of $r^2 > 0.5$ in HapMap CEU to identify such SNPs (see below).

   (b) At least a moderate association under the sex-differentiated tests when the representative SNP is tested using the expanded reference group as controls.

The choices of p-value threshold for the interaction test was slightly arbitrary. The 0.05 value is a commonly used threshold and seemed to be adequate for this analysis. Most of the new regions had very low interaction p-values, so any other plausible threshold would have retained them. Only two had p-values greater than 0.05, but not much greater. Further inspection of these led to the decision to keep just one of these, and choosing a threshold of 0.1 conveniently did this. (None of the strong associations observed in the standard analysis showed a significant interaction.)

There are two ways that a SNP could be a false positive on this analysis: due to an error in genotyping or due to a chance association. The two further pieces of evidence noted for each region are an attempt to detect both of these.

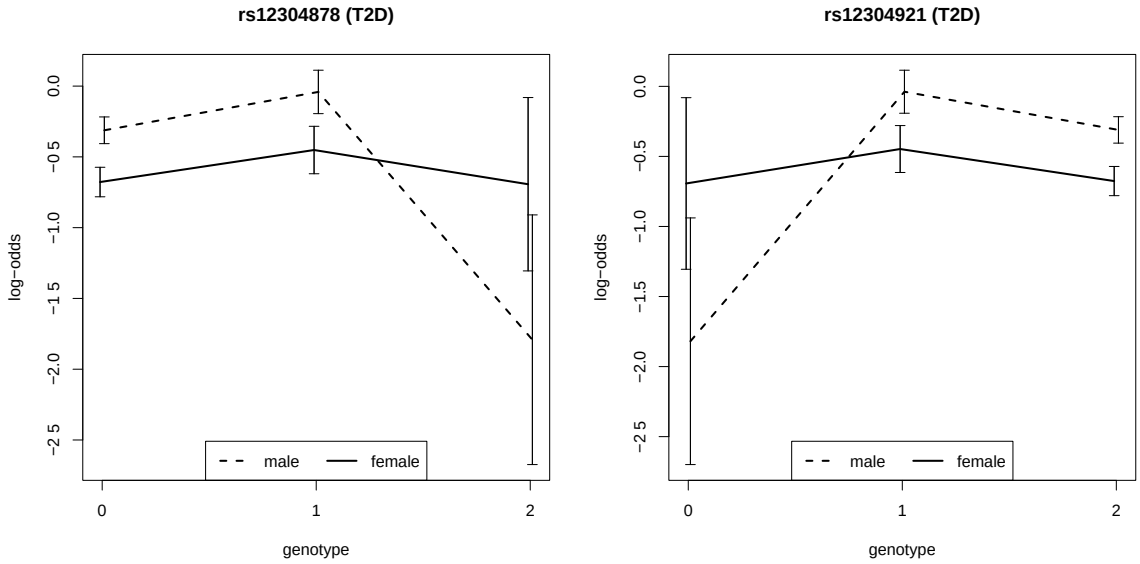

Figure 2.12: **Log-odds plot examples.** Plots of estimated log-odds for each sex-genotype combination with 95% confidence intervals. These two SNPs are from the analysis of T2D on chromosome 12 and have $r^2 = 1$ in HapMap CEU. They show the same type of association signal but with opposite allele codings.

While I already eliminate genotype calling errors by cluster plot inspection, technical artefacts prior to this in the genotyping process (e.g. in the laboratory) will most likely be discovered with the first piece of evidence—if the observed association is due to such an artefact, the immediate haplotype background will most likely show a different type of effect to the SNP. To judge whether SNPs have the same qualitative type of effect, for each SNP I inspected plots of the estimated log-odds for each sex-genotype combination. I will refer to these as *log-odds plots*. Figure 2.12 shows two examples. These make it easy to visually assess the type of effect at each SNP. If all SNPs in moderately high LD displayed similar-looking plots (up to reflection, since the corresponding allele coding might be in reversed in some SNPs), that region was deemed to pass the test. While none of the regions shown here failed this test, some of them had no SNPs in high enough LD to conduct it. The lack of such possibly confirmatory SNPs should not rule out a region, but should be an indication that it has not had the same level of quality control applied to it as the rest.

The second piece of evidence indicates which SNPs are less likely to be chance associations, since increasing the controls is expected to give us greater power to detect an association.

2.6.2 Results

Twenty-nine regions showed at least a moderate sex-stratified association, these are shown in Table 2.8. Eight of these still show a moderate association with the use of the expanded reference group.

The regions show a variety of genetic effects: an effect only in one sex, different types of effects in each sex (e.g. additive vs recessive) and effects in different directions in the two sexes. I examined the log-odds plot for each SNP and in Table 2.9 report the best model from visual assessment and the corresponding p-value.

Two regions showed a strong association and were reported in the WTCCC study:

- rs11761231 (chromosome 7) for RA,
- rs3792048 (chromosome 2) for CD.

The SNP rs11761231 gave a p-value of 3.91×10^{-7} on the sex-differentiated additive test. Figure 2.13 shows the log-odds plot for this SNP. It appears as if the signal is due to a female-only effect: the additive test in females gives a p-value of 5.66×10^{-8} with a RR of 1.35 (1.21–1.51), while in males the p-value is 0.68 with a RR of 1.03 (0.89–1.20). The SNP is surrounded by recombination hotspots on either side and there are no other SNPs in even moderate LD ($r^2 > 0.1$) with this SNP to use as corroborating evidence, so some caution is required. Given that RA is 2–3 times more common in females than males (ADVIWARE PTY LTD 2007), a susceptibility locus that only acts in females is certainly a plausible finding. Unfortunately, three subsequent studies have failed to replicate this association (THOMSON ET AL. 2007, FUNG ET AL. 2009, KORMAN ET AL. 2009).

The SNP rs3792048 showed a p-value of 2.13×10^{-7} on the sex-differentiated general test, and only 0.038 on the sex-differentiated additive test (on the standard tests the corresponding p-values are 0.028 and 0.016). This is due to a strong differential effect in males for the rare homozygote (coded as 0) as compared to the other genotypes. However, this homozygote is very rare and has a sex bias in controls (1 male, 11 females), so this looks like it might be a chance effect. Using the expanded reference group as controls removes the sex bias (33 males, 35 females) and also reduces the signal, as shown in Figure 2.14. Taking just the individuals with the other two genotypes and fitting a full logistic regression model with

Table 2.8: **Sex-differentiated analysis results.** Regions with at least one SNP showing a strong or moderate association on the sex-differentiated additive or general tests and also a significant sex-genotype interaction. A representative SNP is shown for each region. Regions in bold show a strong association and were reported in the WTCCC study. SNPs that were reported in *other* analyses in the WTCCC study are marked with 1–3, representing the strength of the reported association, corresponding respectively to strong–weak. The extra evidence reported for each region are whether there are SNPs in high r^2 that show the same type of association, and whether a moderate association is observed when using the expanded reference group (ERG) as controls. None of these regions contain SNPs identified as having strong geographic variation.

		Pos.		In	p-value		Extra evidence	
Case	**Chr.**	**(Mb)**	**SNP**	**other**	**Additive**	**General**	r^2	**ERG**
BD	2	241.2	rs2953146		4.39×10^{-5}	2.08×10^{-6}	y	
	5	142.9	rs13158686		2.01×10^{-4}	1.56×10^{-6}	y	y
	14	78.3	rs10133425		9.23×10^{-6}	5.63×10^{-5}	y	
CAD	3	161.1	rs1109156		9.19×10^{-6}	5.90×10^{-5}	y	
	5	135.4	rs30756		6.90×10^{-6}	1.43×10^{-5}	y	
CD	**2**	**105.4**	**rs3792048**		3.80×10^{-2}	2.13×10^{-7}	y	
	6	20.8	rs6908425	2	2.54×10^{-5}	6.48×10^{-6}		
	6	32.8	rs7775228		2.82×10^{-5}	8.31×10^{-6}	y	y
	7	90.4	rs879428	3	1.66×10^{-6}	1.66×10^{-6}	y	
	11	115.9	rs12362410	3	1.84×10^{-3}	7.11×10^{-6}	y	
	21	26.4	rs2234988	3	9.42×10^{-6}	1.07×10^{-4}	y	
HT	3	28.7	rs4399848	3	1.24×10^{-6}	3.64×10^{-6}		
	4	16.1	rs4698483		2.40×10^{-6}	6.58×10^{-7}		y
	4	145.5	rs13123791		9.10×10^{-3}	5.40×10^{-6}	y	y
	16	64.1	rs1559344		2.60×10^{-6}	1.69×10^{-5}	y	
	20	42.2	rs878559		8.70×10^{-6}	3.59×10^{-5}	y	
	20	61.2	rs12625378		9.82×10^{-6}	1.21×10^{-4}	y	
RA	**7**	**130.8**	**rs11761231**		3.91×10^{-7}	1.37×10^{-6}		y
	15	96.3	rs923658		2.32×10^{-3}	8.29×10^{-7}	y	
T1D	1	242.9	rs10924730		3.28×10^{-6}	2.36×10^{-5}		
	3	46.3	rs6441961	3	6.79×10^{-6}	2.48×10^{-5}	y	y
	5	32.0	rs6861526		7.80×10^{-3}	6.68×10^{-6}		
	11	116.9	rs4938390	3	9.27×10^{-6}	1.80×10^{-6}		
	17	36.0	rs7221109	3	2.16×10^{-6}	1.93×10^{-5}	y	y
T2D	2	60.5	rs243018	3	6.67×10^{-6}	4.93×10^{-5}	y	
	4	176.1	rs6851555		8.64×10^{-4}	7.12×10^{-6}		
	8	10.1	rs4448276		8.23×10^{-6}	5.35×10^{-5}	y	
	11	1.8	rs7932087		1.61×10^{-4}	2.43×10^{-6}	y	
	12	49.6	rs12304921	2	1.09×10^{-1}	4.28×10^{-6}	y	y

Table 2.9: **Disease effect types for sex-differentiated regions.** The type of disease effect for each SNP in Table 2.8. The p-value corresponds to the sex-model combination that highlights the effect and the description is from visual assessment of the log-odds plot. For SNPs labelled 'm+f', the p-value is for the sex-differentiated test.

Case	Chrom.	Pos. (Mb)	SNP	Sex	Model	p-value	Effect type
BD	2	241.2	rs2953146	f	general	9.39×10^{-7}	dominant in females
	5	142.9	rs13158686	m+f	general	1.56×10^{-6}	opposite recessive: risk in males, protective in females
	14	78.3	rs10133425	m+f	additive	9.23×10^{-6}	opposite additive
CAD	3	161.1	rs1109156	f	additive	7.40×10^{-7}	additive in females
	5	135.4	rs30756	m	additive	9.89×10^{-7}	additive in males
CD	**2**	**105.4**	**rs3792048**	m	general	7.64×10^{-6}	recessive in males
	6	20.8	rs6908425	m+f	general	6.48×10^{-6}	dominant in males, additive in females
	6	32.8	rs7775228	f	additive	5.63×10^{-6}	additive in females
	7	90.4	rs879428	m	additive	1.52×10^{-6}	additive in males
	11	115.9	rs12362410	m+f	general	7.11×10^{-6}	recessive in females, opposite underdominant in males
	21	26.4	rs2234988	f	additive	1.81×10^{-6}	additive in females
HT	3	28.7	rs4399848	f	additive	1.27×10^{-7}	additive in females
	4	16.1	rs4698483	f	additive	9.34×10^{-7}	additive in females
	4	145.5	rs13123791	m+f	general	5.40×10^{-6}	opposite recessive: risk in females, protective in males
	16	64.1	rs1559344	m	additive	4.53×10^{-7}	additive in males
	20	42.2	rs878559	m	additive	1.73×10^{-6}	additive in males
	20	61.2	rs12625378	f	additive	1.50×10^{-6}	additive in females
RA	**7**	**130.8**	**rs11761231**	f	additive	5.66×10^{-8}	additive in females
	15	96.3	rs923658	m	general	9.19×10^{-6}	overdominant in males
T1D	1	242.9	rs10924730	f	additive	1.56×10^{-6}	additive in females
	3	46.3	rs6441961	f	additive	4.13×10^{-6}	additive in females
	5	32.0	rs6861526	m+f	general	6.68×10^{-6}	overdominant in females, underdominant in males
	11	116.9	rs4938390	f	additive	1.27×10^{-6}	additive in females
	17	36.0	rs7221109	f	additive	4.40×10^{-7}	additive in females
T2D	2	60.5	rs243018	m	additive	1.04×10^{-6}	additive in males
	4	176.1	rs6851555	f	general	9.22×10^{-6}	recessive in females
	8	10.1	rs4448276	f	additive	7.76×10^{-6}	additive in females
	11	1.8	rs7932087	m+f	general	2.43×10^{-6}	additive in males, opposite underdominant in females
	12	49.6	rs12304921	m	general	1.38×10^{-6}	overdominant in males

Table 2.10: **Sex-differentiated analysis, strong associations in a single sex.** Regions from Table 2.9 that show a strong association within a single sex group at the representative SNP.

Case	Chrom.	SNP	Sex	Model	p-value	Relative risk
HT	3	rs4399848	female	additive	1.27×10^{-7}	1.69 (1.39–2.07)
HT	16	rs1559344	male	additive	4.53×10^{-7}	1.42 (1.24–1.63)
RA	7	rs11761231	female	additive	5.66×10^{-8}	1.35 (1.21–1.51)
T1D	17	rs7221109	female	additive	4.40×10^{-7}	1.37 (1.21–1.54)

sex and genotype as explanatory variables results in a non-significant interaction term (p-value = 0.41), and a non-significant genotype main effect (p-value = 0.58). Thus, there is no evidence for a genetic signal here apart from the rare homozygote, for which the evidence is sketchy.

For regions where the effect is present only in one sex, Table 2.9 shows the p-value when fitting the best model in only that sex group. This is the correct test if we believe that model, since the data from the other sex will dilute the signal. Four regions show a strong association when calculated this way, all of them with the additive model. These are shown in Table 2.10, with corresponding log-odds plots in Figure 2.15.

The first of these, rs4399848 for HT, suffers the same problem as described for rs3792048 above. Very few individuals have the rare homozygote and there is a sex bias in the controls (8 males, 18 females). The association at this SNP should be more robust to such effects because we are fitting an additive model, since the heterozygotes will contribute to estimating the effect size. Indeed, excluding the rare homozygotes gives a p-value of 2.54×10^{-5} for the additive model and a relative risk estimate of 1.57 (1.27–1.94). While the p-value is still low, it is no longer as compelling.

The second of these, rs1559344 for HT, shows a more convincing association signal, this time in males only. There are no known genes in the region.

The third of these, rs11761231 for RA, was reported in the WTCCC and I have already discussed it above.

The last of these, rs7221109 for T1D, shows some promise. It is the only one of these four that has both of the extra pieces of evidence described earlier—SNPs in high LD that show the same type of signal and a moderate association when using the expanded reference group. It is also the only one that is near a known gene, *SMARCE1*, although this gene does not appear to have been implicated in any autoimmune disorders.

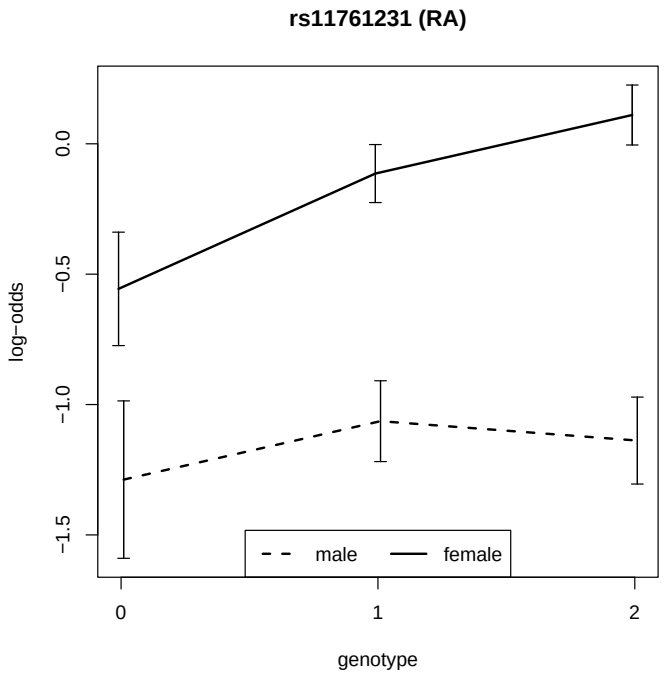

Figure 2.13: **Log-odds plot for rs11761231 (chromosome 7) in RA.** A female-only additive effect is observed. Note: the plotted lines correspond to the general model, allowing for a visual assessment of deviation from an additive model.

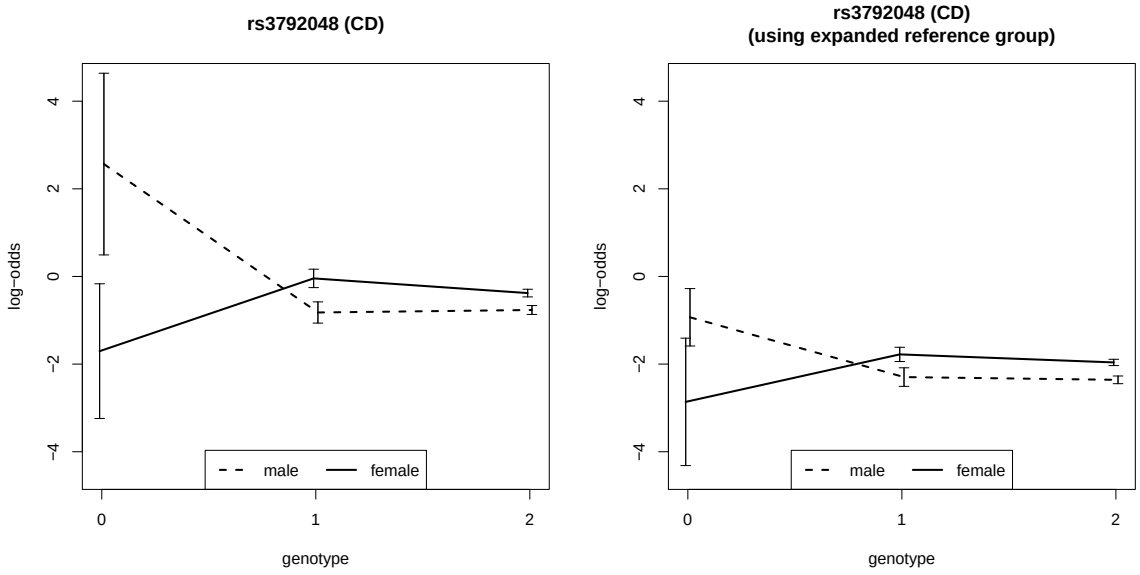

Figure 2.14: **Log-odds plots for rs3792048 (chromosome 2) in CD.** The differential effect for genotype 0 mostly disappears with the use of the expanded reference group.

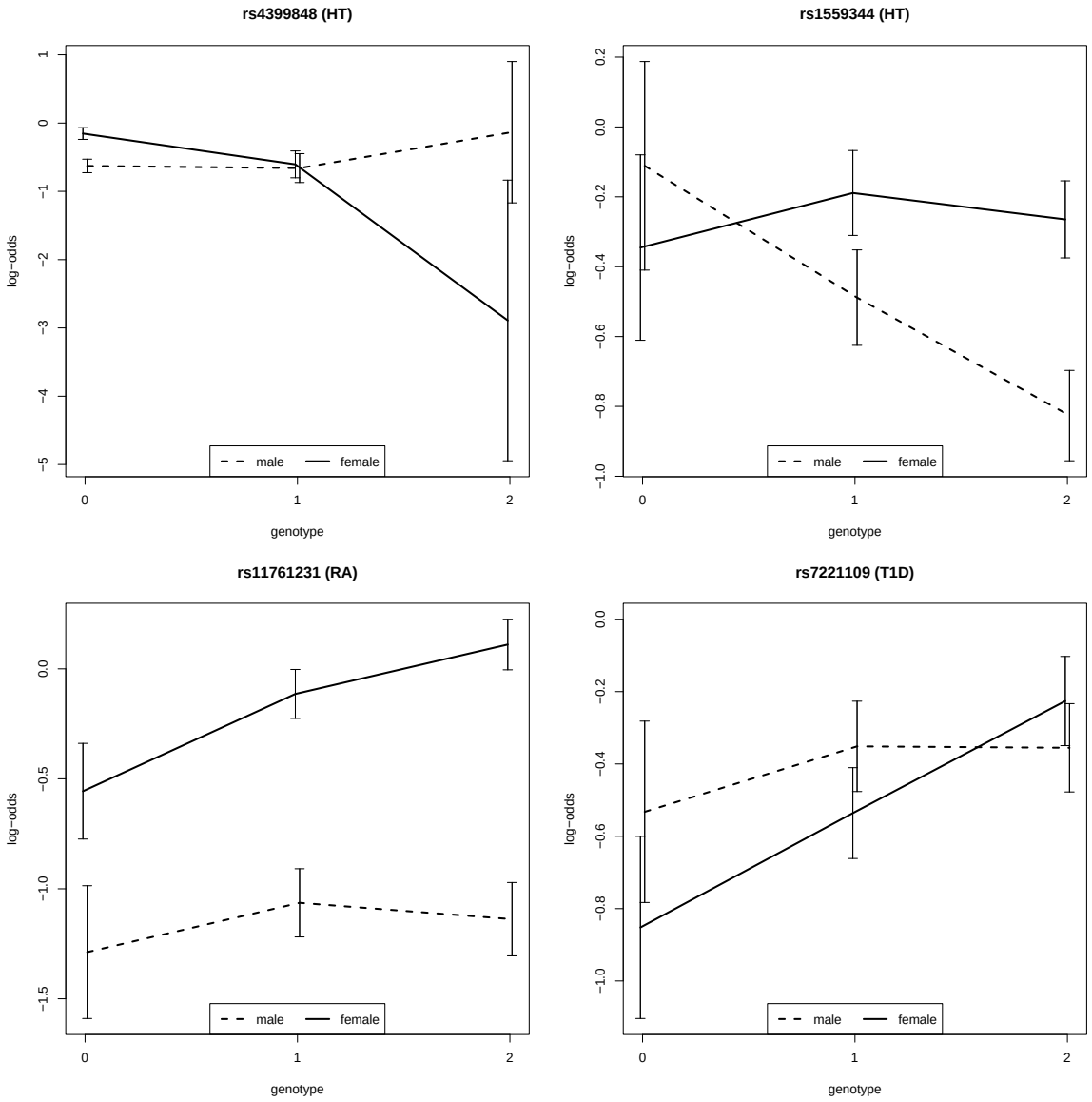

Figure 2.15: **Regions with a strong sex-specific association.** Log-odds plots for SNPs that show a strong association when performing the best association test (here, always the additive test) only on individuals of the sex that shows the signal. For rs159344 (top-right) the effect is in males, for the others it is in females.

To illustrate some effects that are in opposite directions or differ in character in the two sexes, Figure 2.16 shows log-odds plots for four regions I selected as examples. The first, rs1315868 in BD, shows a recessive effect that is harmful for males but protective for females. In HT, rs13123791 shows a similar opposite male/female recessive effect. In T1D, rs6861526 shows an overdominant effect in females but an underdominant one males. Finally, rs7932087 in T2D is an example of an additive effect in males and an underdominant one in females.

2.6.3 Discussion

The sex-differentiated analysis has sought to detect loci which act differently in males and females. While such loci have been observed in animal models, none have yet been discovered for common human diseases. The large samples collected by the WTCCC and subsequent GWAS have allowed such an analysis to be carried out, albeit at a reduced power as compared with effects that act similarly in both sexes.

The analysis has discovered 29 regions with at least moderate evidence of a sex-differentiated effect. While many of these are likely to be false positives, ancillary evidence supports many of the associations. In particular, the use of the expanded reference group maintained a positive result for 8 of these regions. Unfortunately, one of the strongest signals, rs11761231, showed good evidence for a female-only effect in RA but has failed to replicate in subsequent studies. While disappointing, a few other SNPs that also showed strong evidence remain to be explored.

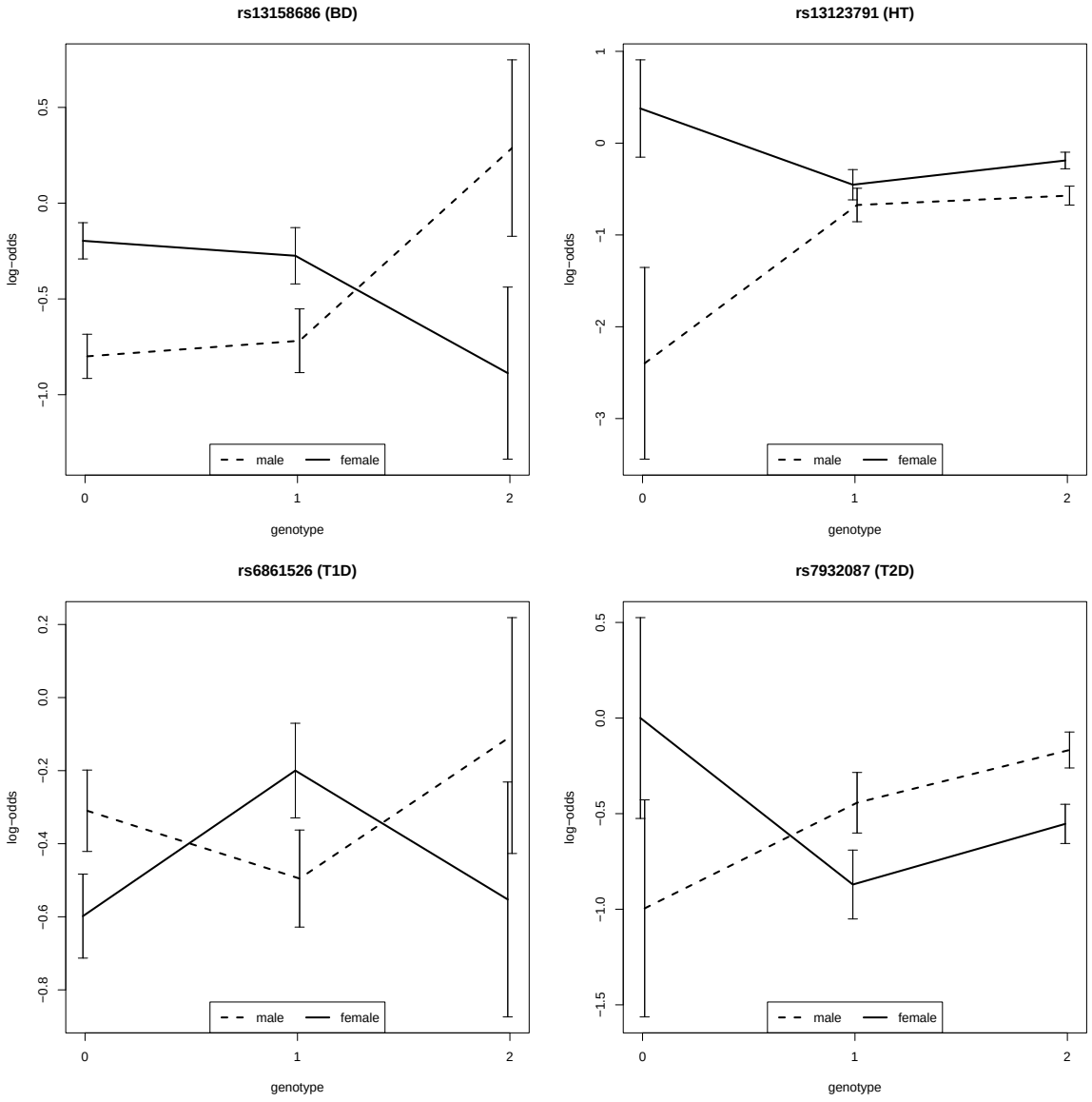

Figure 2.16: **SNPs with differential sex effects.** Log-odds plots for SNPs selected as examples showing different effects in males and females.

Chapter 3

GWAS of a Bipolar Disorder Sub-phenotype

Contents

In the WTCCC study, we observed a relative paucity of signals for bipolar disorder (BD), and also hypertension, compared to the other diseases. A number of possible explanations exist, including: phenotypic heterogeneity, unmeasured environmental factors, poorly tagged causal SNPs or effect sizes too weak for for the sample size used. The first of these explanations is particularly likely to apply for BD, since it is not a single disorder but a collection of mood disorders. Given the range of diagnostic criteria used to include cases in the WTCCC, it seems likely that phenotypic heterogeneity might at least partly explain the lack of convincing findings.

In this chapter, I undertake a GWAS on a subset of the BD cases from the WTCCC. The subset was provided by Professor Craddock and was chosen amongst the various clinically defined subtypes as being one that showed the strongest signal at the candidate gene *GABRB1*. Evidence of genetic effects at other loci were then of interest. With this more homogeneous

case subset, I find a number of loci that now show substantially greater evidence of association, highlighting the utility of using stricter phenotype definitions in order to isolate genetic effects.

This work has partly contributed to CRADDOCK ET AL. (2008), a candidate gene association study using the same subset of BD cases and looking for signals within the $GABA_A$ receptor gene family, of which *GABRB1* is a member.

3.1 Phenotype refinement

Diagnosis of BD involves a psychiatric interview where a trained psychologist or psychiatrist evaluates the patients' experiences and abnormalities in mood and behaviour according to a set of criteria. Patients can display different types of behaviour, satisfying different diagnostic criteria, and still be diagnosed with BD. See CRADDOCK ET AL. (2008) for a more detailed discussion of the criteria.

The BD cases from the WTCCC included a mix of the following subtypes which have been shown to co-aggregate in family studies (WTCCC 2007): bipolar I disorder (71% cases), schizoaffective disorder, bipolar type (SABP, 15% cases), bipolar II disorder (9% cases) and manic disorder (5% cases).

A phenotype optimisation procedure was conducted by the Craddock group, to select a subset of BD cases that are likely to be more biologically homogeneous. The hope is that by analysing only these cases we can detect genetic effects that would otherwise be masked when analysing the whole BD group. The optimisation procedure is described in CRADDOCK ET AL. (2008), but I briefly summarise it here for completeness.

Eleven phenotype subsets, not necessarily mutually exclusive, were considered for selection. These included the ones mentioned above but also alternate phenotype subset definitions from other diagnostic criteria. The SNP rs783021 was chosen as an 'index' SNP on which to base subset selection. This SNP showed a p-value of 6.2×10^{-5} for the additive test in the WTCCC study and lies within *GABRB1*, once of the $GABA_A$ receptor genes. These receptors play a key role in the central nervous system, have been implicated in anxiety and alcohol disorders, and have been hypothesised to also be involved in mood and psychotic

illnesses (CRADDOCK ET AL. 2008). Thus, they are strong candidate genes for BD. Using stepwise variable selection in a logistic regression with phenotype subtypes as predictors and the genotype at the index SNP as the response variable resulted in only one subset being retained: SABP, with 279 cases. The index SNP showed a stronger signal when comparing just this case group against the controls, with a p-value of 3.8×10^{-6} on the additive test.

3.2 Results

I conducted a GWAS that compared the 279 cases (BD SABP) with the 2,938 controls (58C & UKBS) from the WTCCC. I used the autosomal SNP genotypes from the WTCCC and the same quality control exclusion criteria. However, I did not exclude SNPs based on already conducted cluster plot inspections from the WTCCC, but instead examined each SNP afresh for this study. This was to avoid accidentally excluding SNPs that were observed to have genotyping problems only in collections other than the three I use here (58C, UKBS, BD). I performed two genome-scans, one for SNPs on the genotyping chip and one for imputed SNPs. I describe each of these in turn.

3.2.1 Standard analysis

I applied the additive and general tests to all SNPs passing the quality control criteria and took those with a p-value less than 1×10^{-5} on either test. I then inspected cluster plots for each and excluded those with poor genotypes. There were 216 SNPs passing the p-value thresholds and 53 remained after cluster plot inspection.

Most of the identified 53 SNPs clustered into small regions along the genome, but 18 were singletons. To be conservative, I removed the singletons and also those with MAF less than 0.05, leaving 8 regions. The best SNP from each is shown in Table 3.1.

Of all these regions, only the index region on chromosome 4 is highlighted in the WTCCC study. Given that we used it as the basis for selecting this sub-phenotype, its presence on the list is not surprising. However, the strongest signal in this region is at rs6414684, rather than at the index SNP (rs7680321) as would be expected. Interestingly, these two are in very

Table 3.1: **BD SABP associations, standard analysis.** Regions with at least one SNP showing a p-value less than 1×10^{-5} on the additive or general tests. A representative SNP is shown for each region, along with any genes in or near the region. The relative risk (RR) is shown with respect to the risk allele, which is not necessarily the minor allele. *Region contains the index SNP (rs7680321) used to select the SABP sub-phenotype, but this SNP is in low LD with the representative SNP shown here.

	Position		p-value		$\log_{10}(\mathbf{BF})$			
Chromosome	**(Mb)**	**SNP**	**Add.**	**Gen.**	**Add.**	**MAF**	**RR**	**Gene**
3	49.9	rs2352974	3.8×10^{-7}	2.1×10^{-6}	4.86	0.49	1.58 (1.32–1.89)	*TRAIP*
4	46.9	rs6414684*	1.2×10^{-6}	2.0×10^{-6}	4.37	0.50	1.55 (1.30–1.85)	*GABRB1*
4	78.6	rs12644949	6.3×10^{-6}	3.7×10^{-5}	3.62	0.22	1.55 (1.28–1.88)	(CCNG2 protein)
4	108.2	rs7667341	1.2×10^{-2}	2.1×10^{-7}	0.98	0.07	1.45 (1.08–1.95)	*DKK2*
5	76.4	rs13154602	1.5×10^{-6}	5.6×10^{-6}	4.21	0.28	1.55 (1.29–1.86)	*ZBED3*
6	84.3	rs1171115	4.4×10^{-6}	2.0×10^{-5}	3.74	0.27	1.54 (1.29–1.85)	*PRSS35*
15	87.4	rs16942644	6.6×10^{-7}	1.7×10^{-8}	4.34	0.11	1.81 (1.43–2.29)	*ABHD2*
21	40.0	rs4818065	2.3×10^{-7}	5.3×10^{-7}	4.83	0.19	1.68 (1.38–2.05)	*B3GALT5*

low LD ($r^2 < 0.1$ in HapMap CEU). In fact, in the region there are two separate clusters of correlated SNPs, of which these two are their best representatives. In the HapMap CEU, the r^2 between these two SNPs is less than 0.1, and the best r^2 between any two SNPs in the two clusters that were picked out in my analysis is 0.19. Fitting an additive model at the index SNP while including rs6414684 as a covariate gave a p-value of 0.019. Thus, there seem to be two independent signals acting in this region, which also has a strong candidate gene. The signal plots in Figure 3.1 show these two separate signals.

The other regions have not previously been implicated for any common diseases, nor do the genes in each show any clear functional relationship to psychiatric disorders. However, the two showing the strongest evidence of association, rs2352974 (chromosome 3) and rs4818065 (chromosome 21), both have a $\log_{10}(\mathrm{BF})$ greater than 4.8. This places them on par with some of the strongest associations reported in the main WTCCC study (BFs are directly comparable in this context, but p-values are not because sample sizes differ—see discussion in Sections 5.1 and 5.8). In addition, they show stronger signals than what is observed at the index region, which is already a strong candidate.

3.2.2 Imputation analysis

Taking all of the imputed SNPs passing quality control criteria from WTCCC, I applied the additive and general tests using *expected* genotype counts[1] and took those with a p-value less than 1×10^{-5} on either test. I then grouped these together into regions by pairing up SNPs that were less than 500 kb apart. This gave 44 regions (252 SNPs).

I was primarily interested in discovering regions that were not already highlighted by the standard analysis, so I excluded any region that was within 20 kb of a SNP on our genotyping chip that showed a p-value less than 1×10^{-5} on the first scan. This left 10 regions (61 SNPs).

Finally, I needed to check the quality of the genotype calls for each region. Cluster plot inspection is not possible at the majority of imputed SNPs, since they were not on the chip and

[1]Expected genotype counts are obtained by summing over the posterior genotype calls for each individual. This is an alternative to using genotype calls obtained by thresholding the posterior probability at a given value (a threshold of 0.9 was used in the WTCCC study). I used expected counts here in order to mirror the approach used in the WTCCC imputation analysis. However, note that this procedure has since been superseded by missing-data likelihood methods (MARCHINI ET AL. 2007).

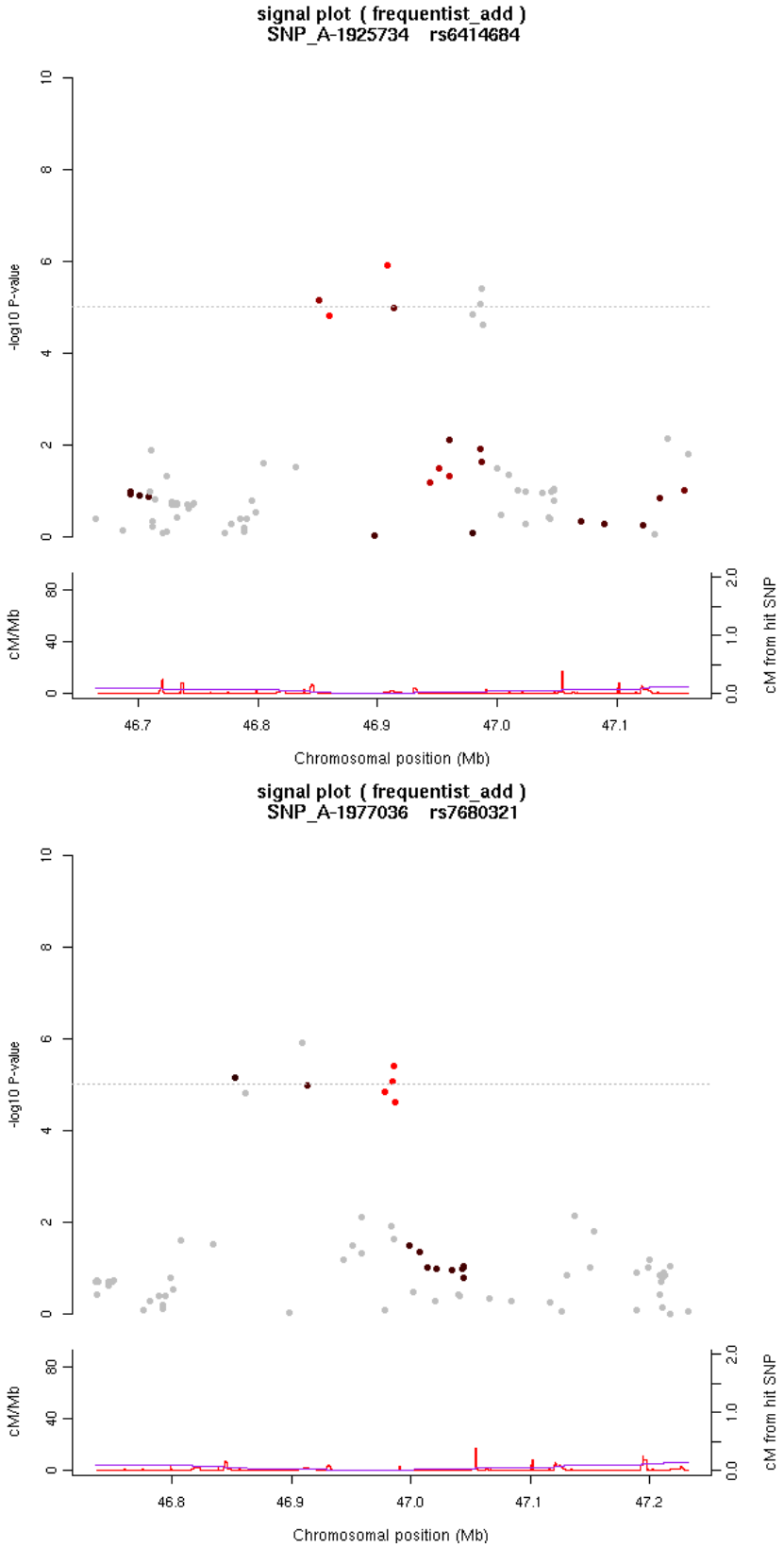

Figure 3.1: **Signal plots at the index region.** The bottom plot is at the index SNP (rs7680321), the top plot is at a nearby SNP (rs6414684) which shows an independent signal. See Figure 2.7 for a detailed description.

thus lack genotyping intensity data required to draw such plots. Instead I examined what I call an *imputation signal plot*; some examples are shown in Figure 3.2. These plots show the p-values for all imputed and non-imputed SNPs in a region, particularly highlighting any differences in p-value at SNPs on the chip between the two analyses. Large differences can indicate either poor initial genotype calls or poor imputed calls, either of which can lead us to doubt any association signal. A particularly telling example is top-right plot in Figure 3.2, where the two strongest SNPs are both on the chip and have 'traded places', one showing a big increase in p-value and the other a big drop. The most likely explanation is spurious signals due to poor genotype calls at the SNP with an initially low p-value. After imputation, the calls improve at this SNP and its p-value no longer shows a signal. Conversely, at the other SNP where the calls are fine initially, the imputed calls are poor due to being based on the artefactual calls of its neighbour, leading to a similarly spurious signal. (Remember that imputation works by, for each SNP, using the information from surrounding SNPs to infer the genotypes at that SNP.)

An imputation signal plot can give an indication of some problems, but is not as definitive as a cluster plot. I complemented it by looking at some other indicators of poor quality data. In particular, for each I region I checked for the presence of each of the following:

- **Low calling confidence.** The average posterior probability of the best genotype across individuals was less than 0.99 in any of the collections for the SNPs in the region.
- **Low MAF.** The MAF was less than 0.05 for the SNPs in the region.
- **Large changes in p-value at SNPs on the chip.** This was done visually from the imputation signal plots, looking for changes greater than about one degree of magnitude, especially near SNPs showing an association signal.
- **Many downward changes in p-value at SNPs on the chip.** Such changes are indicative of poor initial calls, with imputation 'cleaning them up', however they can therefore induce spurious signals at neighbouring SNPs when their poor calls are used as a basis for imputation. This was done visually.
- **'Lone' signals.** Where only one SNP show a signal. This was done visually.
- **Presence of copy number variants in or near the region.** This was done visually using the HapMap browser (`www.hapmap.org`).

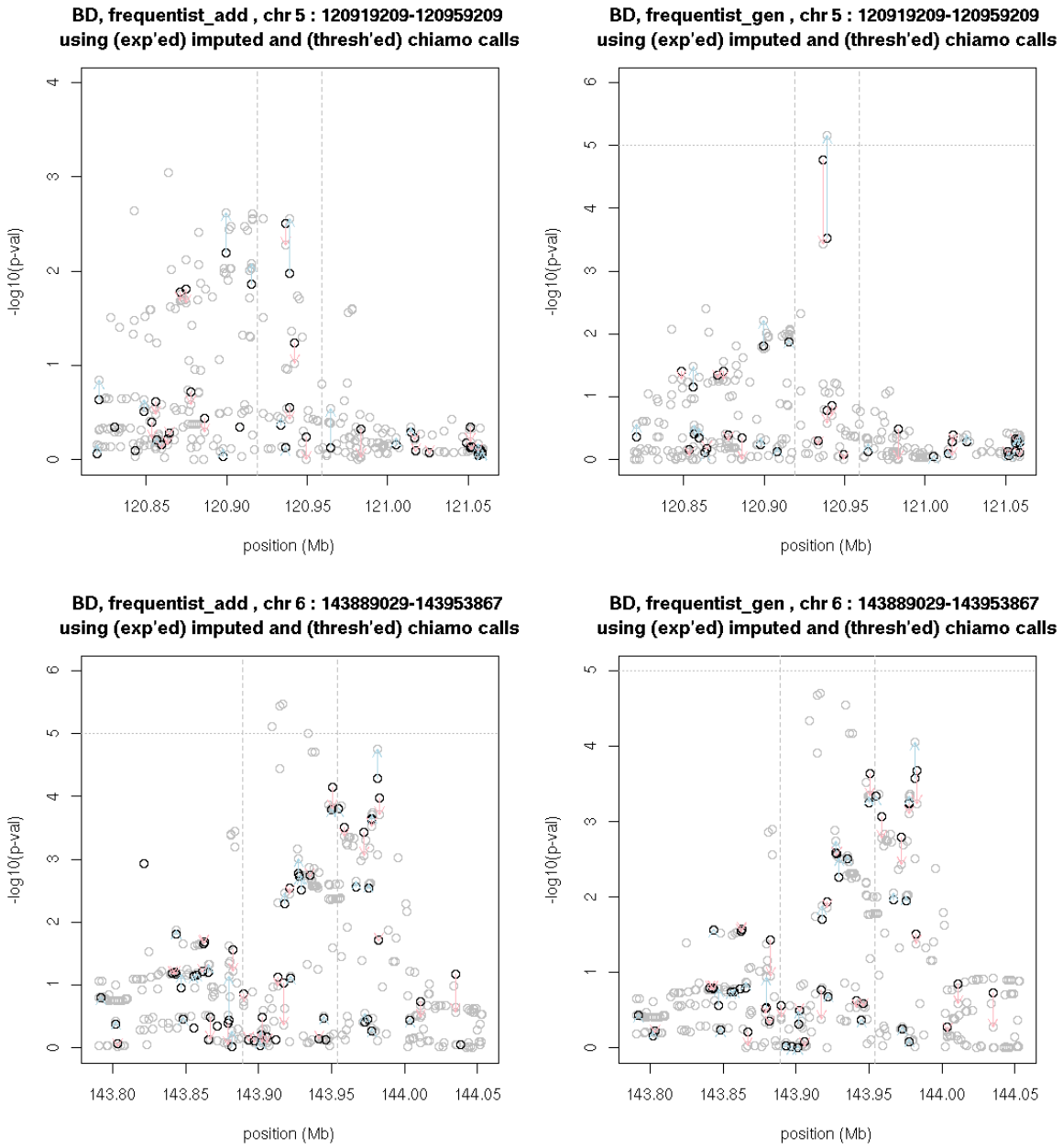

Figure 3.2: **Imputation signal plot examples.** Association test p-values (on a log scale) for two regions under the additive and general tests. Rows correspond to regions and columns to tests. For each, p-values are shown for every SNP in the region (after exclusions) using imputed and/or actual genotype calls, as grey and black points respectively. SNPs on the genotyping chip are shown twice (once with each set of calls) and their points are linked by arrows, coloured blue or red for respectively an increase or decrease in association signal following imputation.

Table 3.2: **BD SABP associations, imputation analysis.** Regions with at least one imputed SNP showing a p-value less than 1×10^{-5} on the additive or general tests but not highlighted by the standard analysis. A representative SNP is shown for each region.

	Pos.		p-value			
Chr.	(Mb)	SNP	Add.	Gen.	MAF	RR
6	143.9	rs9399446	3.4×10^{-6}	2.0×10^{-5}	0.31	1.53 (1.28–1.82)
8	73.7	rs4738269	1.6×10^{-2}	9.5×10^{-6}	0.15	1.32 (1.05–1.65)
8	127.7	rs9297741	2.3×10^{-3}	3.5×10^{-8}	0.05	1.66 (1.20–2.30)
14	23.2	rs1020089	7.5×10^{-1}	4.9×10^{-6}	0.11	1.04 (0.80–1.37)
21	44.3	rs915877	6.6×10^{-6}	3.7×10^{-5}	0.07	1.90 (1.43–2.52)

I excluded any region with two or more of the above indicators. This left 5 regions (20 SNPs). The best SNP from each is shown in Table 3.2.

Despite passing the above procedure, these five regions do not look very encouraging. The p-values are not overly compelling and three of them only show a signal on the general test, which we know is less robust. They should be treated with great caution, especially since these signals are only at imputed SNPs.

The only one of these regions that was highlighted in the WTCCC study is the chromosome 14 region represented by rs1020089. However, here the signal is only on the general test, while in the WTCCC study the signal was an additive one at rs221703. Furthermore, for the sub-phenotype analysis there are no appreciable signals at any of the SNPs on the Affymetrix chip in this region, so these two potential signals are unlikely to be related.

3.3 Discussion

The selection of BD cases in the WTCCC was based on a range of diagnostic criteria, possibly giving rise to phenotypic heterogeneity. This would potentially explain the relative lack of strong association findings when compared to the other diseases studied. The BD sub-phenotype analysis has sought to overcome this difficulty by determining a subset of cases which are more phenotypically similar and show an elevated genetic signal. The hope is that such individuals will be more likely to share the same genetic effects, and that we will be able to detect them.

Various diagnostic subsets of the BD cases were considered, and a particular one (SABP) was selected following a phenotypic refinement procedure based on a chosen 'index' locus. A SNP in *GABRB1* was used for this purpose. This gene showed a promising signal in the main WTCCC analysis and is a strong candidate for involvement with psychiatric disorders.

A genome-scan for SABP highlighted 8 regions with at least a moderate association signal. One of these was in *GABRB1*, containing the index SNP, and this region actually showed two independent signals. Another two regions showed strong evidence comparable to the headline findings from the WTCCC, despite the substantially smaller sample size here. A further 5 regions were identified by analysis of imputed SNPs, although these were not as compelling.

The *GABRB1* gene is part of the $GABA_A$ receptor family. CRADDOCK ET AL. (2008) examined the same subset of cases in other genes in this family and found evidence of association at a number of them. Of these, only the association at *GABRB1* was strong enough to be highlighted in my analysis. They also showed that the remaining BD cases did not show evidence of association at any of these loci.

Together, these results support the hypothesis that SABP represents a more homogeneous category distinct from other types of BD (at least in part, since SABP itself is heterogeneous), and raise the prospect of a more biological characterisation of this psychological condition. This would have important implications for the study of BD. Finally, they also show that the phenotype refinement approach can be useful for dealing with heterogeneity.

Chapter 4

Frequentist Analysis of GWAS: Extensions & Approximations

Contents

In this chapter I collect together a number of results concerning frequentist inference in GWAS. While some of these are of interest on their own, the primary purpose is to set the stage and act as a reference for subsequent chapters.

I first describe a general logistic modelling framework that can encompass a wide variety of scenarios, including multiple SNPs, environmental effects, interactions and incorporation of covariates. This extends the basic approaches outlined in Chapter 1, and also serves as a base from which I develop analogous Bayesian methods in Chapter 5. I then describe a technical adjustment to these models, involving reparameterisation, that removes correlations between some of the parameters. I use this in subsequent derivations and in Chapter 5.

Following this, I focus more closely on the simple additive and general models, deriving approximations for the variance of the disease effect parameters and for the power of tests for association and for deviation from an additive model. My aim is to gain a deeper understanding of these widely used approaches and develop tools for quick and approximate calculation of useful quantities like power. I also use these results frequently in subsequent chapters.

