On genetic variants underlying common disease

Eliana Hechter
St John's College
Department of Statistics
University of Oxford

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

Supervisor: Professor Peter Donnelly

Hilary Term, 2011

Abstract

Genome-wide association studies (GWAS) exploit the correlation in genetic diversity along chromosomes in order to detect effects on disease risk without having to type causal loci directly. The inevitable downside of this approach is that, when the correlation between the marker and the causal variant is imperfect, the risk associated with carrying the predisposing allele is diluted and its effect is underestimated. This thesis explores four different facets of this risk dilution: (1) estimating true effect sizes from those observed in GWAS; (2) asking how the context of a GWAS, including the population studied, the genotyping chip employed, and the use of imputation, affects risk estimates; (3) assessing how often the best-associated SNP in a GWAS coincides with the causal variant; and (4) quantifying how departures from the simplest disease risk model at a causal variant distort the observed disease risk model.

Using simulations, where we have information about the true risk at the causal locus, we show that the correlation between the marker and the causal variant is the primary driver of effect size underestimation. The extent of the underestimation depends on a number of factors, including the population in which the study is conducted, the genotyping chip employed, whether imputation is used, and the strength, frequency, and disease model of the risk allele. Suppose that a GWAS study is conducted in a European population, with an Affymetrix 6.0 genotyping chip, without imputation, and that the causal loci have a modest effect on disease risk, are common in

the population, and follow an additive disease risk model. In such a study, we show that the risk estimated from the most associated SNP is very close to the truth approximately two-thirds of the time (although we predict that fine mapping of GWAS loci will infrequently identify causal variants with considerably higher risk), and that the best-associated variant is very often perfectly or nearly-perfectly correlated with, and almost always within 0.1cM of, the causal variant. However, the strong correlations among nearby loci mean that the causal and best-associated variants coincide infrequently, less than one-fifth of the time, even if the causal variant is genotyped. We explore ways in which these results change quantitatively depending on the parameters of the GWAS study. Additionally, we demonstrate that we expect to identify substantial deviations from the additive disease risk model among loci where association is detected, even though power to detect departures from the model drops off very quickly as the correlation between the marker and causal loci decreases. Finally, we discuss the implications of our results for the design and interpretation of future GWAS studies.

Acknowledgments

My studies at Oxford were made possible by a Rhodes Scholarship.

I am indebted to my supervisor Peter Donnelly, whose exacting scientific standards and incisive mind have shaped me indelibly. In addition, I wish to express my gratitude to Chris Spencer and Damjan Vukcevic, both of whom made intellectual contributions to the project, and to Simon Myers, Gil McVean, and Andrew Morris for their comments on earlier versions of the thesis. I would also like to thank Bryan Howie and Bryan Browning for helpful discussions and advice regarding imputation. Simon Myers and Joel Hirschhorn were particularly generous in agreeing to examine me. I am appreciative of their efforts to make my viva possible, especially Joel, who modified the parameters of a family vacation in order to come to Oxford.

I am grateful to Eric Lander for welcoming me at the Broad Institute as a visiting graduate student during the last eighteen months of my DPhil. My work with him is not represented in this thesis, but I have valued his mentorship in the last stages of my graduate work.

Finally, I wish to thank my family and friends, both in Oxford and back home, for their encouragement and support. I dedicate this thesis to my father, who first brought me to Oxford, and to my mother, who helped me to get there myself.

Contents

1

Introduction

Trait variation in human populations, including disease status, shows correlation between relatives, suggesting an underlying genetic contribution. To the extent that it is known, the number and frequency of alleles contributing to variation – sometimes referred to as the "allelic spectrum" of a trait – varies greatly between diseases (Reich & Lander, 2001). At one end of the spectrum are single-gene Mendelian diseases like Huntington disease, in which a mutant protein in the *HTT* gene caused by too many CAG repeats is necessary and sufficient to cause the disease (Walker, 2007). At the other end of the spectrum are complex diseases, such as breast cancer, Crohn's disease, and type 2 diabetes, which are influenced both by genetic and environmental factors and do not exhibit Mendelian patterns of inheritance in general.