Finally, I explore the consequences of using cohort samples in place of control samples. I quantify the resulting bias on odds ratio estimates and show that the resulting estimators are in fact estimating the relative risk rather than the odds ratio. This has implications for interpretation of GWAS results and for theoretical derivations in Chapter 7.

4.1 General modelling framework

In Section 1.3.2, I described some simple logistic regression models used for the analysis of GWAS data. Such models are easily extended to allow for more complex scenarios, by adding extra terms linearly to the model equation. Depending on what these terms are and how they are added, we can model a variety of situations. For example, inclusion of covariates, stratification correction, models involving multiple SNPs optionally including interactions, environmental effects and gene-environment interactions. In this section I describe a general logistic regression framework encompassing all such models.

While expositions of logistic regression models and related results can be found in numerous sources (e.g. COX & SNELL 1989, MCCULLAGH & NELDER 1983, MYERS ET AL. 2002) and while similar frameworks have also been described specifically for association studies (e.g. SCHAID ET AL. 2002, WAKEFIELD 2007, 2009), for completeness I reprise these well-known results here to act as a reference and provide for consistent notation for the remainder of my thesis.

Logistic regression models are a subset of generalised linear models, using a binomial distribution for the response variable and the logit link function. The logit link function is the canonical link function for the binomial distribution, and is preferred to other link functions because it can be used for both prospective and retrospective data. In particular, a retrospective sample will simply modify the intercept term, which should therefore be treated as a nuisance parameter. This proof of this is shown in MCCULLAGH & NELDER (1983, pp. 111–4) and is repeated in Section 4.6.2 for completeness as part of a different result.

4.1.1 Likelihood equations

Model equation

For the ith individual, let $\boldsymbol{X}_i$ be a vector of disease effect variables and $\boldsymbol{Z}_i$ be a vector of the remaining variables, which includes the baseline, covariates and stratification variables (below I discuss how to set and interpret these for different types of models). As before, let y_i represent the case-control status, and $p_i = \Pr(y_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i)$ be the probability of an individual being a case given the other variables. The logistic regression model is defined by the equation,

$$\operatorname{logit}(p_i) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\,,$$

where $\boldsymbol{\beta}$ is a vector of disease effect parameters and $\boldsymbol{\mu}$ is a vector of the remaining parameters, including the baseline, covariate and stratification effects. Let the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ be k and l respectively.

As written, the model equation looks broadly similar to the additive model. Intuitively, we can think of it in the same way, with $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ acting as more general versions of the intercept and slope in the simpler model. Different types of models are obtained by setting $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ appropriately.

The $\boldsymbol{X}_i$ should be a relevant function of the genotype (or genotypes, for models with multiple SNPs). For example, the additive model is obtained by $\boldsymbol{X}_i^{\mathrm{T}} = [G_i]$ and the general model by $\boldsymbol{X}_i^{\mathrm{T}} = [G_i,\ \mathbf{1}_{G_i=1}]$. A model with more SNPs would include genotypes from each SNP. For example, a two-SNP additive model with no interaction is obtained by $\boldsymbol{X}_i^{\mathrm{T}} = [G_{i1}, G_{i2}]$.

The $\boldsymbol{Z}_i$ should contain at least a vector of 1s for the baseline (intercept) term, and then any covariates and stratification variables of interest. Continuous covariates would generally just be included directly in $\boldsymbol{Z}_i$, for a model where the genetic effect is to be over and above the effect explained by the covariate. Categorical covariates can be incorporated in two ways. One way is to just include them in $\boldsymbol{Z}_i$ as a separate indicator variable for each group, resulting in a different baseline (intercept) for each. This retains a common genetic effect (slope) across the groups and is appropriate for stratification correction. The other way is to also allow a different slope for each group, expanding $\boldsymbol{X}_i$ by creating a copy for each group of each original variable in $\boldsymbol{X}_i$. This can be used to model differential effects amongst the groups (e.g. a sex-differentiated effect).

For more compact notation, define,

$$\boldsymbol{W}_i = \begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{X}_i \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{bmatrix}.$$

The model equation can then be written as,

$$\operatorname{logit}(p_i) = \boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\theta}\,,$$

and the disease risk for the ith individual,

$$p_i = \frac{e^{\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\theta}}}{1 + e^{\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\theta}}}\,.$$

Likelihood and log-likelihood

The likelihood equation for the model is,

$$L = \prod_{i=1}^{N} p_i^{y_i}\,(1-p_i)^{1-y_i}\ ,$$

and the corresponding log-likelihood,

$$\begin{aligned}
\ell &= \log(L) \\
&= \sum \left(y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right) \right) \\
&= \sum \left(y_i \operatorname{logit}\left(p_i\right) + \log\left(1 - p_i\right) \right) \\
&= \sum \left(y_i \boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) \right) \\
&= \left(\sum y_i \boldsymbol{W}_i^{\mathrm{T}} \right) \boldsymbol{\theta} - \sum \log\left(1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) .
\end{aligned}$$

For clarity, I have omitted the summation specification: all sums are evaluated over i, from 1 to N. I now calculate the derivatives of the log likelihood. I use the following identities (which are easily verified) to make this simpler,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \log\left(1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) &= \boldsymbol{W}_i \frac{e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}}{1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}} = \boldsymbol{W}_i p_i \,, \\
\frac{\partial^2}{\partial\left(\boldsymbol{\theta} \boldsymbol{\theta}^{\mathrm{T}}\right)} \log\left(1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) &= \boldsymbol{W}_i \frac{\partial p_i}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \boldsymbol{W}_i \boldsymbol{W}_i^{\mathrm{T}} \frac{e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}}{\left(1 + e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\theta}}\right)^2} = \boldsymbol{W}_i \boldsymbol{W}_i^{\mathrm{T}} p_i \left(1 - p_i\right) .
\end{aligned}$$

Note that the parameters in these equations are contained within p_i.

Score function

The score function is the first derivative of the log-likelihood,

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\theta}} &= \sum \left(y_i \boldsymbol{W}_i - \boldsymbol{W}_i p_i \right) \\
&= \sum \boldsymbol{W}_i \left(y_i - p_i \right) .
\end{aligned}$$

Let the score function be $\boldsymbol{U}_\theta$. This can be subdivided into the score functions for the two types of parameters; let these be $\boldsymbol{U}_\mu$ and $\boldsymbol{U}_\beta$ for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ respectively,

$$\boldsymbol{U}_\theta = \begin{bmatrix} \boldsymbol{U}_\mu \\ \boldsymbol{U}_\beta \end{bmatrix} = \begin{bmatrix} \frac{\partial \ell}{\partial \boldsymbol{\mu}} \\ \frac{\partial \ell}{\partial \boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \sum \boldsymbol{Z}_i \left(y_i - p_i \right) \\ \sum \boldsymbol{X}_i \left(y_i - p_i \right) \end{bmatrix} .$$

Fisher information matrix

The Hessian matrix of the log-likelihood is,

$$H(\ell) = \frac{\partial^2 \ell}{\partial(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}})} = -\sum \boldsymbol{W}_i \boldsymbol{W}_i^{\mathrm{T}} p_i \left(1 - p_i\right) \,.$$

This does not involve the y_i terms, so is equal to its expectation (under a prospective model). Thus, the Fisher information matrix is just its negative, $\mathcal{I}_\theta = \mathbb{E}\left(-H(\ell)\right) = -H(\ell)$. As with the score function, this can be subdivided into blocks corresponding to the $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ parameters,

$$\mathcal{I}_\theta = \begin{bmatrix} \mathcal{I}_{\mu\mu} & \mathcal{I}_{\mu\beta} \\ \mathcal{I}_{\beta\mu} & \mathcal{I}_{\beta\beta} \end{bmatrix} .$$

Note that $\mathcal{I}_\theta$, $\mathcal{I}_{\mu\mu}$ and $\mathcal{I}_{\beta\beta}$ are all symmetric, and that $\mathcal{I}_{\beta\mu} = \mathcal{I}_{\mu\beta}^{\mathrm{T}}$. Using a standard formula for the inverse of a block matrix (e.g. RENCHER 1995, p. 407),

$$\mathcal{I}_\theta^{-1} = \begin{bmatrix} \mathcal{I}_{\mu\mu}^{-1} + \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\mathcal{I}_{\beta|\mu}^{-1}\mathcal{I}_{\beta\mu}\mathcal{I}_{\mu\mu}^{-1} & -\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\mathcal{I}_{\beta|\mu}^{-1} \\ -\mathcal{I}_{\beta|\mu}^{-1}\mathcal{I}_{\beta\mu}\mathcal{I}_{\mu\mu}^{-1} & \mathcal{I}_{\beta|\mu}^{-1} \end{bmatrix} ,$$

where,

$$\mathcal{I}_{\beta|\mu} = \mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\mu}\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} \,.$$

This latter matrix is known as the Schur complement of $\mathcal{I}_{\mu\mu}$ in $\mathcal{I}_\theta$. Asymptotically, it is the conditional variance of $\boldsymbol{U}_\beta$ given $\boldsymbol{U}_\mu$, a fact used below when deriving the score test statistic. Note that the asymptotic *marginal* variance of $\boldsymbol{U}_\beta$ is $\mathcal{I}_{\beta\beta}$. These should not be confused with the corresponding quantities for the MLE, $\hat{\boldsymbol{\beta}}$, for which the situation is 'reversed': $\mathcal{I}_{\beta|\mu}^{-1}$ is the marginal variance and $\mathcal{I}_{\beta\beta}^{-1}$ the conditional variance given $\boldsymbol{\mu}$.

4.1.2 Inference

Association testing

To test for association, we compare the full model as described to one where the genetic effect parameters are zero,

$$H_0 \colon \boldsymbol{\beta} = \mathbf{0} \quad \text{vs} \quad H_1 \colon \boldsymbol{\beta} \neq \mathbf{0} \,,$$

in the presence of unknown confounding parameters $\boldsymbol{\mu}$. As before, using a score test is a convenient and efficient choice for carrying out this test in the GWAS context. The score test statistic is,

$$T_{\text{score}} = \boldsymbol{U}_{\beta}^{\text{T}} \, \mathcal{I}_{\beta|\mu}^{-1} \, \boldsymbol{U}_{\beta} \, , \tag{4.1}$$

evaluated at $\boldsymbol{\beta} = \mathbf{0}$ and where the nuisance parameters are set to their MLE, $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$, under H_0. The test statistic has a χ^2_k distribution under H_0.

Note that $\mathcal{I}_{\beta\beta}$ is the (asymptotic) variance of $\boldsymbol{U}_{\beta}$, but that $\mathcal{I}_{\beta|\mu}$ is its conditional variance given $\boldsymbol{U}_{\mu}$. In the situation above we need to use $\mathcal{I}_{\beta|\mu}$ to adjust for the fact that we are estimating the nuisance parameters, because setting the nuisance parameters to their MLE is equivalent to setting $\boldsymbol{U}_{\mu} = 0$ (SMYTH 2003). If the confounding parameters were known, $\mathcal{I}_{\beta|\mu}$ would be replaced by $\mathcal{I}_{\beta\beta}$.

Effect size estimation

Inference for the genetic effects can be done using the MLE, $\hat{\boldsymbol{\beta}}$, which asymptotically has a multivariate normal distribution,

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} \text{N}_k\Big(\boldsymbol{\beta}, \, \mathcal{I}_{\beta|\mu}^{-1}\Big) \, .$$

4.2 Reparameterisation to remove correlations

The logistic regression models parameterised as shown in Section 1.3.2 and 4.1 have the property that, in general, their parameter estimators will be correlated. This is usually undesirable for both theoretical and numerical calculations. Thus, it would useful to reparameterise the models to remove unnecessary correlations. For clarity, I first show how to do this for the simple additive and general models, which I return to later in the chapter. I then show how to do it in the general modelling framework from the previous section.

4.2.1 Simple models: mean-centering

Considering firstly the additive model. The parameter of interest is that for the additive effect, β, while the baseline parameter, μ, is not of interest in a case-control study. Thus, we

seek a reparameterisation that reduces the correlation between their estimators while retaining the interpretation of the additive effect. The correlation is induced by the distribution of genotypes in the sample. A natural way to reduce it is to mean-centre the genotype labels,

$$\text{logit}\left(p_i\right) = \nu + \beta\left(G_i - \bar{G}\right) ,$$

where $\bar{G}$ is the genotype mean (the average number of B alleles per individual in the sample),

$$\bar{G} = \frac{n_1 + 2n_2}{N} ,$$

and $\nu = \mu + \beta\bar{G}$ is the new baseline. This makes the parameterisation more orthogonal and would completely eliminate the correlation in a general linear model. In fact, this parameterisation is 'null orthogonal' as defined by KASS & VAIDYANATHAN (1992), who show that, in what follows, we may assume that the Fisher information matrix will be approximately diagonal.

Note that the β parameter is unchanged (it only depends on the *differences* between genotypes between individuals and these are unchanged after mean-centering), but μ has been replaced by ν. The net effect is that the linear variation in μ with respect to β has been subtracted out. Thus, the parameterisation (ν, β) satisfies our requirements: the additive parameter is unchanged, and is uncorrelated with the nuisance baseline parameter.

Reparameterisation by mean-centering also works for other models, including the general model and models involving multiple SNPs. The coefficients of each disease effect parameter are mean-centered individually, depending on their mean across the sample. For example, for the general model,

$$\text{logit}\left(p_i\right) = \nu + \beta\left(G_i - \bar{G}\right) + \gamma\left(\mathbf{1}_{G_i=1} - \frac{n_1}{N}\right) ,$$

with n_1/N being the mean of $\mathbf{1}_{G_i=1}$ across the sample, and $\nu = \mu + \beta\bar{G} + \gamma n_1/N$.

4.2.2 More complex models

Consider the general logistic regression model from Section 4.1. We can reparameterise the model to (asymptotically) remove the correlation between the nuisance and disease param-

eters, without changing the interpretation of the disease parameters. Using an idea similar to that of KASS & VAIDYANATHAN (1992), WAKEFIELD (2009) showed how to do this for one-parameter disease models. Here I generalise this result to multi-dimensional disease models.

The reparameterisation involves using $\boldsymbol{\nu} = \boldsymbol{\mu} + \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\boldsymbol{\beta}$ instead of the original nuisance variables $\boldsymbol{\mu}$. This is equivalent to replacing $\boldsymbol{X}_i^{\mathrm{T}}$ with $\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{Z}_i^{\mathrm{T}}\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}$. (The approach of KASS & VAIDYANATHAN (1992) differs only in that the Fisher information is evaluated at the null.) The equivalent representations of the logistic regression equation are,

$$\begin{aligned}
\text{logit}\left(p_i\right) &= \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} \\
&= \boldsymbol{Z}_i^{\mathrm{T}}\left(\boldsymbol{\nu} - \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\boldsymbol{\beta}\right) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} \\
&= \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\nu} + \left(\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{Z}_i^{\mathrm{T}}\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\right)\boldsymbol{\beta}\,.
\end{aligned}$$

The reparameterisation subtracts out the linear contribution of $\boldsymbol{\beta}$ to $\boldsymbol{\mu}$. It is similar to mean-centering as described for the additive and general models above. For example, consider a simple additive model, which has $\boldsymbol{Z}_i = [1]$ and $\boldsymbol{X}_i = [G_i]$. Assuming small effects, such that $p_i(1-p_i) \approx p(1-p)$ for all i, gives,

$$\boldsymbol{Z}_i^{\mathrm{T}}\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} = \frac{\mathcal{I}_{\mu\beta}}{\mathcal{I}_{\mu\mu}} = \frac{\sum G_i p_i\left(1-p_i\right)}{\sum p_i\left(1-p_i\right)} \approx \frac{\sum G_i}{\sum 1} = \frac{n_1 + 2n_2}{N} = \bar{G}\,,$$

with exact equality under H_0.

To see that $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$ are asymptotically uncorrelated, we could calculate the new Fisher information matrix. Instead, I calculate the new asymptotic distribution of the MLE. Let the new parameters be $\boldsymbol{\psi}$ and the transformation matrix be A,

$$\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} + \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta}\boldsymbol{\beta} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} I & \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} \\ 0 & I \end{bmatrix}\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{bmatrix} = A\boldsymbol{\theta}\,.$$

Since $\boldsymbol{\psi}$ is a linear transformation of $\boldsymbol{\theta}$, the MLE of $\boldsymbol{\psi}$ will asymptotically have a multivariate

distribution with mean $A\boldsymbol{\theta} = \boldsymbol{\psi}$ and covariance matrix,

$$\begin{aligned}
\mathcal{I}_{\psi}^{-1} = A\,\mathcal{I}_{\theta}^{-1}\,A^{\mathrm{T}} &= \left(\left(A^{\mathrm{T}}\right)^{-1}\mathcal{I}_{\theta}\,A^{-1}\right)^{-1} \\
&= \left(\begin{bmatrix} I & 0 \\ \mathcal{I}_{\beta\mu}\mathcal{I}_{\mu\mu}^{-1} & I \end{bmatrix}^{-1} \mathcal{I}_{\theta} \begin{bmatrix} I & \mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} \\ 0 & I \end{bmatrix}^{-1}\right)^{-1} \\
&= \left(\begin{bmatrix} I & 0 \\ -\mathcal{I}_{\beta\mu}\mathcal{I}_{\mu\mu}^{-1} & I \end{bmatrix} \begin{bmatrix} \mathcal{I}_{\mu\mu} & \mathcal{I}_{\mu\beta} \\ \mathcal{I}_{\beta\mu} & \mathcal{I}_{\beta\beta} \end{bmatrix} \begin{bmatrix} I & -\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} \\ 0 & I \end{bmatrix}\right)^{-1} \\
&= \left(\begin{bmatrix} \mathcal{I}_{\mu\mu} & \mathcal{I}_{\mu\beta} \\ 0 & \mathcal{I}_{\beta|\mu} \end{bmatrix} \begin{bmatrix} I & -\mathcal{I}_{\mu\mu}^{-1}\mathcal{I}_{\mu\beta} \\ 0 & I \end{bmatrix}\right)^{-1} \\
&= \begin{bmatrix} \mathcal{I}_{\mu\mu} & 0 \\ 0 & \mathcal{I}_{\beta|\mu} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mathcal{I}_{\mu\mu}^{-1} & 0 \\ 0 & \mathcal{I}_{\beta|\mu}^{-1} \end{bmatrix}.
\end{aligned}$$

This is block diagonal, thus $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\nu}}$ are asymptotically uncorrelated. Note that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is the same under either parameterisation.

4.2.3 Correlation between disease model parameters

The above reparameterisations remove correlations between the nuisance parameters and the disease model parameters. That still leaves correlations *between* disease model parameters. These correlations cannot be removed without affecting the definition and interpretation of the parameters. If desired, orthogonal parameterisations can be constructed in a similar way to above, by adding the disease parameters to the model one by one and subtracting out the linear variation in each using the Gram-Schmidt algorithm. However, rather than selecting a parameterisation purely for mathematical convenience and then dealing with the problem of trying to interpret the results, it seems more desirable to retain a parameterisation that is easy to interpret and consider the correlation as part of the inference.

4.3 Simulating SNPs

Simulating case-control samples at a SNP is useful for validating theoretical results and analysis methods. It is relatively simple in principle, essentially being just a special case of multinomial sampling. However, some choices need to be made and there are some common 'traps' encountered in practice.[1] For this reason, and the fact that I use simulated data later in this chapter, I have chosen to document these here.

We are generally interested in retrospective samples. That is, we will have a fixed number of cases and controls, and the genotypes for each individual will be random. If we specify a genotype distribution for cases, $\Pr(G = i \mid Y = 1)$, and for controls, $\Pr(G = i \mid Y = 0)$, we can just take a multinomial sample with the required sample sizes from each directly.

Usually we have a prospective model in mind and are interested in simulating the data with a given set of parameter values under that model. In such a scenario we need to use Bayes' Theorem to derive the correct sampling distribution. Let the penetrances given by the model be $p_i = \Pr(Y = 1 \mid G = i)$. Before we proceed, we also need to specify a genotype distribution in the population. Let this be $g_i = \Pr(G = i)$. Then the case and control genotype distributions are, respectively,

$$\begin{aligned} \Pr(G = i \mid Y = 1) &= \frac{p_i g_i}{\sum p_i g_i}, \\ \Pr(G = i \mid Y = 0) &= \frac{(1 - p_i) g_i}{\sum (1 - p_i) g_i}. \end{aligned}$$

If a cohort sample is desired, then we can just sample directly from the population distribution, g_i.

There are choices to do with both p_i and g_i, and I consider each of these in turn.

The penetrances depend on the disease model, usually a logistic regression model, so it is more fruitful to talk in terms of choices for parameter values. Considering firstly an additive model, we need to choose both the baseline (μ) and additive effect (β) parameter values. The latter is usually the one of main interest and will be determined based on the desired use of the simulated data, often covering a range of values. The former is less important since the retrospective sampling mostly, but not completely, negates its effect. It will be sufficient

[1] In fact, both my colleagues and myself have all fallen into the same traps when trying to simulate SNP data. What we assumed would be quick and straightforward turned out to be much tricker than we first thought.

to set it at a value that results in a plausible disease prevalence, e.g. $\mu = -4.6$ results in a prevalence of about 1% (the actual prevalence depends also on β and the genotype distribution). For more complex models there are more parameter values that need to be selected, but these will be informed by the intended use of the simulations.

The g_i can be specified arbitrarily. With three genotypes, this defines a two-dimensional space of possibilities, giving potentially too much flexibility. Most commonly we are interested in SNPs that are in HWE (or close to it), allowing us to specify the genotype distribution in terms of a single quantity—the allele frequency, f,

$$\begin{aligned} g_0 &= (1-f)^2\,, \\ g_1 &= 2f(1-f)\,, \\ g_2 &= f^2\,. \end{aligned}$$

Now the choice is down to just a single-dimensional quantity, making it easier to run simulations spanning a range of genotype frequencies and also to plot the results of these simulations.

One common trap is to simulate a cohort sample instead of a control sample. Given that actual studies typically employ cohort samples, there is a temptation to do so for simulations as well. This is fine if emulating GWAS data is the goal, but is not appropriate, for example, for validating theoretical results based on the assumption of a case-control sample. One potentially undesirable consequence is that the true effect sizes will no longer be those that were specified for the simulation (this is discussed further in Section 4.6). Unless you take this into account, it can be particularly confusing if you are trying to validate your sampling method by checking whether the simulations give 'correct' effect sizes estimates.

Finally, there may be situations where a prospective sample is desired. This is done similarly to a retrospective sample, except that now the genotype counts are fixed and the disease status is sampled. In this scenario, the μ parameter plays a more important role, since it will primarily determine the prevalence of diseased individuals in the sample.

4.4 Variance approximations

The variance of the disease parameters will generally depend both on quantities we know, such as the sample size, and those that are unobserved, such as the true underlying parameter values. For various applications, including some that I consider in this thesis, it is useful to have approximations that are easy to calculate and only depend on observed quantities. Here I derive such approximations for the parameters in the additive and general models. My approach is to use asymptotic properties of the MLE, the reparameterisation from Section 4.2, and to assume HWE and weak effect sizes. I then use simulated data to assess the accuracy of these approximations, showing that they are very accurate for common SNPs and slightly underestimate the variance at rarer SNPs.

4.4.1 Additive model

Asymptotically, the variance of the MLE is equal to the inverse of the Fisher information matrix. Using the reparameterisation from Section 4.2, asymptotically we only need to consider the submatrix corresponding to the disease parameters (because the covariance matrix is asymptotically block diagonal),

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right) = \mathcal{I}_{\beta\beta}^{-1} \,. \tag{4.2}$$

The additive model has only one disease parameter and so this equation becomes,

$$\operatorname{var}\left(\hat{\beta}\right) = \mathcal{I}_{\beta\beta}^{-1}\left(\nu, \beta\right) \,,$$

where the Fisher information is,

$$\begin{aligned} \mathcal{I}_{\beta\beta}\left(\nu, \beta\right) &= \mathbb{E}\left(-\frac{\partial^2 l}{\partial \beta^2}\left(\nu, \beta\right)\right) \\ &= \sum_{j=0}^{2} n_j a_j^2 \, p_j(1-p_j) \\ &= \sum_{j=0}^{2} n_j a_j^2 \frac{e^{\nu+\beta a_j}}{\left(1+e^{\nu+\beta a_j}\right)^2} \,, \end{aligned} \tag{4.3}$$

where l is the log-likelihood of the reparameterised model and $a_j = j - \bar{g}$ are the mean-

centered values of the genotype labelled j, the sum being over the three genotype labels. In what follows, I use the following two approximations:

1. I approximate the logistic function by a second-order Taylor expansion about 0,

$$\frac{e^x}{1+e^x} \approx \frac{1}{2} + \frac{1}{4}x = \frac{2+x}{4}\,.$$

This is shown in Figure 4.1. Note that this is results in a first-order polynomial because the quadratic coefficient is 0. I will use this approximation on terms of the form

$$\frac{e^{\nu+\beta a_j}}{1+e^{\nu+\beta a_j}}\,,$$

for $j = 0, 1, 2$. It will be accurate whenever $\nu + \beta a_j$ is close to 0. This will certainly happen when both ν and β are close to 0 (i.e. if we have roughly the same number of cases and controls, and small effect sizes). However, because the a_j are fixed and differ by 2 in the extreme ($a_2 - a_0 = 2$), the accuracy of the approximation is more dependent on β. If β is not small, some of the βa_j will be far away from 0 where the approximation is most accurate. On the other hand, if $\beta \approx 0$ then ν can vary away from 0 while still keeping this approximation adequate. For example, for the WTCCC Crohn's disease data where $\nu \approx \log(\frac{1748}{2938}) = -0.52$, I observed the approximation to be adequate. Using the Taylor expansion, we also get,

$$\frac{e^x}{(1+e^x)^2} = \frac{e^x}{1+e^x}\left(1 - \frac{e^x}{1+e^x}\right) \approx \left(\frac{2+x}{4}\right)\left(\frac{2-x}{4}\right) = \frac{4-x^2}{16}\,.$$

2. I assume the sample is roughly in HWE to get the approximation $n_2 \approx \bar{f}^2 N$, where the allele frequency for allele B in the sample is $\bar{f} = \frac{n_1+2n_2}{2N} = \frac{\bar{G}}{2}$. This assumption will be most accurate for SNPs that are in HWE and have small effects. The derived expressions in this section are all symmetric with respect to the allele coding, so without loss of generality we can assume that $\bar{f}$ is the MAF.

I now use these to derive an approximation to $\mathrm{var}\left(\hat{\beta}\right)$. Applying the logistic function ap-

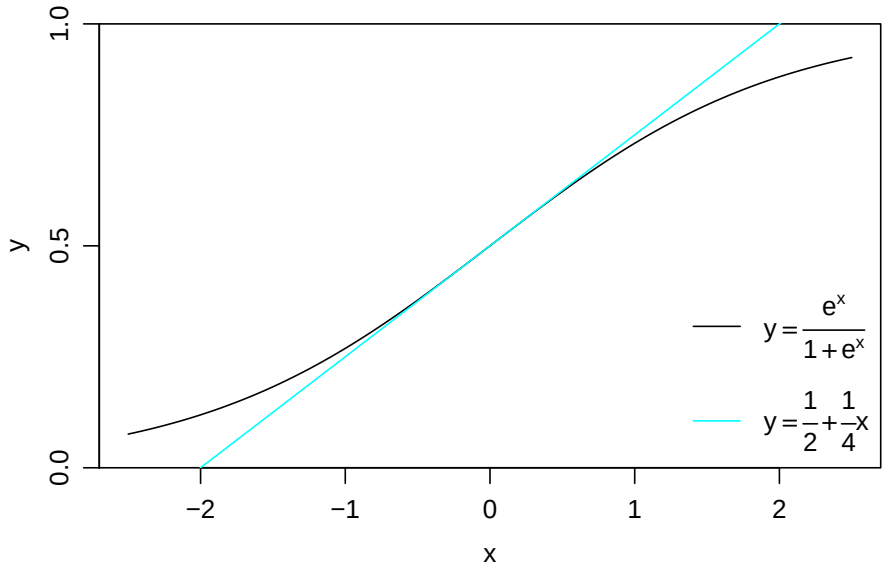

Figure 4.1: **Logistic function approximation.** Using a second-order Taylor expansion about 0, which is a first-order polynomial. The logistic function is in black, the Taylor expansion is in cyan.

proximation to equation (4.3) gives,

$$\begin{aligned}
\mathcal{I}_{\beta\beta}(\nu,\beta) &\approx \sum_{j=0}^{2} n_j a_j^2 \frac{4-(\nu+\beta a_j)^2}{16} \\
&= \frac{1}{16}\sum_{j=0}^{2} n_j a_j^2 \left(4-\nu^2-2\nu\beta a_j-\beta^2 a_j^2\right) \\
&= \frac{1}{16}\left((4-\nu^2)\sum_{j=0}^{2} n_j a_j^2 - 2\nu\beta\sum_{j=0}^{2} n_j a_j^3 - \beta^2\sum_{j=0}^{2} n_j a_j^4\right) .
\end{aligned}$$

The sums in this expression can be rewritten in terms of $\bar{f}$ and n_2. If we apply the HWE assumption, we can re-write them in terms of $\bar{f}$ only (complete derivation not shown),

$$\begin{aligned}
\sum_{j=0}^{2} n_j a_j^2 &= 2N\bar{f}(1-\bar{f}) , \\
\sum_{j=0}^{2} n_j a_j^3 &= 2N\bar{f}(1-\bar{f})(1-2\bar{f}) , \\
\sum_{j=0}^{2} n_j a_j^4 &= 2N\bar{f}(1-\bar{f}) .
\end{aligned}$$

For completeness, note that $\sum_{j=0}^{2} n_j a_j = 0$, due to the definition of a_j as being the mean-

centered value of the genotype. Substituting these into the above expression gives,

$$\mathcal{I}_{\beta\beta}(\nu,\beta) \approx 2N\bar{f}(1-\bar{f})\left(\frac{4-(\nu+\beta)^2+4\nu\beta\bar{f}}{16}\right). \tag{4.4}$$

For common diseases we are primarily interested in the case of small effect size, which is also when the above approximations are most accurate. We can derive an even simpler formula in this case. Assuming that the true OR is small enough that we can take $\beta = 0$ as an approximation, and using $\hat{\nu}$ to approximate μ, gives

$$\text{var}\left(\hat{\beta}\right) = \mathcal{I}^{-1}_{\beta\beta}(\hat{\nu},0). \tag{4.5}$$

Deriving $\hat{\nu}$ from the first derivative of the log-likelihood,

$$\frac{\partial l}{\partial \nu} = S - \sum_{j=0}^{2} n_j \frac{e^{\nu+\beta a_j}}{1+e^{\nu+\beta a_j}}.$$

Setting this to 0 and solving, while also applying the logistic function approximation, gives,

$$\hat{\nu} = \frac{2}{N}(S-R).$$

Substituting this and equation (4.4) into equation (4.5) gives,

$$\text{var}\left(\hat{\beta}\right) \approx \frac{1}{2N\bar{f}(1-\bar{f})\phi(1-\phi)}, \tag{4.6}$$

where $\phi = S/N$ is the proportion of the sample that are cases.

I now use simulated data to evaluate the accuracy of this approximation. Using the procedure described in Section 4.3, I simulated SNPs under a selection of sample sizes, allele frequencies and values for the true OR. In particular, I consider: samples with 500, 2000 and 10,000 cases and the same number of controls; allele frequencies of 0.02, 0.05, 0.5, 0.95 and 0.98; and ORs from 1 to 2 in steps of 0.1. This automatically covers both protective and risk effects. If an allele has an OR less than 1 it is equivalent to the other allele having an OR greater than 1, and I cover the allele frequency spectrum in both directions—that is, common SNPs as well as rare SNPs where the rare allele shows either a risk or a protective effect. I assumed HWE and used $\mu = -4.6$ for the baseline parameter, which gives a prevalence

of 1% under a null model. These choices were made so as to cover a range of scenarios for typical GWAS but also to correspond to similar simulations in Section 4.5 where I evaluate power approximations, and the choices there are also made so as to at least show moderate power.

I simulated 10,000 SNPs for each combination of parameters above. I excluded simulated SNPs that were monomorphic in either the cases or controls, since the MLE is not finite for such datasets. This makes sense, practically speaking, since we cannot estimate the OR in such scenarios, but it means that the empirical estimates will be biased downwards compared to the theoretical result when many such exclusions occur. However, this will only happen at very low allele frequencies, when the asymptotic assumptions break down anyway.

Figure 4.2 compares the standard deviation of $\hat{\beta}$ as evaluated by simulation to the approximation given by the formula in equation (4.6). It shows very good agreement, particularly for common SNPs, with slight underestimation at rare SNPs.

4.4.2 General model

The general model has two disease parameters and equation (4.2) becomes,

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right) = \mathcal{I}_{\beta\beta}^{-1}\left(\nu, \beta, \gamma\right) ,$$

where the Fisher information is,

$$\mathcal{I}_{\beta\beta}\left(\nu, \beta, \gamma\right) = \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} p_i \left(1 - p_i\right) = \begin{bmatrix} \sum n_j a_j^2 & \sum n_j a_j b_j \\ \sum n_j a_j b_j & \sum n_j b_j^2 \end{bmatrix} \frac{e^{\nu + \beta a_j + \gamma b_j}}{\left(1 + e^{\nu + \beta a_j + \gamma b_j}\right)^2} , \quad (4.7)$$

where $a_j = j - \bar{g}$ and $b_j = \mathbf{1}_{j=1} - \frac{n_1}{N}$ are the mean-centered coefficients of the parameters. The sums in the matrix are over j, the three genotype labels, whereas the first sum is over individuals.

I now apply the same approach as for the additive model, but jump directly to the small disease effects assumption. In other words, I assume $\beta \approx 0$ and $\gamma \approx 0$. Along with the

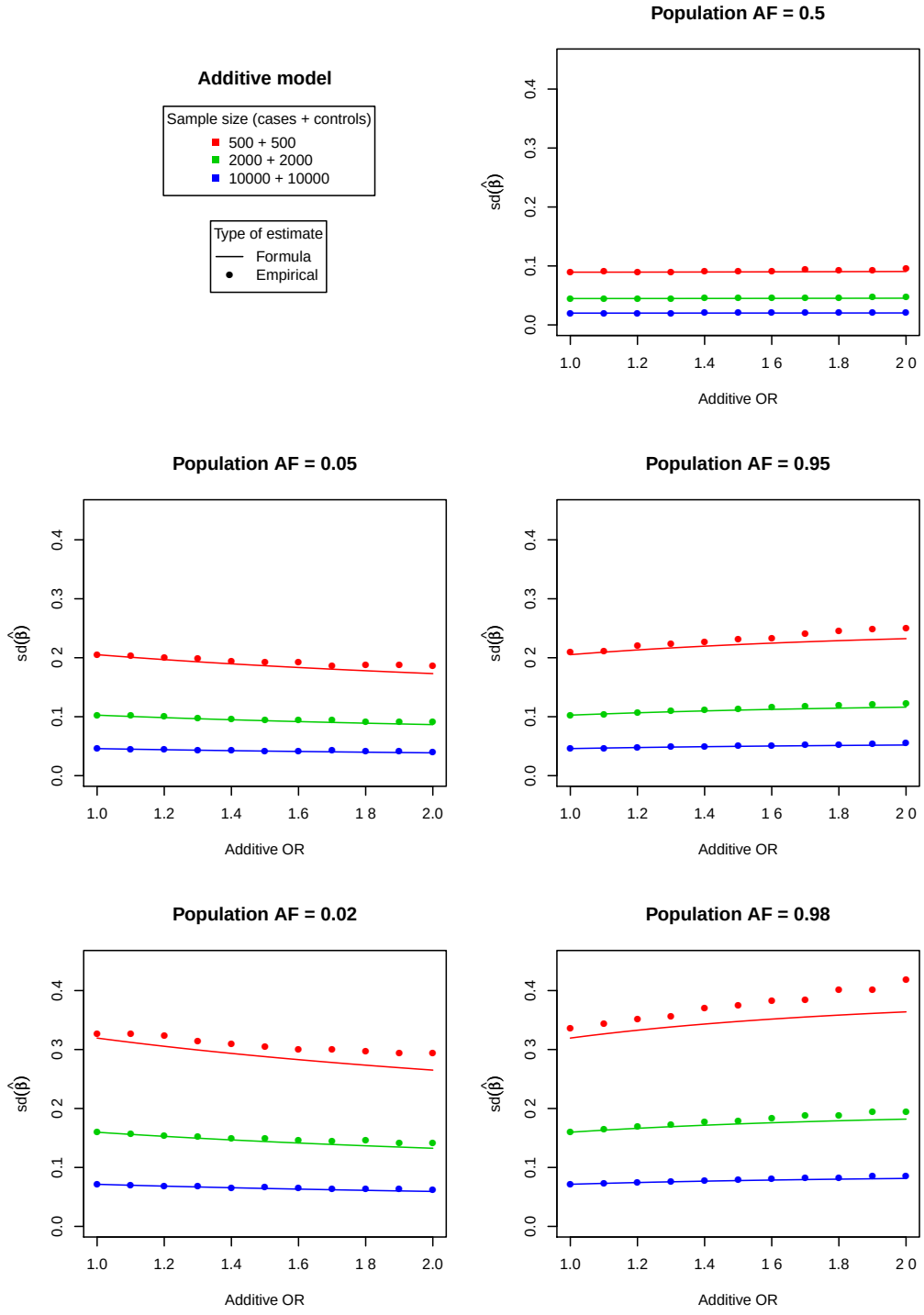

Figure 4.2: **Variance approximation comparison for $\hat{\beta}$ under an additive model.** The lines were calculated using the formula in equation (4.6). The points are empirical estimates based on 10,000 simulated SNPs at each point. Simulations where the SNP was monomorphic in either cases or controls were excluded. Confidence intervals were negligibly tight so were omitted.

logistic function approximation, this gives,

$$\frac{e^{\nu+\beta a_j+\gamma b_j}}{\left(1+e^{\nu+\beta a_j+\gamma b_j}\right)^2} \approx \frac{e^{\nu}}{(1+e^{\nu})^2} \approx \frac{4-\nu^2}{16}\,.$$

Under these assumptions, the MLE of ν is the same as in the derivation for the additive model and we have,

$$\frac{4-\hat{\nu}^2}{16} \approx \phi(1-\phi)\,.$$

The sums in the matrix can be simplified with the HWE assumption (complete derivation not shown),

$$\begin{aligned}
\sum_{j=0}^{2} n_j a_j^2 &= 2N\bar{f}(1-\bar{f})\,,\\
\sum_{j=0}^{2} n_j a_j b_j &= 2N\bar{f}(1-\bar{f})(1-2\bar{f})\,,\\
\sum_{j=0}^{2} n_j b_j^2 &= 2N\bar{f}(1-\bar{f})(1-2\bar{f}+2\bar{f}^2)\,.
\end{aligned}$$

Substituting these into equation (4.7) gives,

$$\mathcal{I}_{\beta\beta} \approx 2N\bar{f}(1-\bar{f})\phi(1-\phi)\begin{bmatrix} 1 & 1-2\bar{f} \\ 1-2\bar{f} & 1-2\bar{f}+2\bar{f}^2 \end{bmatrix}, \tag{4.8}$$

and inverting gives,

$$\begin{aligned}
\operatorname{var}\left(\hat{\beta}\right) &\approx \frac{\bar{f}^2+(1-\bar{f})^2}{4N\bar{f}^2(1-\bar{f})^2\phi(1-\phi)}\,,\\
\operatorname{var}\left(\hat{\gamma}\right) &\approx \frac{1}{4N\bar{f}^2(1-\bar{f})^2\phi(1-\phi)}\,.
\end{aligned} \tag{4.9}$$

From the inverted matrix we also get an approximate formula for the correlation between the two parameters,

$$\operatorname{cor}\left(\hat{\beta},\hat{\gamma}\right) \approx \frac{2\bar{f}-1}{\sqrt{f^2+(1-f)^2}}\,.$$

The correlation is 0 when $\bar{f}=0.5$, and goes to ± 1 as $\bar{f}$ approaches 1 or 0 respectively. In other words, the parameters are strongly correlated when the MAF is low, and the direction of the correlation depends on the allele coding. This is illustrated graphically in Figure 4.3.

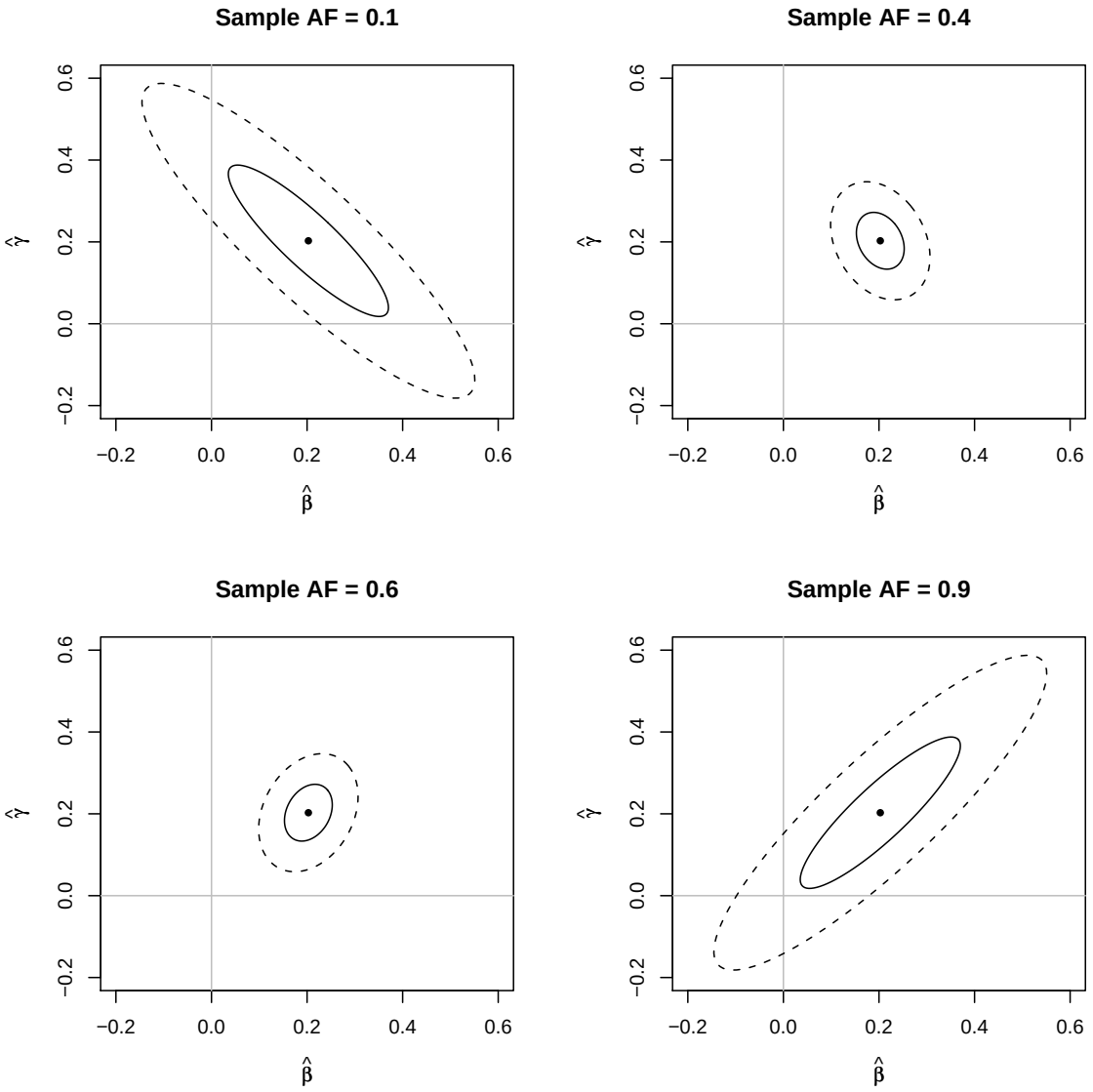

Figure 4.3: **Asymptotic distribution of parameter estimates for the general model.** Bivariate normal probability contours for $\left(\hat{\beta}, \hat{\gamma}\right)$ for a selection of allele frequencies. The true model is a dominant model with OR of 1.5, which corresponds to general model parameter values of $\beta = \gamma = 0.5 \log(1.5)$, shown as a black point on the plots (since the MLE is asymptotically unbiased). The solid lines are 50% probability contours, and the dashed lines are 95% probability contours.

Of particular interest in the general model is $\mathrm{var}\,(\hat{\gamma})$, because γ measures deviation from an additive model and I later use it to approximate the power of a test for such deviations.

I now use simulated SNPs to check the accuracy of this variance approximation. With two parameters to specify, many choices are possible for the true model from which to simulate. To make this simpler, I consider only dominant and recessive models. Both are single-parameter non-additive models that have been observed in Mendelian diseases and are biologically plausible for common diseases. Let the log odds ratio for such a model be α, and let allele B be the risk allele (i.e. assume $\alpha > 0$). There is simple relationship with the disease parameters in the general model. For a dominant model we have,

$$\beta = \gamma = \alpha/2\,,$$

and for a recessive model,

$$\beta = -\gamma = \alpha/2\,.$$

Similarly to the previous section, I simulated 10,000 SNPs for each combination of parameters and excluded those which gave rise to non-finite MLEs. For the general model this entails excluding all SNPs with a zero genotype count in either cases or controls. This is mostly a problem at low allele frequencies, although it will manifest itself at higher allele frequencies than the similar problem did for the additive model.

The simulations I show here are under the same selection of sample sizes and ORs as for the additive model (shown in Figure 4.2), but this time only for minor allele frequencies down to 0.1. This is for two reasons. Firstly, the simulations and theoretical results match very poorly at lower allele frequencies, most likely due to a combination of the bias from exclusions (see previous paragraph) and the related effect of the asymptotic assumptions breaking down. Secondly, I wanted the plots to be comparable to corresponding plots in Section 4.5 where I evaluate power approximations, and the power diminishes almost to zero at lower allele frequencies anyway.

Figures 4.4 and 4.5 compare the standard deviation of $\hat{\gamma}$ as evaluated by simulation to the approximation given by the formula. There is very good agreement for common SNPs, with some underestimation at rarer SNPs.

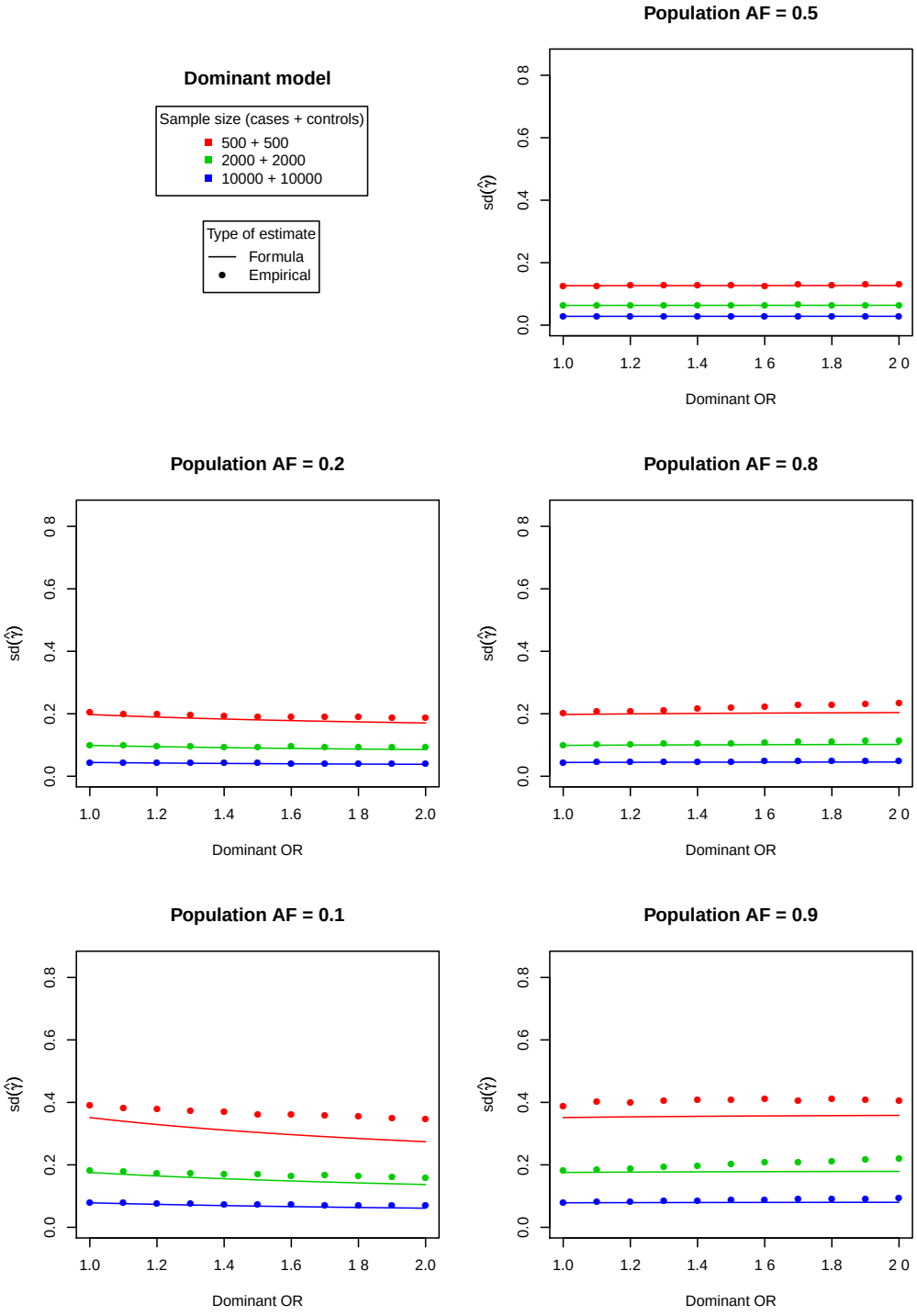

Figure 4.4: **Variance approximation comparison for $\hat{\gamma}$ under a dominant model.** The lines were calculated using the formula in equation (4.9). The points are empirical estimates based on 10,000 simulated SNPs at each point. Simulations where any genotype count was zero in either cases or controls were excluded. Confidence intervals were negligibly tight so were omitted.