Identifying the genetic factors that underlie disease is of intrinsic biological interest, potentially of crucial importance to the development of new treatments (Hamburg & Collins, 2010) and has been an area of intense research activity for the past thirty years. Different approaches have been successful for different sorts of diseases. In order to set the context we briefly summarise these approaches, but for more detail, see Altshuler *et al.* (2008).

Genetic mapping, which correlates genetic and trait variation without the require-

ment of a biological hypothesis, was first developed in model organisms, but was not made possible in humans until the 1980s (Altshuler *et al.*, 2008). Botstein *et al.* (1980) suggested using naturally occurring genetic variation to identify regions tied to disease inheritance. Two loci in the genome are referred to as "linked" if they are inherited together more than would be expected under independence. Genetic linkage maps correlate polymorphic loci with regions of the genome that are linked in this way.

Linkage studies examine affected families looking for regions shared more than would be expected by chance among affected individuals. The statistic that measures the degree of linkage is variable across the genome simply by chance (Lander & Kruglyak, 1995). A lod score, which computes the ratio of probabilities under a model of linkage and under the null hypothesis of no assortment, can be used to determine if a given region is shared more than expected by chance. Lander & Kruglyak (1995) argued that the threshold stringent enough to prevent non-replicable findings yet low enough to yield results was a lod score of 3.6-4, depending on study design. Linkage studies have been very successful in the identification of Mendelian disease genes (Hirschhorn & Daly, 2005)[1], with specific loci identified for over 2800 disorders (McKusick-Nathans Institute of Genetic Medicine & National Center for Biotechnology Information). The results of linkage studies suggested that the candidate gene approach was inadequate, because most identified genes had not been previously suspected (Altshuler *et al.*, 2008).

Despite attempts to extend this method beyond Mendelian diseases, the linkage approach failed in the main to detect many loci associated with complex diseases (Altmüller *et al.*, 2001). Risch & Merikangas (1996) and others pointed out that linkage analysis is less powerful than genome-wide association for detecting common genetic variants with small effects. For a number of years, researchers tried to pursue this observation in candidate gene studies, but these proved largely unsuccessful and often produced putative associations that failed to replicate for a number of reasons, perhaps

[1]The reader is referred to Teare & Barrett (2005) for a fuller review of linkage studies.

most importantly, small sample size and lack of power to detect variants of small effect (Altshuler *et al.*, 2008). Whether common variants would play a role was a matter of some debate. The positive hypothesis is referred to as the common disease common variant (CD/CV) hypothesis. Modest-sized common variants are expected in many complex diseases under a model in which disease arises as a result of an interplay between multiple genetic and environmental factors (Wang *et al.*, 2005). Given that many complex diseases, such as type 2 diabetes, have increased in incidence over a period of a few decades, this model seemed plausible. Additionally, a number of theoretical arguments have used population genetics to justify an allelic architecture in which common variants contribute to disease susceptibility (Pritchard, 2001; Reich & Lander, 2001).

Recently, another type of genetic mapping, called a genome-wide association study, has proved successful at identifying common variants that contribute to complex diseases. Consistent with the CD/CV hypothesis, within the last few years, genome-wide association studies (GWAS) have uncovered a panoply of common genetic variants that predispose to disease (Hindorff *et al.*, 2010; Manolio *et al.*, 2008). At its most basic, an association study seeks to compare allele frequencies in cases and controls. There are approximately 3 billion bases in the human genome, and an estimated 10 million common variants (Altshuler *et al.*, 2008). As it turns out, it is unnecessary to test each variant separately for association with disease, because common variants are correlated with one another. The GWAS method exploits this correlation structure in the genome to test a subset of common variants that serve as proxies for their neighbours. This aspect of the method means that the associated variants are unlikely to be causal themselves, but are correlated with the causal variants. This thesis explores what we can learn about the underlying causal variants on the basis of GWAS findings.

This chapter will focus on the GWAS method in more detail. First we review some results that were important to the development of GWAS. We then discuss methodological considerations, including ways in which allele frequency differences can arise

between cases and controls that are not due to genetic variation contributing to disease. A summary of GWAS findings is presented. Finally, we define and describe statistical methods germane to the GWAS method and to the later chapters of this thesis.

1.1 Linkage disequilibrium

An extension of the idea of linkage is the concept of linkage disequilibrium, the correlation of alleles at nearby loci in the genome. In the same way that two loci are "linked" if they are inherited together more often than expected, two loci are in "linkage disequilibrium" (LD) if particular alleles at these loci appear together on the same chromosome in the population more often than expected under independence.

We will use an example drawn from Donnelly (2006) to illustrate how such correlation might emerge. Figure 1.1 illustrates the following scenario. Suppose a mutation arises at locus L_1 back in time that changes an ancestral A to a T allele. Then suppose, sometime later, at a locus L_2 on a chromosome with an A allele, a novel mutation arises, changing a C allele to a G allele. This mutation then spreads through the population (otherwise we would not observe it today). Unless there is a recombination event between L_1 and L_2 on one of these chromosomes, a G allele at L_2 will be predictive of an A allele at L_1 since the G mutation originally occurred on a chromosome with an A allele. So long as the rate of mutation, the rate of recombination per meiosis, and the number of generations between individuals and their most recent common ancestor are all low, the ancestral haplotype background will be inherited in the way just described, and nearby SNPs will tend to be inherited together (Manolio *et al.*, 2008). Recombination is neither randomly nor uniformly distributed across the genome but occurs especially often at small regions termed "recombination hotspots" (Myers *et al.*, 2005). As a consequence, patterns of LD in the genome are complicated.

A single nucleotide polymorphism (SNP) is a site in the genome at which individuals

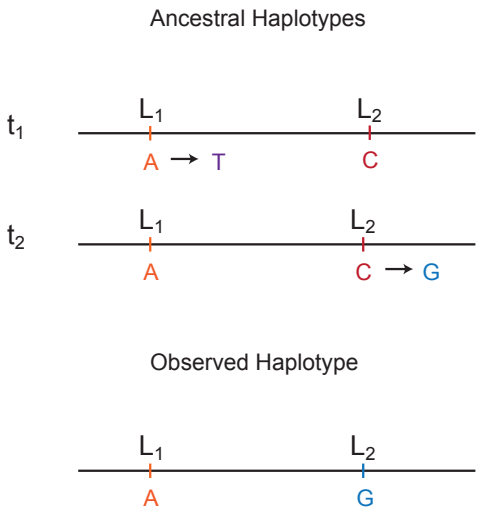


Figure 1.1: Linkage disequilibrium example - The figure shows two mutations at loci L_1 and L_2 on ancestral haplotypes, at times t_1 and t_2. In the absence of a recombination event, a G allele on the observed haplotype at L_2 is predictive of an A allele at L_1.

in a population differ by a single nucleotide base. SNPs are defined as "common" if the frequencies of each of the alleles exceed 1% in the population, and most are biallelic, meaning that there are two observed variants, or alleles, at the site (Manolio *et al.*, 2008). Accordingly, for the purposes of this thesis, we will assume that all SNPs are biallelic. Strong correlations between nearby SNPs mean that commercially available genotyping chips, which assay 300,000 - 1,000,000 SNPs, can capture much of the common variation in the human genome, particularly in Caucasian populations (International HapMap Consortium, 2007). The HapMap project catalogued these correlations systematically (International HapMap Consortium, 2005).