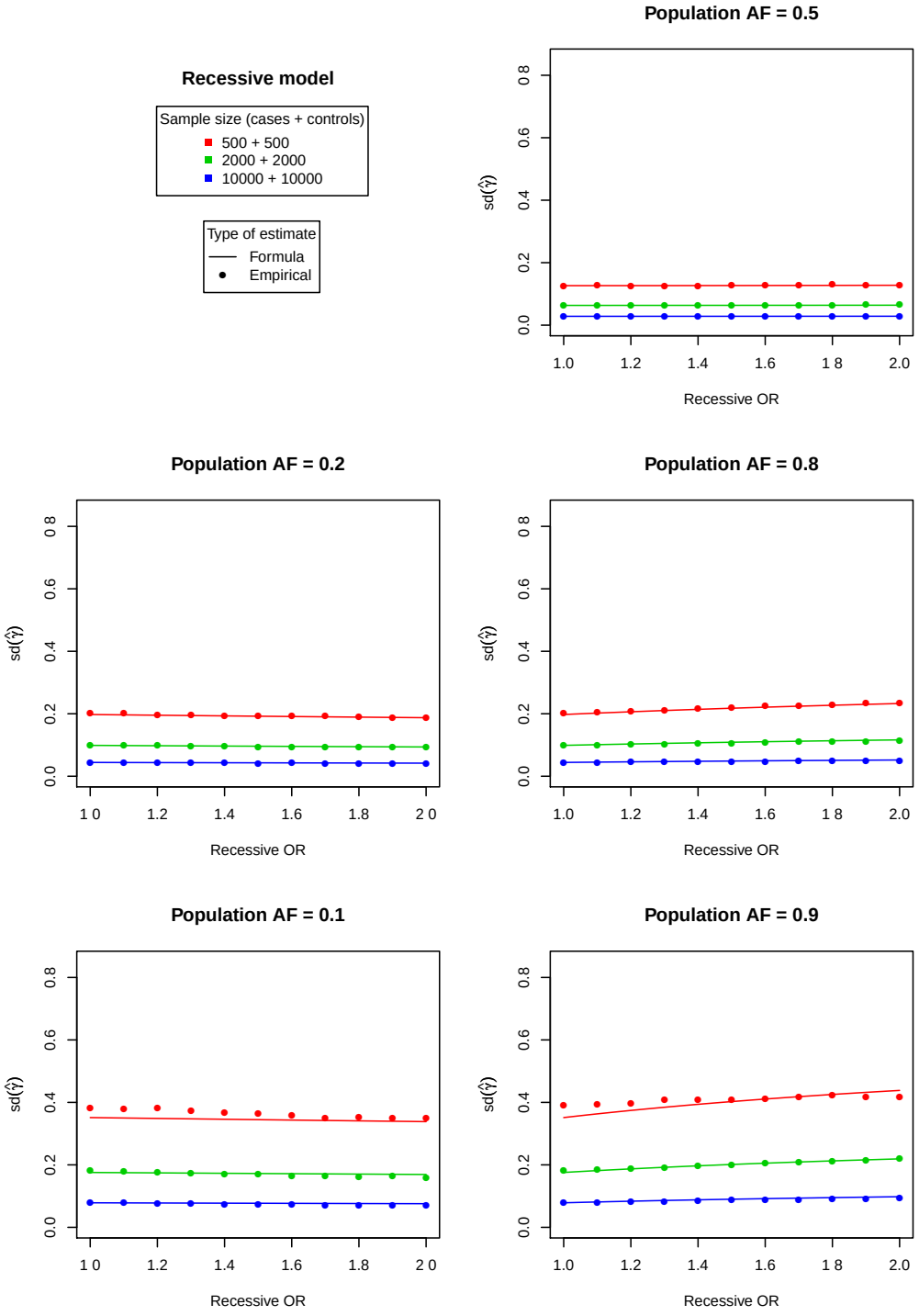

Figure 4.5: **Variance approximation comparison for $\hat{\gamma}$ under a recessive model.** The lines were calculated using the formula in equation (4.9). The points are empirical estimates based on 10,000 simulated SNPs at each point. Simulations where any genotype count was zero in either cases or controls were excluded. Confidence intervals were negligibly tight so were omitted.

4.5 Power approximations

Power calculations are often of interest when designing GWAS and thinking about the range of effect sizes that could be detected. Depending on the exact scenario of interest and the desired accuracy, estimating the power can involve very detailed and computationally intensive calculations (e.g. SPENCER ET AL. 2009). Sometimes we only need an estimate that is approximate but is quick to calculate and only depends on a few simple quantities such as sample size. Here I derive such approximations, based on the variance approximations from the previous section. In addition to being useful for convenient quick power calculations, these results can be used as part of further approximate theoretical derivations. For example, I use them in Chapter 7 when I examine the effect of LD on power.

I derive power approximations for the following three scenarios: (i) testing for association using the additive test; (ii) testing for association using the general test; (iii) testing for deviation from an additive model. The first of these is the standard GWAS testing scenario, while the third is often of interest when doing further analyses of putative associations. My approach for each is to use the appropriate Wald test and combine it with the corresponding variance approximation from the previous section. I then use simulated data to assess the accuracy of these approximations, and show they are very accurate for common SNPs and slightly less so at rarer SNPs.

4.5.1 Association testing with the additive test

It is standard to use the additive test when testing for an association, which is well-powered to detect an additive effect. It is the score test for the additive parameter in the additive model. The test statistic is known (e.g. PRITCHARD & PRZEWORSKI 2001, CHAPMAN ET AL. 2003) to approximately follow a χ^2_1 distribution with non-centrality parameter,[2]

$$\eta_1 = 2N\phi(1-\phi)\frac{(f_1 - f_0)^2}{\bar{f}(1-\bar{f})}\,, \tag{4.10}$$

[2]A non-central χ^2 distribution is related to the usual (central) χ^2 distribution as follows. Let $X_i \overset{\mathrm{d}}{=} \mathrm{N}(\mu_i, \sigma_i)$ be k independent random variables. Then $\sum_i \left((X_i - \mu_i)/\sigma_i\right)^2$ follows a central χ^2_k distribution, or equivalently a non-central χ^2_k distribution with non-centrality parameter 0. In contrast, $\sum_i \left(X_i/\sigma_i\right)^2$ follows a non-central χ^2_k distribution with non-centrality parameter $\eta = \sum_i \left(\mu_i/\sigma_i\right)^2$.

where f_1, f_0 and $\bar{f}$ are the (expected) frequencies of allele B in the cases, controls and the whole sample respectively.

I now derive a similar result using the Wald test. This is primarily for unity with the following sections, where Wald tests are more convenient to consider. While in practice a score test or a MLRT are generally preferred, all three of these are asymptotically equivalent (COX & HINKLEY 1974). Thus, these results will be useful for actual scenarios of interest.

A Wald test is constructed from a parameter of interest in a model and is typically based on the parameter's asymptotic distribution. The Wald test for an additive effect based on the additive model uses the test statistic,

$$z = \frac{\hat{\beta}}{\text{se}\left(\hat{\beta}\right)} .$$

Its square asymptotically follows a χ^2_1 distribution with non-centrality parameter,

$$\eta_2 = \frac{\beta^2}{\text{var}\left(\hat{\beta}\right)} .$$

Using the variance approximation from equation (4.6) gives,

$$\eta_2 \approx 2N\bar{f}(1-\bar{f})\phi(1-\phi)\beta^2 . \tag{4.11}$$

Equations (4.10) and (4.11) are actually similar and I briefly show this for the scenario of a relatively rare disease at a locus with small effect size. Let p be the prevalence of the disease, and p_1 and p_0 be the penetrances of the two alleles if we consider this as a haploid model. By definition, $p = p_1 f + p_0(1-f)$. Bayes' Theorem gives $f_1 = p_1\bar{f}/p$ and $f_0 = (1-p_1)\bar{f}/(1-p)$, which together give the identity,

$$f_1 - f_0 = \frac{f(1-f)}{p(1-p)}(p_1 - p_0) .$$

This allows us to re-write equation (4.10) in terms of effect size parameters,

$$\begin{aligned}\eta_1 &\approx 2N\phi(1-\phi)\frac{1}{\bar{f}(1-\bar{f})}\left(\frac{f(1-f)}{p(1-p)}(p_1-p_0)\right)^2 \\ &= 2N\bar{f}(1-\bar{f})\phi(1-\phi)\left(\frac{f(1-f)}{\bar{f}(1-\bar{f})}\right)^2\left(\frac{e^{\mu}(1-e^{\mu})}{p(1-p)}\right)^2\left(\frac{e^{\beta}-1}{1+e^{\mu}(e^{\beta}-1)}\right)^2.\end{aligned}$$

When the disease is relatively rare, we have $p \approx e^{\mu} \approx 0$, giving

$$\eta_1 \approx 2N\bar{f}(1-\bar{f})\phi(1-\phi)\left(\frac{f(1-f)}{\bar{f}(1-\bar{f})}\right)^2\left(e^{\beta}-1\right)^2.$$

When the genetic effect is small ($\beta \approx 0$), we have $e^{\beta}-1 \approx \beta$. Furthermore, the allele frequencies will be similar in cases and controls, and thus also in the population and sample as well, when the effect is small. As long as the SNP is not too rare, the ratios of these frequencies will then be close to 1, $f/\bar{f} \approx (1-f)/(1-\bar{f}) \approx 1$, giving $\eta_1 \approx \eta_2$.

I now use simulated data to evaluate the accuracy of these approximations, using the same simulation procedure and parameter values as described in Section 4.4.1 but now with 100,000 samples per parameter combination and without excluding any samples. I use a p-value threshold of 5×10^{-7} to determine significance. When a p-value cannot be calculated, I treat it as non-significant (for the trend test this occurs when a SNP is observed to be monomorphic in either cases or controls). The results are shown in Figure 4.6. There is very good agreement between all three power estimates, especially for common SNPs. Understandably, the Wald test approximation (η_2) is slightly less accurate than the one based on the score test (η_1). In particular, the Wald test approximation is a slight overestimate. This corresponds to the variance approximation on which it is based being a slight underestimate, as was observed in Section 4.4.1.

One question that may be asked is whether the difference between the empirical and theoretical results are just due to approximation error or due to comparing different sorts of tests (despite their asymptotic equivalence). Running some simulations using the Wald test gave similar results for the empirical power estimates (data not shown), so it seems that the former is the case here.

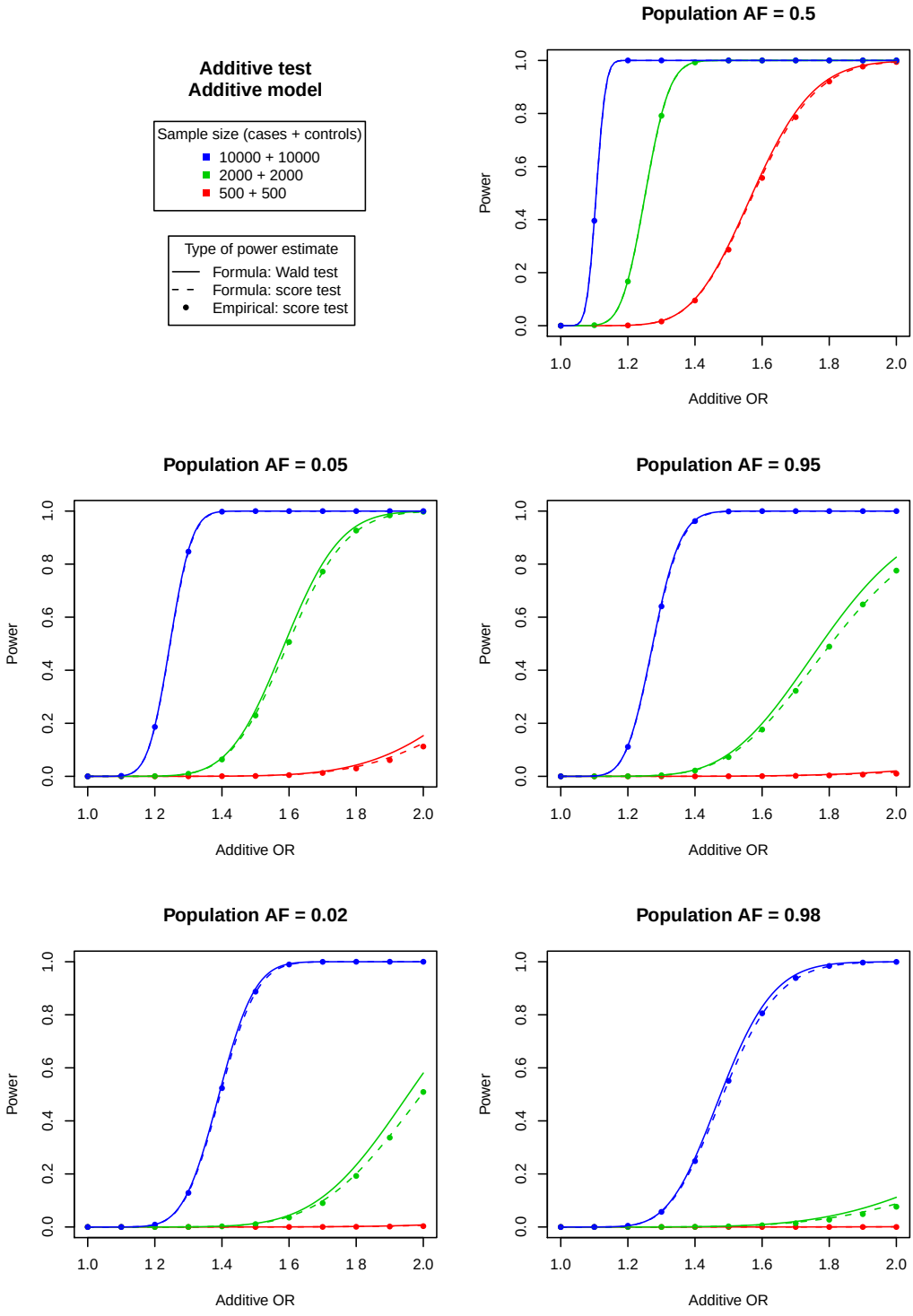

Figure 4.6: **Power approximation comparison: additive test under an additive model.** The lines were calculated using the formulae in equations (4.10) and (4.11). The points are empirical estimates based on 100,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

4.5.2 Association testing with the general test

I derive two approximations to the distribution of the general test statistic. The first is an *ad hoc* approximation in the spirit of the score test approximation for the additive test. The second is based on a bivariate Wald test.

The non-centrality parameter in equation (4.10) can be obtained from the formula for the additive test statistic, shown in equation (1.3), by simply replacing the observed allele frequencies with their expected values (and also assuming HWE). Using the same idea with the general test statistic, shown in equation (1.4), gives the approximate non-centrality parameter,

$$\eta_1 = \frac{N\phi(1-\phi)}{h_0\, h_1\, h_2} \left(h_0 \left(f_1 g_2 - f_2 g_1\right)^2 + h_1 \left(f_0 g_2 - f_2 g_0\right)^2 + h_2 \left(f_0 g_1 - f_1 g_0\right)^2 \right) , \qquad (4.12)$$

where f_i, g_i and h_i are the (expected) frequencies of genotype i in controls, cases and the whole sample respectively. It turns out that this is a very good approximation (see simulations below) and that the score test statistic follows a non-central χ^2_2 distribution.

A bivariate Wald test statistic asymptotically equivalent to the score test is,

$$z^2 = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\, \mathcal{I}_{\beta|\mu}\, \hat{\boldsymbol{\beta}} \,.$$

This will asymptotically have a non-central χ^2_2 distribution. To obtain the non-centrality parameter I replace $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$ and use the approximation to the Fisher information from equation (4.8),

$$\begin{aligned} \eta_2 &\approx \begin{bmatrix} \beta \\ \gamma \end{bmatrix}^{\mathrm{T}} 2N\bar{f}(1-\bar{f})\phi(1-\phi) \begin{bmatrix} 1 & 1-2\bar{f} \\ 1-2\bar{f} & 1-2\bar{f}+2\bar{f}^2 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \\ &= 2N\bar{f}(1-\bar{f})\phi(1-\phi) \left(\beta^2 + 2(1-2\bar{f})\beta\gamma + (1-2\bar{f}+2\bar{f}^2)\gamma^2\right) . \end{aligned} \qquad (4.13)$$

I now use simulated data to evaluate the accuracy of these approximations. I use the same simulation procedure as in the previous section, 100,000 samples and a p-value threshold of 5×10^{-7}, but now with a greater range of models. Figures 4.7, 4.8 and 4.9 show results when simulating from additive, dominant and recessive models respectively. The results are

similar to those for the additive test. There is very good agreement between all three power estimates, especially for common SNPs, and the Wald test approximation (η_2) is slightly less accurate than the one based on the score test (η_1).

4.5.3 Testing for deviation from an additive model

A natural way to test for a deviation from an additive model is to use a MLRT comparing it to the general model, giving a test with one degree of freedom (which I earlier defined this as the *non-additivity* test). Another way is to perform a Wald test on the dominance parameter (γ) in the general model, since it measures deviation away from an additive model. The two tests are asymptotically equivalent. I now derive a power approximation based on the latter.

The Wald test statistic for the dominance effect in the general model is,

$$z = \frac{\hat{\gamma}}{\text{se}\,(\hat{\gamma})}\,.$$

Its square asymptotically follows a χ^2_1 distribution with non-centrality parameter,

$$\eta = \frac{\gamma^2}{\text{var}\,(\hat{\gamma})}\,.$$

Using the variance approximation from equation (4.9) gives,

$$\eta \approx 4N\bar{f}^2(1-\bar{f})^2\phi(1-\phi)\gamma^2\,. \tag{4.14}$$

I now use simulated data to evaluate the accuracy of these approximations, using the same simulation procedure as described in Section 4.4.2. I use a p-value threshold of 5×10^{-7} to determine significance. The results are shown in Figures 4.10 and 4.11, for simulations using the dominant and recessive models respectively. The approximation is very accurate for common SNPs, but sometimes less accurate at rarer SNPs.

4.5.4 Using a very large control sample

With increasing amounts of GWAS data becoming available, an attractive possibility is to use control samples from other studies to help boost power in your own study. In addi-

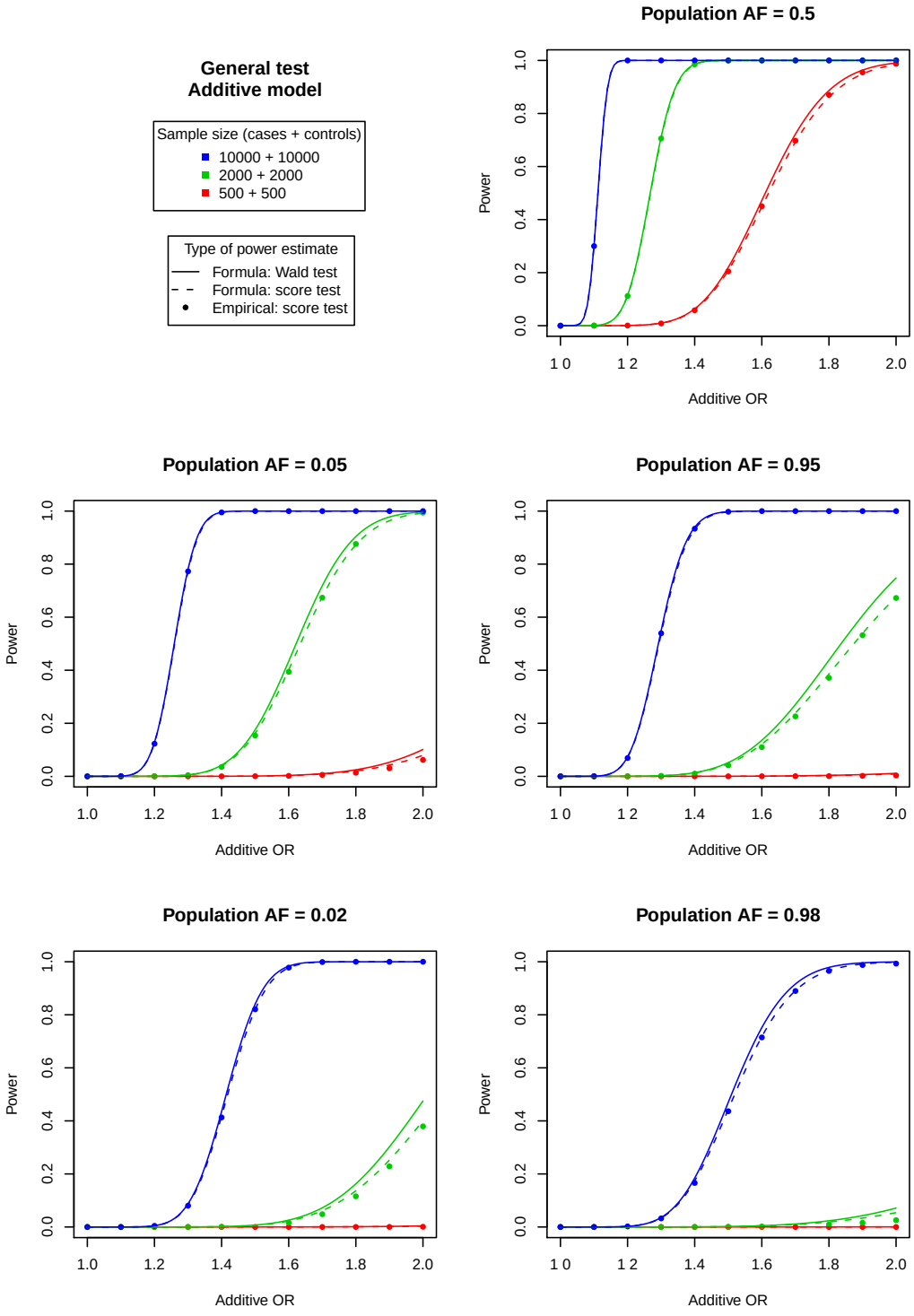

Figure 4.7: **Power approximation comparison: general test under an additive model.** The lines were calculated using the formulae in equations (4.12) and (4.13). The points are empirical estimates based on 100,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

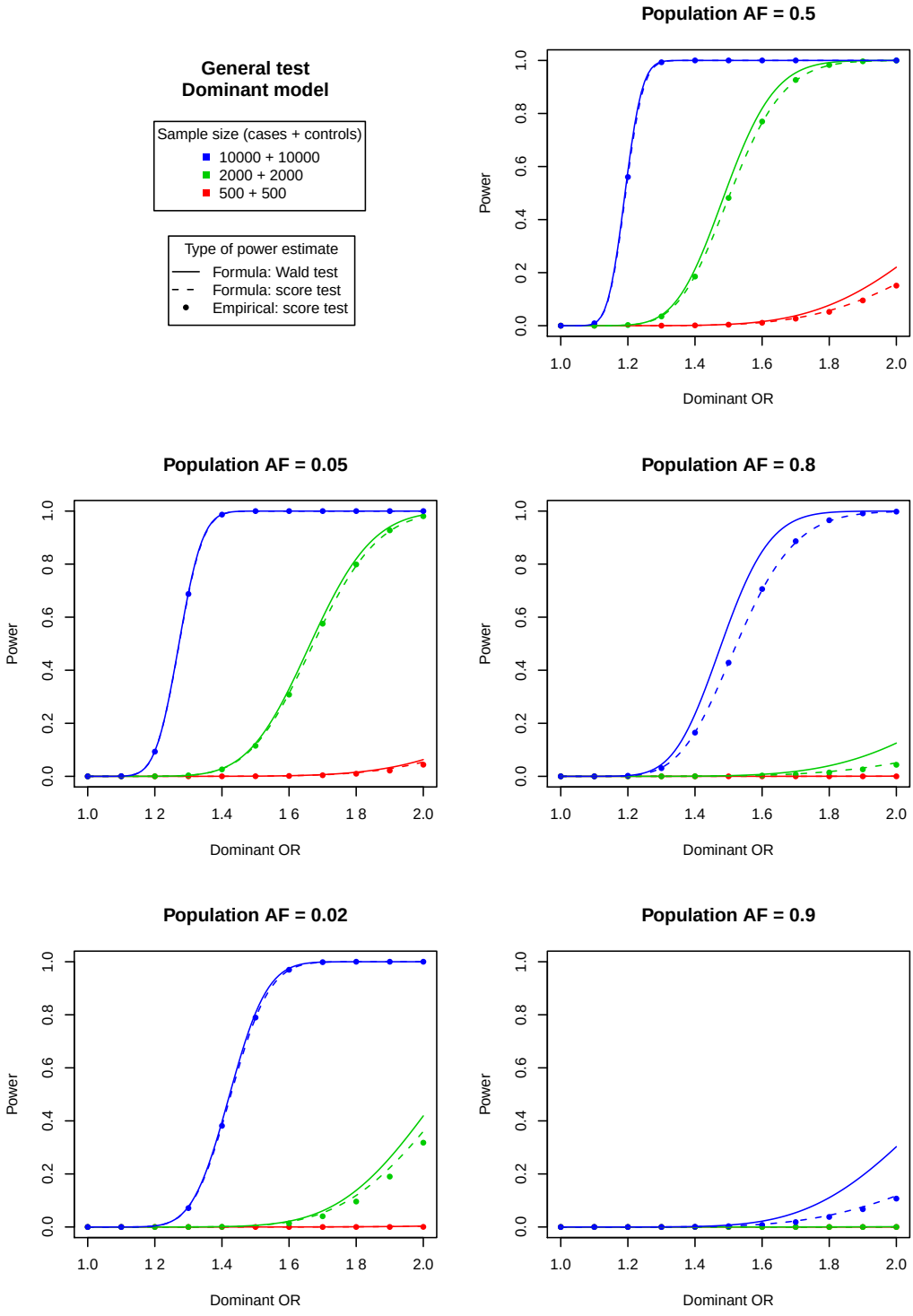

Figure 4.8: **Power approximation comparison: general test under a dominant model.** The lines were calculated using the formulae in equations (4.12) and (4.13). The points are empirical estimates based on 100,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

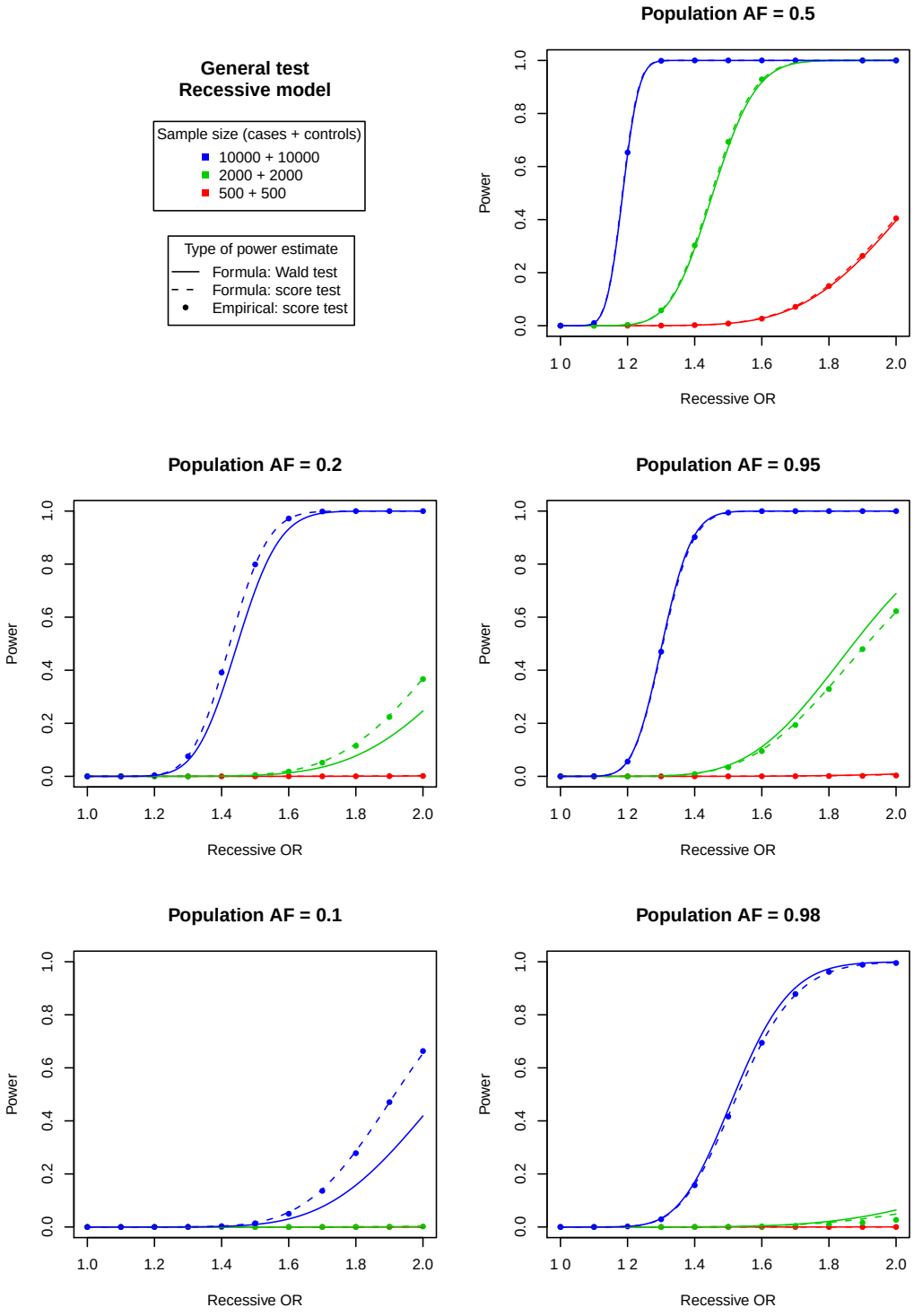

Figure 4.9: **Power approximation comparison: general test under a recessive model.** The lines were calculated using the formulae in equations (4.12) and (4.13). The points are empirical estimates based on 100,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

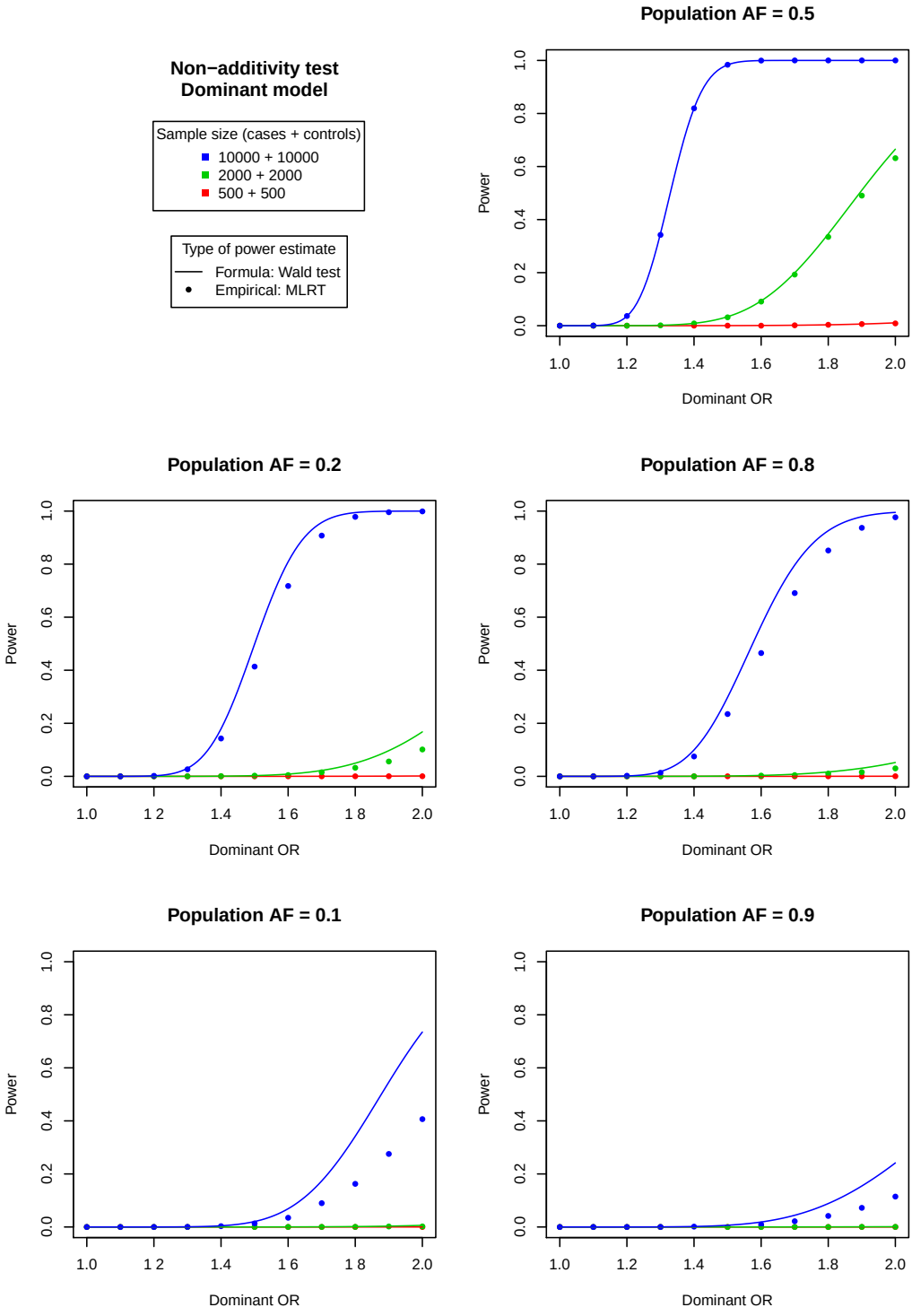

Figure 4.10: **Power approximation comparison: non-additivity test under a dominant model.** The lines were calculated using the formula in equation (4.14). The points are empirical estimates based on 10,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

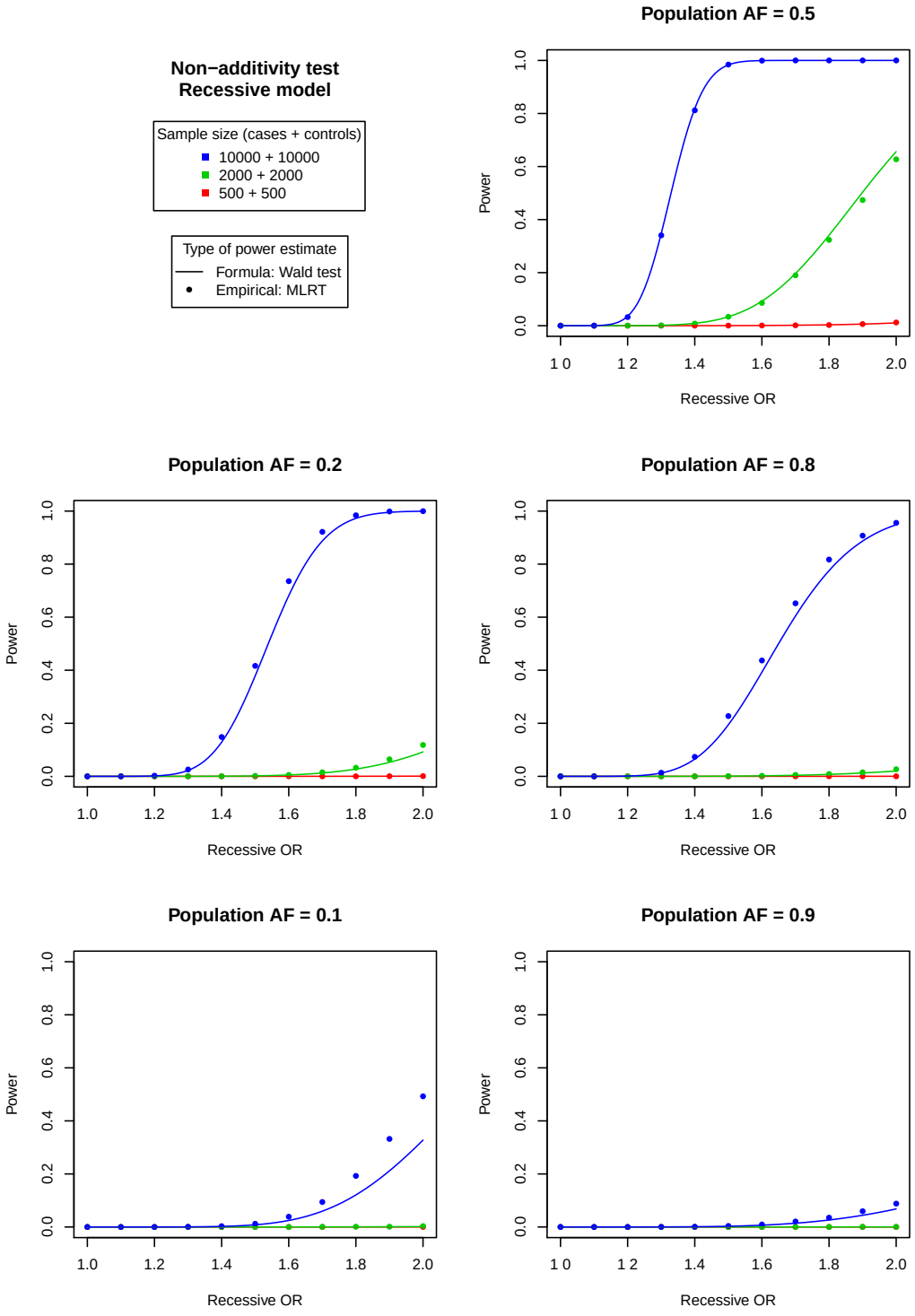

Figure 4.11: **Power approximation comparison: non-additivity test under a recessive model.** The lines were calculated using the formula in equation (4.14). The points are empirical estimates based on 10,000 simulated SNPs at each point, using a p-value threshold of 5×10^{-7}.

tion, there is the possibility of using case samples from unrelated diseases as further control samples (e.g. WTCCC 2007). A natural question to ask is, how much power can we get out of an ever expanding control sample? While both approaches need to be done with some care and simply pooling the data is not advised, we can imagine taking a control sample of infinite size as a way to determine the limit to what we can get out of a given case sample.

Recall that $\phi = S/N$. Therefore, as $R \to \infty$,

$$N\phi(1-\phi) = N\frac{S}{N}\frac{R}{N} = \frac{RS}{R+S} = \left(\frac{1}{R} + \frac{1}{S}\right)^{-1} \to S\,.$$

In other words, we can calculate the limiting power using any of the previous approximation formulae simply be replacing $N\phi(1-\phi)$ with S. The other terms, which specify allele frequencies and effect sizes, will not be affected.

4.6 Consequences of using cohort 'controls'

It is typical in GWAS to use cohort samples in place of proper control samples. That is, instead of using individuals known to be free from the disease being studied, a random sample from the population, or at least a sample that is deemed to be representative of the population at large, is used instead. This is done for convenience and also because it is costly to phenotype thousands of individuals, most of whom are not expected to have the disease being studied. Also for convenience, the resulting data are usually analysed under the assumption that these cohort samples are proper controls. This is expected to lead to a slight loss of power and a bias in estimates, although not significantly enough to affect the outcome of most GWAS.

In this section, I characterise the outcome of this assumption in two ways. Firstly, I quantify the downward bias on OR estimates under a haploid (additive) model, showing that it is small unless the disease is fairly prevalent in the population. Secondly, I show a proof that analysing the data in this way implies that estimates of OR parameters actually estimates of their RR equivalents, which provides an alternative interpretation of the bias. While these results are not fully novel, they tend to be either overlooked or are unknown amongst researchers conducting GWAS. The latter result in particular is useful, since it is often the

RR that is desired. We now know that with a case-cohort sample we get it 'for free', without needing to invoke the rare disease assumption and OR-RR equivalence as would normally be done.

4.6.1 Bias in effect size estimates

We can think of the situation involving cohort 'controls' as just a standard case-control scenario but with a different OR. I now derive a formula for the new OR under this assumption, using the haploid model, and calculate it for some typical GWAS scenarios.

Let the prevalence of the disease be $p = \Pr(Y = 1)$. Let the allele frequency of haplotype 1 be f_1 in cases, f_0 in controls, and $f = f_0(1-p) + f_1 p$ in the population. The formula for the OR given in equation (1.5) is in terms of penetrances. Using Bayes' rule, we can rewrite it in terms of allele frequencies,

$$\mathrm{OR} = \frac{\Pr(H=1 \mid Y=1)}{\Pr(H=0 \mid Y=1)} \frac{\Pr(H=0 \mid Y=0)}{\Pr(H=1 \mid Y=0)} = \frac{f_1(1-f_0)}{f_0(1-f_1)}\,.$$

This is the true OR at the SNP. Now suppose that we have a cohort sample and a case sample. If we treat the cohort sample as controls, that is equivalent to generating data where the effective control allele frequency has now become f. This will give an effective OR different to the true underlying OR, and we can relate the two,

$$\begin{aligned} \mathrm{OR}' &= \frac{f_1(1-f)}{f(1-f_1)} \\ &= \frac{f_1(1-f_0)\left(1 + p\frac{f_0-f_1}{1-f_0}\right)}{f_0(1-f_1)\left(1 + p\frac{f_1-f_0}{f_0}\right)} \\ &= \mathrm{OR}\frac{1 + p\left(\frac{R}{\mathrm{OR}} - 1\right)}{1 + p\,(R-1)}\,, \end{aligned} \tag{4.15}$$

where $\mathrm{R} = f_1/f_0$ is a ratio of the allele frequency of haplotype 1 in cases to controls. When $\mathrm{R} \approx 1$, we have a simpler approximate formula,

$$\mathrm{OR}' = \mathrm{OR}(1-p) + p\,,$$

which says that the effective OR shrinks to 1 linearly with p.

Table 4.1: **Effective odds ratio (OR') when using cohort 'controls'.** As given by equation (4.15), for a range of values for the true underlying OR and the prevalence of the disease.

	Prevalence (%)		
OR	**1**	**5**	**10**
1.1	1.10	1.09	1.09
1.2	1.20	1.19	1.18
1.3	1.30	1.28	1.26
1.5	1.49	1.47	1.43
2	1.98	1.92	1.84

These equations show the bias induced by using a cohort sample in place of a proper control sample. Table 4.1 shows the effective OR for a range of effect sizes and prevalences that span the typically expected values for common diseases. When p is small, $\mathrm{OR}' \approx \mathrm{OR}$. In other words, for rare diseases the effect will be negligible, which is expected since a cohort sample will consist of mostly individuals without the disease, so will be very close to a true control sample. As the prevalence increases we see a downwards bias on the OR, caused by the dilution of the 'control' sample by diseased individuals diminishing the disease signal.

4.6.2 OR estimates are RR estimates

Using Bayes' rule to rewrite the RR formula given in equation (1.6) in terms of allele frequencies gives,

$$\mathrm{RR} = \frac{\Pr(H=1 \mid Y=1)}{\Pr(H=0 \mid Y=1)} \frac{\Pr(H=0)}{\Pr(H=1)} = \frac{f_1(1-f)}{f(1-f_1)}\,.$$

This is the same as the effective OR in the previous section. In other words, assuming our cohort sample is a control sample leads to our OR estimator actually estimating the RR (the OR estimator for the haploid model is just the well-known cross-product estimator, having the same form as the above equation). This is an alternative way to interpret the bias—just think of it as an RR estimate rather than a biased OR estimate.

From the above equations this is clear for the haploid model and has been previously described (e.g. SZKLO ET AL. 2001). In fact, a version of this is true in general for any disease model (including haploid and diploid models) at a multi-allelic locus. I describe this and show a proof below, but note that SCHOUTEN ET AL. (1993) have previously derived essentially the same result. In their formulation, they treat a scenario where the cases are identified from the cohort sample and so are counted twice when fitting the model (the cases are

included in the cohort sample as well as forming the case sample). I treat the scenario where there are separate case and cohort samples. However, the key logic in both approaches is largely equivalent.

Suppose we have a locus with k alleles, $A_0, A_2, \ldots, A_{k-1}$. Let the penetrance at allele i be $d_i = \Pr(Y = 1 \mid A_i)$. We usually consider a logistic regression model. Write a general such model as,

$$\operatorname{logit}(d_i) = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \,, \tag{4.16}$$

where the x_{ij} coefficients will depend on the allele. For example, the additive model over SNP genotypes will have $k = 3$, $l = 1$ and $x_{i1} = i$; the general model will have $k = 3$, $l = 2$, $x_{i1} = i$ and $x_{i2} = \mathbf{1}_{i=1}$. A related model is a log risk regression model,

$$\log(d_i) = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \,. \tag{4.17}$$

The two models are related in the same way that the OR and RR are related. In a haploid model, the additive parameter in the logistic regression model is a log odds ratio, whereas the corresponding parameter in the log risk regression model is a log relative risk. The result I will show is that what is calculated as $\hat{\alpha}_j$ is actually equal to $\hat{\beta}_j$ when cohort samples are used in place of control samples.

Let the frequency of the ith allele be f_i in controls, g_i in cases and h_i in the population. Let the prevalence of the disease be $p = \Pr(Y = 1)$. Using Bayes' Theorem we can write d_i in terms of p and allele frequencies,

$$d_i = \Pr(Y = 1 \mid A_i) = \frac{\Pr(A_i \mid Y = 1)\Pr(Y = 1)}{\Pr(A_i)} = \frac{g_i p}{h_i} \,,$$

and the same for the odds of disease at allele i,

$$\frac{d_i}{1 - d_i} = \frac{\Pr(Y = 1 \mid A_i)}{\Pr(Y = 0 \mid A_i)} = \frac{\Pr(A_i \mid Y = 1)\Pr(Y = 1)}{\Pr(A_i \mid Y = 0)\Pr(Y = 0)} = \frac{g_i p}{f_i(1 - p)} \,.$$

Suppose we collect a case-control dataset. Let the number of case and control individuals having allele i be s_i and r_i respectively. Analysing this using a logistic regression model

involves maximising the (prospective) likelihood function,

$$L = \prod_i p_i^{s_i} (1 - p_i)^{r_i} \,,$$

over the parameters defined by the model equation,

$$\text{logit}\,(p_i) = \gamma_0 + \sum_{j=1}^{l} \gamma_j x_{ij} \,,$$

to get estimates $\hat{\gamma}_j$. This model is actually equivalent to that in equation (4.16), except for the γ_0 parameter. This can be seen by determining the true value of p_i. Let the case sampling proportion be $\phi = \Pr(\text{case})$. Then in our retrospectively sampled dataset we have,

$$\begin{aligned} \Pr(A_i) &= \Pr(A_i \mid \text{case}) \Pr(\text{case}) + \Pr(A_i \mid \text{control}) \Pr(\text{control}) \\ &= g_i \phi + f_i (1 - \phi) \,, \end{aligned}$$

from which we can calculate p_i using Bayes' Theorem,

$$p_i = \Pr(\text{case} \mid A_i) = \frac{\Pr(A_i \mid \text{case}) \Pr(\text{case})}{\Pr(A_i)} = \frac{g_i \phi}{g_i \phi + f_i (1 - \phi)} \,.$$

This gives,

$$\frac{p_i}{1 - p_i} = \frac{g_i \phi}{f_i (1 - \phi)} = \frac{g_i p}{f_i (1 - p)} \frac{\phi (1 - p)}{p (1 - \phi)} = \frac{d_i}{1 - d_i} k \,,$$

where k is a constant. Comparing to equation (4.16),

$$\begin{aligned} \text{logit}\,(p_i) &= \text{logit}\,(d_i) + \log k \\ \gamma_0 + \sum \gamma_j x_{ij} &= \alpha_0 + \sum \alpha_j x_{ij} + \log k \\ \gamma_0 + \sum \gamma_j x_{ij} &= \alpha_0' + \sum \alpha_j x_{ij} \,, \end{aligned}$$

where $\alpha_0' = \alpha_0 + \log k$. We can match the coefficients termwise, giving $\gamma_j = \alpha_j$ for $j = 1, \ldots, l$. Thus, we see that apart from the intercept term, the (prospective) logistic regression model applied to the retrospective dataset gives the correct parameter estimates. This is essentially the same derivation as shown in McCullagh & Nelder (1983, pp. 111–4).

Now suppose that we have a case-cohort dataset. Let the case and cohort counts for allele i be s_i and r_i as above. We fit the same logistic regression to the data to get estimates $\hat{\gamma}_j$. The difference now is that the p_i, and thus γ_j, have different underlying true values. In particular,

$$\begin{aligned}\Pr(A_i) &= \Pr(A_i \mid \text{case})\Pr(\text{case}) + \Pr(A_i \mid \text{cohort})\Pr(\text{cohort}) \\ &= g_i\phi + h_i(1-\phi)\,,\end{aligned}$$

which gives,

$$p_i = \Pr(\text{case} \mid A_i) = \frac{\Pr(A_i \mid \text{case})\Pr(\text{case})}{\Pr(A_i)} = \frac{g_i\phi}{g_i\phi + h_i(1-\phi)}\,.$$

Also,

$$\frac{p_i}{1-p_i} = \frac{g_i\phi}{h_i(1-\phi)} = \frac{g_i p}{h_i}\frac{\phi}{p(1-\phi)} = d_i k\,,$$

where k is a constant (but different to the k in the case-control scenario). From this we see that fitting a logistic regression model actually corresponds to a log risk regression model. Comparing to equation (4.17),

$$\begin{aligned}\operatorname{logit}(p_i) &= \log(d_i) + \log k \\ \gamma_0 + \sum \gamma_j x_{ij} &= \beta_0 + \sum \beta_j x_{ij} + \log k \\ \gamma_0 + \sum \gamma_j x_{ij} &= \beta_0' + \sum \beta_j x_{ij}\,,\end{aligned}$$

where $\beta_0' = \beta_0 + \log k$. Matching the coefficients termwise gives $\gamma_j = \beta_j$ for $j = 1, \ldots, l$. Thus, apart from the intercept term which we treat as a nuisance parameter, the fitted model parameters are actually RRs.

4.6.3 Discussion

I have described how the use of cohort samples in place of control samples has two effects.

Firstly, it reduces power by decreasing the size of the additive effect, but that this reduction is negligible unless the disease prevalence is high. For diseases that occur in less than 1% of the population, this effect may be ignored. However, even when the bias is non-negligible,

it is probably cheaper to make up for the reduced power by collecting a larger sample than to spend the money in phenotyping the cohort sample to remove the diseased individuals.

Secondly, the effect size that *is* estimated is actually a RR. This provides an alternative, probably more insightful, way to interpret what might otherwise be thought of as a bias in the effect size estimate. Many researchers consider RRs easier to interpret than ORs anyway, so would consider the fact that our default procedures result in such estimates to be actually both helpful and convenient.

Given this second result, wherever I show effect size estimates from actual studies in this thesis and the estimates are of primary interest, I will refer to them as relative risks. However, where I derive theoretical results or use data simply to illustrate and validate such results, I will label them as odds ratios. This is to maintain consistency with the underlying mathematical assumptions and other theoretical work.

Interestingly, the case-cohort scenario is a reversal of the usual situation with regard to estimability of the OR/RR. The usual scenario, a case-control sample, allows estimation of the OR, with the RR only estimable if the prevalence is known or assumed. The reverse is true for a case-cohort sample.