In its first phase, the HapMap Project aimed to genotype at least one common SNP in every 5 kilobases (kb) in the genome (International HapMap Consortium, 2005). In addition, a representative collection of ten regions, each approximately 500kb in length, was examined as part of the ENCODE Project (ENCODE Project, 2004). These

regions were resequenced in 48 individuals in order to get a fuller picture of common variation in these regions, and the chromosomes were subsequently genotyped at all newly discovered SNPs as well as those already in the dbSNP database. The ENCODE regions confirmed a view of the genome in which nearby common variants are highly correlated, separated by recombination hotspots (International HapMap Consortium, 2005). Figure 1.2 shows a graphical depiction of LD in two of the ten ENCODE regions, which illustrates the concept of highly correlated 'blocks' of LD broken by recombination hotspots. Quite a lot of data exists on empirical patterns of LD in the human genome; for a review of this literature see, for example, Pritchard & Przeworski (2001), as well as International HapMap Consortium (2005) and Myers *et al.* (2005).

Here we introduce two commonly used measures of the LD between two nearby SNPs: r^2 and D'. Let us return to our previous example of loci L_1 and L_2. Unless there is a recombination event between L_1 and L_2 then the G allele at L_2 is predictive of an A allele at L_1. Now suppose that a recombination event does occur. Then the prediction will no longer be perfect. The r^2 measure lets us quantify the precision of the prediction. Considering haploid chromosomes separately, if we think of SNPs as associated to random variables (taking the values 0 if the minor allele is not present and 1 if it is) at two sites in the genome, then r^2 is the squared correlation coefficient between these random variables (Chen *et al.*, 2006). Suppose two SNPs A and B have minor alleles a and b at frequencies p and q. Then r^2 can be calculated as,

$$r^2(A,B) = \frac{(\Pr(a,b) - pq)^2}{pq(1-p)(1-q)}\,,$$

where $\Pr(a,b)$ is the probability of alleles a and b appearing together on the same chromosome in the population. Note that the earlier choice of the minor allele was arbitrary.

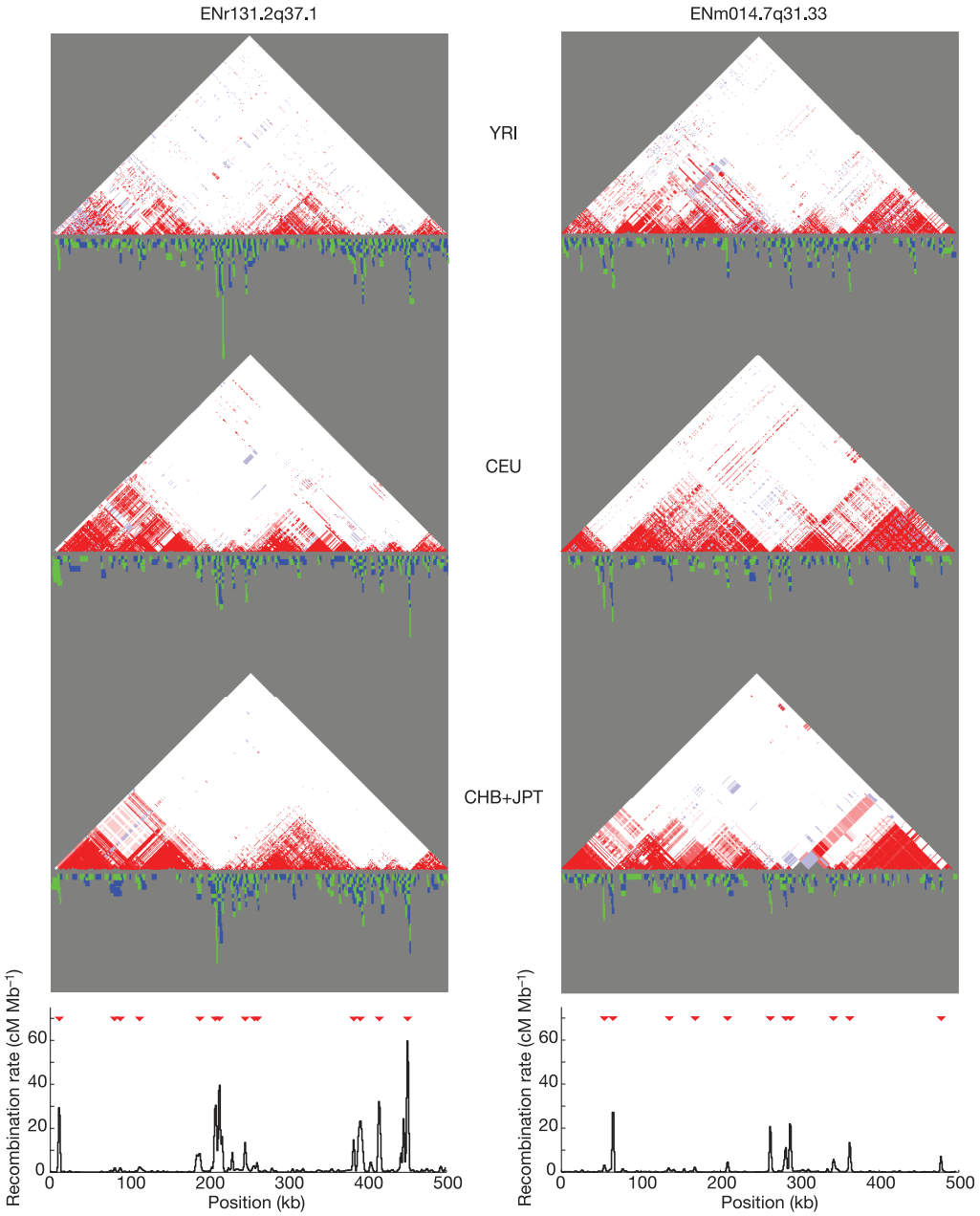


Figure 1.2: Linkage disequilibrium in two ENCODE regions - This figure, from International HapMap Consortium (2005), shows a comparison of LD (measured by D') and recombination. Red indicates no evidence of recombination. Below each of the plots, blue and green lines indicate intervals where recombination events must have occurred. Finally, the bottom plots show estimated recombination rates, with hotspots depicted as red triangles.

The D' statistic, defined in Lewontin (1964), is given by,

$$D'(A,B) = \begin{cases} \frac{(\Pr(a,b)-pq)}{\min((1-p)(1-q),pq)}, & \text{if } \Pr(a,b) - pq < 0\,, \\ \frac{(\Pr(a,b)-pq)}{\min((1-p)q,p(1-q))}, & \text{if } \Pr(a,b) - pq \geq 0. \end{cases}$$

As discussed in Pritchard & Przeworski (2001), D' and r^2 behave very differently. As an example of how the two measures can differ, let's return to our earlier example and suppose again that there is no recombination between loci L_1 and L_2, so that they share the same genealogical tree. Whereas the r^2 between L_1 and L_2 will be 1 if and only if the mutations $A \to T$ and $C \to G$ occur on the same branch of the tree, D' will still be 1 even if the mutations occur on different branches. We choose to describe our results with the r^2 statistic because it arises naturally in the association study context when comparing the information captured at the marker SNP to the information at the causal SNP (Pritchard & Przeworski, 2001).

1.2 Methodological considerations in GWAS

Genome-wide association studies (GWAS) seek to identify loci harbouring genetic variants that affect disease susceptibility by comparing allele frequencies in healthy and sick individuals at common genetic polymorphisms. Since GWAS test differences in allele frequencies in case and control populations, any systematic differences between these samples, aside from disease status, have the potential to create artificial association signals. Here we review some of the sources of artefacts and point to references for methods developed to circumvent them.

1.2.1 Avoiding artefacts

Some of the potential sources of false positives due to systematic differences include population stratification and differences in genotyping between cases and controls (Hirschhorn

& Daly, 2005). Population stratification can create false association signals, either because of the manner in which controls are sampled, or if one ancestry group carries higher disease risk and is therefore over-represented in cases (Marchini *et al.*, 2004; Wellcome Trust Case Control Consortium, 2007).