Chapter 5

Bayesian Analysis of GWAS I: Methods & Theoretical Comparisons with Frequentist Approaches

Contents

Most analyses of GWAS to date have adopted the frequentist statistical paradigm. In particular, a hypothesis testing framework is commonly used, reporting p-values from χ^2 tests for association at each SNP. While this approach is familiar to the majority of researchers, it can present some difficulties which I briefly outline at the start of this chapter. The Bayesian paradigm offers an alternative approach, one that is becoming increasingly popular for analysing GWAS. It provides a number of advantages, including the ability to incorporate prior information and easier interpretation of results. In addition, there is evidence that they can be more powerful than equivalent frequentist approaches (BALDING 2006, SERVIN & STEPHENS 2007, WAKEFIELD 2008, GUAN & STEPHENS 2008).

In this chapter, I describe a Bayesian approach for analysing single SNPs for association with disease in the context of a case-control study. This is an analogue of the widely used frequentist approach described above. I use the well-known Bayes factor (BF) as a natural summary of the evidence of association and discuss suitable choices of prior distributions in the GWAS context. Some of the difficulties of the frequentist approach are naturally dealt with in the Bayesian framework and I discuss these throughout.

A natural question to ask is how the Bayesian and frequentist methods compare. I address this question both theoretically and, in the following chapter, empirically using data from real GWAS. The theoretical comparison throws additional light on the frequentist approach by showing that it is equivalent to a particular Bayesian implementation, but one with prior assumptions allowing for large effect sizes when the MAF is low and/or when the sample size is small.

While detecting association with disease is my primary focus, estimating the size of the genetic effect at a SNP is also naturally handled in a Bayesian framework by using the posterior distributions on the parameters in the model. I briefly discuss this, showing how the effect size estimate depends on the amount of information in the data through what is known as a 'shrinkage' effect.

5.1 Interpretation of p-values

Even under a frequentist perspective, interpretation of a p-value is not straightforward without also having some sense of the power of the experiment. While this has been discussed extensively in the literature (e.g. STERNE & DAVEY SMITH 2001, THOMAS & CLAYTON 2004, WTCCC 2007, WAKEFIELD 2008), I illustrate the point informally by considering the following rather artificial example. Suppose a GWAS is undertaken with 6,000 cases and 6,000 controls for a particular disease and three different statistical methods are used for its analysis:

1. Ignore the data and simply generate a p-value by choosing a random number distributed uniformly between 0 and 1.

2. Only examine the first 10 cases and the first 10 controls and apply the additive test.

3. Use all of the data and apply the additive test.

Note that all three approaches are valid statistical tests. Suppose that it happens that the p-value from method 1 is 8×10^{-6}, and that this is also true for method 2 (for example, if the 10 cases are AA and the 10 controls are BB) and method 3. Most people would put no weight on the evidence from method 1—it ignores the data and so has essentially no power to reject the null hypothesis even if it is false. Many people would put limited weight on the analysis from method 2 because, under plausible assumptions about effect sizes, it also has very limited power. On the other hand, the analysis from method 3 is likely to be quite persuasive. Although contrived, it seems clear from the example that the same p-value can mean rather different things depending on the study design and analysis.[1]

As shown in the previous chapter, the power to detect a true association at a SNP will vary depending on the sample size, the MAF and the size of the genetic effect. Even in large samples, power will vary substantially across SNPs with different MAFs. To use an often quoted example, assuming the true OR at a particular SNP is 1.3 and using a study with 6,000 cases, 6,000 controls and a p-value threshold of 1×10^{-6}, the power will be 3% and 94% if the MAF is 0.02 and 0.1 respectively (WANG ET AL. 2005).

[1] For comparison, corresponding Bayesian analyses for these same examples are shown in Section 5.7.

A natural question then arises as to how to set an appropriate p-value threshold. Given the large number of tests that are conducted in a GWAS, a related question is whether to, and how to, correct for these 'multiple comparisons'. Some standard approaches, such as Bonferroni correction, can be misleading in the GWAS context, which I also illustrate with a hypothetical example. Suppose we undertake a GWAS, but our our genotyping process fails and we only have data for our samples on chromosome 1. We analyse it anyway and find a SNP which passes our significance threshold. Later we re-do the genotyping and have data for all chromosomes. After correcting for the now much greater number of tests, our initial finding is no longer classed as significant. This seems nonsensical since the new data does not tell us anything new about our initial SNP. Thus, a naive application of multiple testing correction does not accord with a natural intuition about the information in the data.

While these problems are not insurmountable in a frequentist analysis, there is great scope for misunderstanding and misinterpretation when tackling them. In particular, the multiple testing formulation is inherently misleading because it is not the number of tests that is usually of concern, but rather the low prior expectation that any particular SNP in the genome is causative (WTCCC 2007).

5.2 Historical review

Bayesian methods have steadily gained popularity in genetics over the last decade. This follows the trend of increased computing power and also the desire to incorporate prior knowledge in the analysis of new data (BEAUMONT & RANNALA 2004). Such methods have been applied to a wide variety of problems in genetics, for example in the detection of population structure (PRITCHARD ET AL. 2000a). However, it is not until recently that Bayesian methods tailored for use in large GWAS have been developed. In fact, in a recent review of GWAS methods, BALDING (2006) highlighted the fact that Bayesian methods do not yet play a prominent role in GWAS analysis, although he suggested that this would change in the future and briefly outlined a possible single-SNP approach.

Large data sets have only relatively recently become the norm for association studies, both in the number of samples and in the coverage of the genome. Therefore it is understandable that previous methods have focused on smaller data sets. A variety of such approaches

have been developed (e.g. BALL 2001, 2005, 2007, CASELLAS & PIEDRAFITA 2006, JOHNSON 2007, MOLITOR ET AL. 2003, MORRIS 2005, 2006, TACHMAZIDOU ET AL. 2007, VERZILLI ET AL. 2006, 2008), and they tend to share the following features: focus on analysing data from candidate genes/regions, designed for quantitative trait mapping, often use MCMC or similar computationally-intensive approaches, usually employ a variable selection strategy and generally use diffuse priors. Some more recent methods also attempt variable selection but on a more genome-wide scale (FRIDLEY 2008, HOGGART ET AL. 2008).

The aim of a GWAS is usually hypothesis generation. For this purpose, methods that are too eager to 'throw away' potentially associated loci are undesirable, and variable selection approaches tend to be of this ilk. They would be more appropriate in follow-up studies that try narrow down the list of loci, but initially what is required is a way to evaluate each potential hypothesis (i.e. each locus) quickly and efficiently. The WTCCC (2007) introduced a practical single-SNP Bayesian method to do this, in the context of a case-control study. In contrast to the above approaches, it is simple and fast, the emphasis being on analysing a large amount of data approximately rather than a small amount of data in more detail. Another innovation is the use of priors based on prior knowledge of the range of disease effect sizes for common human diseases. Similar approaches have been subsequently used by others (MARCHINI ET AL. 2007, WAKEFIELD 2007, SERVIN & STEPHENS 2007, GUAN & STEPHENS 2008). It is this approach that I describe and develop further in this chapter.

A number of other recent Bayesian methods are worth mentioning. The 'semi-Bayesian' approach of SCHRODI (2005) uses a complicated retrospective model that does not require a prior on the effect size. SALANTI ET AL. (2007) model the deviation from HWE as part of a meta-analysis, while LEE ET AL. (2008) attempt to predict phenotypes using genome-wide SNP data and a reversible-jump MCMC method; both place diffuse priors on effect sizes. The method of LEWINGER ET AL. (2007) uses a regression framework to incorporate multiple sources of prior information to use as a prior.

Some methods that are related to GWAS but not directly concerned with quantifying the evidence of association include that of CHAPMAN ET AL. (2009), which is concerned with determining sub-phenotypes using GWAS data, and KUSTRA ET AL. (2008), which uses a Bayesian method to speed up the calculation by permutation of a p-value in a candidate gene study.

Motivated by difficulties with classical approaches to multiple testing correction in the GWAS context, interest has been shown in methods based on *false discovery rates* (STOREY & TIBSHIRANI 2003). These quantities have a Bayesian flavour, in the sense that they try to quantify the probability of disease status given a significant finding (i.e. given the data), rather than (as is done more classically) the probability of the data given disease status. Some have gone further and assign weights to SNPs depending on prior knowledge (ROEDER ET AL. 2006, GREENWOOD ET AL. 2007), bringing them even closer to a proper Bayesian method.

5.3 The Bayes factor

In this section I describe the well-known Bayes factor and related results, motivating its use for the analysis of association studies.

Consider an arbitrary SNP in our GWAS. We are interested in determining the evidence for association at this SNP. Similar to the hypothesis testing framework from Section 1.3.3, we posit two possible models:

H_0 **Null model.** There is no association between the SNP and the disease.

H_1 **Disease model.** There is an association between the SNP and the disease. Different disease models are possible and I discuss these in Section 5.4.

In the Bayesian framework, we are interested in $\Pr(H_1 \mid \text{data})$. By Bayes' Theorem,

$$\Pr(H_1 \mid \text{data}) = \frac{\Pr(\text{data} \mid H_1)\Pr(H_1)}{\Pr(\text{data} \mid H_1)\Pr(H_1) + \Pr(\text{data} \mid H_0)\Pr(H_0)} .$$

This calculation involves the prior probabilities $\Pr(H_1)$ and $\Pr(H_0)$. We can reformulate this in terms of odds. Let $\text{odds}(\cdot) = \Pr(\cdot)/\left(1 - \Pr(\cdot)\right)$, and noting that $\Pr(H_1) + \Pr(H_0) = 1$, Bayes' Theorem can be expressed as,

$$\text{odds}(H_1 \mid \text{data}) = \text{BF} \times \text{odds}(H_1) ,$$

where the *Bayes factor* (BF) is defined as,

$$\text{BF} = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} .$$

The BF measures how much prior beliefs about the odds of association for this SNP should be changed after observation of the data, making it a natural summary of the strength of evidence.

An advantage of the BF is that it is separate from these prior beliefs about the particular SNP in question (but not from beliefs about effect sizes at associated SNPs). For example, a researcher may wish to put more weight on SNPs that alter gene products or gene regulation. This is naturally accommodated by simply changing the prior odds for particular SNPs or categories of SNPs, but does not change the BF.

Calculation of the numerator and denominator of the BF requires the specification of prior distributions on the parameters in the models. For example, take H_1 to be the additive model,

$$\text{logit}\,(p) = \mu + \beta G\,.$$

Let $\pi(\mu, \beta)$ be the prior for μ and β under the additive model and $\pi(\mu)$ is the prior for μ under the null model ($\beta = 0$ under the null). Then the BF is calculated using,

$$\text{BF} = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} = \frac{\iint L(\mu, \beta)\, \pi(\mu, \beta)\, d\mu\, d\beta}{\int L(\mu, 0)\, \pi(\mu)\, d\mu}\,,$$

where $L(\mu, \beta)$ is the likelihood of the data. The quantities in the numerator and denominator are called *marginal likelihoods*. They are weighted averages of the likelihood under the two models, where the averaging is over the prior distributions of the parameters. This highlights a key difference between Bayesian and frequentist approaches. For large sample sizes, the additive test statistic, T_{add}, is approximately equal to the classical maximum likelihood ratio test statistic (COX & HINKLEY 1974),

$$T_{\text{LR}} = \frac{\max_{H_1} L(\mu, \beta)}{\max_{H_0} L(\mu, \beta)} = \frac{\max_{\mu,\beta} L(\mu, \beta)}{\max_{\mu} L(\mu, 0)}\,.$$

Instead of averaging over the unknown parameters, the frequentist approach uses the maximum over the set of possible parameters. We will see later that this difference is most marked when the likelihood is quite flat, which typically occurs when there is limited information in the data (in the GWAS context, for small sample sizes or at rare SNPs).

5.4 Models & priors

I will use the same logistic regression models as for the frequentist analyses, described in Section 1.3.2 and 4.1. I describe how to implement such models in general, but focus primarily on the additive model given its prominent usage and central importance to GWAS.

A case-control study involves a retrospective sample of individuals, but these models assume it to be prospective. As noted in Section 1.3.2, it has been shown that this assumption does not make a difference when inference concerns odds ratios, so in what follows we can proceed as if the sampling were prospective.

5.4.1 Reparameterisation

An important step in a Bayesian analysis is the specification of priors on the parameters. For example, the additive model has two parameters: β, the parameter of interest, and μ, a nuisance parameter. One approach, which has worked well in practice, is to specify independent priors on both parameters, ensuring that the prior on μ is diffuse and has negligible impact on the analysis (WTCCC 2007, MARCHINI ET AL. 2007). I now describe a different approach, based on reparameterising the model, that does not require the prior on μ to be specified. In fact, I show more generally that under this approach we also do not require a prior for any confounding parameters. This idea was developed independently by WAKEFIELD (2009) for one-parameter disease models and is similar to that of KASS & VAIDYANATHAN (1992).

Briefly, the idea is to reparameterise the model as described in Section 4.2, which then leads to the marginal likelihood factorising into components attributable to each parameter. Assuming independent priors on the parameters then also gives independent posteriors and calculation of the BF will only involve the priors on the disease parameters.

Using the general modelling framework from Section 4.1 and the reparameterisation described in Section 4.2, the BF is a ratio of the following marginal likelihoods,

$$\Pr(\text{data} \mid H_1) = \iint L(\boldsymbol{\nu}, \boldsymbol{\beta})\, \pi(\boldsymbol{\nu}, \boldsymbol{\beta})\, d\boldsymbol{\nu}\, d\boldsymbol{\beta}$$
$$\Pr(\text{data} \mid H_0) = \int L(\boldsymbol{\nu}, \mathbf{0})\, \pi(\boldsymbol{\nu})\, d\boldsymbol{\nu}\,.$$

For large sample sizes, the likelihood will be asymptotically Gaussian up to a scale factor (COX & HINKLEY 1974). Furthermore, the nuisance and disease parameters are asymptotically uncorrelated after reparameterisation. Thus, asymptotically the likelihood factorises into a product of two Gaussian distributions and a scale factor, $L(\boldsymbol{\nu}, \boldsymbol{\beta}) = k f_0(\boldsymbol{\nu}) f_1(\boldsymbol{\beta})$. If we place independent priors on the two sets of parameters, the joint prior also factorises, $\pi(\boldsymbol{\nu}, \boldsymbol{\beta}) = \pi(\boldsymbol{\nu})\pi(\boldsymbol{\beta})$. The marginal likelihoods thus factorise into contributions due to $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$,

$$\begin{aligned}
\Pr(\text{data} \mid H_1) &= k \left(\int f_0(\boldsymbol{\nu})\, \pi(\boldsymbol{\nu})\, d\boldsymbol{\nu} \right) \left(\int f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta} \right) , \\
\Pr(\text{data} \mid H_0) &= k \left(\int f_0(\boldsymbol{\nu})\, \pi(\boldsymbol{\nu})\, d\boldsymbol{\nu} \right) f_1(\mathbf{0}) .
\end{aligned}$$

In calculating the BF, the $\boldsymbol{\nu}$ contributions cancel (providing the same prior is used for $\boldsymbol{\nu}$ under both H_0 and H_1, which is natural), leaving a k-dimensional integral that only depends on the prior for $\boldsymbol{\beta}$ and (a component of) the likelihood,

$$\text{BF} = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} = \int \frac{f_1(\boldsymbol{\beta})}{f_1(\mathbf{0})}\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta} . \tag{5.1}$$

This approach neatly avoids the need to specify a prior on $\boldsymbol{\nu}$, which includes the baseline effect and any covariates. This is useful since they are nuisance parameters and we would generally have little prior knowledge on the effect of covariates in combination with a genetic effect. The assumption of independence is not overly restrictive. Even if a prior on $\boldsymbol{\nu}$ were required to be specified, such an assumption is likely to be made in practice. For example, this was done in the WTCCC (2007) where such a prior was also chosen to be diffuse.

Another advantage of this approach is that it involves lower-dimensional integrals, which allows for easier, faster and more accurate computation. This is likely to be more important when there are many covariates, since without them we only save on a single dimension (the baseline parameter).

In practice, in the scenario with no covariates and using an additive model, the reparameterisation approach gives very similar results to using a relatively diffuse prior for μ. Comparing the two approaches on subsamples of the WTCCC data gave $\log_{10}(\text{BF})$ values to within

0.005 for sample sizes down to 100 cases and 100 controls, and to within 0.02 for samples as small as 10 cases and 10 controls (data not shown).

5.4.2 Effect size estimation

A Bayesian analogue to the MLE is the posterior mode, $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$, the value of $\boldsymbol{\beta}$ that maximises its posterior probability density function. Other point estimates are also natural in the Bayesian setting (BERNARDO & SMITH 1994), like the posterior mean $\mathbb{E}(\boldsymbol{\beta} \mid \mathrm{data})$, but many of these will give nearly identical estimates in our context because the posterior distribution of $\boldsymbol{\beta}$ will be approximately normally distributed (see Section 5.6.2).

The Bayesian analogue to a confidence interval is a probability interval calculated from the posterior distribution, called a *credible interval*.

5.4.3 Priors on model parameters

The priors we specify on disease model parameters should capture our knowledge of plausible effect sizes for complex human diseases. While such knowledge is still developing, it is at least clear that we expect to observe mostly weak to moderate effects at common SNPs. With this in mind, I now suggest some priors for the additive and general models.

A subtle point to remember is that these methods are primarily aimed at analysing SNPs on genome-wide SNP arrays. These cover a large number of SNPs, but certainly far from all of them. Hence, we are generally searching for SNPs that are correlated with causal loci, rather than causal SNPs themselves. This has implications for our choices for priors, especially with regard to the general model (see below).

Here I only consider priors for which the effect size does not depend on the allele frequency. I later discuss MAF-dependent priors in Section 5.10.

Additive Model

There is now quite an extensive extensive range of confirmed associations for many common diseases (MANOLIO ET AL. 2008, HINDORFF ET AL. 2009), showing risk odds ratios in the range 1–2, with most of them in the range 1–1.5. (With respect to the the protective allele, the

corresponding odds ratio intervals are 0.5–1 and 0.67–1 respectively.) This suggests a prior centred on an odds ratio of 1, with most of its weight on odds ratios in the interval 0.67–1.5, and very little weight outside of 0.5–2. An example of such a prior (using an additive model) is $\beta \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma^2)$ with $\sigma = 0.2$, as was used by the WTCCC (2007). A very similar one was suggested by WAKEFIELD (2007). This prior places probability 0.96 on the odds ratio for the risk allele being less than 1.5, and 0.999 on it being less than 2.

It is important to realise that there is not just a single, natural prior to use for β. Different priors can be used, depending on the assumptions that one wishes to make. In particular, less is known about true effects at rare SNPs, giving a large scope to plausible assumptions. Whatever the choice of prior, ideally it should have some theoretical or empirical motivation.

In order to illustrate the effect of different priors, I use three different values of σ in later analyses, $\sigma = 0.2,\ 0.5,\ 1.0$. Prior probabilities and 95% probability intervals for these priors are given in Table 5.1, in terms of the risk OR. This is the OR of the risk-conferring allele, and can be defined in terms of the additive parameter as follows,

$$\text{Risk OR} = \max\left(e^{\beta}, e^{-\beta}\right) .$$

By definition, the risk OR has an implied lower bound of 1 (otherwise it would be the OR of the protective allele). As it is the most conservative of these three priors, for convenience I will refer to the prior with $\sigma = 0.2$ as the *conservative* prior.

Increasing σ corresponds to an expectation of larger effect sizes. While $\sigma = 1$, and maybe even $\sigma = 0.5$, seem unrealistic for common SNPs for many common human diseases, some researchers may feel that a larger range of effect sizes is plausible for rarer alleles.

General Model

Most loci discovered by GWAS show little deviation from an additive model.[2] This may well be due to our currently low power to detect such deviations, particularly when we do not test the causal SNP directly (I show this in Chapter 7). Consequently, not much is known

[2]Actually, many studies and reviews (e.g. HINDORFF ET AL. 2009) only report additive effects, but even where studies look for deviations from an additive model (e.g. WTCCC 2007) they generally do not find them.

Table 5.1: **Probabilities and 95% intervals for three priors under the additive model.** Intervals are for the risk odds ratio and are displayed as '$< x$' to mean '$1 <$ risk OR $< x$' since they have an implied lower bound of 1 (see explanation in the text). The middle columns show the probabilities of the corresponding intervals for each prior. The priors are specified by the standard deviation (σ) for the log odds ratio (the additive parameter, β), which is normally distributed with mean 0.

	$\Pr$(Risk OR interval)			
σ	< 1.3	< 1.5	< 2	95% Probability interval
0.2	0.81	0.96	0.999	< 1.48
0.5	0.40	0.58	0.83	< 2.66
1.0	0.21	0.31	0.51	< 7.10

about general true disease effect sizes. However, since we expect to be mainly analysing SNPs that are correlated with causal loci, and that both empirical and theoretical evidence points to the fact that such loci will tend to be closer to an additive model (see Chapter 7), using a prior centered on the additive model seems appropriate.

While various priors are possible, as a starting point I suggest using $\beta \overset{\mathrm{d}}{=} \mathrm{N}(0, 0.2^2)$ as in the additive model, and independently the same distribution also for γ. This choice centres the general model on the additive model, and models deviations from additivity in a way that is symmetric with respect to homozygotes and heterozygotes. In particular, under the additive model, the log-odds of disease for either of the homozygotes differs from that of the heterozygote by β; under the general model, the log-odds for the heterozygote varies from its value under the additive model by γ, which has the same distribution as β under the proposed prior.

Note that my parameterisation of the general model, and hence also my suggested prior, is different to that used in the WTCCC (2007). Mine has the extra parameter, γ, incorporated linearly as is standard in a regression model, whereas in the WTCCC it scales the homozygote log-odds multiplicatively (see equation (1.2)). I believe my parameterisation is easier to interpret, with γ modelling deviation from an additive model in an additive way similar to β. This more naturally suggests a suitable prior, as described above. In contrast, in the WTCCC a diffuse prior was used for the extra parameter. Another advantage of my parameterisation is that, due to its linearity, it has better numerical properties. Using the Newton-Raphson method and a Laplace approximation to calculate the BF (see Section 5.5), I found that it was often difficult to find the mode with the WTCCC parameterisation, and

the shape of the likelihood did not appear to be close to Gaussian, meaning that the Laplace approximation is likely to be inaccurate (data not shown).

5.4.4 Prior on odds of association

The BF measures the evidence of association at a SNP given the data and our prior on effect sizes. To determine if the evidence is convincing, we need to combine it with our prior odds that the SNP is associated. For an arbitrary SNP in the genome, this is likely to be very low. Our knowledge of common diseases is not yet mature enough to give a definite answer, so this is still an open question. Suggestions in the literature place prior odds of association on the order of 10^{-5} to 10^{-4} (BALDING 2006, WTCCC 2007, WAKEFIELD 2009). These are based on arguments that speculated on the total number of 'independent' genomic regions and how many of them will show an appreciable genetic effect. For example, an expectation of 10 regions amongst 1,000,000 gives prior odds of 10^{-5}. This means that a $\log_{10}(\mathrm{BF})$ of 5 is required to have a posterior probability of association of 0.5 (i.e. even odds).

Clearly, plausible values can vary by an order of magnitude or two, and so it is not surprising that researchers generally do not commit to a particular prior odds, preferring to use the BF directly. Another advantage to this is that some SNPs may be deemed to be more or less likely to be associated, for example if they are located near genes thought to be relevant to the disease. Reporting a BF thus allows different researchers to incorporate their own possibly different prior beliefs, whether this is done formally by specifying different prior odds, or informally by getting more excited about SNPs with moderate BFs that are in regions suspected to harbour causal variants.

The above derivation of a prior has the flavour of a crude multiple testing correction, but crucially it is independent of the number of tests actually conducted. It would be applicable whether we are testing just a single SNP or all SNPs on a genotyping chip.

5.5 Implementation

I follow the Laplace approximation approach of the WTCCC (2007) and MARCHINI ET AL. (2007), but extend it to the general modelling framework and simplify it using reparame-

terisation as described in Section 5.4. Others have also used a Laplace approximation for similar calculations (GUAN & STEPHENS 2008).

To calculate the BF we need to evaluate the marginal likelihoods, which are integrals over the parameters in the models we are comparing. For large sample sizes the likelihood will be asymptotically Gaussian, suggesting that these integrals can be accurately calculated using a Laplace approximation (KASS & RAFTERY 1995). This involves approximating the logarithm of the integrand by a quadratic curve fitted at its mode (i.e. a Gaussian-shaped curve). For the d-dimensional integral,

$$I = \int h(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,,$$

where h has mode $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{H}$ is the Hessian matrix of $\log(h)$ at $\hat{\boldsymbol{\theta}}$, the Laplace approximation gives,

$$\log(I) \approx \log h(\hat{\boldsymbol{\theta}}) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|-\boldsymbol{H}|\ .$$

Using the reparameterisation, we only need to calculate the single integral given by equation (5.1). Since the likelihood only factorises asymptotically, I actually calculate the integral using the full likelihood evaluated at the maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$, as follows. Firstly, using the Newton-Raphson method I maximise $L(\boldsymbol{\mu}, \boldsymbol{\beta})\pi(\boldsymbol{\beta})$ to simultaneously find both the MLE, $\hat{\boldsymbol{\mu}}$, and the posterior mode, $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$. I then apply a Laplace approximation at $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$ to calculate the integral,

$$\mathrm{BF} = \int \frac{L(\hat{\boldsymbol{\mu}}, \boldsymbol{\beta})}{L(\hat{\boldsymbol{\mu}}, \mathbf{0})}\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta}\,.$$

One difference to the WTCCC implementation is that I do not need to average over the possible allele codings when fitting the additive or general models. Previously the prior was not symmetric with respect to the (arbitrary) allele coding. This was undesirable and was solved by calculating the BF for both possible codings (either counting the A allele or the B allele) and then averaging them. The reparameterisation makes this unnecessary. To see this, consider the reparameterised additive model,

$$\mathrm{logit}\,(p_i) = \nu + \beta\left(G_i - \bar{G}\right)\ .$$

Let G_i' be the genotype under the alternative allele coding. Thus, we have $G_i' = 2 - G_i$. The

same relationship will hold for the new genotype mean,

$$\bar{G}' = \frac{2n_0 + n_1}{N} = \frac{2N - (n_1 + 2n_2)}{N} = 2 - \bar{G}\,.$$

Therefore, the model under the alternative allele coding is,

$$\text{logit}\,(p_i) = \nu + \beta\left(G'_i - \bar{G}'\right) = \nu - \beta\left(G_i - \bar{G}\right)\,.$$

The two codings are equivalent when using a prior symmetric around 0 (which is natural). A similar derivation holds for the general model.

If there are missing or uncertain genotype calls, the extra uncertainty can be taken into account by integrating over the possible genotypes. For this purpose we require posterior distributions on the genotype calls, as would be produced by a Bayesian genotype calling algorithm such as CHIAMO (`http://www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html`). If data is completely missing, then it may be possible to impute it using an algorithm such as IMPUTE (MARCHINI ET AL. 2007). The integration involves summing over the genotype posteriors for each individual. Since we assume individuals are independent when calculating the likelihood (and its derivatives), this extra summation can be done on a per-individual basis and so is computationally feasible. In practice, it is more efficient and convenient to simply threshold the posterior calls to a moderately high value (such as 0.9, as in the WTCCC study) and exclude any calls not reaching the threshold. For SNPs where this leads to high missing data rates, either the genotype calls or the raw data are likely to be poor quality. By themselves, such SNPs will not usually be convincing enough to discover new associations and will often be excluded after downstream QC anyway.

The WTCCC implementation is available in the software package SNPTEST (`http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html`). I created a custom version of this where I incorporated the reparameterisation and my version of the general model.

5.5.1 Accuracy of the Laplace approximation

In order to check the accuracy of the Laplace approximation, I compared it to two other numerical integration techniques: importance sampling and adaptive quadrature (EVANS & SWARTZ 1995). For the former, I took 1,000 samples from a Gaussian proposal distribution centered on the MLE and with the inverse of the Fisher information as its covariance matrix. For the latter, I used the quadrature routines as implemented in the functions `integrate` and `adapt` (for 1-dimensional and 2-dimensional integrals respectively) in the R software package (R DEVELOPMENT CORE TEAM 2007). Given the relatively large sample sizes, the likelihood was highly peaked and required a transformation of variables in order for these routines to converge successfully. I used a modal transformation based on a multivariate t-distribution as described by GENZ & KASS (1997). After some experimentation, I found that a transformation with 12 degrees of freedom and a scale factor of $\delta = 1.4$ worked well. Using the WTCCC version of the additive model (with a diffuse prior on μ), and applying all three approaches to subsamples of the WTCCC data, I observed differences in $\log_{10}(\mathrm{BF})$ to be less than 5×10^{-4} even for samples as small as 10 cases and 10 controls (data not shown). Thus, as well as being fast to compute, the Laplace approximation is highly accurate for this application.

5.6 Asymptotic results

Using a normally distributed prior and the fact that the MLE is asymptotically normally distributed, we can derive an asymptotic expression for the BF. I show how to do this in the general modelling framework and also give a simple formula for one-parameter disease models. I then explore and visualise the relationship between the BF and other quantities in the context of the additive model, to better understand its behaviour.

5.6.1 Asymptotic BF

From Section 4.1, we know that $\hat{\boldsymbol{\beta}}$ asymptotically has a multivariate normal distribution,

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathrm{N}_k(\boldsymbol{\beta},\, V) \ ,$$

where $V = \mathcal{I}^{-1}_{\beta|\mu}$. Let the prior on $\boldsymbol{\beta}$ also have a multivariate normal distribution,

$$\boldsymbol{\beta} \stackrel{\mathrm{d}}{=} \mathrm{N}_k(\mathbf{0},\, W)\ .$$

Consider the integral in equation (5.1). Asymptotically, the integrand involves only multivariate normal density functions and can be simplified as follows.

Define the following quantities,

$$\begin{aligned} \Sigma^{-1} &= V^{-1} + W^{-1}\,, \\ R &= \Sigma V^{-1}\,, \\ \boldsymbol{\alpha} &= R\hat{\boldsymbol{\beta}}\,. \end{aligned}$$

From these we have the following identity, which can be verified by expanding the brackets,

$$\begin{aligned} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} V^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \,+\, (\boldsymbol{\beta} - \mathbf{0})^{\mathrm{T}} W^{-1} (\boldsymbol{\beta} - \mathbf{0}) \\ = \hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}} \,-\, \boldsymbol{\alpha}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\alpha} \,+\, (\boldsymbol{\beta} - \boldsymbol{\alpha})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\alpha})\,. \end{aligned}$$

This identity allows us to re-write the product of the two Gaussian densities in equation (5.1) as just a single density,

$$\begin{aligned} &\int f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta} \\ &\quad \to \int \frac{1}{(2\pi)^{\frac{k}{2}}\, |V|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} V^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \frac{1}{(2\pi)^{\frac{k}{2}}\, |W|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{0})^{\mathrm{T}} W^{-1} (\boldsymbol{\beta} - \mathbf{0})}\, d\boldsymbol{\beta} \\ &\quad = \int \frac{1}{(2\pi)^{k}\, |V|^{\frac{1}{2}}\, |W|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} V^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\boldsymbol{\beta} - \mathbf{0})^{\mathrm{T}} W^{-1} (\boldsymbol{\beta} - \mathbf{0})\right)}\, d\boldsymbol{\beta} \\ &\quad = \frac{|\Sigma|^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}\, |V|^{\frac{1}{2}}\, |W|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\alpha}\right)} \int \frac{1}{(2\pi)^{\frac{k}{2}}\, |\Sigma|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\alpha})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\alpha})}\, d\boldsymbol{\beta}\,. \end{aligned}$$

The integral on the last line is equal to 1 since it integrates a Gaussian density, giving,

$$\begin{aligned} \int f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta} \;&\to\; \frac{|\Sigma|^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}\, |V|^{\frac{1}{2}}\, |W|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\alpha}\right)} \\ &= \frac{|\Sigma|^{\frac{1}{2}}}{|W|^{\frac{1}{2}}} e^{\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\alpha}}\, \frac{1}{(2\pi)^{\frac{k}{2}}\, |V|^{\frac{1}{2}}} e^{-\frac{1}{2}\hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}}} \\ &= \frac{|\Sigma|^{\frac{1}{2}}}{|W|^{\frac{1}{2}}} e^{\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\alpha}}\, f_1(\mathbf{0})\,. \end{aligned}$$

Substituting into equation (5.1) asymptotically gives,

$$\mathrm{BF} = \int \frac{f_1(\boldsymbol{\beta})}{f_1(\mathbf{0})}\,\pi(\boldsymbol{\beta})\,d\boldsymbol{\beta} = \frac{|\Sigma|^{\frac{1}{2}}}{|W|^{\frac{1}{2}}} e^{\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\alpha}}\,.$$

This is more simply expressed on a log scale,

$$2\log \mathrm{BF} = \log\frac{|\Sigma|}{|W|} + \boldsymbol{\alpha}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\alpha}\,.$$

The two terms in this formula both have intuitive interpretations, relating to the posterior distribution (see below for details of the posterior). The first compares the variances of the prior and posterior, having a range from $-\infty$ (posterior has zero variance, the limiting case of the data being very informative) to 0 (posterior and prior have equal variance, the limiting case of the data being completely uninformative). The second is the Mahalanobis distance of the prior mode (the origin) from the posterior mode (MAHALANOBIS 1936). It has a range from 0 (no disease effect) to ∞ (limiting case of a very strong/definite disease effect). Together, they combine to give the BF, measuring the evidence of association. An approximate intuition is that the second term measures the strength of association, while the former modifies it to take some uncertainty into account. However, it is hard to see how the two terms interact when viewed in this way. I return to this question later below when I consider single-parameter models, and also again in Section 5.7 where I focus particularly on understanding the effect on the BF of quantities relevant to GWAS, like the OR, MAF, sample size and choice of prior.

An alternate simple formula for the BF, which will later prove useful, can be derived as follows. Firstly, re-write the above expression to get,

$$2\log \mathrm{BF} = \log\left|\Sigma W^{-1}\right| + \hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1}\Sigma V^{-1}\hat{\boldsymbol{\beta}}\,.$$

Since $\Sigma W^{-1} + \Sigma V^{-1} = \Sigma\left(W^{-1} + V^{-1}\right) = I$, we have that $\Sigma W^{-1} = I - R$. This gives the simple expression,

$$2\log \mathrm{BF} = \log|I - R| + \hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} R\hat{\boldsymbol{\beta}}\,. \tag{5.2}$$

5.6.2 Asymptotic effect size posterior

Similarly, we can derive a simple expression for the asymptotic posterior distribution of $\boldsymbol{\beta}$ under H_1. From above,

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \text{data}) &= \frac{\int L(\boldsymbol{\nu}, \boldsymbol{\beta})\, \pi(\boldsymbol{\nu}, \boldsymbol{\beta})\, d\boldsymbol{\nu}}{\int L(\boldsymbol{\nu}, \boldsymbol{\beta})\, \pi(\boldsymbol{\nu}, \boldsymbol{\beta})\, d\boldsymbol{\nu}\, d\boldsymbol{\beta}} \\
&= \frac{k \left(\int f_0(\boldsymbol{\nu})\, \pi(\boldsymbol{\nu})\, d\boldsymbol{\nu} \right) f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})}{k \left(\int f_0(\boldsymbol{\nu})\, \pi(\boldsymbol{\nu})\, d\boldsymbol{\nu} \right) \left(\int f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta} \right)} \\
&= \frac{f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})}{\int f_1(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta})\, d\boldsymbol{\beta}} \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}\, |\Sigma|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\alpha})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\beta}-\boldsymbol{\alpha})}\,.
\end{aligned}$$

This is a multivariate normal density with mean $\boldsymbol{\alpha} = R\hat{\boldsymbol{\beta}}$ and covariance matrix $\Sigma = RV$. In other words,

$$\hat{\boldsymbol{\beta}} \mid \text{data} \xrightarrow{d} \mathrm{N}_k\Big(R\hat{\boldsymbol{\beta}},\, RV\Big)\,. \tag{5.3}$$

5.6.3 Single-parameter models & shrinkage

In order to better understand these results it is helpful to consider single-parameter disease models ($k = 1$). For such models the BF can be further simplified,

$$2 \log \mathrm{BF} = \log\left(\frac{V}{V+W}\right) + \left(\frac{W}{V+W}\right)\frac{\hat{\beta}^2}{V}\,.$$

This is equivalent to the asymptotic BF of WAKEFIELD (2009). Following the example of WAKEFIELD (2007), let $r = W/\left(V+W\right)$ and $z = \hat{\beta}/\sqrt{V}$, to give,

$$2 \log \mathrm{BF} = \log\left(1-r\right) + rz^2\,. \tag{5.4}$$

This allows us to gain further insight into the BF. The quantity $z = \hat{\beta}/\sqrt{V}$ is the Wald test statistic, the square of which will be asymptotically equal to the score test statistic. It has a one-to-one relationship with the p-value, $|z| = \Phi^{-1}\left(1 - p/2\right)$. The quantity r takes a value between 0 and 1, measuring the relative contributions of the prior and likelihood to the inference. Values of r closer to 1 indicate a larger contribution from the likelihood (i.e. the

data), which is slightly clearer if we rewrite it in terms of the Fisher information,

$$r = \frac{V^{-1}}{V^{-1} + W^{-1}} \,.$$

Compared with the general results above, we can see that r is the one-dimensional version of R. Thus, the asymptotic posterior distribution of β is $\mathrm{N}\Big(r\hat{\beta}, rV\Big)$. We see that both the mean and variance of the posterior have been 'shrunk' by the factor r as compared to the distribution of the MLE. For this reason I will refer to r as the *shrinkage factor*.

Figure 5.1 shows this shrinkage effect on the posterior mode,

$$\hat{\beta}_{\mathrm{MAP}} = r\hat{\beta} \,,$$

for two SNPs. The SNP in panel A has much flatter likelihood than that in panel B. In other words, there is less information in the data for the first SNP. Correspondingly, it shows greater shrinkage (using the same prior for both SNPs).

The degree of shrinkage can be quite noticeable when the variance of the MLE is comparable to, or greater than, the variance of the prior. Conversely, shrinkage will be minimal when the MLE variance is much smaller. This will occur, for example, when the data is very informative or when using a diffuse prior.

5.6.4 Usage in calculations

Given the large sample sizes in GWAS, these asymptotic results will be very good approximations to the BF at all but very rare SNPs. For this reason, it has been suggested that they be used directly for calculation of the BF (WAKEFIELD 2007, 2009), avoiding the need for procedures that are more computationally intensive. This is a sensible suggestion given the adequacy of the approximation and the large number of SNPs that need to be analysed in a GWAS.

Our implementation using a Laplace approximation (see Section 5.5) is similar to using the asymptotic formula since: (i) both approaches require maximising the likelihood/posterior of the parameters; (ii) the calculations involved in the Laplace approximation and the asymptotic formula are roughly equivalent and relatively minimal; and (iii) they both rely on the asymptotic normality of the MLE for their accuracy.

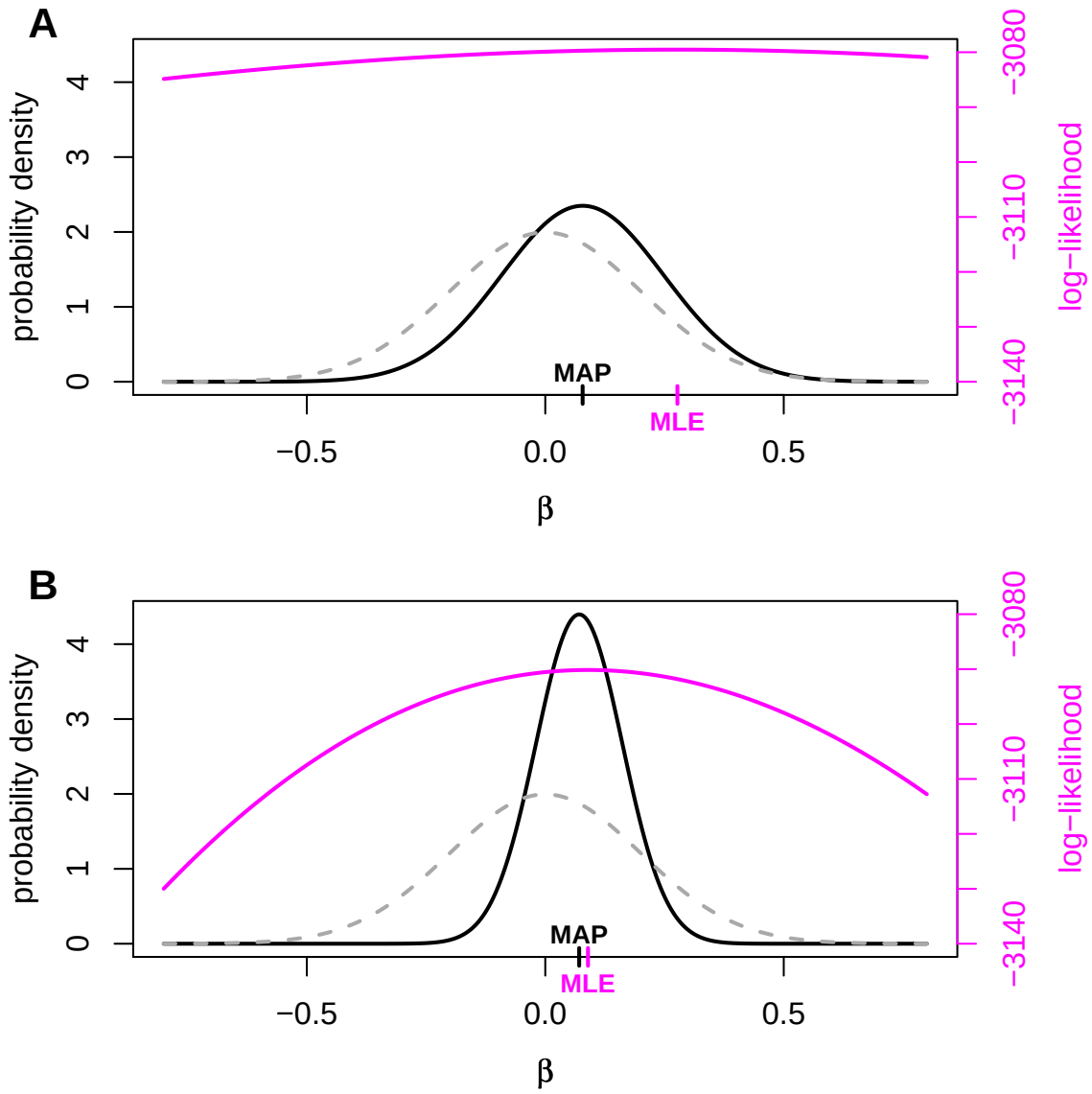

Figure 5.1: **Shrinkage demonstration.** The prior (grey), posterior (black) probability densities and conditional log-likelihood (magenta) for the additive effect, for two SNPs from the WTCCC data. The sample included approximately 2,000 cases and 3,000 controls. Calculations are under the conservative prior ($\sigma = 0.2$). The SNP in panel A has MAF = 0.44%, giving it a shrinkage factor of $r = 0.28$; the SNP in panel B has MAF = 4.9%, giving it a shrinkage factor of $r = 0.80$. The maxima of both the posterior and the log-likelihood (MAP and MLE, respectively) are marked on the x-axis, highlighting the greater shrinkage at the first SNP.

One difference and potential advantage of our implementation is improved numerical stability when the likelihood is very flat, which would occur for SNPs with low allele frequencies. Using the formula directly as suggested by Wakefield involves first finding the MLE. Doing so when the likelihood is very flat would lead to poor convergence. In contrast, in my implementation I maximise the posterior instead, which would be regularised by the prior and thus converge successfully. Having said this, in practice the SNPs for which this might be a problem are likely to be excluded from the study due to allele frequencies that are too low.

5.7 Visualising & understanding the BF

I now explore the behaviour of the BF under different scenarios, using the asymptotic formulae from the previous section. I look at the impact of different factors on the BF and aim to give an intuitive explanation of its behaviour. I focus primarily on the additive model but much of the intuition extends to BFs under more general models. The key idea in each instance is that the BF summarises the evidence between two given models, and will reflect the amount of information in the data for distinguishing between them.

For a given study, the sample size is common across SNPs (up to exclusions after QC), while the MAF and OR are expected to vary. I first look at how the BF relates to these two quantities, and then later examine the effect of changing the sample size and the prior.

Figure 5.2 shows how the BF relates to the MAF and OR at a SNP in a sample. The two plots show two different views of essentially the same information. Recall that a positive $\log_{10}(\text{BF})$ value represents evidence in favour of the disease model, whereas a negative value is evidence in favour of the null model. A value of zero represents ambivalence, having no effect on our prior odds. We can observe the following features of the BF:

- **The BF increases with the OR.** This is as expected, since the BF represents the evidence that the OR is not equal to 1. Moreover, this happens no matter what the MAF is, although it happens differently for different MAFs.

- **The spread of BFs, for the same range of ORs, increases with the MAF.** Common SNPs are more informative than rarer SNPs, and the BF reflects this with a greater abil-

ity to discriminate between the two models. As the MAF drops to 0, the information in the data slowly vanishes and the BF range diminishes correspondingly.

- **Increasing the MAF can make the BF go up *or* down.** This is an important point and perhaps not immediately intuitive. It reflects the last point about common SNPs being more informative, which can mean that they show stronger evidence of being associated *or* of being null. It also reminds us that the OR by itself does not capture all the information in the data.

- **The BF prefers the null model for very low MAFs.** Because the prior is centered on the null model, some minimum amount of evidence is required in order for the BF to show a preference for the disease model.

- **For a given OR, the BF is not necessarily monotone with respect to MAF.** This is a consequence of the last two observations, and shows how the BF balances the information provided by both the MAF and OR. When the OR is very close to 1 (essentially null), the BF steadily drops as the MAF increases. However, for larger ORs the BF starts to increase again when the MAF is large enough. This occurs sooner, and quicker, for larger ORs.

- **There is a minimum OR that is required for the BF to prefer the disease model.** The BF prefers the null model when the OR is 1, and steadily increases as the OR increases. At some OR value it starts to prefer the disease model. The exact value that the OR breaks out of this 'null zone' depends on the MAF, with a smaller OR required as the MAF increases. The lowest it can be is when the MAF is 0.5, which in this figure is at an OR slightly above 1.1.

This last point might seem slightly counter-intuitive, given that we have defined the null model to be strictly when OR $= 1$. We can imagine having a situation where the observed OR is seemingly not close to 1, while the BF can be telling us there is good evidence in favour of the null. Remember that the BF is weighing up the evidence between the two models, and that distinguishing between an OR of 1 and an OR near 1 requires substantially informative data. In that respect, it may be helpful to think of the null model as being OR ≈ 1, rather than the point null OR $= 1$ actually used in calculations, and that we are just using the latter as a convenient approximation to the former.

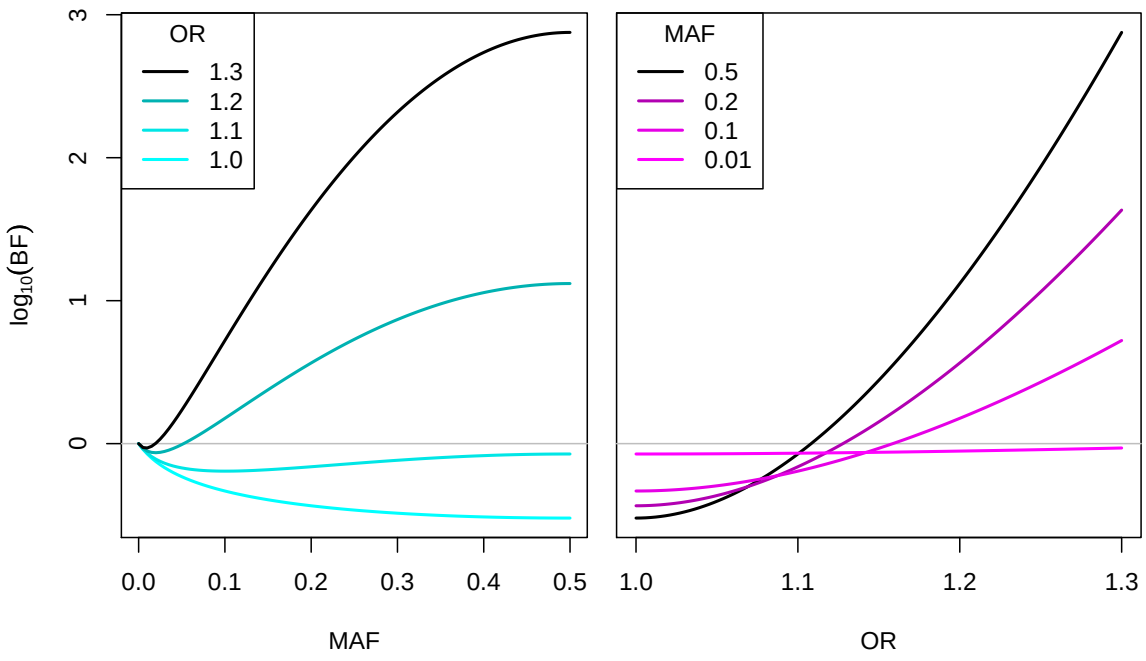

Figure 5.2: **Asymptotic BF relationships.** Effect of the sample MAF ($\bar{f}$) and OR ($e^{\hat{\beta}}$) on the BF (shown on a log scale). Calculated for an additive model using the asymptotic formulae in equations (4.6) and (5.4), under the conservative prior ($\sigma = 0.2$) and for a sample with 1,000 cases and 1,000 controls.

Figure 5.3 shows the effect of changing from the conservative prior ($\sigma = 0.2$) to one which allows a greater range of effect sizes ($\sigma = 1$). We can see that the second prior requires greater ORs before preferring the disease model. It also has a stronger preference for the null model at low ORs than does the conservative prior. Looking at the asymptotic formula (equation (5.4)) helps to explain these effects. The second prior, which is more diffuse and exhibits less shrinkage, will have a greater r value at each SNP. For small ORs, the BF is dominated by the $\log(1-r)$ term, which will be lower. For greater ORs, the BF is dominated by the rz^2 term, which means that the lines actually increase at a faster rate and will eventually 'overtake' the corresponding lines for the conservative prior (not visible in this figure). The exact location of where the lines cross the x-axis depends on the interplay between these two terms, but it is easy to show, by setting the BF to 0 and rearranging the asymptotic formula, that this follows a monotonically increasing relationship with r.