Several methods, including the model-based clustering program STRUCTURE (Pritchard *et al.*, 2000) and the use of principal components, were developed to detect population stratification on the basis of markers spread throughout the genome. Suggestions have been put forward for how to correct for structure, when it exists, in the GWAS context (see, for example, Devlin & Roeder (1999), which describes the method of genomic control, and Price *et al.* (2006), which proposes a correction based on principal components analysis). In practice, population structure has not been a major problem in many GWAS studies (Wellcome Trust Case Control Consortium, 2007).

Finally, differences in genotyping technologies, nonrandom genotyping failure, or differential bias in genotyping scoring, as reported in Clayton *et al.* (2005), can create false positive associations. Ideally, cases and controls would be typed at the same facility with the same technology on the same plates to minimize the potential for this bias. Checking cluster plots visually is another way to guard against these effects (Wellcome Trust Case Control Consortium, 2007).

Other challenges to the GWAS method, which will not be detailed here, include phenotypic heterogeneity, and differing environmental backgrounds (McCarthy *et al.*, 2008; Newton-Cheh & Hirschhorn, 2005). Due to the large number of sites tested in a GWAS, the potential for false positive results must be taken into account. Stringent p-values, together with large numbers of cases and controls and replication studies, reduce the probability that an identified locus is a false positive.

1.3 Variants discovered by GWAS have small effects

Most loci discovered by GWAS have small effects (Hindorff *et al.*, 2010; Manolio *et al.*, 2008). Figure 1.3 shows the distribution of effect sizes from the NHGRI GWAS Catalogue. The dip in the histogram near to 1 is expected to be a result of the limitation in power to detect even smaller effects due to current sample sizes (Gibson, 2010). The effect size distribution of loci discovered in GWAS contrasts with that of loci discovered in linkage studies, and Mendelian disease genes more generally.

As an example of the contrast between Mendelian disease gene effects and those discovered by GWAS, consider the frequencies and effect sizes of rare mutations known to confer susceptibility to breast cancer with GWAS variants. The mutations in the *BRCA1* and *BRCA2* genes that predispose to a Mendelian breast cancer subtype have frequencies of less than 1% and those who have these mutations are approximately 8 times more likely than someone without the mutation to develop the disease in their lifetimes (Easton, 1999). GWAS have discovered 18 common loci with mean frequency 31% in the population with a mean increase in risk of 1.13-fold over the population prevalence of 1 in 9 (Turnbull *et al.*, 2010). In contrast to the setting of breast cancer, in which the *BRCA* genes contribute significantly to the explained familial inheritance, known common variants contribute more than known rare variants to the genetic variance of hypertriglyceridemia (Johansen *et al.*, 2010).

What to make of these results has been the subject of intense debate in the past few years, with a wide range of views. At one end of this spectrum, there is great scepticism about the role of common variants in disease risk (Goldstein, 2009; McClellan & King, 2010). At the other end of the spectrum, the number and variety of new biological clues for disease mechanism is celebrated (Hirschhorn, 2009; Manolio *et al.*, 2008). While this debate is beyond the scope of this thesis, we note that it provides impetus for trying to understand the true effect sizes of those variants causing disease association signals.

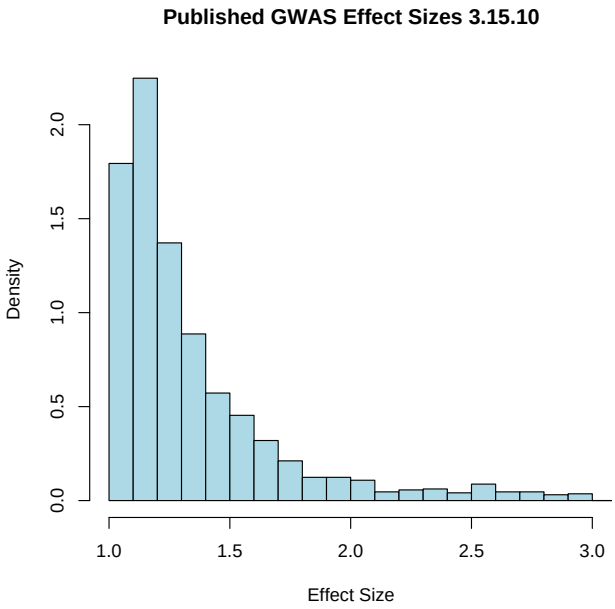


Figure 1.3: Effects discovered by GWAS - The distribution of effect sizes from GWAS studies published prior to March 2010. Effect sizes were culled from the NHGRI GWAS Catalog, available at `www.genome.gov/gwastudies` (see Hindorff *et al.* (2010) for a discussion of the catalog).