Figure 5.4 shows the effect of increasing the sample size. This has the effect of decreasing the variance of the MLE, V, which increases both r and z. We thus see less shrinkage, similar to what we observed with a more diffuse prior, but this time because the data is

Table 5.2: **Genotype counts for an example SNP.**

	AA	AB	BB
Cases	1,500	3,000	1,500
Controls	1,332	2,990	1,678

more informative rather than our prior less restrictive. In addition, we also have greater power to detect smaller effect sizes, with the lines crossing the x-axis at lower OR values (this is driven by the rz^2 term). Note that increasing the sample size has little effect when the MAF is very low. From equation (4.6), we see that this is because for rare SNPs the MAF dominates and results in a very large variance. In particular, letting the minor allele count be n, the variance becomes approximately proportional to $1/n$, whereas it would generally otherwise be proportional to $1/N$. Thus, at rare SNPs the minor allele count takes on the role of the sample size. Therefore, any increase in the number of samples will only substantially affect the BF if it significantly boosts the minor allele count.

Bear in mind that here I have only examined priors independent of MAF. It is possible to use priors where the assumed range of effect sizes varies according to MAF. I discuss using such priors in Section 5.10.

With a thorough understanding of BFs now under our belt it is worth re-visiting the examples from Section 5.1. Three scenarios were postulated, all giving the same p-value of 8×10^{-6} but with very different numbers of samples. In the first scenario, there was no data and the p-value was generated at random. The corresponding Bayesian analysis gives $\log_{10}(\text{BF}) = 0$, indicating there is no evidence for or against association in the (non-existent) data. In the second scenario, there are 10 cases all with the AA genotype, and 10 controls all with the BB genotype. Under the conservative prior these give $\log_{10}(\text{BF}) = 0.7$. The third scenario features 6,000 cases and 6,000 controls. While a large range of possible genotype counts consistent with this sample size give rise to the required p-value, I show a particular example in Table 5.2. Under the conservative prior these gives $\log_{10}(\text{BF}) = 3.4$, much larger than for either of the previous scenarios. These differences in the BF mirror the level of 'persuasiveness' of the same p-value in the corresponding scenarios. This illustrates how the BF automatically takes into account the relative information in the data and provides a more interpretable summary of the evidence of association.

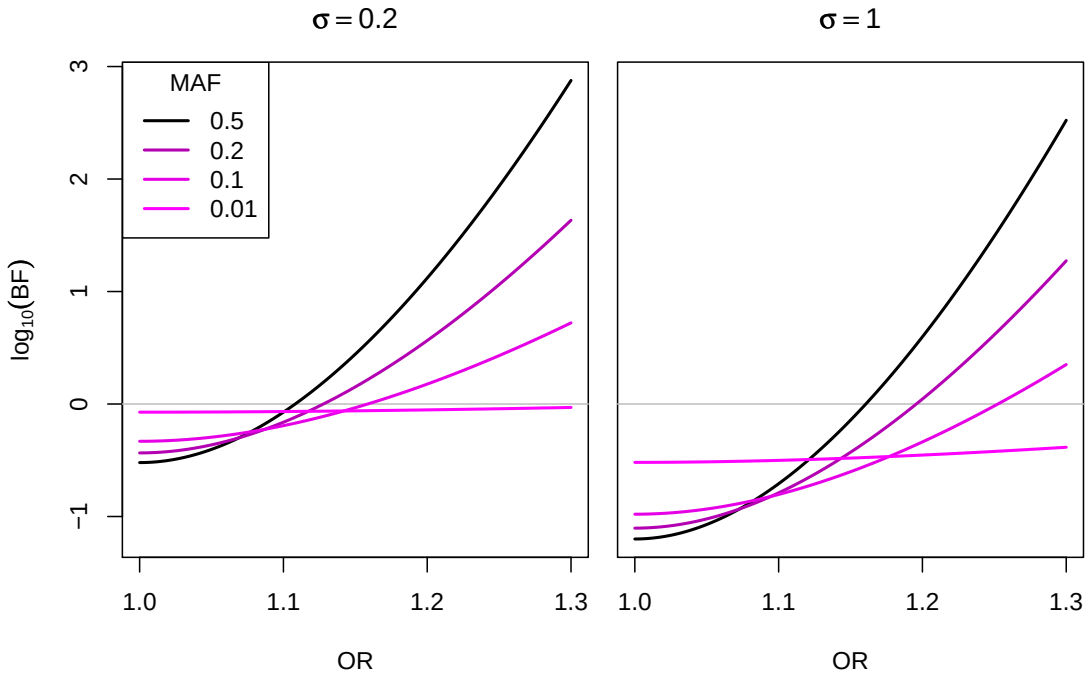

Figure 5.3: **Effect of using a different prior.** The right-hand plot from Figure 5.2 repeated for two different priors (as labelled) for a sample with 1,000 cases and 1,000 controls.

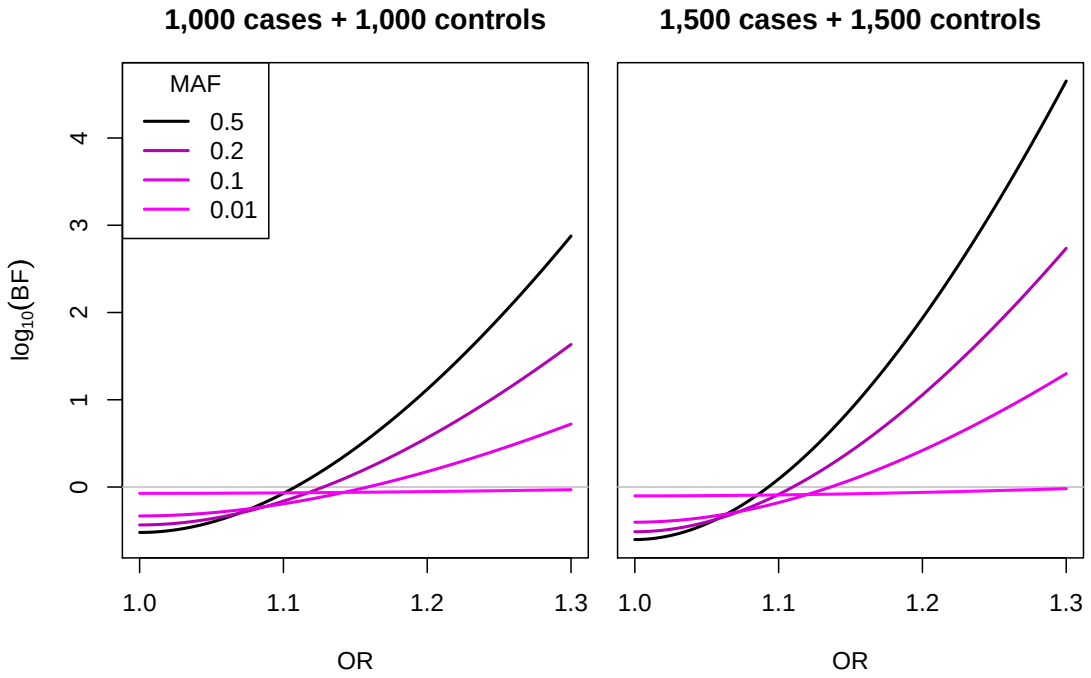

Figure 5.4: **Effect of sample size.** The right-hand plot from Figure 5.2 repeated for two different sample sizes (as labelled) under the conservative prior.

5.8 Equivalence of rankings under a g-prior

Given the popularity of the p-value as a measure of evidence, and having described the BF as an alternative, a natural question to ask is whether there exists a prior for $\boldsymbol{\beta}$ that gives a one-to-one relationship between them? Such a prior would give equivalent rankings of SNPs in a given study. There is indeed such a prior; here I describe it and show how it gives us further insight into the assumptions underlying frequentist approaches to GWAS analysis. In particular, I show that it is equivalent to a Bayesian procedure which assumes larger effect sizes at rarer SNPs. WAKEFIELD (2009) independently derived this same prior in the context of single-parameter models. My exposition here differs in that I show the result for multi-parameter models, and use the variance approximations from Section 4.4 to give an intuitive description of these priors in context of the additive model.

To motivate the prior, examine equation (5.4). This shows how the BF relates to the p-value (via z) for a single-parameter disease model. For a one-to-one relationship to exist we would need r to be constant across all SNPs in the study. If we just set r to be a constant, we effectively set the prior variance to be a constant multiple of the variance of the MLE,

$$W = g^{-1}V\,,$$

where g is a positive constant. WAKEFIELD (2009) independently derived this same result and a similar prior was discussed in the context of a normal model by COX & HINKLEY (1974, pp. 395–9). These are all examples of the so-called g-prior (ZELLNER 1986, SMITH & SPIEGELHALTER 1980, KASS & WASSERMAN 1995), which is a Gaussian prior where the mean can be set arbitrarily (typically to a null model value) and a covariance matrix that is a scalar multiple of the covariance matrix of the MLE. Thus, the variance of the g-prior can be interpreted as being the variance of the MLE after some number of pseudo-observations of the data (not necessarily an integer). Increasing g corresponds to increasing the number of pseudo-observations, making the prior more informative. The special case where $g = 1/N$ is equivalent to one pseudo-observation and is called the *unit information prior* (KASS & WASSERMAN 1995).

Using equation 5.2, I show that a g-prior also gives a one-to-one relationship for multi-parameter models. Letting $W^{-1} = gV^{-1}$ gives $\Sigma^{-1} = (g+1)V^{-1}$ and $R = \Sigma V^{-1} = I/(g+1)$.

Therefore,

$$2 \log \mathrm{BF} = \log\left(\frac{g}{g+1}\right) + \left(\frac{1}{g+1}\right) \hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}} \,.$$

The term on the right, $\hat{\boldsymbol{\beta}}^{\mathrm{T}} V^{-1} \hat{\boldsymbol{\beta}}$, is the multivariate Wald test statistic (see Section 4.5), which asymptotically follows a χ^2_k distribution under the null model and will have a one-to-one relationship with a p-value. Since g is constant, the BF will also have such a one-to-one relationship.

Notice that g-priors have a variance that depends both on the data (in our context this being mainly on the MAF and sample size) and also on an arbitrary constant (g). The one-to-one relationship between p-values and BFs holds for any value of g, so it is actually a family of priors that result in this correspondence.

To put these results in context, for a given study a Bayesian using a g-prior would rank SNPs in terms of strength of evidence for association in exactly the same way as a frequentist who uses p-values. Since the Bayesian paradigm makes various underlying assumptions explicit, this can be a helpful device for shedding further light on the frequentist approach. I now explore this further in the context of the additive model.

Let the g-prior on β be distributed as $\mathrm{N}\big(0, \sigma_g^2\big)$. In other words, $\sigma_g^2 = g^{-1} V$. Using the variance approximation from equation (4.6), we can express σ_g as a function of the MAF and sample size,

$$\sigma_g^2 = \frac{g^{-1}}{2N\bar{f}(1-\bar{f})\phi(1-\phi)} \,. \tag{5.5}$$

It is instructive to see how σ_g^2 depends on the MAF. Figure 5.5 illustrates this relationship. I have chosen g such that the g-prior has the same variance as the conservative prior ($\sigma = 0.2$) for $\bar{f} = 0.5$ and for given fixed values of N and ϕ; I will refer to this as the '0.2 g-prior'. Table 5.3 shows 95% probability intervals for this g-prior for a selection of MAFs. From these we see that the g-prior acts similarly across common SNPs but assumes a progressively larger range of effect sizes as the MAF decreases. Indeed, from equation (5.5) we see that $\sigma_g^2 \to \infty$ as $\bar{f} \to 0$. In other words, a Bayesian procedure which behaves the same as the trend test assumes larger effect sizes at rarer SNPs, with the assumed effect sizes becoming very much larger as the MAF approaches 0. This will be true irrespective of the actual value chosen for g.

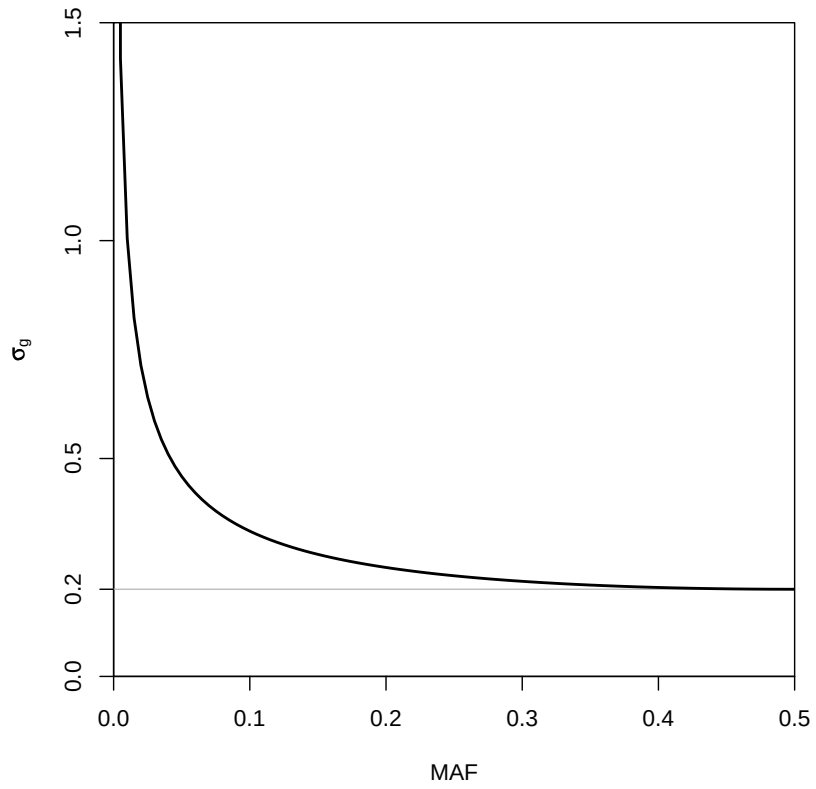

Figure 5.5: **The 0.2 g-prior.** The prior standard deviation (σ_g) against MAF ($\bar{f}$) for the g-prior scaled to have a standard deviation of 0.2 when $\bar{f} = 0.5$.

An intuitive way to see why this might be so is to remember that p-values are uniformly distributed. For this to be true at all SNPs, we require that the value of the MLE at which we deem an effect to be 'significant' to scale with its variance. For the Bayesian procedure to match the frequentist one, the prior must scale with the MLE variance correspondingly. Since the variance at rarer SNPs is greater, stronger effects must be assumed at such SNPs.

The dependence of σ_g on the sample size illustrates an important and subtle point. Above, I chose g in order to obtain a certain prior variance at a particular MAF. Due to the dependence on N, doing the same for studies with different sample sizes implies a different value of g for each. That is, if we fixed the prior variance of effect sizes for a particular MAF to reflect prior beliefs, and considered two GWAS experiments with different sample sizes, the use of p-values under the trend test corresponds to *different* Bayesian procedures for the two experiments. This reinforces the point that the interpretation of p-values depends (amongst other things) on the sample size in the experiment, and that direct comparison of p-values between different experiments is not straightforward. It also demonstrates that specifying a fixed p-value threshold for 'genome-wide significance' (DUDBRIDGE & GUSNANTO 2008,

Table 5.3: **The 0.2 g-prior.** Shows the prior standard deviation (σ_g) and 95% probability intervals on the risk OR for given MAFs.

MAF	σ_g	95% Probability interval
0.5	0.2	< 1.48
0.1	0.33	< 1.92
0.05	0.46	< 2.46
0.01	1.01	< 7.17

PE'ER ET AL. 2008) is flawed—while the idea that the threshold should be based on properties of the genome rather than the actual number of tests is helpful, advocating the use of the same threshold for all studies is problematic. In particular, smaller studies will often require a more extreme p-value than larger studies (for a SNP with the same MAF) for the evidence to be as compelling (WTCCC 2007). In practice, the situation is further complicated by the fact that, even in one study, the actual sample sizes will differ from SNP to SNP due to missing data (although this should be a small effect for most SNPs). Thus, p-values will strictly speaking not even be comparable within a study.[3] BFs do not suffer from this disadvantage, by nature they will always be directly comparable across SNPs and across studies.

In the context of studies using imputed data, GUAN & STEPHENS (2008) have recently pointed out that under the g-prior the assumed effect size also depends on the imputation accuracy, since this will also affect the variance of the MLE. In other words, in this scenario the frequentist approach implicitly assumes that SNPs which are harder to impute will have larger effect sizes. This is counterintuitive and further illustrates the difficulties in comparing any two given p-values.

One apparent difference between the Bayesian and frequentist approaches is that additional assumptions are needed to calculate the BF. In particular one needs to make assumptions, encapsulated in a prior distribution, about the likely sizes of the genetic effect. On closer examination this difference between the approaches is somewhat illusory. To paraphrase I. J. Good, Bayesian approaches need to be explicit about the assumptions they make, whereas many of the assumptions underlying frequentist approaches are often implicit (GOOD 1976). More precisely, recall that GWAS are often conducted as part of a design where the SNPs are ranked and a fixed number (depending on funds) from the top of this list are followed up in a replication study. We could choose to rank by the BF or the p-value.

[3]Furthermore, this all assumes that one actually believes the g-prior is sensible. Otherwise, even when sample sizes do not differ, p-values will only be comparable between SNPs with the same MAF.

Under such a design, the correspondence under the g-prior highlights some assumptions inherent in the frequentist approach, such as that it implicitly allows much larger effect sizes for rarer SNPs than for common SNPs.

Given these inherent assumptions, an obvious question to ask is whether a g-prior is appropriate for studying complex diseases? To answer this, we need a good grasp of the range of effect sizes at rare variants. There is an expectation amongst many researchers and some evidence that larger effect sizes are more likely to be observed at rarer SNPs (BODMER & BONILLA 2008). Despite this, the g-prior's rapid increase in assumed effect size as the MAF approaches 0 seems a reasonably strong and quite specific assumption, which some authors have deemed implausible (e.g. GREENLAND 2008). It should be possible to answer the question empirically. While our current knowledge is limited, in Chapter 6 I use confirmed loci from two common diseases to partially tackle this question, by comparing the performance of p-values and BFs using the priors I discussed earlier in Section 5.4.3. I find that BFs tend to perform better, suggesting that the g-prior is not the best at capturing the true effect size distribution. Whatever set of priors are deemed to be the most satisfactory, the above discussion makes clear that an important advantage of the BF over the p-value is clarity of presentation and easier interpretation.

5.9 Frequentist properties of the BF

In the previous section I showed an equivalence between BFs and p-values when using a g-prior, which would result in them having equivalent frequentist properties for a given sample size. That is, if we were to use the BF as a test statistic, where we specify a BF threshold for declaring a 'significant' association, the resulting test would have the same operating characteristics (power and false positive rate (FPR)). However, we are not limited to using the g-prior. In this section, I explore the operating characteristics of the BF when using a prior independent of MAF, such as the conservative prior. I only consider the BF for the additive model. My exposition is particularly relevant to those more familiar with frequentist methods, providing an intuitive hook to gain further insight into the Bayesian approach.

WAKEFIELD (2008) has previously suggested using the asymptotic BF relationships as a basis to explore operating characteristics (as I do below), and has shown power calculations based on the posterior odds for a range of priors, sample sizes and allele frequencies. My exposition here differs in that I explore the operating characteristics of the BF, explaining its general behaviour and how it relates to the trend test.

The asymptotic formula in equation (5.4) relates the BF to the trend test statistic (approximately equal to z^2), which we know asymptotically follows a non-central χ^2_1 distribution. We also have accurate approximations of the non-centrality parameter (see Section 4.5), giving us a complete description of the distribution of the BF. We can use this result to calculate the operating characteristics of the resulting test. Let the BF threshold be t. Rearranging equation (5.4), this corresponds to the z score,

$$z = \sqrt{\frac{1}{r}\log\left(\frac{t^2}{1-r}\right)}\,.$$

Using the two-tailed Wald test formulation gives,

$$\text{Power} = \Phi\left(-z + \frac{\beta}{\sqrt{V}}\right) + \Phi\left(-z - \frac{\beta}{\sqrt{V}}\right), \tag{5.6}$$

$$\text{FPR} = 2\,\Phi\left(-z\right). \tag{5.7}$$

This highlights an important difference between the BF-based test and one based on a p-value threshold. In the latter we control the FPR directly. For the former the FPR depends on r which can vary across SNPs (and studies), so thresholding the BF does not control FPR directly. However, controlling the worst-case FPR is possible as I show below.

In Section 5.8, we saw that equation (5.4) gives a one-to-one relationship between the BF and p-value, but only for a fixed value of r. For a fixed prior and sample size, r is only fixed if the MAF is also fixed. In other words, the BF and p-value share the same operating characteristics *for a given MAF*. How they perform overall depends on the allele frequency distribution of truly associated SNPs.

An example helps to make this clearer. Figure 5.6 shows the power of both types of tests for various MAFs. I have set the p-value threshold at 0.001, and the BF threshold to the value that gives equivalent characteristics when the MAF is 0.05. This choice of thresholds then

entails different characteristics at other MAFs. In particular, in this example we see that the BF has higher power and false positive rate at more common SNPs (MAF above 0.05), and vice versa for less common SNPs.

This difference in behaviour reflects the assumptions encoded in the prior. In particular, the same distribution of effect sizes is assumed at all SNPs. The rarer SNPs are inherently less informative and so are discounted.

Another way to look at this comparison is using receiver operating characteristic (ROC) curves. Figure 5.7 shows examples of such curves for a true OR of 1.3 and a range of MAFs and samples sizes. Remember that for a given MAF the two approaches have the same characteristics so will draw out the same ROC curve (grey). Now consider what happens when we specify fixed thresholds and vary the MAF. A fixed p-value threshold will select a point on each grey line with the same x-coordinate, keeping the false positive rate constant across MAFs. Plotting this point for all possible MAFs traces out the line in magenta. A fixed BF threshold selects a different set of points and gives the line in cyan. (Points at which the curves intersect correspond to MAFs at which the two thresholds have equivalent characteristics.)

For the smaller sample sizes we see what we have already observed—the BF threshold operates at a higher power and false positive rate as the MAF increases. However, we can also observe another effect. As the sample size increases, the BF-based test starts to 'retreat' back to operating at a lower power and false positive rate at the common SNPs, eventually meeting and 'crossing over' the operating characteristic of the p-value–based test.

From the plots, it appears as if each these curves (in cyan and magenta) trace out the same path on the different plots, with different samples sizes simply allowing them to extend by different amounts. In other words, it looks like each of them traces out a portion of a characteristic curve that does not depend on sample size. This is indeed the case, as I now show.

The statement is clearly true for the magenta curves, which trace out portions of a vertical line segment of unit length. The cyan curves trace out portions of a curve which starts at the origin, follows the curves shown in the figure, and then 'loops back' to the point $(0, 1)$. To

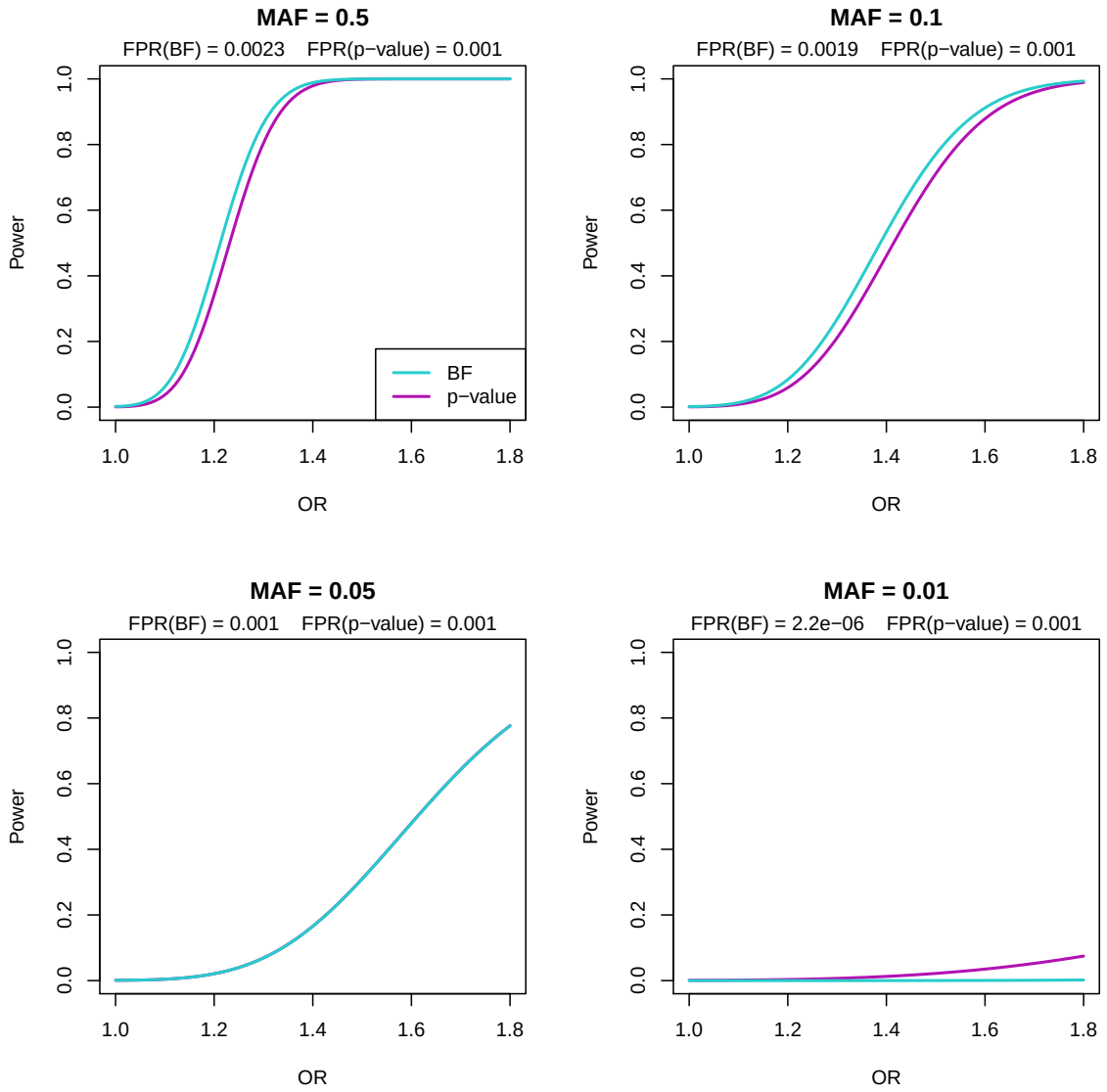

Figure 5.6: **Power curves for different MAFs for the additive test.** Based on a sample of 1,000 cases and 1,000 controls. The magenta curve is for a p-value threshold of 0.001, and the cyan curve for a $\log_{10}(\text{BF})$ threshold of 1.31 using the conservative prior ($\sigma = 0.2$). The false positive rate (FPR) for both curves is shown in the title of each plot. The two curves overlap exactly when the MAF is 0.05 due to a deliberate choice of thresholds.

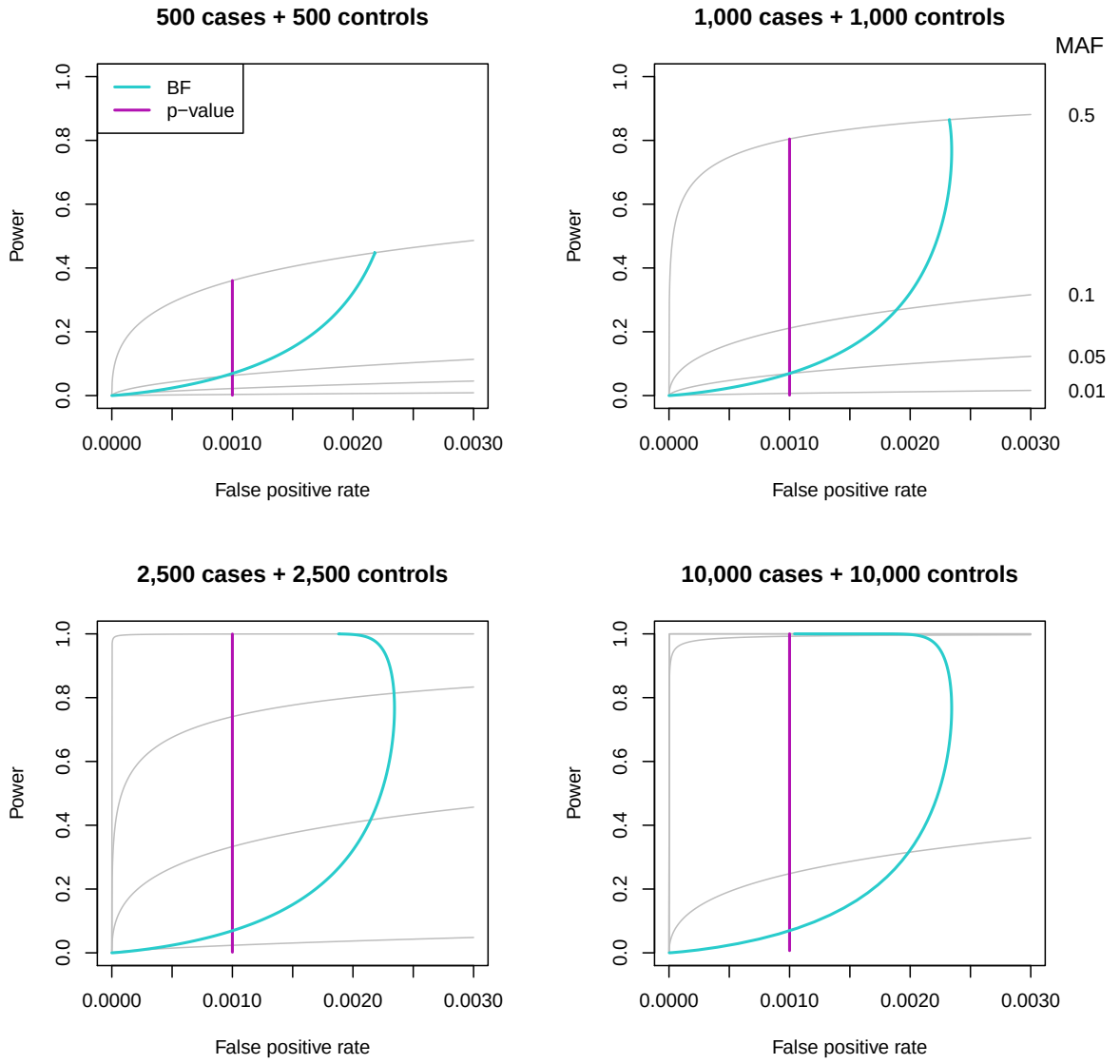

Figure 5.7: **ROC curves for different sample sizes for the additive test.** Based on a true OR of 1.3. Grey lines show ROC curves for the following MAFs: 0.5, 0.1, 0.05, 0.01 (from top to bottom in each plot, as labelled in the top-right plot). These apply to tests based on either the BF or the p-value. However, given a specific threshold on both, the actual operating characteristics vary across MAFs. The coloured lines show this for a p-value threshold of 1×10^{-3} and a $\log_{10}(\text{BF})$ threshold of 1.31 using the conservative prior ($\sigma = 0.2$).

see this, re-write equation (5.6),

$$\text{Power} = \Phi\left(-z + \frac{\beta}{\sqrt{W}}\sqrt{\frac{r}{1-r}}\right) + \Phi\left(-z - \frac{\beta}{\sqrt{W}}\sqrt{\frac{r}{1-r}}\right).$$

This defines the y-coordinate of points on the curve, and the x-coordinate is defined by equation (5.7). Both of these are parameterised by r, which is the only quantity that varies as the sample size changes (note that z is a function of r). Thus, the points trace out a one-dimensional curve. The limiting points of the curve can be derived by considering what happens as r approaches 0 or 1, and the worst-case FPR can be calculated from equation (5.7) using differential calculus (derivation not shown). Changing the prior (W) or the true OR (β) will only shift this curve in the vertical direction, because only the y-coordinate depends on these factors other than through r. Shifts in the horizontal direction, in particular a shift in the worst-case FPR, only occurs with a change in the BF threshold (t).

At least two observations come from this analysis. Firstly, it shows that the FPR for any given BF threshold is bounded, so it is possible to control for the (worst-case) FPR if desired. Secondly, when used as a classifier we see that the BF acts in a complex way given the various factors at play. In fact, it is acting in quite an intuitive and adaptive way, which can be described as follows:

- When the power is inherently low, be stringent, since otherwise we would get mostly false positives.
- When the power is respectable, be more liberal, since we now have a good chance of capturing some true positives.
- When the power is overwhelming, be more stringent again, since we are likely to capture the true positives anyway.

Another analogy is to think about how to set the correct volume on a sound amplifier. When the noise is stronger than the signal, you would turn down the volume because the noise is unpleasant. When the signal is fine, you would turn it up to hear it. When the signal is very strong, you can afford to turn it down and still be able to hear it.

I have so far only described the properties of a BF-based frequentist test and tried to give an intuitive explanation of its behaviour. The question of how to set an appropriate BF

threshold still remains. One way is to control the worst-case false positive rate, akin to a traditional frequentist approach. A more principled way is to use decision theory (e.g. BERGER 1993), which takes into account the relative 'costs' of making false positive and false negative decisions. For example, it may be deemed more important not to miss true associations than it is to follow up false leads (i.e. false negatives have higher 'cost' than false positives). WAKEFIELD (2008) shows how to do this in the GWAS testing context.

Here I have only discussed the behaviour of the BF when using priors independent of the MAF. In general, the BF can behave very differently with a different prior (as observed earlier with the g-prior). We can interpret these differences as, by choosing the prior appropriately, placing 'bets' on what kind of effect sizes we want to be powered to detect.

5.10 MAF-dependent priors

As discussed in Section 5.8, there is evidence that larger effect sizes are more likely to be observed at rarer SNPs (BODMER & BONILLA 2008), suggesting the use of priors that depend on the MAF. There is currently little empirical evidence to guide us on how to choose such a prior. Furthermore, current studies will generally have insufficient power for low MAFs, where such priors are most likely to differ from the MAF-independent ones suggested earlier. Nevertheless, we can think about convenient ways to implement such a dependency. Here I briefly describe some possible MAF-dependent priors. I do not use these for any analyses in this thesis, but they serve as a base for the further work discussed in Section 6.5.

The nature of non-additive effects in common diseases is still poorly understood, let alone how they depend on allele frequency, so I limit my discussion to the additive model. However, in principle the implementations below could be applied just as readily to the dominance parameter as for the additive parameter.

WAKEFIELD (2009) suggested a simple family of priors where the variance of the additive effect varies exponentially with MAF,

$$\sigma^2 = ce^{-df} ,$$

where c and d are constants. These constants can be set by, for example, specifying a de-

sired standard deviation at two different MAFs. When $f = 0.5$ (common SNPs), we could set it to 0.2 as before. Then we could specify a value when $f = 0$, which would be a limiting maximum value that is applicable for very rare SNPs. This would completely specify the prior, with intermediate MAF values having a spread of effect sizes consistent with the above exponential relationship.

Other functions of the MAF could also be used in this way, where we specify the limiting variances. For example,

$$\sigma^2 = ce^{-df(1-f)} .$$

This is similar to Wakefield's prior but has faster decay to the $f = 0.5$ variance.

The g-prior could also be considered as a potential choice.[4] It has quite a different functional form to the previous examples, with only one constant to set and the variance going to infinity at $f = 0$.

Examples of all three of these priors are compared visually in Figure 5.8. Which of these priors might be most appropriate for modelling effect sizes in common diseases, and what is an appropriate limiting variance for rare SNPs, are open questions. In Section 6.5 I give a brief description of a method to infer a good prior, although this is still work in progress.

All of the above suggestions are symmetric with respect to risk versus protective effects. It is possible to break this symmetry and have a different prior distribution of effect sizes depending on whether the allele in question increases or decreases the susceptibility to disease.

Implementing a MAF-dependent prior is complicated slightly by the fact that the true MAF (at the SNP being tested) is unknown, although is observed with some uncertainty. A proper analysis would involve specifying a prior on the MAF and taking this uncertainty into account. Given the large sample sizes of GWAS (and much prior information on most SNPs) it seems that we should be able to simply condition on the observed MAF. Whether this is adequate depends on the relative information in the MAF and effect size parameters. If the posterior for the MAF is highly peaked, and the effect size prior is relatively flat over the corresponding small range of MAFs, then conditioning will be adequate. For the priors

[4] This is not equivalent to simply using the p-value, despite the ranking equivalence described in Section 5.8. The interpretation of the BF differs to that of a p-value, and the ranking equivalence only holds for a fixed sample size.

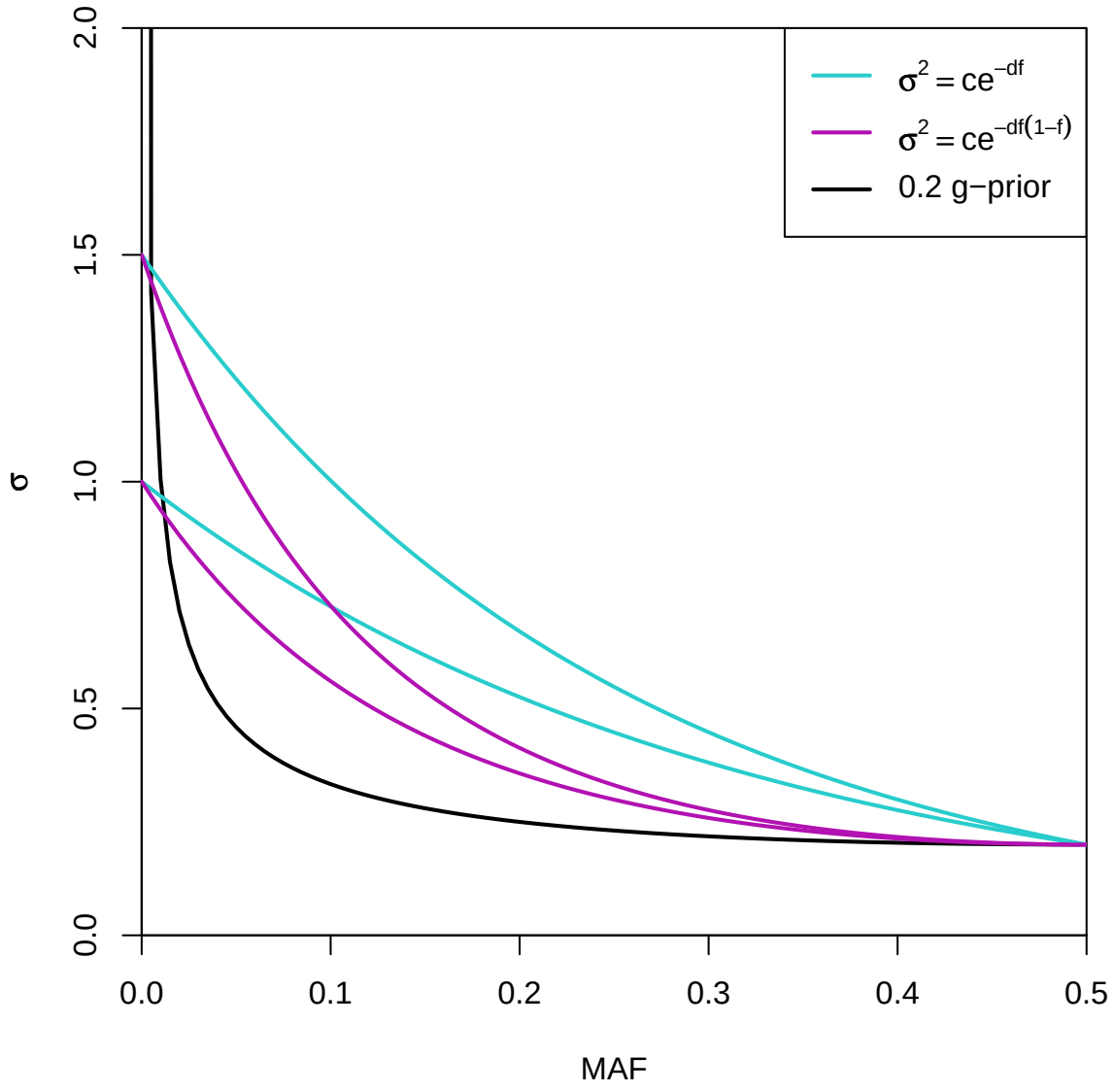

Figure 5.8: **MAF-dependent priors.** The prior standard deviation (σ) against MAF (f) for examples from three families of MAF-dependent priors. All are chosen to have $\sigma = 0.2$ when $f = 0.5$. This gives only one possible g-prior. Two example are plotted from each of the other two families, chosen to have $\sigma = 1, 1.5$ when $f = 0$.

considered above, this should suffice in all situations except for very rare SNPs when using the g-prior (where it is far from flat).

5.11 Extensions & alternative approaches

In this chapter I have introduced a Bayesian approach for convenient analysis of GWAS data. Focusing primarily on the single-SNP testing scenario, I discussed choices of prior distributions, derived asymptotic results for the BF and posterior, and explored connections with the frequentist approach. There is clearly much scope for further exploration, development and refinement in many directions. I comment just briefly on some possibilities in the single-SNP, case-control context.

Interval null prior. The BF approach I considered is equivalent to using a prior that is a mixture of a point mass on the null value (corresponding to H_0) and a distribution around this value (corresponding to H_1). This is sometimes referred to as a 'slab and spike' prior. The mixing proportions are defined by the prior odds of association. Use of a BF allows us separate these two aspects of the prior—the mixing proportions and the shape of the alternative distribution—which is convenient because we often only want to set the latter explicitly. However, depending on one's view of disease effects, the point mass on the null may be considered slightly unnatural. An alternative is to specify a prior on a small interval around the null value. This would correspond to a view that 'null' SNPs can have a very small, likely undetectable, effect on disease risk. This may have some benefits in terms of interpretation and robustness, although it does entail extra choices to be made in the specification of the prior. Furthermore, the two approaches can often lead to similar BFs (BERGER & DELAMPADY 1987, ROUSSEAU 2007), so we can think of each approach as being an approximation to the other. In that respect, use of the point null prior may be convenient even where it is not considered natural.

Missing phenotype model. The fact that cohort samples are generally used in place of proper control samples suggests using a missing phenotype model. We usually know the prevalence of diseases fairly well, so in principle incorporating this as a prior and summing over disease status for the cohort samples should allow a slight boost in power.

Meta-analyses. With many GWAS datasets now available, meta-analyses have begun to be published (e.g. BARRETT ET AL. 2008). Currently, none of these feature Bayesian analyses. Apart from the general advantages of the Bayesian approach, BFs naturally accommodate combining information from multiple studies (WAKEFIELD 2009). Given that meta-analyses are likely to be more important in the future, there is utility in developing methodology that also considers all the complications inherent in GWAS meta-analyses, such as population effects.

Chapter 6

Bayesian Analysis of GWAS II: Empirical Comparisons with Frequentist Approaches

Contents

In the previous chapter I introduced a Bayesian approach to GWAS analysis and showed how it compares to standard frequentist approaches theoretically. I now turn to comparisons that use data from actual GWAS and replication studies.

I adopt a practical perspective, asking how the approaches differ from the point of view of someone carrying out a GWAS. Taking replicated loci from two diseases, I compare the rankings of these loci in their respective original studies. The question of interest is which method ranks true associations higher. I show that the Bayesian approach offers a slight, but definite, improvement in ranking, particularly at smaller sample sizes. Following this, I look at how BFs and p-values compare across all SNPs in a study to gain further insight into the differences between the approaches. I explore the effect of sample size, allele frequency

and choice of prior distribution.

An important question that has great bearing on these sorts of comparisons is the choice of prior, and in particular how the effect size distribution depends on the allele frequency. At the end of this chapter, I sketch out a method to estimate the joint distribution of these two quantities using replicated GWAS loci.

6.1 Data & methods

My comparisons use data from studies of two different diseases: coeliac disease and Crohn's disease. GWAS and replication studies have recently been published for each of these diseases, all of them finding and confirming multiple disease loci.[1] The availability of true disease loci allows for more realistic comparisons of how various methods perform on actual data, without necessitating the assumptions required if using simulated data.

For each disease, I use the full GWAS data and the list of confirmed loci from the respective replication study. I describe the data sets in detail below. For each, I exclude SNPs and individuals according to the QC criteria of the original study. In addition, for some data sets and particular comparisons, I exclude further SNPs with very low MAF (details below). This is common in GWAS analysis, both because genotyping errors are more prevalent for rarer SNPs and because the asymptotic assumptions underpinning both the frequentist and Bayesian methods begin to fail at very low MAFs, where power is very low anyway. For simplicity, I only consider the autosomes (none of the confirmed loci were on the X chromosome).

While there are a range of possible approaches for analysing GWAS, many of which have been described earlier, I have focused on the approach taken by the majority of published GWAS—namely, single-SNP analyses of case-control studies using the additive model. In other words, I used the trend test p-value for the frequentist approach, and the BF as described in the previous chapter for the Bayesian approach. In addition, I focus on ranking of SNPs for prioritisation for follow-up. The studies above all used such an approach when selecting SNPs for replication. I now describe each data set in turn.

[1] Replication studies confirming multiple loci for many other disease have also been published since the completion of this work.

Coeliac disease. I used GWAS data from VAN HEEL ET AL. (2007). After QC filtering, the data set consisted of 778 cases and 1,422 controls typed at 301,658 autosomal SNPs. The study excluded SNPs having MAF less than 1%.

A subsequent replication study followed up 1,020 SNPs in a further 1,643 cases and 3,406 controls (HUNT ET AL. 2008). Nearly all of these SNPs were selected by taking those with the lowest p-value, with a few others chosen because they were non-synonymous. The human leukocyte antigen (HLA) genes, located in the major histocompatibility complex (MHC) on chromosome 6, are known to play a role in the disease (VAN HEEL & WEST 2006), so only non-HLA SNPs were considered in the replication study. The study identified 21 SNPs, in 8 genomic regions, as being convincingly associated with the disease.

Given that the HLA loci were ignored for replication, for my comparisons I excluded the 1,580 SNPs in the MHC from the initial study data set.

Crohn's disease. I used GWAS data from the WTCCC (2007). After QC filtering, the data set consisted of 1,748 cases and 2,938 controls typed at 456,502 autosomal SNPs. To explore the effect of sample size on the comparisons, I also emulated smaller association studies by subsampling from the full Crohn's disease data set. In particular, I created 12 subsamples each with the number of cases and controls both set to either 500, 1,000 or 1,500.

This study did not use a simple MAF filter as in coeliac disease above. In my analyses, I noticed that many of the extremely rare SNPs are obvious outliers in scatter plots of BFs against p-values. I applied a MAF filter to exclude these, choosing an appropriate threshold by gradually increasing it until all the outliers were removed. For the full sample of individuals, I used a MAF threshold of 0.25%, which excludes 51,252 SNPs. For the subsamples, I used thresholds of 0.7%, 0.5%, and 0.33% respectively, for the three sizes of subsample in increasing order. The number of SNPs so excluded was on average 52,709, 55,162, and 57,437 respectively, and varied slightly for each subsample (by no more than 300 SNPs from the respective mean).

A subsequent replication study used a meta-analysis of three GWAS, one of which was the WTCCC, to select SNPs for follow-up in a replication sample (BARRETT ET AL. 2008). The meta-analysis data consisted of 3,230 cases and 4,829 controls at 635,547 SNPs, with imputation methods (MARCHINI ET AL. 2007, LI ET AL. 2006) used to combine SNPs genotyped

on different platforms. Interest was focused on SNPs with a meta-analysis p-value less that 5×10^{-5}. There were 526 such SNPs, in 74 genomic regions. Replication was attempted for at least one SNP in each region, with 129 SNPs in total, in a new sample of 2,325 cases, 1,809 controls and 1,339 parent-parent-affected offspring trios. Including already known loci, 32 regions were thus identified as being associated with Crohn's disease.

6.2 Comparing rankings

The SNPs of most interest in a GWAS are those showing the strongest evidence of association, so I first focus on the question of choosing SNPs for follow-up. To compare the frequentist and Bayesian approaches, I applied them to the initial GWAS data then looked to see how they compare for SNPs which have subsequently been successfully replicated. I show that, under the conservative prior, the majority of such loci have somewhat higher ranks in the initial GWAS when ranked by their BF as compared to their rank based on p-value. That is, it would have been slightly more efficient to choose SNPs for follow-up using BFs rather than p-values. For now, I only use the conservative prior ($\sigma = 0.2$) when calculating BFs. In the following section I explore the relationship between BFs and p-values more generally, including the effects of different priors.

WAKEFIELD (2008) compared rankings in the context of SNPs simulated independently, a few of which were chosen to have a genetic effect that is weak to moderate and the rest chosen to have no effect. He also chose the effect size to be independent of allele frequency and used a prior of the same form. Under these conditions, he found that BFs gave better rankings than p-values. There is now enough real data available to allow for comparisons using results from replication studies, so I use real data throughout. My findings are consistent with Wakefield's conclusions based on simulated data.

In a typical study, SNPs in a GWAS would be ranked and the top ones followed up in a replication study, taking as many SNPs as funds allow, with the possibility of also including SNPs from lower down the list where biological candidacy is strong. In practice, many of the top SNPs would cluster together in the same small genomic regions due to LD, and only a few from each region would generally be selected for follow up. For example, this was true with the Crohn's disease replication study, but not for coeliac disease, where researchers

opted for the simpler approach of just taking all highly-ranked SNPs.[2]

We could try to take such clustering into account by splitting the genome into regions and then ranking those instead of SNPs, where we take the best SNP to represent each region. While this might be more representative of some actual replication studies, there are various problems with such an approach. Firstly, despite the fact that recombination hotspots induce a block-like haplotype structure in human populations (HAPMAP 2005), defining exactly where region boundaries should be drawn is not clear. Secondly, since replication is conducted on a per-SNP basis, only single SNPs are truly comparable in terms of representing the same genetic effect(s) and having the same probability of replication. While a replicated effect at a SNP allows us to conclude that the region contains a genetic effect, we cannot conclude that attempting replication at *any* SNP from that region will result in detecting it (other SNPs may not be in sufficiently high LD, or might even be surrogates for a different genetic effect in the same region). This disallows the use of different SNPs as representatives for the same region, as would be necessary when different measures of evidence select different best SNPs. Imposing a particular choice of representative SNPs for the purposes of comparison might therefore act to bias the comparison in favour of one of the methods.

These problems are not necessarily insurmountable, and further thought on overcoming them might be of value to make the comparisons more realistic. However, to keep the comparisons simple I ignore clustering of highly-ranked SNPs and just consider rankings of SNPs sequentially down the list. In terms of comparing the performance of the Bayesian and frequentist approaches, this should be of little consequence. Of interest is the general behaviour of the two approaches, seeing which one improves rankings of true associations, and this will be apparent whether or not I adjust for clustering.

Ignoring the representability issue described above, the effect on the high rankings of any adjustment would be essentially to thin them (removing strong correlated associations). Such thinning would nonetheless preserve any relative advantage.

Coeliac disease Figure 6.1 compares the rankings in the initial study of the 21 confirmed associated SNPs for coeliac disease. Nearly all SNPs show an improved ranking under the

[2]By a *high* rank I mean a rank closer to 1, the top rank, as opposed to a *low* rank by which I mean a rank further down the list away from 1. I use this convention throughout.

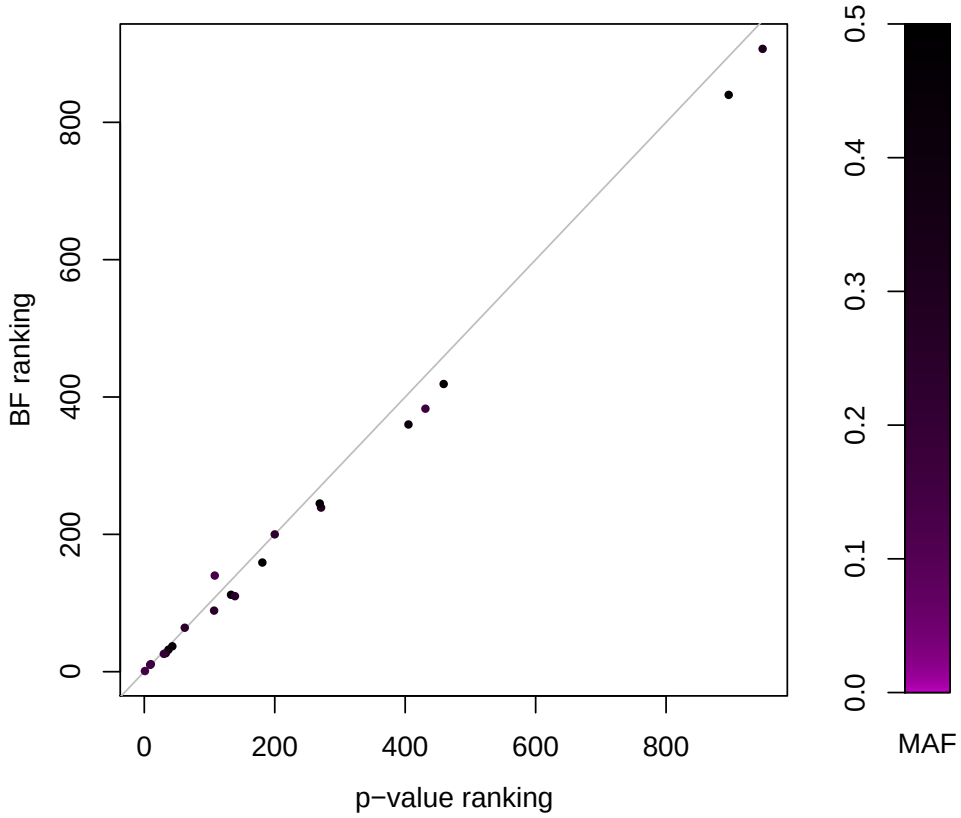

Figure 6.1: **Ranking comparison of replicated loci for coeliac disease.** Each point shows the rank, in the initial GWAS of coeliac disease, of each SNP shown to be associated with the disease in the replication study, under both the p-value and BF using the conservative prior ($\sigma = 0.2$). Points are coloured according to the MAF in the initial study, as shown in the legend on the right.

BF, with the SNPs ranked in the 800s showing an improvement of about 50 places. From this we can conclude that if fewer SNPs were followed up, depending on exactly how many, some of these replicated SNPs might have been missed if the p-value ranking were used for prioritisation but captured if the BF ranking was used instead. Conversely, since true loci that were discovered seem to come mostly from SNPs with higher ranks under the BF, then if the replication study was prioritised by such ranks we might expect to have observed more true loci from the same number of SNPs followed up.

The loci that successfully replicated had a range of allele frequencies. Figure 6.2 shows the replication success rate for the followed-up loci, split into five MAF classes. While not particularly conclusive, it shows that the more common loci were more likely to replicate (as would be expected, due replication power decreasing with MAF), with none of the loci with MAF less than 0.1 successfully replicating.

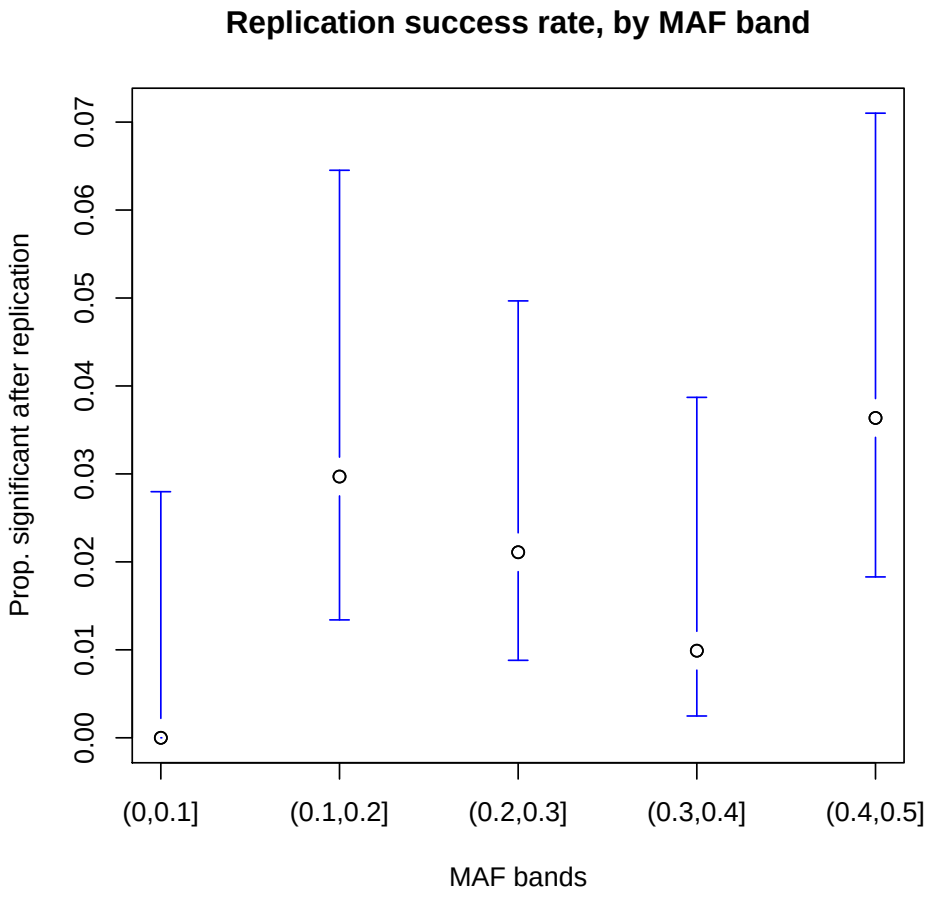

Figure 6.2: **Replication success rate by MAF.** The proportion of successful replications and 95% confidence intervals for the 1,020 SNPs followed up in the coeliac disease replication study, split into five MAF classes.

Crohn's disease Comparisons for Crohn's disease are less straightforward, owing to the more complicated SNP selection procedure used for follow-up. For simplicity, I treat the WTCCC study as being the initial study and use it to compare rankings. While it is only one of the three GWAS in the meta-analysis for prioritising SNPs, it represents the majority of the samples used and allows for a simpler comparison by avoiding issues to do with imputation.

For each successfully replicated genomic region, I selected a SNP to represent it when considering ranks. Where there was an overlap between SNPs in the replication study and those in the WTCCC data (i.e. the SNP taken to replication was on the genotyping chip), for each region I took the SNP with the strongest signal of association—this was always the same SNP under both p-value or BF. Five regions did not have any such SNPs. I represented these instead by the best tag SNP genotyped in the WTCCC study which showed some association signal. For the purposes of this analysis, I defined a tag SNP to be one having $r^2 > 0.5$ (in the HapMap CEU) with a SNP from the replication study.

I compared the rankings of these representative SNPs in the WTCCC data. I also compared rankings in subsamples of the WTCCC data, to see the effect of varying sample size (hence, power), using the same SNPs for all comparisons.

Figure 6.3 compares the rankings for the 32 confirmed loci. There is one plot for the full WTCCC data and one each for a subsample with different sample sizes. In each case I focus on SNPs with a rank of down to 5,000, both to allow comparison across the different sample sizes and since it is mainly the highest rankings that are of interest (some comparisons for lower rankings are shown in later figures). The differences in power with increasing sample sizes can be indirectly observed by the fact that the loci have progressively higher rankings (by both p-value and BF) as the sample size increases. We can see that differences in rankings are most pronounced when sample sizes are smaller, i.e. when power is lower, and that rankings become more similar as the power of the study increases, with essentially little difference at the WTCCC sample size. These conclusions are robust to choice of subsample—ranking comparisons for 12 replicates of each subsample size are shown in Figures 6.4–6.6. Where there are differences, rankings are generally improved under the BF. Occasionally the p-value gives a better ranking, notably for rare SNPs which replicate (for example, in the bottom-left plot of Figure 6.3, the point well above the diagonal corresponds to a SNP

with a MAF of 1.7%).

There is an ascertainment effect here which works against the BF. For both of the diseases I have considered, SNPs were chosen for follow-up based on p-values. When taking a fixed number of the most highly ranked SNPs, there will be SNPs that are only included if ranked by p-value, and an equal number that are only included if ranked by BF. If such SNPs replicate, they will favour the use of p-values and BFs respectively. In using replication data from p-value–based rankings, we exclude any SNPs in the second category. It is interesting that even with this effect, the BF generally performs better in these comparisons.

The same conclusions apply for Crohn's disease as for coeliac disease: using the BF to prioritise SNPs for follow-up is expected to lead to a greater number of successfully replicated loci, particularly when the study is not well-powered. Another conclusion that can be made is that replication should be attempted for more than just a small number of highly-ranked SNPs. The loci identified in these studies had ranks well in to the thousands, under both p-value or BF, suggesting that targeting loci ranked far down the list is likely to lead to new discoveries. Combining studies through meta-analysis is clearly also a fruitful strategy.

6.3 Comparing BFs & p-values

I now focus on the relationship between BFs and p-values more generally, to gain further insight into the two approaches.

Figure 6.7 shows a scatter plot of the BFs and p-values for all SNPs in the Crohn's disease data set, using the conservative prior ($\sigma = 0.2$) and with points coloured according to the MAF of each SNP. BFs and p-values are measured on different scales so it makes little sense to compare their numerical values directly. I instead focus on them as measures of evidence and look to see whether SNPs with stronger evidence of association in one measure also have stronger evidence in the other measure.

We see a broad agreement between BFs and p-values for common SNPs, and differences for rarer SNPs. Several SNPs show strong evidence of association on both measures (top-right section of the plot) and these all tend to be common SNPs, while the vast majority of SNPs show moderate to no evidence of association (bottom-left section of the plot). Consistent

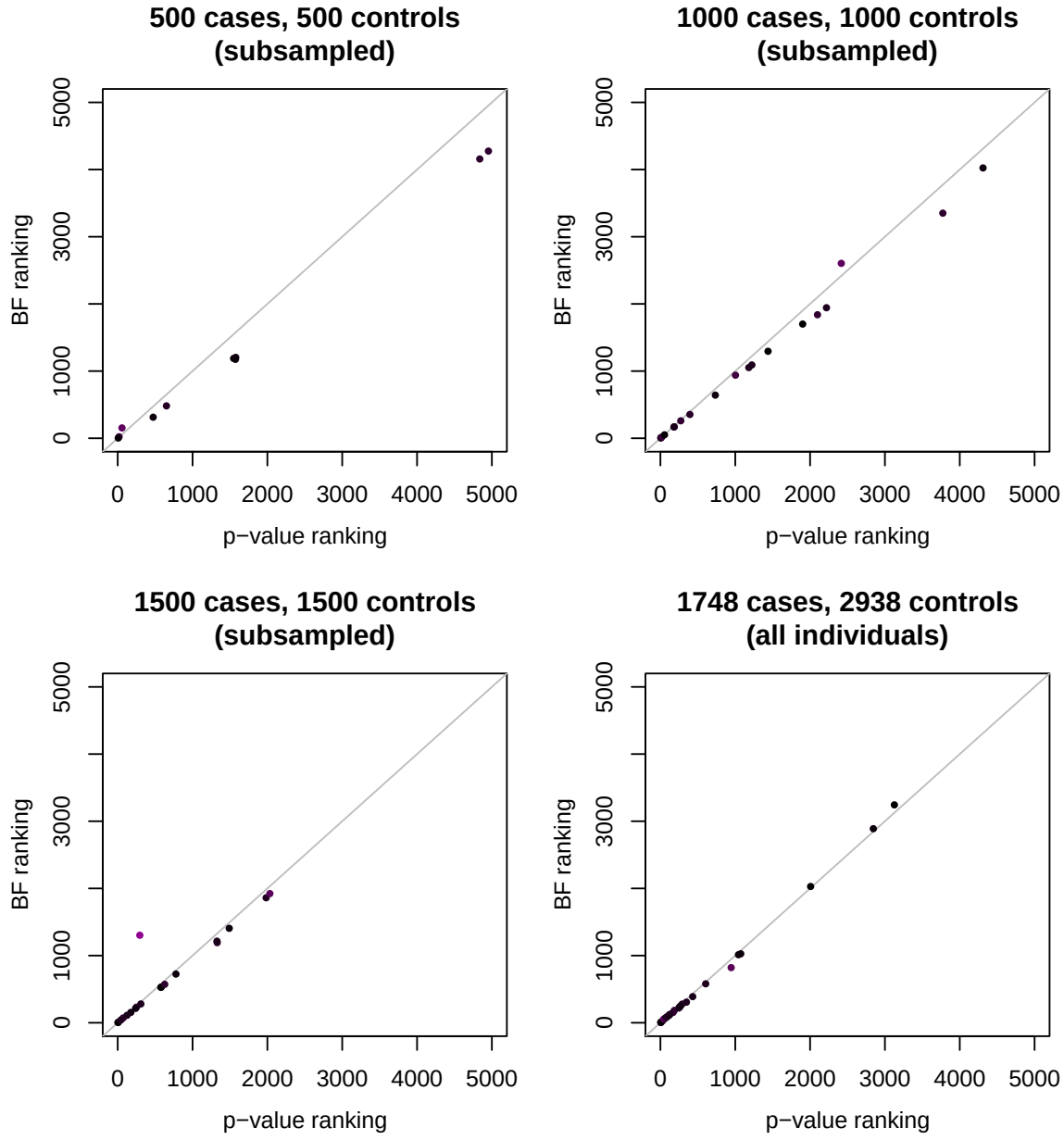

Figure 6.3: **Ranking comparison of replicated loci for Crohn's disease.** Each point shows the rank, in the WTCCC study or a subsample of it, of representative SNPs for each genomic region shown to be associated with the disease in the replication study, under both the p-value and BF using the conservative prior ($\sigma = 0.2$). Only SNPs ranked 5,000 or higher are shown. Each point is coloured according to the MAF of the study in which it is being ranked, according to the legend in Figure 6.1.

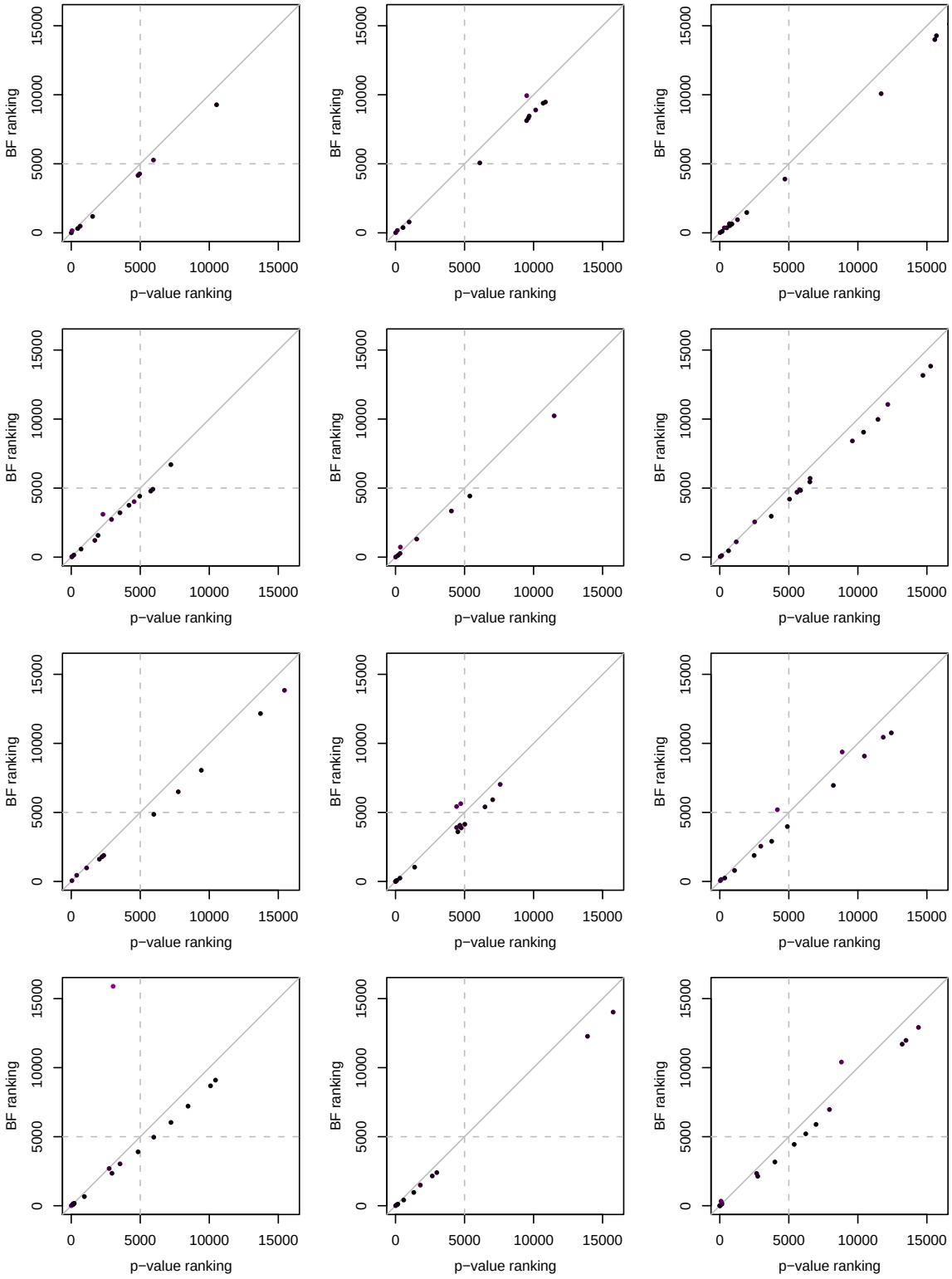

Figure 6.4: **Ranking comparison of replicated loci for Crohn's disease in subsamples.** Same as in Figure 6.3, but for 12 replicate subsamples of the WTCCC data with 500 cases and 500 controls. Note that for some subsamples, SNPs that were ranked higher than 5,000 under one measure did not do so under the other measure; the axis scaling is adjusted to accommodate them. The dashed lines show a rank of 5,000 for reference.

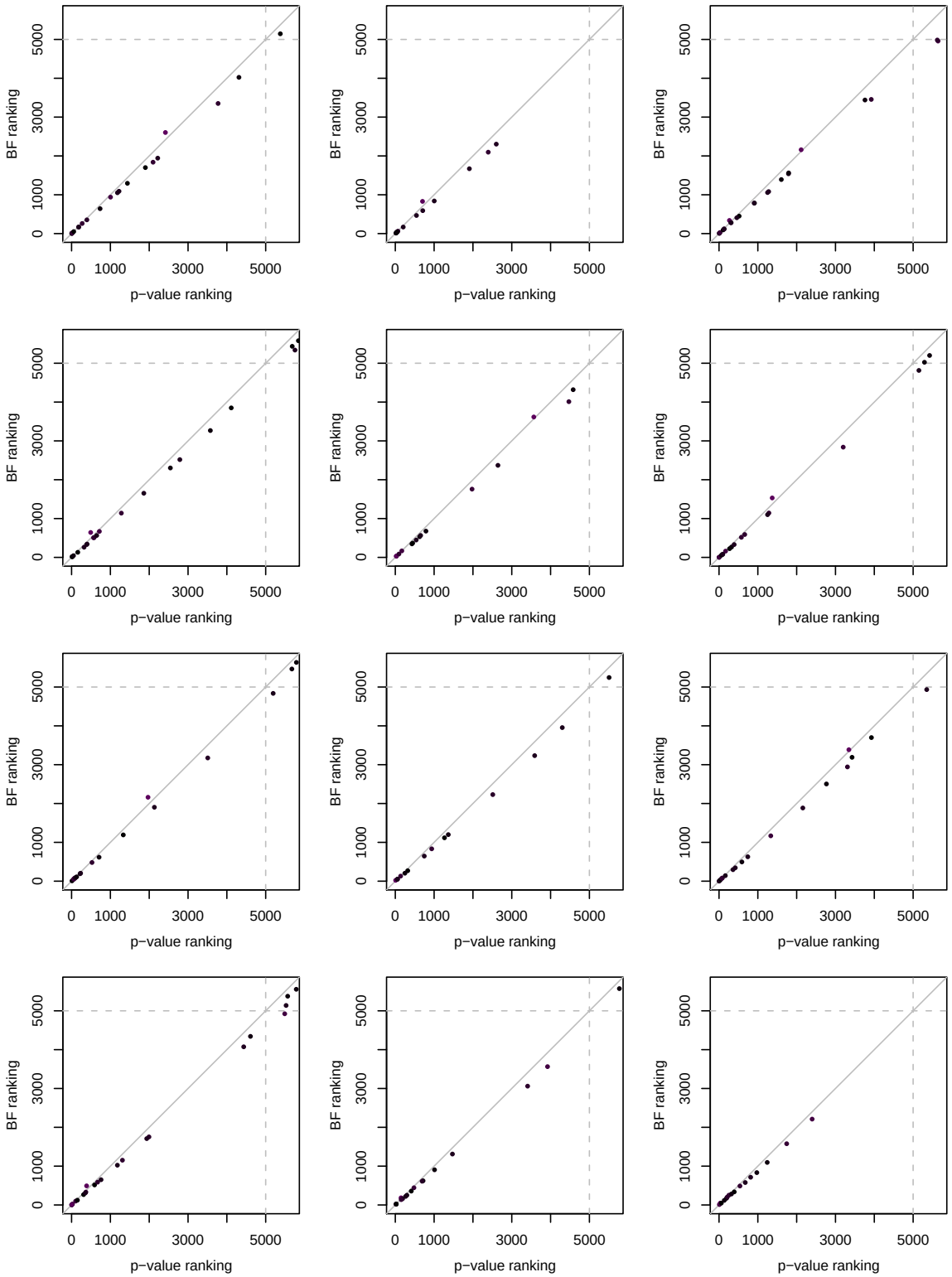

Figure 6.5: **Ranking comparison of replicated loci for Crohn's disease in subsamples.** Same as for Figure 6.4, but with 1,000 cases and 1,000 controls.

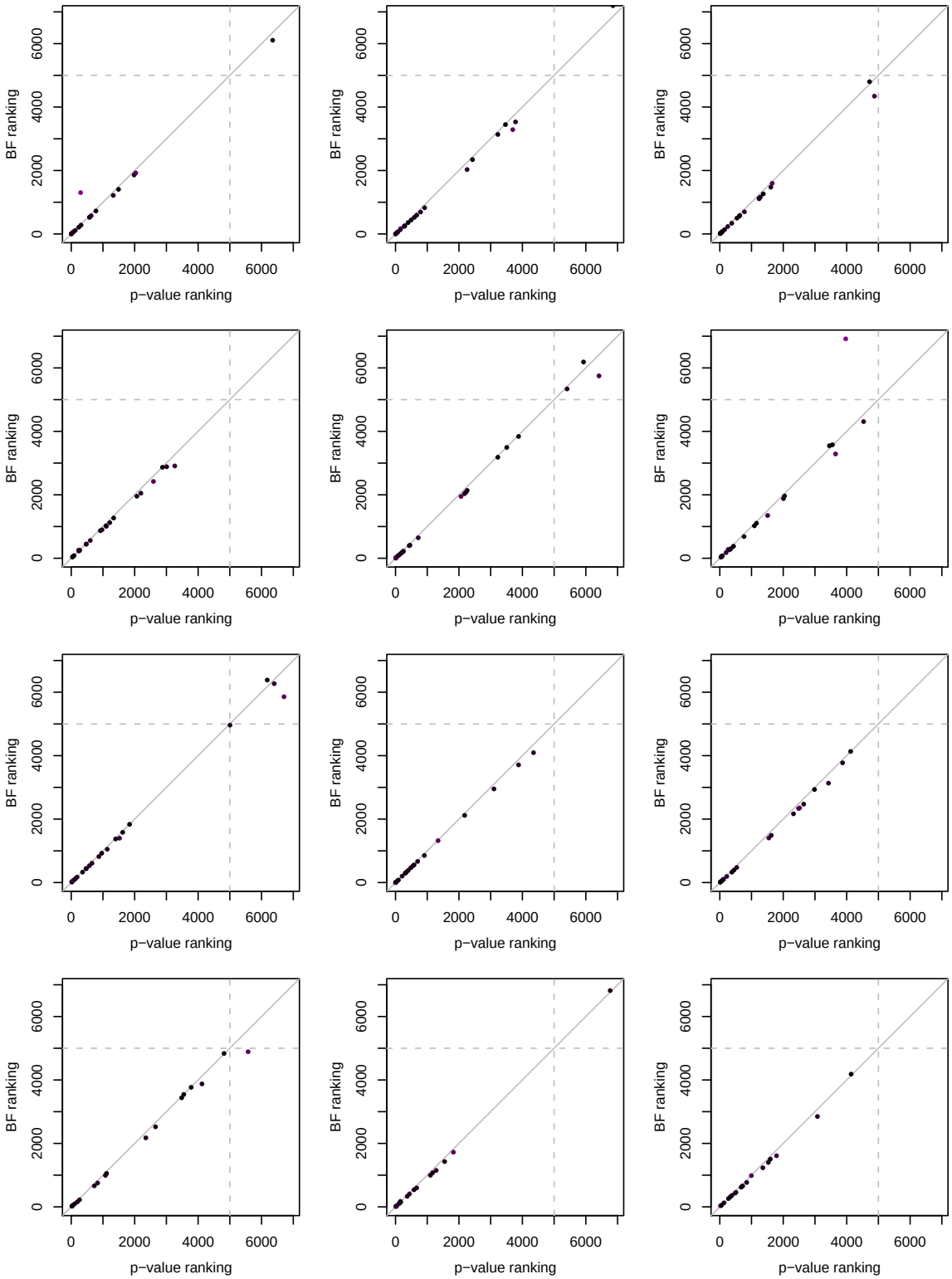

Figure 6.6: **Ranking comparison of replicated loci for Crohn's disease in subsamples.** Same as for Figure 6.4, but with 1,500 cases and 1,500 controls.

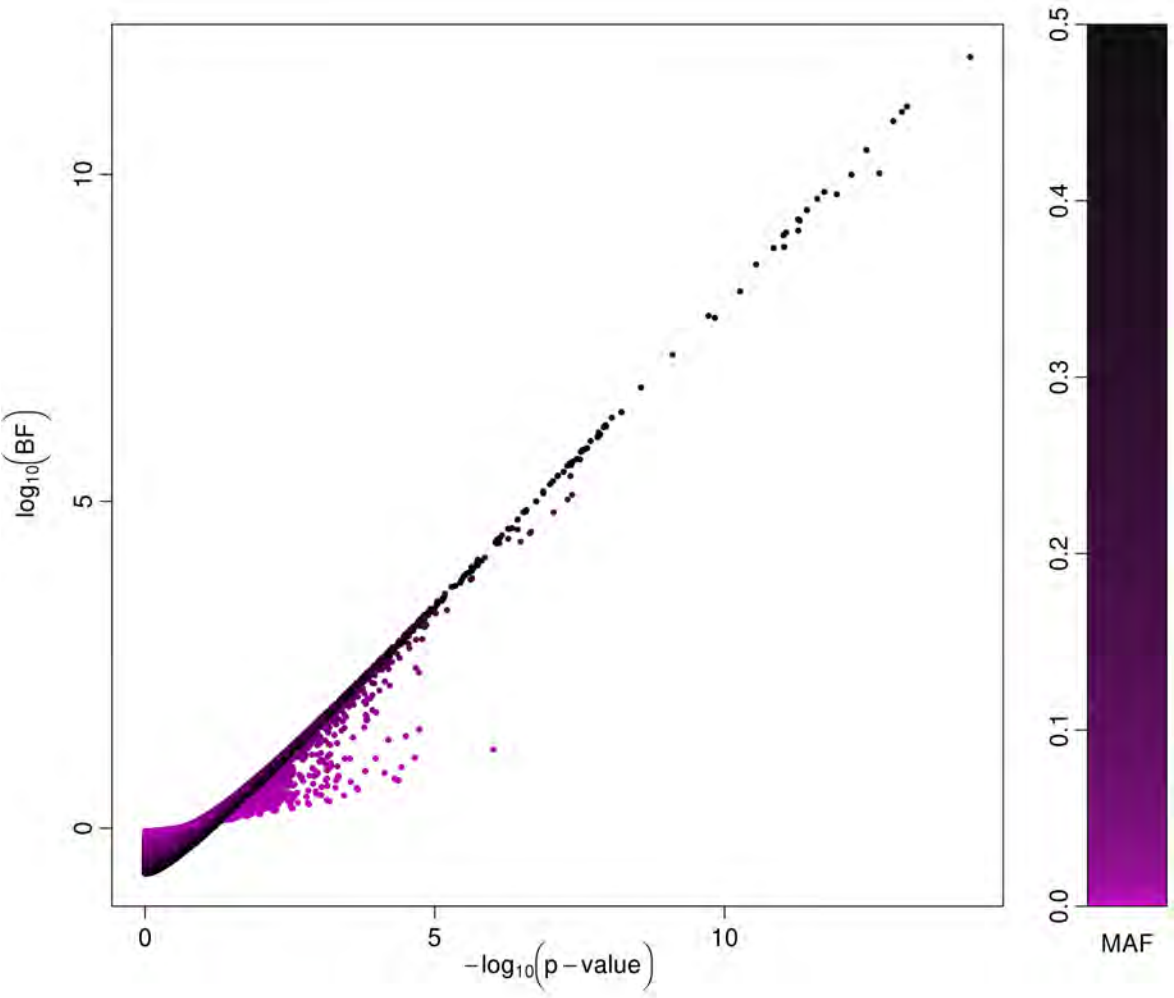

Figure 6.7: **BF versus p-value for Crohn's disease.** Each point represents a SNP from the WTCCC data. BFs are calculated under the conservative prior ($\sigma = 0.2$). Points are coloured according to the MAF, as shown in the legend on the right.

with our theoretical results, the range of BFs for common SNPs is quite wide whereas the BFs for rarer SNPs are more clustered. Intuitively, the common SNPs are more informative and can show stronger evidence of being either associated or null than can rare SNPs. This is akin to the *shrinkage* effect described earlier for effect size estimates. Informally, when there is not much evidence in the data (corresponding to a flat likelihood), the Bayesian approach will prefer smaller effect sizes (in line with the prior), resulting in less strong evidence of association. Put another way, with a prior expectation of small effects, the data at a SNP will need to be both informative and convincing to provide strong evidence of association.

To understand the relationship between BF and p-value in a bit more detail, we can focus our attention on a particular p-value and ask how the BF varies with the MAF. We see three types of relationships:

1. For SNPs with a very low p-value (right-hand side of the plot), common SNPs will have a high BF, and as the MAF decreases BFs become smaller, corresponding to less evidence against the null. (It is only when we reach more moderate p-values that we see SNPs with low MAF.)

2. On the opposite end of the spectrum, for SNPs with a p-value closer to 1 (left-hand side of the plot), common SNPs have BFs favouring the null hypothesis ($\log_{10}(\mathrm{BF})$ is negative) whereas the BF for rarer SNPs with the same p-value provides less strong support for the null (the $\log_{10}(\mathrm{BF})$ is less negative). Intuitively, the common SNPs here show greater evidence of being null than do the rare SNPs, since they are more informative.

3. Between these two extremes, BFs move towards ambivalence with decreasing MAF (that is, towards $\log_{10}(\mathrm{BF}) = 0$), but not necessarily monotonically. This is the same effect observed in Section 5.7, where we saw that changing the MAF can make the BF go up or down depending on other factors, such as the OR and the sample size.

Figure 6.8 shows plots for three different priors, both for the full WTCCC Crohn's disease data, and for an artificial data set created by sampling 500 cases and 500 controls from the full data. The focus is on observing the differences between BFs and p-values, so for visual clarity I only show SNPs with p-value greater than 1×10^{-5}. The cyan lines show the theoretical relationship for different MAF values, as a guide to display the nature and extent

of shrinkage. Again, we see a broad agreement between the BFs and p-values for common SNPs, with various levels of shrinkage for rarer SNPs. In particular, shrinkage is most apparent when the MAF is low or the sample size is smaller, both being situations where the SNPs are less informative (or, equivalently, where we have less power to detect association), and also when the prior has smaller variance.

The p-values for each SNP are the same across all panels using the same data set, only the BFs differ. Although it is not visible on these plots, the p-values are approximately uniformly distributed on the interval $(0, 1)$ independently of the MAF, as expected under the null hypothesis, in contrast to BFs whose range varies depending on the MAF.

When using the most diffuse of the three priors, $\sigma = 1$, it appears as if there is actually no shrinkage towards 0 for the SNPs showing stronger evidence. Indeed, this prior is so diffuse that even the rarest SNPs left in the sample after filtering are not rare enough to display appreciable shrinkage. This is less so for the subsampled data, since the prior has more influence when there is less data.

Note that lack of shrinkage is not the same as equivalent rankings. The latter implies a one-to-one relationship irrespective of MAF, which is not the case for any of the priors displayed here. In Section 5.8, I showed that equivalent rankings are given by g-priors. Figure 6.9 compares the conservative prior with the 0.2 g-prior empirically. As expected from the theory, rankings under the g-prior closely correspond with those using p-values from the trend test. Small deviations from the theoretical relationship are due to poorer approximation by the asymptotic results for SNPs with lower MAF. (For this comparison, I excluded SNPs with MAF less than 0.9%.)

6.4 Discussion

There has been an explosion of interest recently in GWAS, as they have proved to be an effective tool for learning about the genetic basis of complex diseases. Most analyses of GWAS data have involved frequentist approaches. There has been growing interest in the use of Bayesian approaches to the analysis of genetics data, but there does not seem to have been a systematic comparison of the approaches in the GWAS context. Here I have undertaken such a comparison, based on both theoretical considerations and on the empirical analysis of large GWAS data sets.

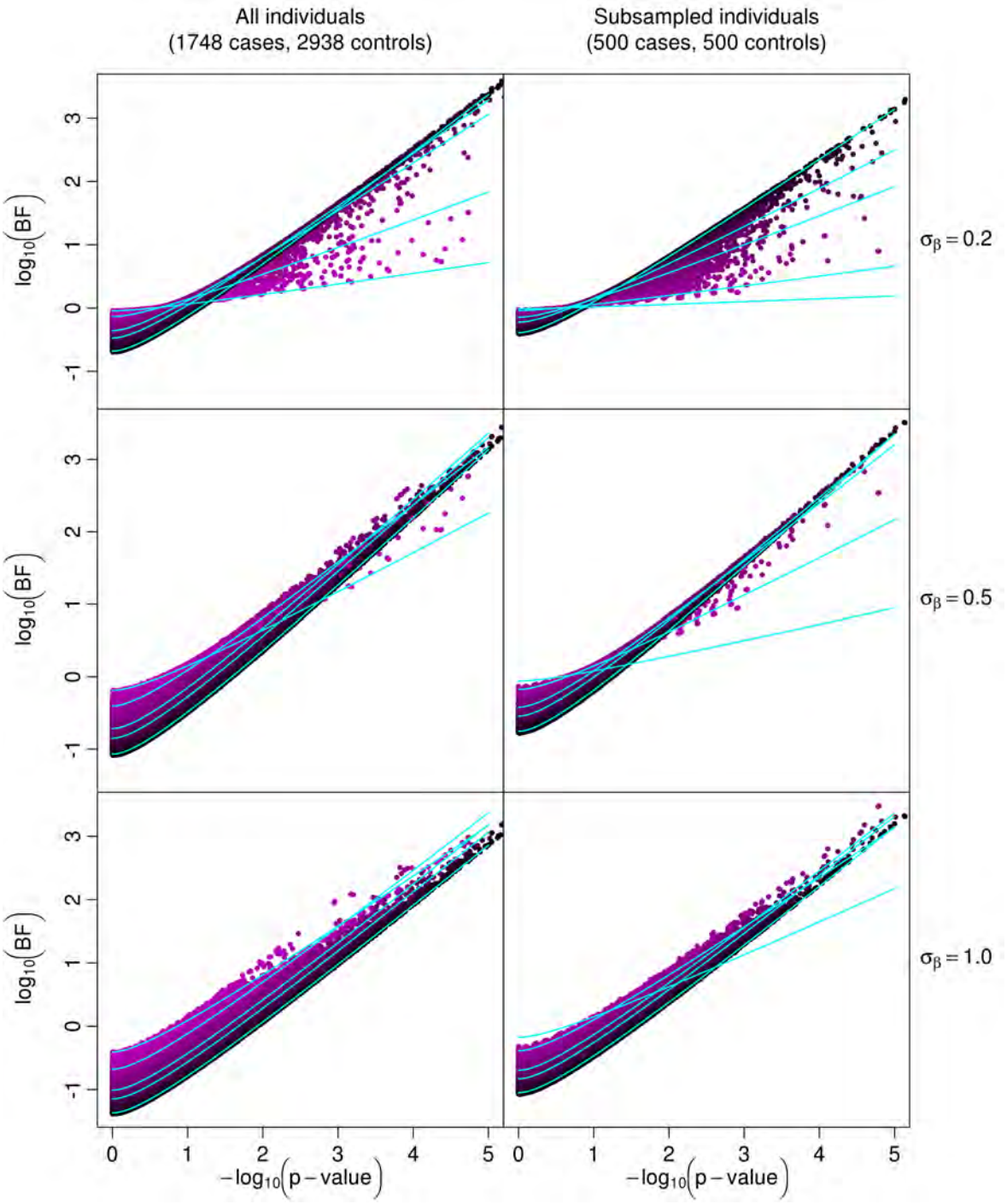

Figure 6.8: **BF versus p-value for different priors & sample sizes.** Each point represents a SNP from a WTCCC Crohn's disease data set. Each panel is labelled by which prior and data set it represents. Points are coloured according to the MAF, as in Figure 6.7. The cyan lines show the theoretical relationships, given by equations (4.6) and (5.4), for the following MAF values: 0.0025, 0.01, 0.05, 0.1, 0.5 (top to bottom, on the left side of each plot). Only SNPs with p-value greater than 1×10^{-5} are displayed, to show more clearly the effect of the different priors.

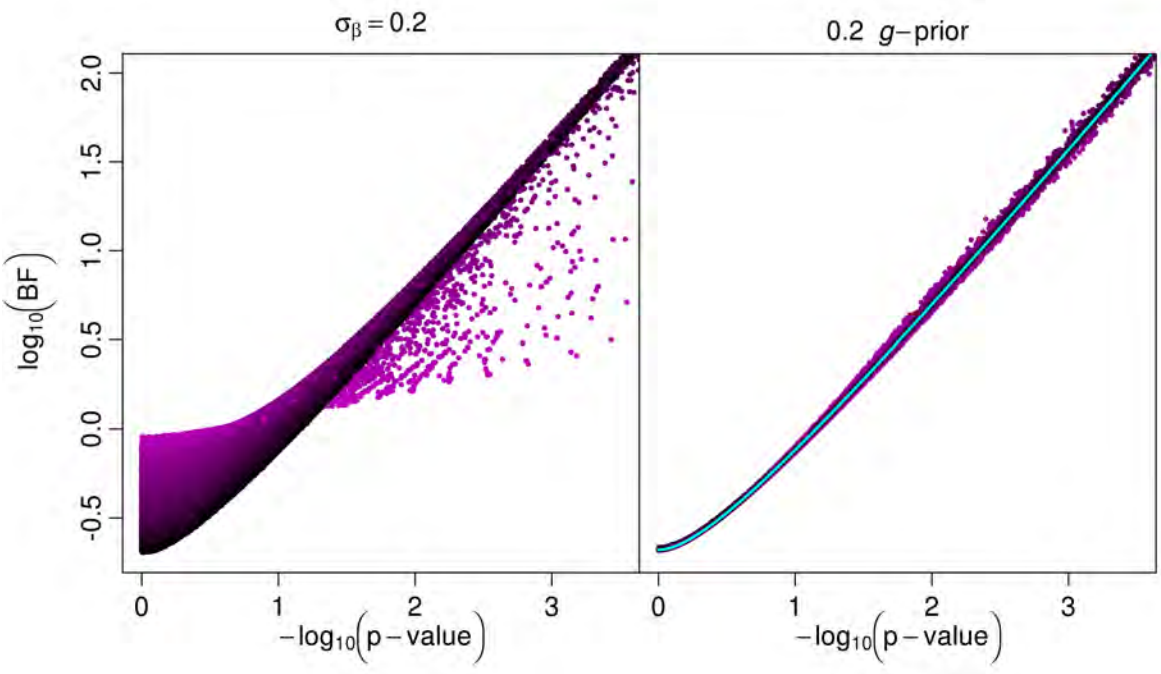

Figure 6.9: **BF versus p-value, comparing the conservative prior ($\sigma = 0.2$) with the 0.2 g-prior.** The left panel reproduces a subset of the plot from Figure 6.7. The right panel shows the same data but under the 0.2 g-prior. The cyan line shows the theoretical relationship according to equations (5.4) and (5.5).

I have focused on what is a major question for GWAS analysis, namely which SNPs should be chosen for a replication study. There are other analysis and quality control issues faced by investigators for which use of frequentist statistics is simple, practicable, and valuable. Examples include use of p-values for exclusion of SNPs failing HWE tests and, in particular, QQ plots which can be extremely helpful for recognising systematic departures from underlying assumptions, whether subsequent analyses are frequentist or Bayesian, or both.

Several broad conclusions follow from these results. The first is that for common SNPs there is not much practical difference between frequentist and Bayesian approaches to inference. On the other hand there can be a marked difference for rare SNPs. The MAF at which this difference becomes apparent will differ between studies: for larger studies the correspondence between BF-based and p-value–based analyses will extend down to lower MAFs than for smaller studies. Informally, unless one thinks large effect sizes are likely at rare SNPs, there is very little power to detect a genetic effect when the MAF is low. The BF automatically takes this into account and discounts the evidence, whereas even when there is virtually no power to detect an effect, the p-value can be small just by chance.

I showed from two large GWAS data sets that choosing SNPs for follow-up using BFs performs better than choosing them using p-values. The difference between the approaches is not huge but in our examples the Bayesian approach performed better.

The thinking behind choice of GWAS SNPs for follow-up is not usually documented well. One suspects that many investigators will upweight evidence at, say, non-synonymous SNPs, and some may also downweight evidence from p-values at rare SNPs, or ignore such SNPs altogether. In practice this informal perturbation to the frequentist approach will bring it closer (and probably quite close) to a Bayesian analysis. Nonetheless, it remains *ad hoc*. Given the slight but definite improvement in the performance of the Bayesian approach, and its other advantages, there is good case for its more routine use in GWAS analyses.

There are other advantages to the Bayesian approach, which have also been pointed out elsewhere. BFs have a direct interpretation, and when calculated with the same prior can be compared between SNPs and between studies in a way that is not straightforward with p-values. If one wishes to undertake a formal analysis that gives different weights to different types of SNPs, for example assuming that a SNP which changes a gene product is, *a priori*, more likely to be causative than a non-coding SNP, or that a SNP which tags many other SNPs is more likely to be due to a true association than one which tags few other SNPs, then this is straightforward in the Bayesian framework provided one specifies the relative probability associated with the different types of SNPs. Finally, under the Bayesian paradigm any 'multiple testing correction' exhibits itself in a more intuitive way—the relevant information is encapsulated in the prior odds of association, rather than by the number of tests carried out. It remains the case that strong evidence will be needed to be convincing for a particular SNP, but in the Bayesian framework this is because it is very unlikely, *a priori*, that the SNP is causative (or in LD with a causative SNP); but this would be true for each SNP, whether there was data on one SNP or one million SNPs.

6.5 Further work: updating our priors

In this chapter, I have compared the performance of the p-value and the BF under a particular set of priors. Given the equivalence between the Bayesian and frequentist approaches described in Section 5.8, this can be interpreted as simply comparing one set of priors with

another. Thus, we can set this problem in a more general framework by asking which priors give good ranking performance. They will be ones that best model the observed disease effects, particularly in regard to how the effect sizes relate to allele frequency.

In principle, there are now enough replicated loci available to set this up as a formal inference problem. However, there are also some key challenges. Firstly, these loci are ascertained differently, coming from studies with different designs and sample sizes, and most crucially with differing power to detect them. This will need to be taken into account. Secondly, the power to detect a locus depends greatly on the MAF, meaning that replicated loci which are rare will tend to be enriched for greater effect sizes. This is a major difficulty when it comes to joint modelling of effect size and allele frequency. For example, ILES (2008) investigated such a joint relationship via simulations but did not take differential power into account, admitting this was a key weakness. Finally, there is the question of how and whether to combine results from different diseases. Very few diseases have a substantial number of known loci (Crohn's disease is a notable example). It seems desirable to be able to generalise to many diseases anyway, especially if one of our aims is to develop a good default prior to use for analyses. However, different diseases will have been studied with more or less fervour, and discovered loci will consequently be more representative of some types of diseases more than others, adding yet another ascertainment bias into the mix. Not withstanding these difficulties, I suggest an in-principle approach. For now, I duck the question of which diseases to include and assume that a choice has been made. Using ideas developed in this thesis, I give a brief, high-level overview of a first attempt at tackling this problem.

We are fortunate to have publicly available an extensive and effectively complete catalogue of discovered disease loci from association studies of common human diseases (HINDORFF ET AL. 2009). This lists the ORs, allele frequencies and sample sizes for a large number of loci and their respective studies. The ORs are generally for an additive model, allowing us to use these data to model how additive effects relate to allele frequency. I suggest constructing a model by starting with one of the MAF-dependent priors described in Section 5.10 and placing hyper-priors on the parameters in that prior. I would choose hyper-priors that are at least relatively diffuse, and optionally centered on values we currently believe in (e.g. the conservative prior).

We must deal with ascertainment bias due to differential power. For this purpose, assume that all studies conducted a standard frequentist analysis with a specific p-value threshold. We need to model both the power of detection and also the censoring effect of the threshold (which varies by MAF). The approximate formula derived in equation (4.11) can be used to do both, and should be simple enough to be conveniently incorporated into any model. It may be simplest to assume a particular p-value threshold was applied for all studies. While unrealistic, it is convenient, and can be dealt with by seeing how sensitive the inference is for a range of plausible values.

Different studies will have used different genotyping chips, and each may or may not have used an imputation analysis. Both of these may impact on the inference but may be prudent to ignore on a first pass.

Finally, there is a question of how to interpret the output of such an analysis. One could think of it as a model of true disease effects, of which it is probably an imperfect approximation due to LD effects which have not been modelled. An alternative perspective is to regard it as an estimate of the observed effects at marker SNPs. This has two advantages. Firstly, it reflects the primary purpose of this inference—to define a good set of priors to use for GWAS analysis, which are necessarily conducted on marker SNPs. Secondly, it indirectly justifies focusing exclusively on additive models by appealing to results from Section 7.1 (in the following chapter), where I show that marker SNPs will tend to exhibit disease models closer to additive as compared to the causal SNPs they represent.

Chapter 7

Linkage Disequilibrium, Marker SNPs & Fine Mapping

Contents

The previous chapters have mainly focused on issues and methods from the perspective of applying them genome-wide. I now turn to questions relevant to studying specific regions, for example those identified by GWAS as showing strong evidence of association. I cover

three things: (i) theoretical results on the effect of LD on observed disease models; (ii) a description of methods for quick and convenient Bayesian inference for fine mapping studies; and (iii) some analysis results from a recent fine mapping project.

7.1 Disease models at marker SNPs: effect of LD

GWA studies take advantage of LD across the genome to capture a large proportion of the genetic variation using a relatively moderate number of marker loci. We therefore expect them to highlight markers correlated with causal loci, rather than the causal loci themselves. Depending on the level of LD, the apparent disease effect at the marker will be an imperfect representation of the true disease effect. Here I explore this relationship analytically.

This effect of LD has been well characterised for additive models. ZONDERVAN & CARDON (2004) show that the 'marker OR' is completely described by four parameters: the true OR, the frequencies of the causal and marker alleles, and the LD between them. Furthermore, the effect on the power to detect an association has also been described (PRITCHARD & PRZEWORSKI 2001, CHAPMAN ET AL. 2003), with the association signal diminishing with decreasing LD. Measuring the LD using the squared correlation (r^2), a rule of thumb is that a sample size of N/r^2 is required at a marker in order to have the same power to detect an association as a study with sample size N that types the causal SNP (PRITCHARD & PRZEWORSKI 2001).

Here I extend the investigation to also consider non-additive models. I show that, as LD decreases, the deviation from an additive model generally decays quadratically while the additive effect decays only linearly. This faster decay distorts the disease effects to make them look closer to being additive, and potentially explains the relative lack of observed non-additive effects in reported findings. It also supports the idea that using association tests that are well-powered to detect additive models (such as the trend test, or BFs using the additive model) is a sensible default approach for single-SNP analysis of GWAS (ILES 2008). My findings can be conveniently summarised by a rule of thumb similar to the above: when testing for non-additivity at a marker, the sample size required for equivalent power is roughly N/r^4.

At this point, it is worth mentioning how the following two sources of bias relate: the *win-*

ner's curse (see Section 1.3.4) and the *surrogate's shortfall* (as defined by SPENCER ET AL. 2008). Both describe mechanisms by which the observed effect sizes from studies may not be indicative of true effects. The former is due to sampling variation and thresholding to a given significance level. The latter describes what I investigate in this section, the fact that marker (surrogate) loci tend to show a weaker effect than the truth due to imperfect correlation. Note that these two biases act independently and in tandem, the former being mainly a property of the variance of the disease effect estimators and the latter a property of the mean.

7.1.1 LD & disease models

Let A and B be a pair of biallelic SNPs and code the alleles at each as 0 and 1. In the situations that I examine, A will be a causal SNP and B will be a marker SNP. I will refer to an allele specific to a SNP using a subscript, e.g. 0_A refers to allele 0 at SNP A. Let the frequency of allele 1_A in the population be f_A, and that of 1_B be f_B. Since the allele codings are arbitrary, we can set them such that 1 codes for the minor allele.