1.4 Statistical methods

We now briefly review some of the statistical methods used in GWAS; for a fuller discussion of these and other methods, see Balding (2006). This is by no means a comprehensive overview, since it covers only those methods which we employ later to mimic a standard GWAS design. For further information on Bayesian methods for association testing, see Stephens & Balding (2009). In addition to the tests we will discuss below, Purcell *et al.* (2007) discuss identity-by-state and identity-by-descent information and how to exploit it in association analysis. Association testing in concert with haplotype phasing has also been performed in the Icelandic population (Kong *et al.*, 2009), and many other methods for detecting association have been proposed. We review those methods that have been widely adopted, in addition to introducing a few others that we will use throughout the thesis.

1.4.1 Disease risk models

A disease risk model relates the disease risk to the number of copies of a putative disease-causing allele. Let p be the probability of developing the disease and let G be genotype, expressed in terms of the number of copies of the allele in question. One common parameterisation of the disease risk model is then,

$$\log(\frac{p}{1-p}) = \mu + \beta G\,, \tag{1.1}$$

which writes the log-odds of disease as a linear function of the number of allele copies. We will refer to this parameterisation as the *additive* model. In a case-control study, μ primarily reflects the proportion of cases in the sample. Thus, in most contexts, μ is a nuisance parameter and β is the parameter of interest, since under this model, e^{β} is a measure of the effect size and we test against the null hypothesis $\beta = 0$. Although the model assumes prospective sampling of individuals, while a GWAS samples

retrospectively, there is theory showing that assuming prospective sampling does not make a difference when inferring β so long as the regression coefficients in the disease risk model are small enough to approximate odds ratios by relative risks (see §1.4.4) (Prentice & Pyke, 1979).

More complex disease models can be considered. At the expense of an extra degree of freedom, we can introduce another parameter, γ, into the model,

$$\log(\frac{p}{1-p}) = \mu + \beta G + \gamma\chi_1(G)\,, \tag{1.2}$$

and $\chi_1(G)$ is an indicator function that takes the value 1 for heterozygotes (i.e. $G = 1$) and 0 for homozygotes (i.e. $G = 0$ or $G = 2$). We will refer to this parameterisation as the *general* model. In place of true controls, in GWAS it is standard to use (unphenotyped) cohort or population samples in place of true controls, but analyse the results as a typical case-control study using logistic regression. Schouten *et al.* (1993) show that this is actually equivalent to fitting a log risk regression model, which we define and utilize in Chapter 6.

Figure 1.4 helps guide an intuitive understanding of the parameterisation at 1.2, which is similar to that of Balding (2006). The γ parameter measures the deviation from the simplest additive model, represented as linear on the log risk scale of the Figure. We assume that $0 < \gamma < \beta$, which restricts the risk model space considered. (For example, it does not cover the case $\gamma > 0$ and $\beta = 0$, in which the heterozygote but not the risk allele homozygote confers risk.) As shown in the figure, a recessive model, in which $G = 2$ is the only risk-conferring allele, can be achieved by setting $\gamma = -\beta$, and a dominant model, in which $G = 1$ and $G = 2$ have the same increased risk, is satisfied when $\gamma = \beta$. These disease models will be explored further in Chapter 6.