Consider the population distribution of the four possible haplotypes formed by these two SNPs. Three parameters are necessary to define an arbitrary distribution. Together with f_A and f_B, I use the population correlation coefficient, ρ, to fully parameterise this distribution. The square of this is a commonly used measure of LD, usually denoted as r^2 (e.g. ZONDERVAN & CARDON 2004). I use the latter notation throughout (and use r in place of ρ).

Define the following conditional probabilities,

$$q_0 = \Pr(1_A \mid 0_B)\,,$$
$$q_1 = \Pr(1_A \mid 1_B)\,.$$

These allow the following representation of the haplotype distribution,

$$\begin{aligned}
\Pr(0_A 0_B) &= (1-q_0)(1-f_B)\,,\\
\Pr(0_A 1_B) &= (1-q_1)f_B\,,\\
\Pr(1_A 0_B) &= q_0(1-f_B)\,,\\
\Pr(1_A 1_B) &= q_1 f_B\,,
\end{aligned}$$

and give the identity,

$$f_A = q_0(1-f_B) + q_1 f_B\,.$$

The correlation coefficient can also be expressed in terms of these quantities,

$$r = (q_1 - q_0)\sqrt{\frac{f_B(1-f_B)}{f_A(1-f_A)}}\,.$$

By solving these last two equations for q_0 and q_1, we can see that the haplotype distribution is fully and uniquely specified by f_A, f_B and r (when they are consistent with a haplotype distribution—i.e. where the above two linear equations are simultaneously solvable).

As is well known, the range of r is bounded depending on the allele frequencies. Suppose that f_A and f_B are MAFs and that $f_A \leq f_B$. By considering the possible values of q_0 and q_1, it can shown that the range of r is,

$$-\sqrt{\left(\frac{f_A}{1-f_A}\right)\times\left(\frac{f_B}{1-f_B}\right)} \;\leq\; r \;\leq\; \sqrt{\left(\frac{f_A}{1-f_A}\right)/\left(\frac{f_B}{1-f_B}\right)}\,, \tag{7.1}$$

and that the roles of f_A and f_B swap if $f_A > f_B$. From this we can see that to get a high positive correlation we need to have $f_A \approx f_B$, and for a high negative correlation we need f_A and f_B to both be large.[1] A correlation in either direction will suffice for the marker to be a good surrogate. Thus, we can conclude that in situations where one of the SNPs is rare (either the marker or the causal SNP), the ability to detect associations will be impaired unless the other SNP is also rare and highly correlated.

I use the term *diplotype* to mean a pair of haplotypes belonging to an individual. Let $\binom{1_A 0_B}{1_A 1_B}$ represent the diplotype comprising of the two haplotypes $1_A 0_B$ and $1_A 1_B$ (so, having geno-

[1] This (apparent) asymmetry is due to the choice of f_A and f_B as being the *minor* allele frequencies.

type 2 at SNP A and genotype 1 at SNP B). To obtain a diplotype distribution I assume HWE for haplotypes. In other words, haplotypes combine at random from the population, which I assume to be infinite (equivalent to sampling with replacement). For example,

$$\begin{aligned}\Pr\binom{1_A0_B}{1_A1_B} &= 2\Pr(1_A0_B)\ \Pr(1_A1_B)\\ &= 2q_0q_1f_B(1-f_B)\,.\end{aligned}$$

There are 10 possible diplotypes but only 9 distinguishable pairs of genotypes—the genotype pair $(1,1)$ corresponds to the two diplotypes $\binom{1_A0_B}{0_A1_B}$ and $\binom{0_A0_B}{1_A1_B}$. I only consider analyses using genotypes, so will sum over these two phases when considering this genotype combination.

My derivations will relate disease models by comparing penetrances at marker and causal SNPs. For this purpose it proves convenient to consider log risk regression models rather than logistic regression. However, for convenience I will retain similar terminology depending on the parameterisation. In particular, I refer to the following as the additive model,

$$\log(p) = \mu + \beta G\,,$$

and the following as the general model,

$$\log(p) = \mu + \beta G + \gamma \mathbf{1}_{G=1}\,.$$

Recall from Section 4.6.2 that these are related to their logistic regression counterparts in the same way that the OR and RR are related, and that when using cohort samples as controls, as is standard in GWAS, we effectively use a log risk model. Furthermore, the two models will be approximately equivalent when the disease prevalence is relatively rare.

To distinguish between disease effects at different SNPs, I label the model parameters with a subscript, e.g. β_A for the additive parameter at SNP A.

7.1.2 Effect of LD on disease parameters

Additive effects (haploid model)

A convenient way to study additive effects in a diploid organism is to use a haploid model and assume HWE. This should be adequate for common diseases since we do not expect significant deviations from HWE. The same approach has been taken by previous authors (PRITCHARD & PRZEWORSKI 2001, CHAPMAN ET AL. 2003, ZONDERVAN & CARDON 2004).

Let SNP A be causal and SNP B be a marker. Define the following disease penetrances:

$$\begin{aligned} a_0 &= \Pr(Y = 1 \mid 0_A)\,, & b_0 &= \Pr(Y = 1 \mid 0_B)\,, \\ a_1 &= \Pr(Y = 1 \mid 1_A)\,, & b_1 &= \Pr(Y = 1 \mid 1_B)\,. \end{aligned}$$

Using the LD model we can relate the penetrances at the two SNPs,

$$\begin{aligned} b_0 &= a_0(1 - q_0) + a_1 q_0\,, \\ b_1 &= a_0(1 - q_1) + a_1 q_1\,. \end{aligned}$$

Taking the difference gives a convenient summary of the relationship,

$$b_1 - b_0 = (a_1 - a_0)(q_1 - q_0)\,.$$

We can re-write this in terms of the disease model parameters, allele frequencies and LD,

$$\begin{aligned} \frac{b_1}{b_0} - 1 &= \left(\frac{a_1}{a_0} - 1\right) \frac{a_0}{b_0}(q_1 - q_0)\,, \\ e^{\beta_B} - 1 &= \left(e^{\beta_A} - 1\right) \frac{e^{\mu_A}}{e^{\mu_B}} r \sqrt{\frac{f_A(1 - f_A)}{f_B(1 - f_B)}}\,. \end{aligned}$$

We can derive a simpler expression when effect sizes are small. Using the approximation $e^x - 1 \approx x$, and also $\mu_A \approx \mu_B$ (which is equivalent to saying the penetrances at allele 0 are similar at the two SNPs), we have,

$$\beta_B \approx \beta_A\, r \sqrt{\frac{f_A(1 - f_A)}{f_B(1 - f_B)}}\,. \tag{7.2}$$

We see that the additive effect at the marker SNP decreases linearly with r as the LD becomes

weaker. This is a key result. It gives an intuitive and convenient relationship between the parameters of interest. Furthermore, the relationship later derived for the effect of LD on power follows directly from it. In this formulation, this result appears to be novel.

ZONDERVAN & CARDON (2004) derive a similar formula, but expressed in terms of different parameters. They parameterise LD in terms of the disequilibrium coefficient ($D = \Pr(1_A 1_B) - \Pr(1_A)\Pr(1_B)$) instead of r, and use the OR instead of the RR (recall that I am using the log risk regression model),

$$\mathrm{OR}_B - 1 = \frac{D\,(\mathrm{OR}_A - 1)}{f_B\left[(1 - f_B) + ((1 - f_B)f_A - D)\,(\mathrm{OR}_A - 1)\right]}\,.$$

Our formulation above is simpler and more directly usable, for example as shown later in the derivation of power relationships.

For reference, it should be remembered that both β_A and β_B are the true parameter values, with samples yielding realisations around these as means.

Non-additive effects

Let i_A with $i \in \{0, 1, 2\}$ now represent the diploid genotypes at SNP A, and likewise for SNP B. Define the following disease penetrances:

$$\begin{aligned} a_0 &= \Pr(Y = 1 \mid 0_A)\,, & b_0 &= \Pr(Y = 1 \mid 0_B)\,, \\ a_1 &= \Pr(Y = 1 \mid 1_A)\,, & b_1 &= \Pr(Y = 1 \mid 1_B)\,, \\ a_2 &= \Pr(Y = 1 \mid 2_A)\,, & b_2 &= \Pr(Y = 1 \mid 2_B)\,. \end{aligned}$$

Using the LD model,

$$\begin{aligned} b_0 &= a_0(1 - q_0)^2 + a_1 2q_0(1 - q_0) + a_2 q_0^2\,, \\ b_1 &= a_0(1 - q_0)(1 - q_1) + a_1\left(q_0(1 - q_1) + q_1(1 - q_0)\right) + a_2 q_0 q_1\,, \\ b_2 &= a_0(1 - q_1)^2 + a_1 2q_1(1 - q_1) + a_2 q_1^2\,. \end{aligned}$$

The expression $b_1^2 - b_0 b_2$ is a measure of the deviation from additivity (it is exactly 0 for an

additive model), and has a simple form that relates the marker and causal SNP penetrances,

$$b_1^2 - b_0 b_2 = \left(a_1^2 - a_0 a_2\right)(q_1 - q_0)^2 \ .$$

We can re-write this in terms of the disease model parameters, allele frequencies and LD,

$$\begin{aligned} \frac{b_1^2}{b_0 b_2} - 1 &= \left(\frac{a_1^2}{a_0 a_2} - 1\right) \frac{a_0 a_2}{b_0 b_2} (q_1 - q_0)^2 \,, \\ e^{2\gamma_B} - 1 &= \left(e^{2\gamma_A} - 1\right) \frac{e^{2(\mu_A + \beta_A)}}{e^{2(\mu_B + \beta_B)}} \frac{f_A(1-f_A)}{f_B(1-f_B)} r^2 \,. \end{aligned}$$

When the deviation from additivity is small, we can derive a simpler expression using the approximations $e^x - 1 \approx x$ and $\mu_A + \beta_A \approx \mu_B + \beta_B$,

$$\gamma_B \approx \gamma_A \frac{f_A(1-f_A)}{f_B(1-f_B)} r^2 \,. \tag{7.3}$$

We see that the dominance effect at the marker SNP decreases quadratically with r as the LD becomes weaker. Analogous to equation (7.2), this is a key result and in this formulation appears to be novel. SHAM ET AL. (2000) derive a similar result relating variance components in models of quantitative traits; my derivation here relates parameters in models of case-control data. The formula gives an intuitive and convenient relationship between the parameters of interest, and the relationship later derived for the effect of LD on power follows directly from it. Crucially, this result contrasts with that for the additive parameter, with the dependence on LD being through r^2 rather than r.

As in the previous section, it should be remembered that both γ_A and γ_B are the true parameter values, with samples yielding realisations around these as means.

Figures 7.1 and 7.2 show the effect of LD and allele frequency on the disease model parameters, for dominant and recessive models respectively. The relative effect sizes at the marker SNP were calculated using,

$$\beta_B = \log\left(\sqrt{\frac{b_2}{b_0}}\right) \,, \qquad\qquad \gamma_B = \log\left(\frac{b_1}{\sqrt{b_0 b_2}}\right) \,. \tag{7.4}$$

The full procedure is as follows:

1. Select the effect size parameters at the causal SNP (β_A and γ_A) and allele frequencies for the causal and marker SNPs (f_A and f_B).

2. Determine the range of possible values of r using equation (7.1).

3. For each value of r, calculate the corresponding values of q_0 and q_1, and then use equations (7.4) to calculate the corresponding parameter values at the marker SNP. While it may seem that we also require a value for μ_A to do this, so that the a_i are fully specified, it turns out that this cancels in the calculation.

From the figures we see that the dominance effect decays faster than the additive effect, approximately quadratically versus linearly. We can also see the inability of the marker SNP to reliably reproduce the correct disease model when it has a substantially different allele frequency to the causal SNP.

Another and perhaps more natural way to see the effect of LD is to plot the two disease parameters against each other. I will refer to this as a *model space* plot, since each point corresponds to a particular disease model and all possible models can be represented in this way (up to the value of μ). Figure 7.3 shows such a plot, with curves for each of the eight scenarios shown in Figures 7.1 and 7.2. The subspace of additive models is shown by the horizontal line, and the null model is at the origin. The curves trace out the theoretical disease model at the marker SNP, with lower LD corresponding to points closer to the origin along these curves. We can now clearly see how LD acts to make the observed model more additive—notice that the curves 'bend' towards the horizontal line.

Additive effects (diploid model)

I have explored the effect of LD on the parameters in the additive and general models when both of these are true. We could also ask what happens if we fit an additive model when the underlying model is *not* additive. Despite the model mis-specification, fitting the additive model will generally pick up some of the association signal. For a given disease model and allele frequency, there will be an underlying mean value for the additive parameter which represents the 'equivalent' additive effect for that model. Let this be β'. Thus, we can distinguish two separate questions of interest. Firstly, how does β' relate to the true disease model at a SNP? Secondly, what is the impact of LD on β' at marker SNPs?

To answer the first question, we would fit the additive model to the population genotype frequencies (i.e. pretend they are counts) for cases and controls. This calculation must be

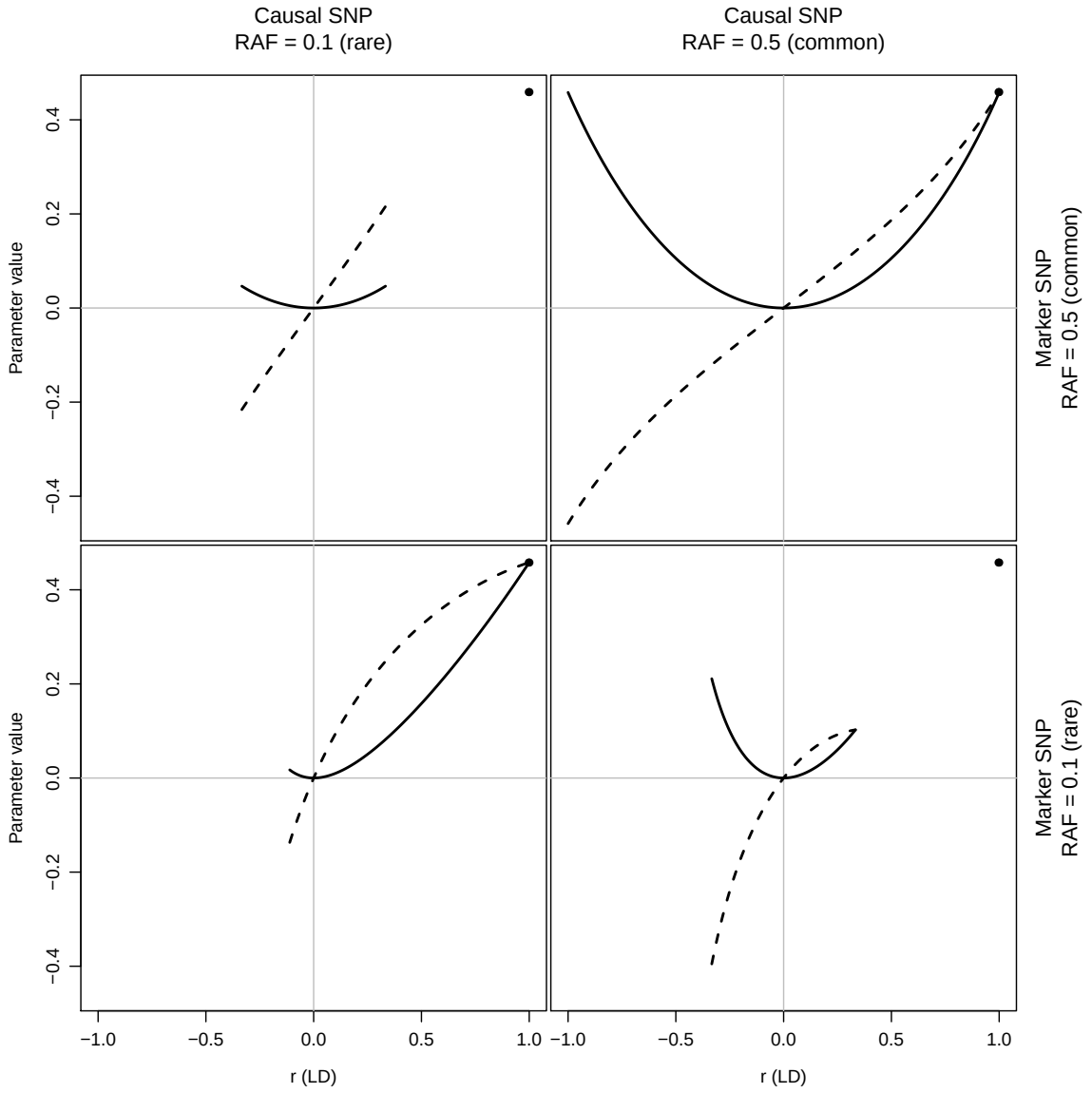

Figure 7.1: **Effect of LD on disease model parameters: dominant model.** Parameter values as functions of r, for a selection of risk allele frequencies (RAF). A dominant model with a RR of 2.5 at the causal SNP is assumed, corresponding to general model parameter values of $\beta_A = \gamma_A = 0.5 \log 2.5$. The solid line shows the dominance parameter (γ_B) and the dashed line the additive parameter (β_B) at the marker SNP. The parameter values at the causal SNP are shown by (in this case, overlapping) points at $r = 1$. Plots in each row correspond to a given marker SNP RAF and columns to a given causal SNP RAF, as labelled. The range of possible values of r depends on the allele frequencies, as shown by equation (7.1).

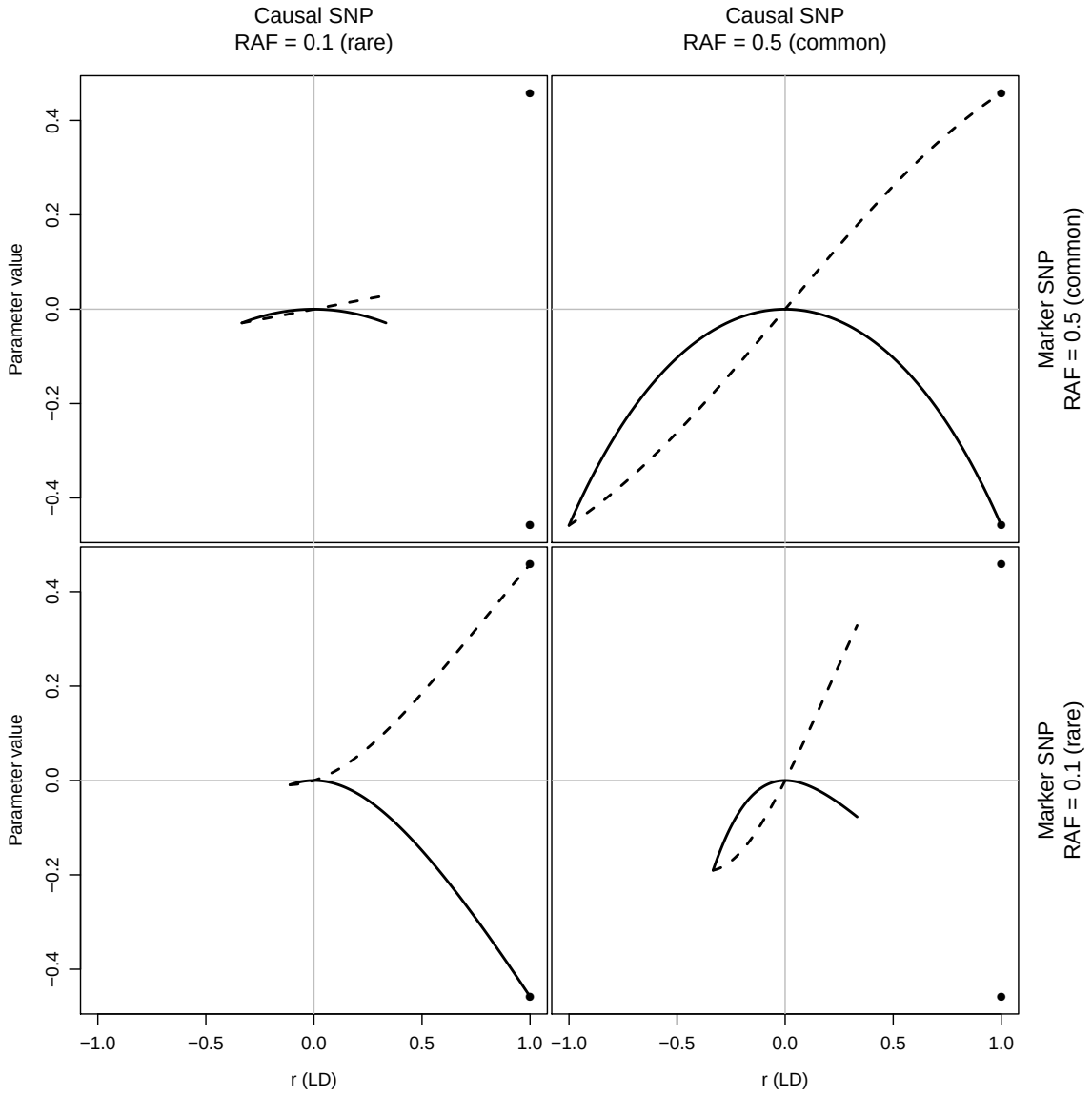

Figure 7.2: **Effect of LD on disease model parameters: recessive model.** Same as in Figure 7.1, but now for a recessive model with a RR of 2.5, corresponding to general model parameter values of $\beta_A = -\gamma_A = 0.5 \log 2.5$.

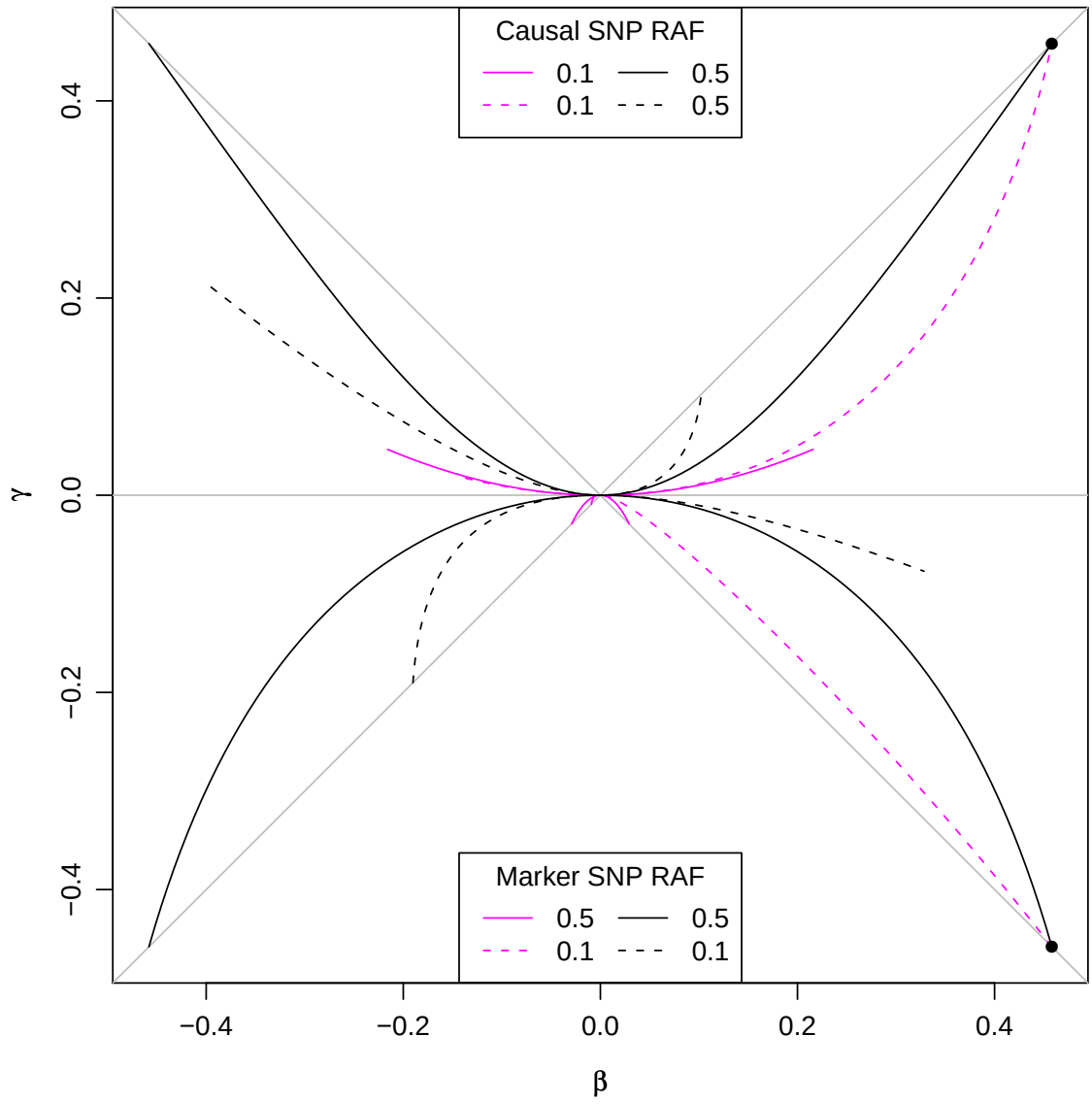

Figure 7.3: **Model space plot, showing decay of non-additive effects.** The two disease parameters (dominance vs additive; γ vs β) plotted against each other, showing the full space of models up to the value of the baseline parameter (μ). The horizontal grey line shows the subspace of additive models. The grey lines above the horizontal show the subspace of dominant models, and those below show the subspace of recessive models. Curves and points trace out the models for the scenarios shown in Figures 7.1 and 7.2, lying above and below the horizontal respectively. Curves are drawn in different styles to show the causal and marker SNP RAF they correspond to, as shown by the two legends. The two points represent the true disease models at the causal SNP.

done numerically. Alternatively, we can derive a closed-form approximation by collapsing the genotypes into alleles (i.e. collapse a 3×2 table of genotype frequencies to a 2×2 table of allele frequencies), averaging the penetrances accordingly and using the haploid model as described earlier. I do the latter below. The following derivation holds for any SNP, so I drop subscripts denoting a specific SNP.

Let a_0, a_1 and a_2 be the penetrances under the diploid model and a'_0 and a'_1 the penetrances under the haploid model after collapsing. It is easy to show that,

$$a'_0 = a_0(1-f) + a_1 f\,,$$
$$a'_1 = a_1(1-f) + a_2 f\,.$$

Let β' be the additive effect under this haploid model. We can now relate it to the true (general) disease parameters,

$$e^{\beta'} = \frac{a'_1}{a'_0} = \frac{a_1(1-f)+a_2 f}{a_0(1-f)+a_1 f} = e^{\beta}\left(\frac{e^{\gamma}(1-f)+e^{\beta}f}{(1-f)+e^{\beta}e^{\gamma}f}\right)\,.$$

Thus,

$$\beta' = \beta + \log\left(e^{\gamma}(1-f)+e^{\beta}f\right) - \log\left((1-f)+e^{\beta}e^{\gamma}f\right)\,.$$

As would be expected, β' depends on the allele frequency (f) as well as the true model (β, γ). To see this intuitively, imagine that one of the alleles is rare. Then the homozygote for that allele would be rarely observed and β' will be mainly driven by the heterozygote and common homozygote. Indeed, the above expression returns such relationships as limiting cases,

$$\lim_{f\to 0}\beta' = \beta + \gamma\,,$$
$$\lim_{f\to 1}\beta' = \beta - \gamma\,.$$

Respectively, these are the differences in log-risk between genotypes 0 & 1 and 1 & 2.

To answer the second question we can now just apply the results from the haploid case. Equation (7.2) shows that β' follows an approximately linear relationship in r when comparing the effect at a marker SNP to the causal SNP. Figure 7.4 shows this for a few examples, under both a dominant and a recessive model.

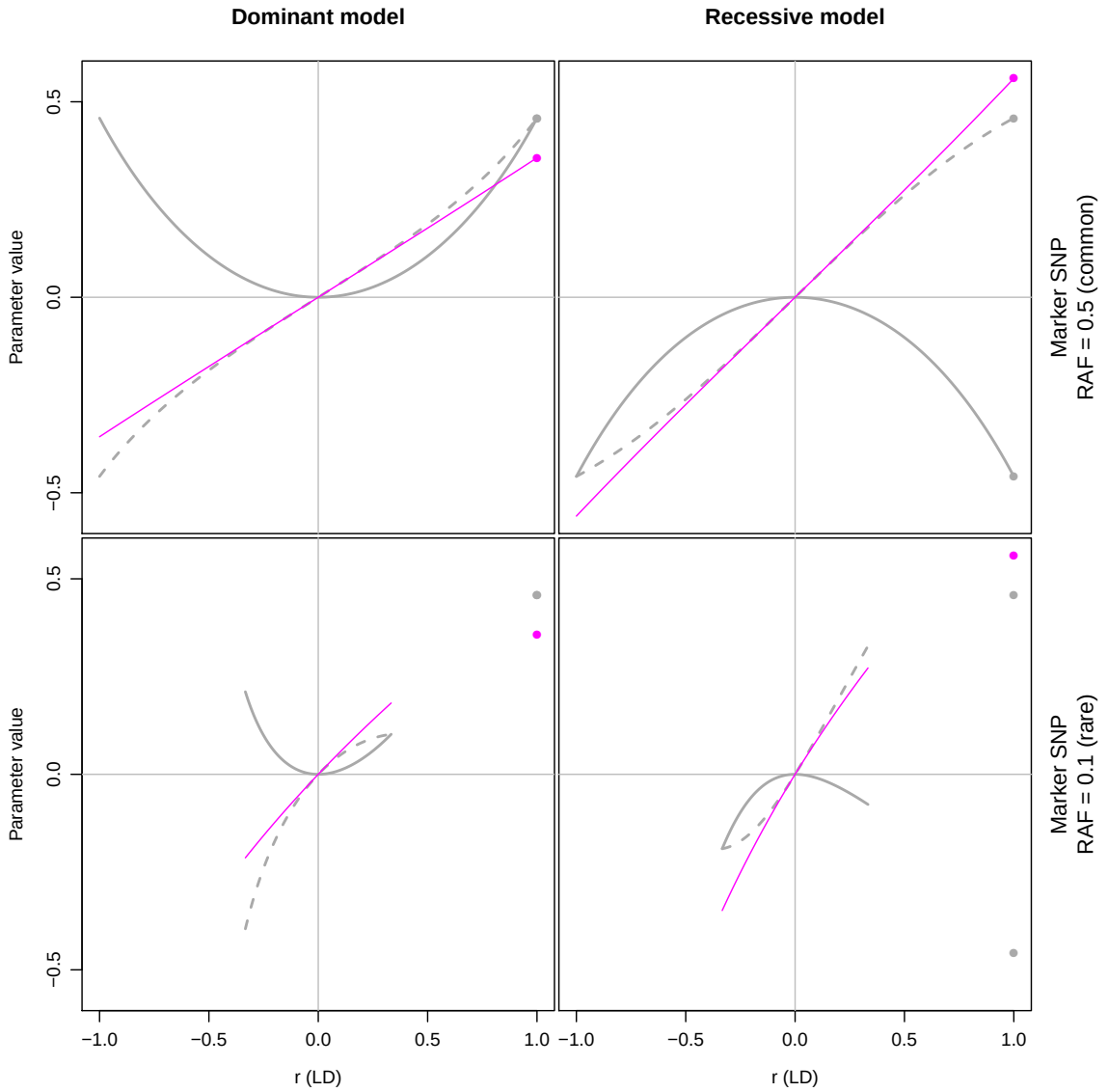

Figure 7.4: **Effect of LD on the equivalent additive parameter value.** Similar to Figures 7.1 and 7.2, but now showing the equivalent additive effect parameter (β') in magenta, with the general model parameter values underlaid in grey for reference. Plots in each row correspond to a given marker SNP RAF and columns to a given disease model (with RR of 2.5), as labelled. The causal SNP RAF is assumed to be 0.5.

7.1.3 Effect of LD on power

Above I have described the effect of LD on disease effects. I now turn to the question of how this impacts on the power for various tests: I consider the additive test for association and the Wald test for departures from additivity. The strategy I employ is to combine the parameter–LD relationships from above with the power approximations from Section 4.5, to give approximate expressions for the power at a marker SNP. One technical difference between the two is that in the former I used log risk models and population MAFs, while in the latter I used logistic regression models and sample MAFs. For the purposes of this section I will consider these to be equivalent. This will be adequate whenever the disease prevalence is low and effect sizes are small, which is often true in practice. Furthermore, from Section 4.6.2 we know that if cohort samples are used in place of controls, as is standard in GWAS, we effectively use a log risk model, so the results above will be compatible in terms of choice of model. Where the truth deviates from these assumptions, the expressions derived here are likely to be less accurate. However, they are primarily meant to act as a guide, so will still retain that utility.

Suppose we have a case-control sample of size N_A that types the causal SNP and also one of size N_B that types a marker SNP. From equation (4.11), an additive test at the causal SNP has non-centrality parameter,

$$\eta_1 \approx 2N_A f_A(1-f_A)\phi(1-\phi)\beta_A^2\,.$$

Applying equation (7.2), the same test at the marker SNP has non-centrality parameter,

$$\begin{aligned}\eta_2 &\approx 2N_B f_B(1-f_B)\phi(1-\phi)\beta_B^2 \\ &\approx 2N_B f_A(1-f_A)\phi(1-\phi)\beta_A^2 r^2\,,\end{aligned}$$

Thus, a sample size of $N_B = N_A/r^2$ is required to achieve the same power as typing the causal SNP directly. This is essentially the same derivation as shown in PRITCHARD & PRZEWORSKI (2001), but here based on the Wald test.

Applying the same idea for the test for deviation from additivity, now using equations (4.14)

and (7.3), gives the non-centrality parameters,

$$\eta_1 \approx 4N_A f_A^2 (1 - f_A)^2 \phi(1 - \phi)\gamma_A^2 \,,$$

and

$$\begin{aligned} \eta_2 &\approx 4N_B f_B^2 (1 - f_B)^2 \phi(1 - \phi)\gamma_B^2 \\ &\approx 4N_B f_A^2 (1 - f_A)^2 \phi(1 - \phi)\gamma_A^2 r^4 \,. \end{aligned}$$

Thus, a sample size of $N_B = N_A/r^4$ is required to achieve the same power as typing the causal SNP directly. The power approximations are most accurate for common SNPs, so correspondingly the results here will be most accurate when the marker SNP is common.

7.2 Bayesian fine mapping: region BFs & posteriors on SNPs

A convenient aspect of Bayesian single-SNP analyses, such as the approach described in Chapter 5, is the ease with which inference can then be done across SNPs within a short genomic region. This is of particular interest for fine mapping studies, where the genetic variants in a previously identified region are examined more exhaustively and in greater detail. In such a study we hope to actually type the causal variant(s). Under the assumption that we do, and the further assumption that there is only a single causal SNP, I derive two relevant quantities that are easily calculated from BFs for single-SNPs: a *region* BF that measures the evidence that there is a causal SNP in the region, and the posterior probability that a given SNP is the causal SNP.

Similar derivations have been described previously in the context of specific types of models (MARCHINI ET AL. 2007, SERVIN & STEPHENS 2007). My derivation here is more general, showing exactly the assumptions under which the results are valid.

7.2.1 Disease model

Take a sample of individuals typed at k SNPs in a genomic region and measured at a phenotype of interest (e.g. disease status). Let $\boldsymbol{Y}$ be the measured phenotypes, $\boldsymbol{X}_i$ be the genotypes at SNP i (where $i = 1, 2, \ldots, k$) and $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k]$.

Let θ be the set of model parameters. I consider models that posit up to one causal SNP in the region and that it is amongst those typed. Let M_0 represent the null model (no causal SNPs) and M_i the model where SNP i is causal. For the latter, I consider models where the phenotype and non-causal genotypes are conditionally independent given the causal genotype,

$$\Pr(\boldsymbol{Y}, \boldsymbol{X}, \theta \mid M_j) = \Pr(\boldsymbol{Y} \mid \boldsymbol{X}_j, \theta, M_j)\ \Pr(\boldsymbol{X} \mid \theta, M_j)\ \Pr(\theta \mid M_j)\,.$$

For the null model, I assume the phenotype is independent of all the genotypes,

$$\Pr(\boldsymbol{Y}, \boldsymbol{X}, \theta \mid M_0) = \Pr(\boldsymbol{Y} \mid \theta, M_0)\ \Pr(\boldsymbol{X} \mid \theta, M_0)\ \Pr(\theta \mid M_0)\,.$$

I make the further simplification that the marginal genotype distribution is independent of the disease model and parameters, $\Pr(\boldsymbol{X} \mid \theta, M_\cdot) = \Pr(\boldsymbol{X})$. This is reasonable, since without reference to the phenotypes there is no reason to distinguish between individuals.

Let M represent the model where exactly one SNP in the region is causal. In other words, it is the union of all non-null models described above,

$$M = M_1 \cup M_2 \cup \cdots \cup M_k\,.$$

7.2.2 Inference

Single-SNP BFs. Let BF_i be the BF that compares M_i and M_0. I will refer to these as the single-SNP BFs. They measure the evidence of association at SNP i,

$$\begin{aligned}
\mathrm{BF}_i &= \frac{\Pr(\text{data} \mid M_i)}{\Pr(\text{data} \mid M_0)} \\
&= \frac{\Pr(\boldsymbol{Y}, \boldsymbol{X} \mid M_i)}{\Pr(\boldsymbol{Y}, \boldsymbol{X} \mid M_0)} \\
&= \frac{\int \Pr(\boldsymbol{Y}, \boldsymbol{X}, \theta \mid M_i) d\theta}{\int \Pr(\boldsymbol{Y}, \boldsymbol{X}, \theta \mid M_0) d\theta} \\
&= \frac{\int \Pr(\boldsymbol{Y} \mid \boldsymbol{X}_i, \theta, M_i) \Pr(\theta \mid M_i) d\theta}{\int \Pr(\boldsymbol{Y} \mid \theta, M_0) \Pr(\theta \mid M_0) d\theta}\,.
\end{aligned}$$

In the last step the $\Pr(\boldsymbol{X})$ terms have cancelled, leaving only the genotypes at SNP i in the formula.

Region BF. Let $\mathrm{BF}_{\mathrm{reg}}$ be the BF that compares M and M_0. I will refer to this as the region BF. It measures the evidence that there is exactly one causal SNP in the region. We can write it in terms of the single-SNP BFs,

$$\begin{aligned}\mathrm{BF}_{\mathrm{reg}} &= \frac{\Pr(\mathrm{data} \mid M)}{\Pr(\mathrm{data} \mid M_0)} \\ &= \frac{\sum_{i=1}^{k} \Pr(\mathrm{data} \mid M_i) \Pr(M_i \mid M)}{\Pr(\mathrm{data} \mid M_0)} \\ &= \sum_{i=1}^{k} \mathrm{BF}_i \Pr(M_i \mid M) \,.\end{aligned}$$

Assuming a uniform prior on any particular SNP in the region being the causal SNP,

$$\Pr(M_i \mid M) = \frac{1}{k} \,,$$

results in the region BF as being simply the mean of the single-SNP BFs,

$$\mathrm{BF}_{\mathrm{reg}} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{BF}_i \,.$$

Posteriors on SNPs. Under the assumption that there is exactly one causal SNP, I show that the posterior that a given SNP is causal is proportional to its BF. By Bayes' Theorem,

$$\begin{aligned}\Pr(M_i \mid \mathrm{data}, M) &= \frac{\Pr(\mathrm{data} \mid M_i, M) \Pr(M_i \mid M)}{\Pr(\mathrm{data} \mid M)} \\ &= \frac{1}{k} \frac{\Pr(\mathrm{data} \mid M_i)}{\Pr(\mathrm{data} \mid M)} \\ &= \frac{1}{k} \frac{\Pr(\mathrm{data} \mid M_i) / \Pr(\mathrm{data} \mid M_0)}{\Pr(\mathrm{data} \mid M) / \Pr(\mathrm{data} \mid M_0)} \\ &= \frac{\mathrm{BF}_i}{k \mathrm{BF}_{\mathrm{reg}}} \\ &\propto \mathrm{BF}_i \,.\end{aligned}$$

7.2.3 Discussion

Calculating region BFs from single-SNP BFs has been previously described in certain contexts (MARCHINI ET AL. 2007, SERVIN & STEPHENS 2007). The above formulation makes explicit that this is facilitated by the conditional independence assumption.

These derivations apply irrespective of the correlation amongst SNPs. In the situation of significantly associated but also highly correlated SNPs, the correct conclusion is that any of these could be causal but without necessarily identifying which one (SERVIN & STEPHENS 2007). This will manifest itself through high single-SNP and region BFs, but with the posterior distributed fairly evenly across multiple SNPs. Notice that in the situation where a set of SNPs are all perfectly correlated but only one of which is causal, fine mapping can only narrow the effect down to that set of SNPs but cannot determine which one is causal.

I have assumed that the causal SNP is typed in the study, which might be reasonable for a sufficiently thorough investigation of the variation in a region. Where this does not hold, the above methods are still applicable if there is a good surrogate SNP for the true effect. The conditional independence assumption still applies, but the conclusions can only extend as far as identifying the presence/location of a surrogate SNP rather than the causal SNP.

In the presence of multiple causal SNPs, these methods are no longer optimal. They will tend to pick out the SNP with the best marginal effect, which may or may not be one of the causal SNPs.

7.3 WTCCC fine mapping analyses

The WTCCC has recently carried out a fine mapping pilot project for three diseases at a selection of regions, many of which were identified in its initial genome-wide study. Samples were genotyped at dense set of SNPs across these regions, with the aim of determining the nature of the genetic signal in each and hopefully identifying the causal variants. Detailed analyses of these data are still underway, but here I present findings from specific analyses that I contributed to the project. In particular, I look for effects that are non-additive or secondary to the main SNP in each region and find evidence of both.

7.3.1 Data & methods

The project examined three diseases: autoimmune thyroid disease (AITD), coronary artery disease (CAD) and type 2 diabetes (T2D). These were compared to a set of common controls coming from both 58C and UKBS from the genome-wide study. Each sample was genotyped

Table 7.1: **Regions targeted by the WTCCC fine mapping pilot project.** Genomic coordinates are for build 36. Numbers of SNPs are after QC exclusions.

	Position (bp)		No. of		Disease of interest		
Chr.	Begin	End	SNPs	Name	AITD	CAD	T2D
1	109,539,481	109,645,000	232	PSRC1		•	
1	155,805,552	156,085,552	606	FCRL3	•		
1	220,784,001	221,044,001	356	1q41		•	
2	204,377,740	204,525,740	292	CTLA-4	•		
2	226,730,739	226,907,739	368	2q36		•	
6	20,632,000	20,838,000	510	CDKAL1			•
7	28,006,285	28,227,285	339	JAZF1			•
9	21,921,100	22,133,000	535	CDKN2A		•	•
10	6,057,797	6,171,000	425	CD25	•		
10	44,014,000	44,150,000	366	CXCL12		•	
10	94,189,000	94,491,000	573	HHEX			•
10	114,707,000	114,820,000	156	TCF7L2			•
16	52,350,000	52,410,000	209	FTO			•

Table 7.2: **Samples sizes for the WTCCC fine mapping pilot project.** Samples sizes are after QC exclusions.

Collection	Sample size
Common controls	1,930
Autoimmune thyroid disease	1,963
Coronary artery disease	1,934
Type 2 diabetes	2,037

on a custom Illumina iSelect assay, designed to cover the 13 regions shown in Table 7.1. One of these regions (CDKN2A) was investigated in two diseases, so it is more accurate to regard the study as having investigated 14 regions.

The genotypes were called using Illuminus (TEO ET AL. 2007) and phased using BEAGLE (BROWNING & BROWNING 2007). Some SNPs and individuals were excluded following standard QC procedures; I received a version of the data with these already applied. A total of 4,967 SNPs were in my data set, with the breakdown by region shown in Table 7.1. The sample size for each collection after exclusions is shown in Table 7.2.

The primary analyses of these data are being handled by other investigators and are still underway. These include single-SNP analyses to identify the SNPs that show the strongest evidence of association using the approach I described in Section 7.2. Here I only focus on two specific analyses that I carried out myself. The first evaluated the evidence of departure

from an additive model at highly associated SNPs. The second looked for effects additional to the one observed at the best SNP. Below, I describe each of these in turn and present my findings.

To take advantage of the full sample of individuals, for all my analyses I use the *expanded reference group* strategy from the initial WTCCC study. That is, for each disease I compare cases against the controls supplemented with cases from unrelated diseases. In this study, only CAD and T2D are considered to be related for this purpose.

7.3.2 Non-additive effects

As previously discussed, most associations discovered by GWAS show an additive signal. In Section 7.1 I showed that this could be due to insufficient power to detect deviations from an additive model when the discovery is at a marker locus not highly correlated with the causal variant. The availability of dense SNP data in regions with known associations allows us to investigate such effects with hopefully greater power.

For each SNP in each region, I calculated a BF which compares the additive and general models. Conveniently, this is just a ratio of the BFs used for association testing,

$$\begin{aligned}\mathrm{BF}_{\text{non-additive}} &= \frac{\Pr(\text{data} \mid M_{\text{general}})}{\Pr(\text{data} \mid M_{\text{additive}})} \\ &= \frac{\Pr(\text{data} \mid M_{\text{general}}) / \Pr(\text{data} \mid M_{\text{null}})}{\Pr(\text{data} \mid M_{\text{additive}}) / \Pr(\text{data} \mid M_{\text{null}})} \\ &= \frac{\mathrm{BF}_{\text{general}}}{\mathrm{BF}_{\text{additive}}}.\end{aligned}$$

I use the implementation described in Chapter 5, including the priors from Section 5.4.3. We are only interested in comparing models at SNPs which actually show an obvious association signal, so I restrict my attention to the best SNPs in each region. In particular, I considered all SNPs with $\log_{10}(\mathrm{BF}_{\text{general}}) > 3$ and then looked for those that also had $\log_{10}(\mathrm{BF}_{\text{non-additive}}) > 0$. Three regions had SNPs with this property: CDKN2A in both CAD and T2D, and FTO in T2D.

Figures 7.5–7.7 show a collection of plots for these three regions. I show BFs for all SNPs in the region, and colour the SNPs with high BFs for easy comparison across plots. I also plot parameter estimates for these SNPs on a model space plot (similar to Figure 7.3), to see the nature of the non-additive signal.

The CDKN2A region in CAD gives the clearest signature of what we might expect from a non-additive effect. The best SNP shows some evidence of departure from additivity, while the surrounding SNPs regress to a more additive signal, giving a crescent shape on the BF–BF plot and a corresponding drift to the x-axis on the model space plot. Nevertheless, the evidence against additivity here is only suggestive, with $\log_{10}(\mathrm{BF}_{\text{non-additive}}) \approx 0.5$ at the best SNP.

The FTO region shows a somewhat similar picture, although this time the best SNPs look to be more additive than some of the SNPs further down the list. This may be just a chance deviation at these SNPs, or a sign that there are more complex effects in this region. While the SNPs shown on the model space plot seem to cluster together, this is slightly misleading because of the high levels of LD in this region—that is, we expect most of these SNPs to be showing similar effects. In the end, as for CDKN2A the evidence against additivity is best described as only suggestive.

The last of the three regions, CDKN2A in T2D shows very inconsistent effects. It is possible that there are multiple effects at work here—the signal plot seems to show effects in two separate sub-regions, and the SNPs shown from each lie in different parts of the model space plot. However, with the evidence coming from so few SNPs and being generally weak, it is not compelling.

For comparison, Figures 7.8–7.10 show plots for three regions that did not show any evidence of departure from an additive model (more accurately, they show evidence in *favour* of an additive model).

It is interesting to note that for the regions shown here, the evidence for a departure from additivity is often inconsistent with that observed in the original study (WTCCC 2007). In particular, none of the three regions highlighted above showed significant departures from an additive model previously. Furthermore, the only region that did show significant departures previously, CDKAL1 for T2D, here shows a definite additive effect. This is mostly a reflection of the fact that the evidence of departure in both studies is fairly weak, and thus likely to be unreliable.

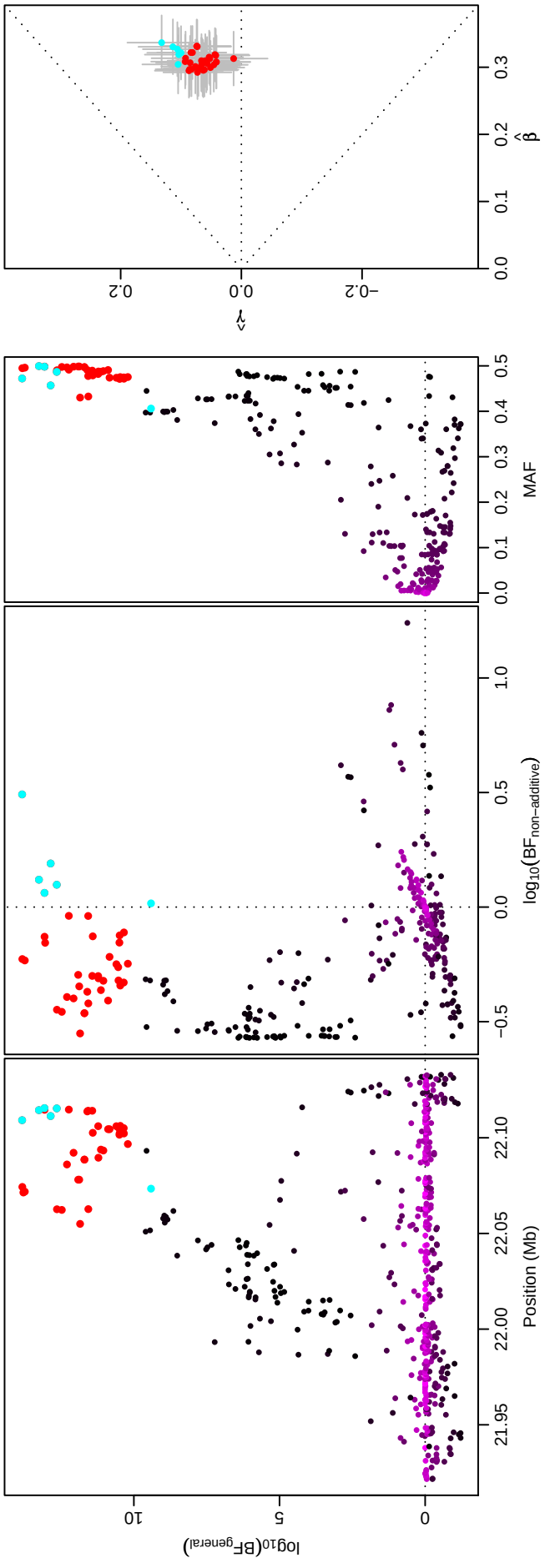

Figure 7.5: **Non-additivity analysis: CAD, CDKN2A.** The group of three plots on the left show the BF for the general model (on a log scale) against, in order, genomic position (i.e. a Bayesian signal plot), the non-additivity BF (also on a log scale) and the sample MAF. Each point is for a SNP in the region. The second of these three plots is the main focus; the others are there to give context. SNPs with a high general model BF are coloured in red or cyan to distinguish them in the different plots; those in cyan have $\log_{10}(\mathrm{BF}_{\text{non-additive}}) > 0$. The other SNPs are coloured on a gradient from black to magenta according to MAF, as in Figure 6.7. The plot on the right shows MLEs of parameters for the SNPs previously coloured in red or cyan (it is a model space plot, like in Figure 7.3 but now showing parameter estimates), with the grey cross-hairs extending for one standard error in each direction.

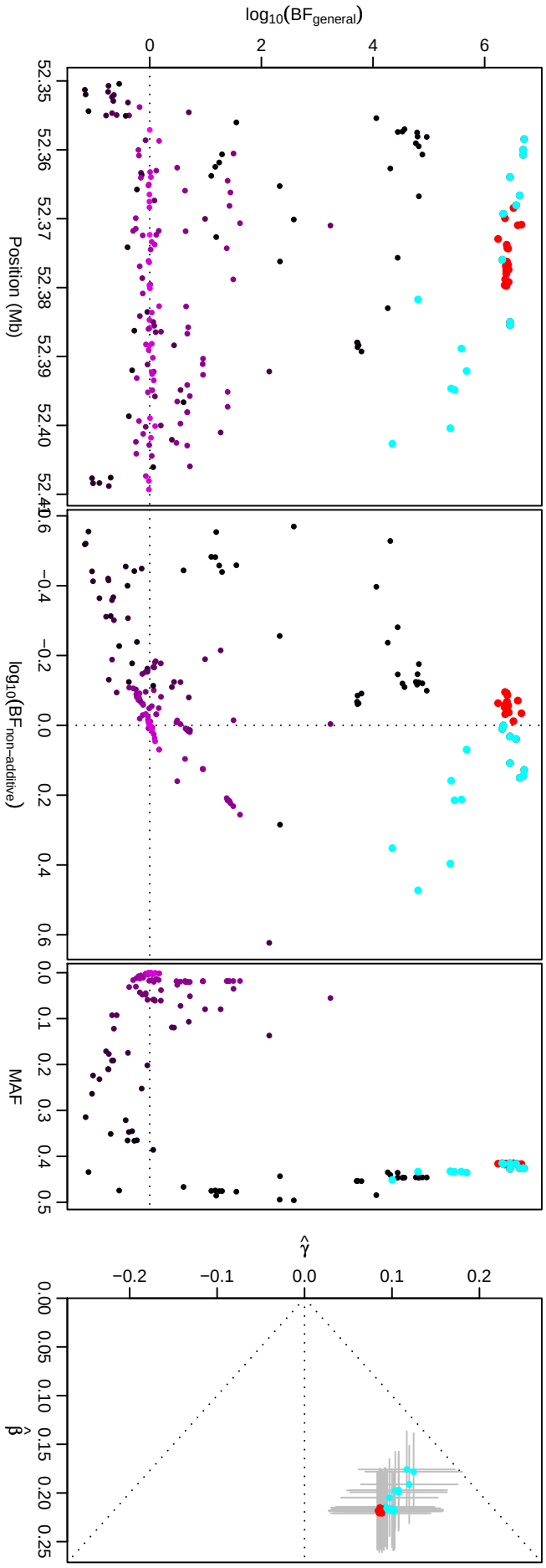

Figure 7.6: **Non-additivity analysis: T2D, FTO.** See Figure 7.5 for a description.

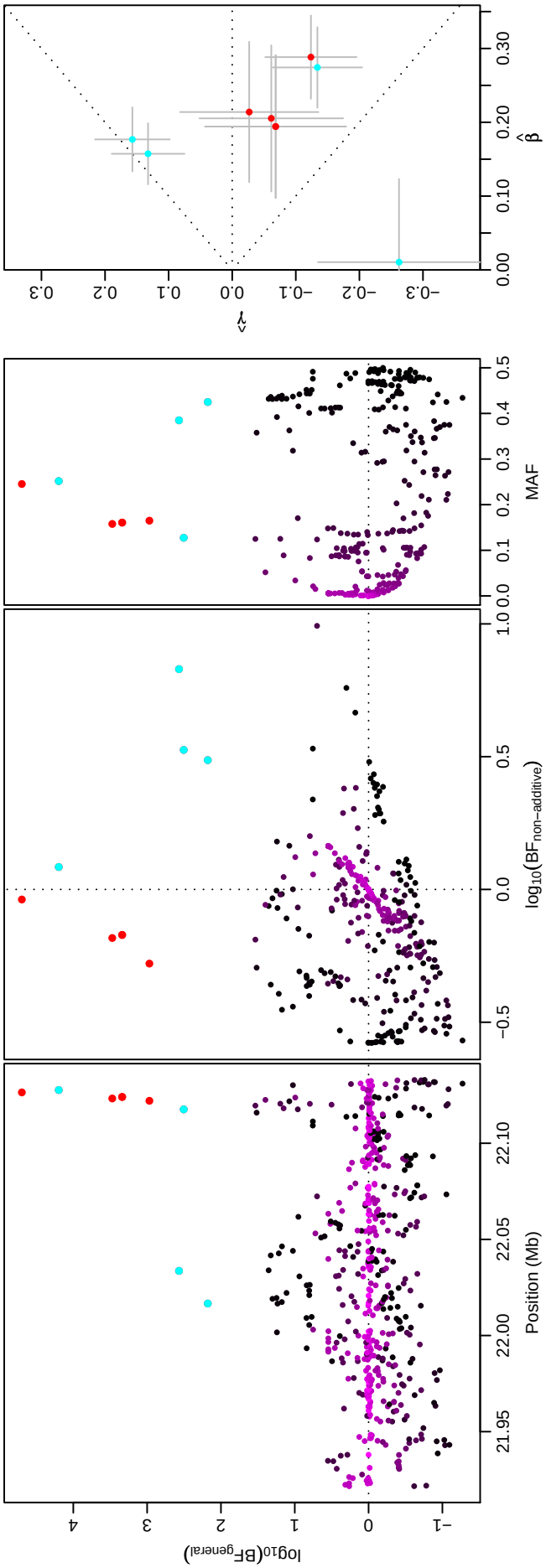

Figure 7.7: **Non-additivity analysis: T2D, CDKN2A.** See Figure 7.5 for a description.

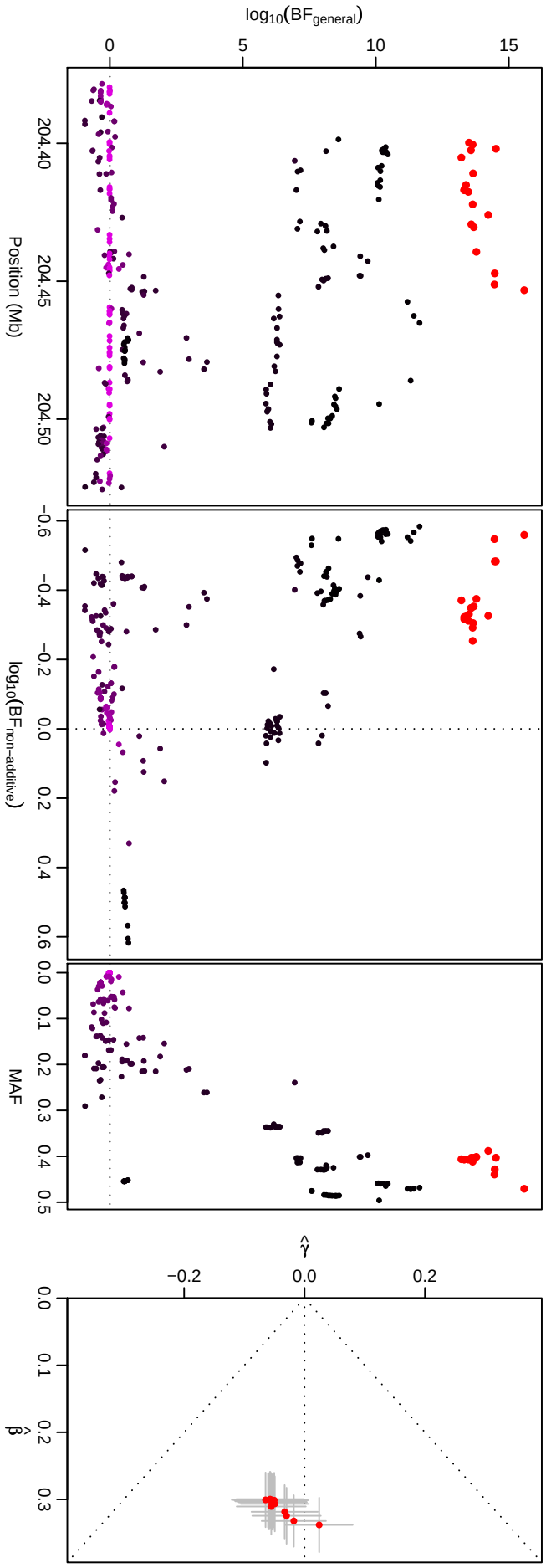

Figure 7.8: **Non-additivity analysis: AITD, CTLA-4.** See Figure 7.5 for a description.

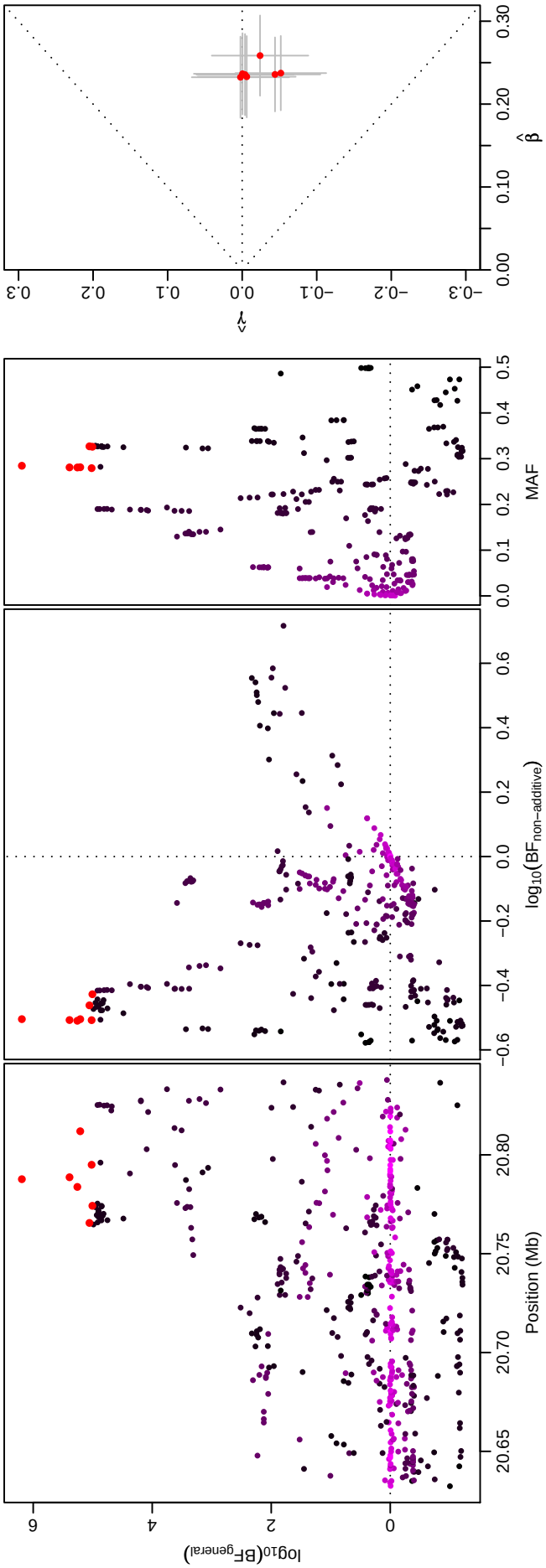

Figure 7.9: **Non-additivity analysis: T2D, CDKAL1.** See Figure 7.5 for a description.

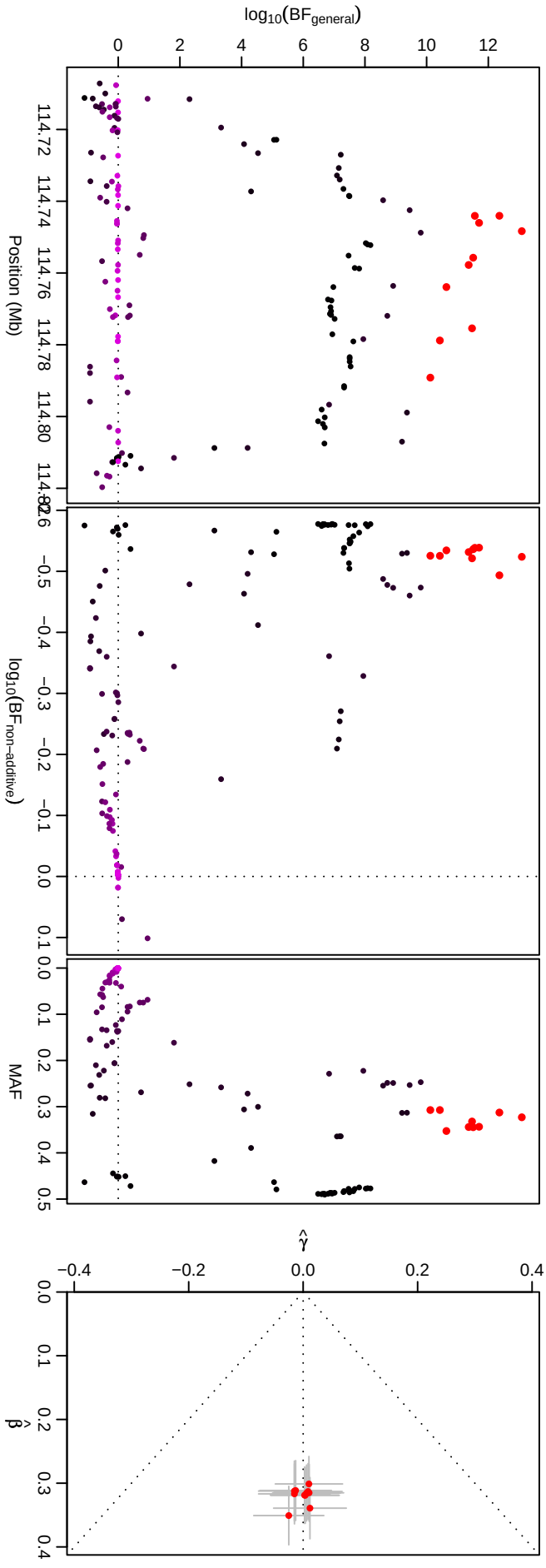

Figure 7.10: **Non-additivity analysis: T2D, TCF7L2.** See Figure 7.5 for a description.

7.3.3 Secondary signals & haplotypic effects

In addition to non-additive effects, these dense SNP data also allow us to more effectively search for signals that are present in addition to the effect at the best SNP. For example, there might be multiple causal variants in the region, or an effect that is tagged well by a particular haplotype background but not a single SNP.

There are many possible approaches to search for such secondary effects. I consider the simple approach of fitting the additive model at each SNP while using as a covariate the genotype calls at the best SNP in the region. In other words, I compare a two-SNP additive model against the best one-SNP additive model, thus testing whether the new SNP provides significant additional explanatory power. I will refer to these as *conditional* tests. For each SNP, I exclude individuals that have missing data at either that SNP or the best SNP.

Using a p-value threshold of 1×10^{-3}, three regions showed evidence of secondary effects, all in T2D: CDKAL1, CDKN2A, FTO. I explored each of these further to characterise the nature of the effect.

T2D, CDKAL1

Figure 7.11 shows the results of the conditional tests for this region. The best SNP in this region is rs7756992 and the best SNP on the conditional tests is rs6456360 with a p-value of 8.7×10^{-5}. Taken together, all 9 possible genotype pairs are observed in the data.

Comparing the two-SNP additive model to the two-SNP general model (here, 3 parameters vs 9 parameters) gives a p-value of 0.8, so the additive model provides an adequate fit.

With all possible genotype pairs present, all 4 possible underlying haplotypes must be present and thus up to 10 possible diplotypes. Indeed, in the phased data we observe all 10. The only phase ambiguity occurs where we observe heterozygotes at both SNPs—this can correspond to either of the diplotypes $\{01,10\}$ or $\{00,11\}$. Using the phased data allows us to resolve this ambiguity and test whether there are haplotype effects masked by phase.

Comparing an additive model on haplotypes (4 parameters, one for each possible haplotype) to a saturated model (10 parameters, one for each possible haplotype pair) gives a p-value of 0.8. Furthermore, comparing the additive model on haplotypes to that of SNPs

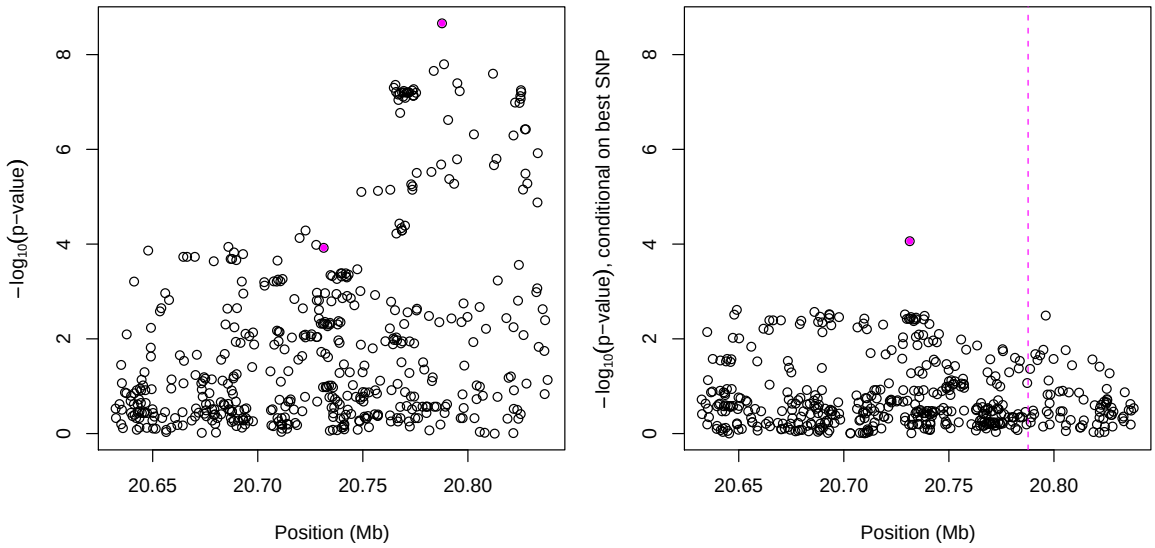

Figure 7.11: **Conditional analysis: T2D, CDKAL1.** On the left, the signal plot under the additive test. On the right, the signal plot under the additive test after conditioning on the best SNP in the region (according to the unconditional additive test). The best SNP in each analysis is highlighted in magenta. It is also highlighted in the other plot respectively, but the best SNP in the unconditional analysis is not present in the conditional analysis, so is only shown by a vertical line denoting its genomic position.

Table 7.3: **Two-SNP additive disease model for T2D, CDKAL1.**

SNP	MAF	Relative risk
rs7756992	0.28	1.28 (1.18–1.40)
rs6456360	0.49	1.17 (1.08–1.26)

(4 parameters vs 3 parameters) gives a p-value of 0.9. Thus, we can confidently say that the SNPs on their own adequately describe the effect. The relative risks for the two-SNP additive model are shown in Table 7.3. Figure 7.12 shows an interaction plot which clearly shows the adequacy of the additive SNP model.

T2D, CDKN2A

Figure 7.13 shows the results of the conditional tests for this region. The best SNP in this region is rs12555274 and the best SNP on the conditional tests is rs10965250 with a p-value of 4.1×10^{-5}.

It turns out that a haplotypic effect has previously been observed in this region, with three haplotype backgrounds conferring different disease risks (ZEGGINI ET AL. 2007, SU 2008). The two SNPs, rs10811661 and rs10217762, together distinguish these haplotypes well, and

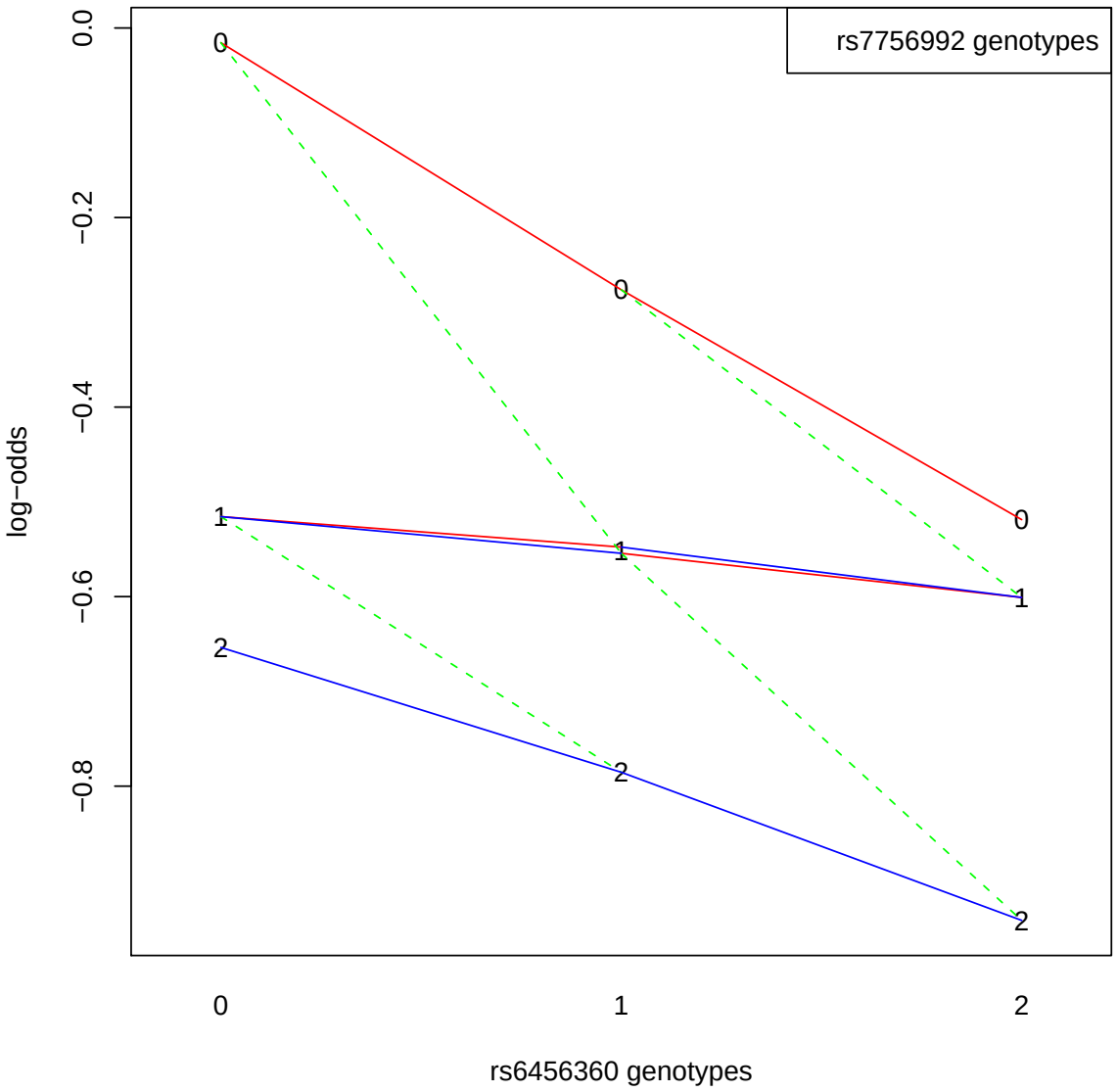

Figure 7.12: **Haplotype model interaction plot for T2D, CDKAL1.** Endpoints of line segments correspond to the observed log-odds of a diplotype (there are 10 diplotypes). Digits correspond to genotypes of a SNP, one on the x-axis and the other on the plot. The genotype combination 11 corresponds to two diplotypes, {01,10} and {00,11}, which here have nearly identical observed log-odds. Coloured lines represent effects which would be described by the same term in an additive haplotype model. Parallel lines of the same colour indicate that an additive haplotype model is a good fit. Parallel lines that join the same SNP genotypes (here, red and blue) further indicate that an additive SNP genotype model is a good fit.

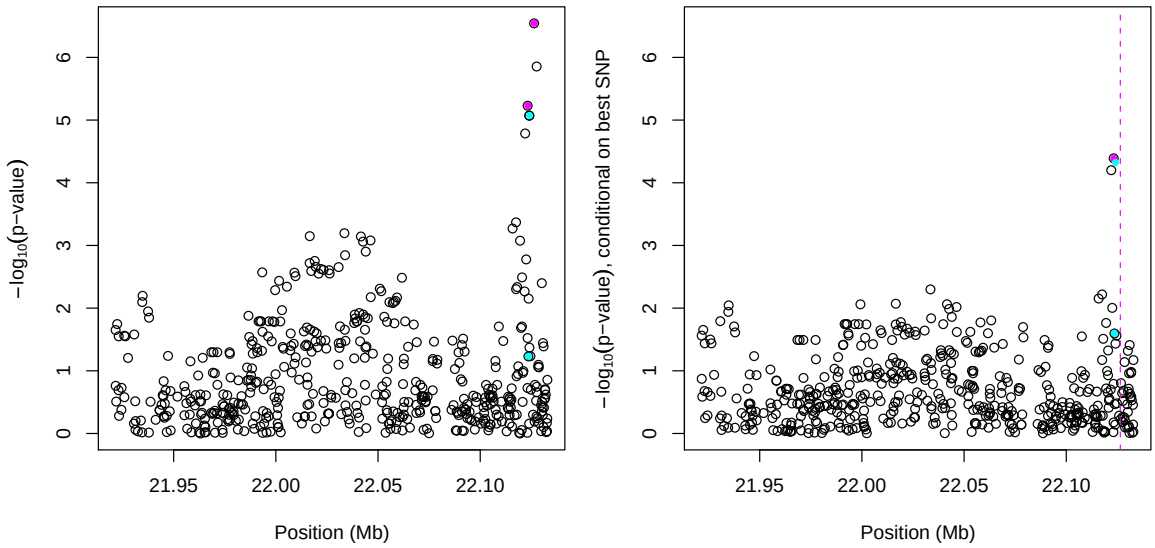

Figure 7.13: **Conditional analysis: T2D, CDKN2A.** Same as Figure 7.11. In addition, the two SNPs that together tag the haplotypes of differing risk are shown in cyan.

Table 7.4: **Haplotype disease model for T2D, CDKN2A.** The haplotypes are defined by rs10811661 and rs10217762 and are ordered by increasing risk.

Haplotype	Frequency	Relative risk
1	0.16	1.00 (ref.)
2	0.59	1.19 (1.06–1.34)
3	0.25	1.49 (1.31–1.69)

are coloured cyan on the plots in the above figure. The SNPs highlighted by the conditional analysis capture essentially the same effects, although not quite as well as the pair just given. This best SNP manages to distinguishes the high-risk haplotype fairly well, but blurs the distinction between the other two somewhat (in a way that allows it to capture a part of the risk difference), so is not an optimal choice when part of a pair of SNPs.

Taken together, rs10811661 and rs10217762 result in only 6 different genotype-pair combinations in the data. This therefore corresponds to 3 underlying haplotypes, for which phase is unambiguous.

Comparing the two-SNP additive model to the two-SNP general model (here, 3 parameters vs 6 parameters) gives a p-value of 0.5, so the additive model provides an adequate fit. This allows us to write the model in terms of the relative risks of the underlying haplotypes, as shown in Table 7.4.

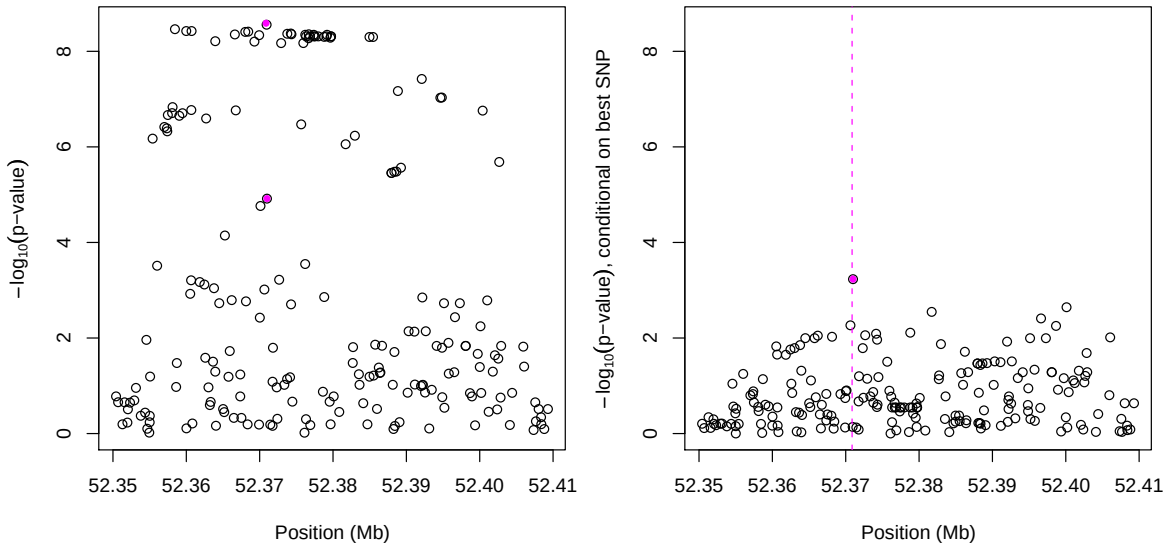

Figure 7.14: **Conditional analysis: T2D, FTO.** See Figure 7.11 for a description.

Table 7.5: **Two-SNP additive disease model for T2D, FTO.**

SNP	MAF	Relative risk
rs17817449	0.42	1.23 (1.14–1.33)
rs8063946	0.055	1.37 (1.14–1.64)

T2D FTO

Figure 7.14 shows the results of the conditional tests for this region. The best SNP in this region is rs17817449, and the best SNP on the conditional tests is rs8063946 with a p-value of 5.9×10^{-4}. Taken together, only 6 different genotype-pair combinations are observed in the data. This therefore corresponds to 3 underlying haplotypes, for which phase is unambiguous.

Comparing the two-SNP additive model to the two-SNP general model (here, 3 parameters vs 6 parameters) gives a p-value of 0.4, so the additive model provides an adequate fit. The relative risks for the two-SNP additive model are shown in Table 7.5. Note that the second SNP is much rarer in the population, with a MAF of ∼5% compared to ∼40%, so some caution is warranted when interpreting this SNP.

7.3.4 Discussion

Fine mapping studies, which seek to elucidate and characterise causal variants in targeted regions, are natural follow-ons from GWAS. I have presented results of particular analyses from a recent such study conducted by the WTCCC. The primary analyses, which are being conducted by others and will be reported elsewhere, seek to localise the causal variants in each region. In contrast, my analyses attempted to find effects that deviate from single-SNP additive models.

I investigated two types of effects in particular. Firstly, departures from an additive model at a single SNP. While I found some evidence of such departures, the evidence was not particularly compelling and was often inconsistent with the original study. Secondly, I looked for disease effects that are more adequately described by more than one SNP. On this front the results look a little more promising. Nevertheless, it is notable that only a few of all the regions studied showed significant results.

A number of possibilities may explain the general lack of convincing findings. It could be that, despite typing these regions on high-density SNP arrays, the causal variants may not have been captured, may not actually be in the defined regions, or may just be poorly tagged by SNPs in general. Another possibility is that individual disease loci may actually be best modelled as single-locus additive effects, and that any departure from this is too small to be observed in our study.

Finally, it is worth noting that even where the disease effect seems to be best described by more than one SNP in the region or as a haplotype effect, it is still possible that a single SNP may be the causal variant and just happens to be poorly tagged. Thus, it is difficult to conclusively show a genuine multi-locus effect is at work.

Given these findings, in fine mapping we may be seeing the limit of relatively simple statistical approaches. Further studies may need to make much greater use of relevant biological knowledge.

List of abbreviations

58C	1958 Birth Cohort
AITD	Autoimmune thyroid disease
BD	Bipolar disorder
BF	Bayes factor
CAD	Coronary artery disease
CD	Crohn's disease
CDCV	Common disease/common variant (hypothesis)
CEU	The HapMap Caucasian sample, consisting of 30 parent-offspring trios from Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection
CNV	Copy number variant
GWA/GWAS	Genome-wide association [study/studies]
HapMap	The International HapMap Consortium
HLA	Human leukocyte antigen
HT	Hypertension
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MHC	Major histocompatibility complex
MLRT	Maximum likelihood ratio test
NHGRI	National Human Genome Research Institute
PAR	Population attributable risk
QC	Quality control (procedures/criteria)
RA	Rheumatoid arthritis
RR	Relative risk (or risk ratio)
SABP	Schizoaffective disorder, bipolar type
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TDT	Transmission disequilibrium test
UKBS	UK Blood Services
WTCCC	The Wellcome Trust Case Control Consortium

Bibliography

ADVIWARE PTY LTD (2007). Wrong Diagnosis (medical information website). http://www.wrongdiagnosis.com/. Website accessed in Jul 2007.

AFFYMETRIX (2006). BRLMM: an improved genotype calling method for the GeneChip®Human Mapping 500K Array Set. Tech. rep., Affymetrix, http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf.

AHN K, HAYNES C, KIM W ET AL. (2007). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann Hum Genet*, 71(2):249–261.

ALTMÜLLER J, PALMER LJ, FISCHER G ET AL. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet*, 69(5):936–950.

ARDLIE KG, LUNETTA KL & SEIELSTAD M (2002). Testing for population subdivision and association in four case-control studies. *Am J Hum Genet*, 71(2):304–311.

ARMITAGE P (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

BALDING DJ (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–791. Bayesian approaches are discussed in the Supplementary Information.

BALL RD (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics*, 159:1351–1364.

—— (2005). Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics*, 170:859–873.

—— (2007). Quantifying evidence for candidate gene polymorphisms: Bayesian analysis combining sequence-specific and quantitative trait loci colocation information. *Genetics*, 177:2399–2416.

BARRETT J, HANSOUL S, NICOLAE D ET AL. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, 40(8):955–962.

BARRETT JC & CARDON LR (2006). Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–662.

BEAUMONT MA & RANNALA B (2004). The Bayesian revolution in genetics. *Nat Rev Genet*, 5:251–261.

BEGOVICH AB, CARLTON VEH, HONIGBERG LA ET AL. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet*, 75(2):330–337.

BERGER JO (1993). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York. Corrected reprint of the second (1985) edition.

BERGER JO & DELAMPADY M (1987). Testing precise hypotheses. *Statist Sci*, 2(3):317–352. With comments and a rejoinder by the authors.

BERNARDO JM & SMITH AFM (1994). *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

BODMER W & BONILLA C (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 40:695–701.

BOTSTEIN D, WHITE RL, SKOLNICK M & DAVIS RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–331.

BOTTINI N, MUSUMECI L, ALONSO A ET AL. (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet*, 36(4):337–338.

BRAND OJ, LOWE CE, HEWARD JM ET AL. (2007). Association of the interleukin-2 receptor alpha (IL-2Ralpha)/CD25 gene region with Graves' disease using a multilocus test and tag SNPs. *Clin Endocrinol (Oxf)*, 66(4):508–512.

BROWNING BL & BROWNING SR (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84:210–223.

BROWNING SR & BROWNING BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81:1084–1097.

CASELLAS J & PIEDRAFITA J (2006). Bayes factor for testing the genetic background of quantitative threshold traits. *J Anim Breed Genet*, 123:301–306.

CHANOCK SJ, MANOLIO T, BOEHNKE M ET AL. (2007). Replicating genotype-phenotype associations. *Nature*, 447:655–660.

CHAPMAN JM, COOPER JD, TODD JA & CLAYTON DG (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56:18–31.

CHAPMAN JM, ONNIE CM, PRESCOTT NJ ET AL. (2009). Searching for genotype-phenotype structure: using hierarchical log-linear models in Crohn disease. *Am J Hum Genet*, 84:178–187.

CONCANNON P, GOGOLIN-EWENS KJ, HINDS DA ET AL. (1998). A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat Genet*, 19(3):292–296.

CORDER EH, SAUNDERS AM, STRITTMATTER WJ ET AL. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923.

COX DR & HINKLEY DV (1974). *Theoretical statistics*. Chapman and Hall, London.

COX DR & SNELL EJ (1989). *Analysis of binary data*, vol. 32 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2nd edn.

CRADDOCK N, JONES L, JONES IR ET AL. (2008). Strong genetic evidence for a selective influence of $GABA_A$ receptors on a component of the bipolar disorder phenotype. *Mol Psychiatry*. **(I am an author on this paper)**.

DEVLIN B & ROEDER K (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

DUDBRIDGE F & GUSNANTO A (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32:227–234.

EPSTEIN MP & SATTEN GA (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet*, 73:1316–1329.

EVANS M & SWARTZ T (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist Sci*, 10(3):254–272.

FREEDMAN ML, HAIMAN CA, PATTERSON N ET AL. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci USA*, 103:14068–14073.

FREEDMAN ML, REICH D, PENNEY KL ET AL. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–393.

FRIDLEY BL (2008). Bayesian variable and model selection methods for genetic association studies. *Genet Epidemiol*, 33:27–37.

FUNG EY, SMYTH DJ, HOWSON JM ET AL. (2009). Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immun*, 10:188–191.

GART JJ & TARONE RE (1983). The relation between score tests and approximate UMPU tests in exponential models common in biometry. *Biometrics*, 39(3):781–786.

GENZ A & KASS RE (1997). Subregion-adaptive integration of functions having a dominant peak. *J Comput Graph Statist*, 6(1):92–111.

GOOD IJ (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In HARPER WL & HOOKER CA (eds.), *Foundations of probability theory, statistical inference, and statistical theories of science (Proc. Internat. Res. Colloq., Univ. Western Ontario, London, Ont., 1973)*, vol. 2 of *Univ. Western Ontario Ser. Philos. Sci., Vol. 6*, pp. 125–174. Reidel, Dordrecht.

GORLOV IP, GORLOVA OY, SUNYAEV SR ET AL. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*, 82:100–112.

GREENLAND S (2008). Multiple comparisons and association selection in general epidemiology. *Int J Epidemiol*, 37:430–434.

GREENWOOD CM, RANGREJ J & SUN L (2007). Optimal selection of markers for validation or replication from genome-wide association studies. *Genet Epidemiol*, 31:396–407.

GUAN Y & STEPHENS M (2008). Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12):e1000279.

GUSELLA JF, WEXLER NS, CONNEALLY PM ET AL. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–238.

HALL JM, LEE MK, NEWMAN B ET AL. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689.

HAPMAP (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

—— (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

HELGADOTTIR A, THORLEIFSSON G, MANOLESCU A ET AL. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316(5830):1491–1493.

HINDORFF LA, SETHUPATHY P, JUNKINS HA ET AL. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*, 106:9362–9367.

HIRSCHHORN JN & DALY MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108.

HOGGART CJ, WHITTAKER JC, DE IORIO M & BALDING DJ (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*, 4:e1000130.

HUGOT JP, CHAMAILLARD M, ZOUALI H ET AL. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411(6837):599–603.

HUNT K, ZHERNAKOVA A, TURNER G ET AL. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet*, 40:395–402.

ILES MM (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet*, 4:e33.

JIMENEZ-SANCHEZ G, CHILDS B & VALLE D (2001). Human disease genes. *Nature*, 409(6822):853–855.

JOHNSON GC, ESPOSITO L, BARRATT BJ ET AL. (2001). Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–237.

JOHNSON T (2007). Bayesian method for gene detection and mapping, using a case and control design and DNA pooling. *Biostatistics*, 8:546–565.

KASS RE & RAFTERY AE (1995). Bayes factors. *J Amer Statist Assoc*, 90(430):773–795.

KASS RE & VAIDYANATHAN SK (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J Roy Statist Soc Ser B*, 54(1):129–144.

KASS RE & WASSERMAN L (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Amer Statist Assoc*, 90(431):928–934.

KEREM B, ROMMENS JM, BUCHANAN JA ET AL. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080.

KORMAN BD, SELDIN MF, TAYLOR KE ET AL. (2009). The chromosome 7q region association with rheumatoid arthritis in females in a British population is not replicated in a North American case-control series. *Arthritis Rheum*, 60:47–52.

KRUGLYAK L & NICKERSON DA (2001). Variation is the spice of life. *Nat Genet*, 27(3):234–236. News.

KRYUKOV GV, PENNACCHIO LA & SUNYAEV SR (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*, 80:727–739.

KUSTRA R, SHI X, MURDOCH DJ ET AL. (2008). Efficient p-value estimation in massively parallel testing problems. *Biostatistics*, 9:601–612.

LAIRD NM & LANGE C (2006). Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*, 7(5):385–394.

LEE SH, VAN DER WERF JH, HAYES BJ ET AL. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet*, 4:e1000231.

LEWINGER JP, CONTI DV, BAURLEY JW ET AL. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*, 31:871–882.

LI Y, DING J & ABECASIS GR (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference [abstract program number 2290]. Presented at the annual meeting of The American Society of Human Genetics, 12 Oct 2006, New Orleans, Louisiana. Available from `http://www.ashg.org/genetics/ashg06s/index.shtml`.

LITT M & LUTY JA (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet*, 44(3):397–401.

LOHMUELLER KE, PEARCE CL, PIKE M ET AL. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33(2):177–182.

MACDONALD ME, AMBROSE CM, DUYAO MP ET AL. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, 72(6):971–983.

MACKAY TFC & ANHOLT RRH (2006). Of flies and man: Drosophila as a model for human complex traits. *Annu Rev Genomics Hum Genet*, 7:339–367.

MAHALANOBIS PC (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 2(1):49–55.

MANOLIO TA, BROOKS LD & COLLINS FS (2008). A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*, 118(5):1590–1605.

MARCHINI J, DONNELLY P & CARDON LR (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37:413–417.

MARCHINI J, HOWIE B, MYERS S ET AL. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913.

MCCARROLL SA & ALTSHULER DM (2007). Copy-number variation and association studies of human disease. *Nat Genet*, 39:37–42.

MCCARTHY M, ABECASIS G, CARDON L ET AL. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9:356–369.

MCCULLAGH P & NELDER JA (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

MCPHERSON R, PERTSEMLIDIS A, KAVASLAR N ET AL. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316(5830):1488–1491.

MCVEAN GA, MYERS SR, HUNT S ET AL. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584.

MOLITOR J, MARJORAM P & THOMAS D (2003). Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genet Epidemiol*, 25:95–105.

MORRIS AP (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genet Epidemiol*, 29:91–107.

—— (2006). A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am J Hum Genet*, 79:679–694.

MUSTELIN T, VANG T & BOTTINI N (2005). Protein tyrosine phosphatases and the immune response. *Nat Rev Immunol*, 5(1):43–57.

MYERS RH, MONTGOMERY DC & VINING GG (2002). *Generalized linear models*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York. With applications in engineering and the sciences.

NEEL JV (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet*, 14:353–362.

OGURA Y, BONEN DK, INOHARA N ET AL. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603–606.

PARKES M, BARRETT J, PRESCOTT N ET AL. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet*, 39:830–832.

PAYNTER NP, CHASMAN DI, BURING JE ET AL. (2009). Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med*, 150:65–72.

PE'ER I, YELENSKY R, ALTSHULER D & DALY MJ (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*, 32:381–385.

PENG B & KIMMEL M (2007). Simulations provide support for the common disease-common variant hypothesis. *Genetics*, 175(2):763–776.

PRENTICE RL & PYKE R (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

PRITCHARD JK (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–137.

PRITCHARD JK & COX NJ (2002). The allelic architecture of human disease genes: common disease-common variant... or not? *Hum Mol Genet*, 11(20):2417–2423.

PRITCHARD JK & PRZEWORSKI M (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14.

PRITCHARD JK, STEPHENS M & DONNELLY P (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

PRITCHARD JK, STEPHENS M, ROSENBERG NA & DONNELLY P (2000b). Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–181.

R DEVELOPMENT CORE TEAM (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.r-project.org/`.

REDON R, ISHIKAWA S, FITCH KR ET AL. (2006). Global variation in copy number in the human genome. *Nature*, 444:444–454.

REICH D, PATTERSON N, DE JAGER PL ET AL. (2005). A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet*, 37(10):1113–1118.

REICH DE & LANDER ES (2001). On the allelic spectrum of human disease. *Trends Genet*, 17(9):502–510.

RENCHER AC (1995). *Methods of multivariate analysis*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

RIORDAN JR, ROMMENS JM, KEREM B ET AL. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245:1066–1073.

RISCH N & MERIKANGAS K (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

RISCH NJ (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856.

ROEDER K, BACANU SA, WASSERMAN L & DEVLIN B (2006). Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet*, 78:243–252.

ROMMENS JM, IANNUZZI MC, KEREM B ET AL. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245:1059–1065.

ROUSSEAU J (2007). Approximating interval hypothesis: p-values and Bayes factors. In *Bayesian statistics 8*, Oxford Sci. Publ., pp. 417–452. Oxford Univ. Press, Oxford.

SALANTI G, HIGGINS JP, TRIKALINOS TA & IOANNIDIS JP (2007). Bayesian meta-analysis and meta-regression for gene-disease associations and deviations from Hardy-Weinberg equilibrium. *Stat Med*, 26:553–567.

SASIENI PD (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53:1253–1261.

SATTEN GA & EPSTEIN MP (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol*, 27:192–201.

SCHAID DJ, ROWLAND CM, TINES DE ET AL. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, 70:425–434.

SCHEET P & STEPHENS M (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–644.

SCHOUTEN EG, DEKKER JM, KOK FJ ET AL. (1993). Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. *Stat Med*, 12:1733–1745.

SCHRODI SJ (2005). A probabilistic approach to large-scale association scans: a semi-Bayesian method to detect disease-predisposing alleles. *Stat Appl Genet Mol Biol*, 4:Article31.

SEAMAN SR & RICHARDSON S (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika*, 91(1):15–25.

SERVIN B & STEPHENS M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114.

SHAM PC, CHERNY SS, PURCELL S & HEWITT JK (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet*, 66:1616–1630.

SHERRY ST, WARD MH, KHOLODOV M ET AL. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.

SMITH AFM & SPIEGELHALTER DJ (1980). Bayes factors and choice criteria for linear models. *J Roy Statist Soc Ser B*, 42(2):213–220.

SMITH MW & O'BRIEN SJ (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet*, 6(8):623–632.

SMYTH D, COOPER JD, COLLINS JE ET AL. (2004). Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes*, 53(11):3020–3023.

SMYTH GK (2003). Pearson's goodness of fit statistic as a score test statistic. In *Statistics and science: a Festschrift for Terry Speed*, vol. 40 of *IMS Lecture Notes Monogr. Ser.*, pp. 115–126. Inst. Math. Statist., Beachwood, OH.

SPENCER C, HECHTER E & DONNELLY P (2008). The allelic architecture of common diseases and its consequences [abstract program number 66]. Presented at the annual meeting of The American Society of Human Genetics, 13 Nov 2008, Philadelphia, Pennsylvania. Available from `http://www.ashg.org/2008meeting/abstracts/fulltext/`.

SPENCER CC, SU Z, DONNELLY P & MARCHINI J (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5:e1000477.

SPIELMAN RS, MCGINNIS RE & EWENS WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52(3):506–516.

STERNE JA & DAVEY SMITH G (2001). Sifting the evidence—what's wrong with significance tests? *BMJ*, 322:226–231.

STOREY JD & TIBSHIRANI R (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 100:9440–9445.

SU Z (2008). *Statistical Methods for the Analysis of Genetic Association Studies*. D.Phil. thesis, Balliol College, University of Oxford.

SYVÄNEN AC (2005). Toward genome-wide SNP genotyping. *Nat Genet*, 37 Suppl:5–10.

SZKLO M, NIETO FJ & MILLER D (2001). Epidemiology: beyond the basics.

TACHMAZIDOU I, VERZILLI CJ & DE IORIO M (2007). Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet*, 3:e111.

TEO YY, INOUYE M, SMALL KS ET AL. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23:2741–2746.

THOMAS DC (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, New York.

THOMAS DC & CLAYTON DG (2004). Betting odds and genetic associations. *J Natl Cancer Inst*, 96:421–423.

THOMSON W, BARTON A, KE X ET AL. (2007). Rheumatoid arthritis association at 6q23. *Nat Genet*, 39:1431–1433.

TODD J, WALKER N, COOPER J ET AL. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*, 39:857–864.

VAN HEEL D, FRANKE L, HUNT K ET AL. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet*, 39:827–829.

VAN HEEL DA & WEST J (2006). Recent advances in coeliac disease. *Gut*, 55:1037–1046.

VELLA A, COOPER JD, LOWE CE ET AL. (2005). Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am J Hum Genet*, 76(5):773–779.

VERZILLI C, SHAH T, CASAS JP ET AL. (2008). Bayesian meta-analysis of genetic association studies with different sets of markers. *Am J Hum Genet*, 82:859–872.

VERZILLI CJ, STALLARD N & WHITTAKER JC (2006). Bayesian graphical models for genomewide association studies. *Am J Hum Genet*, 79:100–112.

VOGEL F & MOTULSKY AG (1982). *Human Genetics: Problems and Approaches*. Springer, Berlin.

WACHOLDER S, ROTHMAN N & CAPORASO N (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst*, 92(14):1151–1158.

WAKEFIELD J (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*, 81(2):208–227.

—— (2008). Reporting and interpretation in genome-wide association studies. *Int J Epidemiol*, 37(3):641–653.

—— (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol*, 33(1):79–86.

WANG W & PIKE N (2004). The allelic spectra of common diseases may resemble the allelic spectrum of the full genome. *Medical Hypotheses*, 63(4):748–751.

WANG WYS, BARRATT BJ, CLAYTON DG & TODD JA (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6(2):109–118.

WEBER JL & MAY PE (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*, 44(3):388–396.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678. **(I am an author on this paper)**.

YAMANOUCHI J, RAINBOW D, SERRA P ET AL. (2007). Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. *Nat Genet*, 39(3):329–337.

YUSUF S, HAWKEN S, OUNPUU S ET AL. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet*, 364:937–952.

ZEGGINI E, WEEDON M, LINDGREN C ET AL. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341.

ZELLNER A (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques*, vol. 6 of *Stud. Bayesian Econometrics Statist.*, pp. 233–243. North-Holland, Amsterdam.

ZHU X, LUKE A, COOPER RS ET AL. (2005). Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet*, 37(2):177–181.

ZÖLLNER S & PRITCHARD JK (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet*, 80(4):605–615.

ZONDERVAN KT & CARDON LR (2004). The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5:89–100.