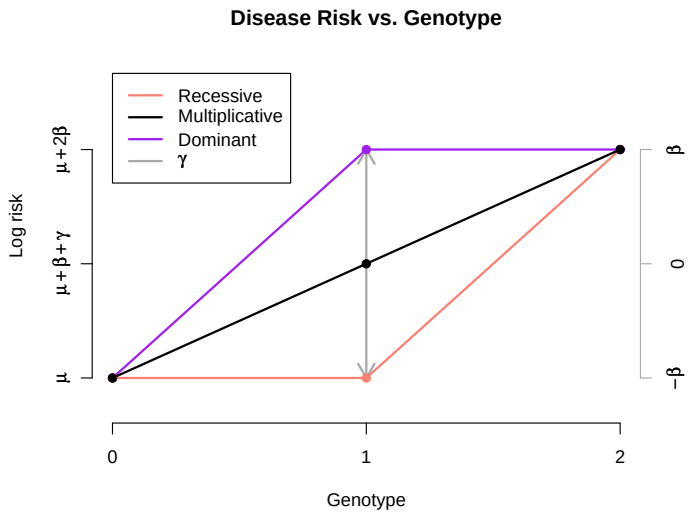


Figure 1.4: Representation of three models in the log risk regression - The grey arrows show how the γ parameter acts to perturb the model at the heterozygote away from its value under a additive model.

1.4.2 The trend test

As reviewed in Balding (2006), there are many possible tests of association, including the Pearson test with 2 degrees of freedom, the Fisher exact test, and the Cochran-Armitage test, also referred to as the trend test. There is no "best" test, but we adopt the trend test because it makes no assumption about the distribution of allele frequencies in the population and has favourable properties in terms of power under an additive disease risk model (Balding, 2006). For a systematic comparison of the different possible testing approaches, see Sasieni (1997).

The Cochran Armitage test statistic (Armitage, 1955) is given by,

$$\frac{N}{RS}\frac{(S(r_1+2r_2)-R(s_1+2s_2))^2}{N(n_1+4n_2)-(n_1+2n_2)^2},$$

for $i \in (0,1,2)$, where the r_i are counts of the three genotypes in controls, s_i are counts of the three genotypes in cases, $n_i = r_i + s_i$, and $N = \sum_i n_i$, $R = \sum_i r_i$, and $S = \sum_i s_i$. Under Hardy-Weinberg equilibrium, this is asymptotically equivalent to a score test of

the null hypothesis ($\beta = 0$) under the disease risk model given by 1.1 (Sasieni, 1997) and has a χ^2_1 distribution under the null. All the GWAS we consider in this thesis have large sample sizes, so we use the asymptotic result as though it was exact.

Since many hundreds of thousands of SNPs are tested for association in a GWAS, the question arises (in the frequentist setting) what p-value threshold to use to assign "genome-wide significance". In particular, an investigator wishes to reduce the number of false positive associations that might arise by chance. The more stringent the p-value threshold, the more cases and controls necessary to achieve power; the less stringent the threshold, the more likely that false associations will be reported. Because the follow-up studies to GWAS, including fine mapping and biological experiments assessing the relevance of nearby genes, tend to be costly, there is a strong incentive to limit false positive associations. The issue is further complicated by the fact that each of the single-SNP tests are not independent because some SNPs will be in LD.

A conservative solution is to employ a Bonferroni correction, which for a significance level α, requires a p-value threshold of α/n where n is the number of tests performed in the association study. For a SNP chip with 1 million SNPs and the standard $\alpha = 0.05$, this gives the value 5×10^{-8} adopted in many studies. Balding (2006) discusses the Bonferroni correction and compares it with some Bayesian approaches that he endorses; again, since we seek to model the results of current association studies, we confine ourselves to the more common frequentist approach. As discussed further in Chapter 2, we use a higher p-value threshold of 10^{-6} and subject each association to further replication to gain a more accurate approximation of the ascertainment implicit in reported GWAS associations.

1.4.3 The general test

We can test the general model given in Equation 1.2 against the null hypothesis that the β and γ parameters are both 0, with score test statistic given by,

$$\frac{N}{RSn_0n_1n_2}(n_0(r_1s_2 - r_2s_1)^2 + n_1(r_0s_2 - r_2s_0)^2 + n_2(r_0s_1 - r_1s_0)^2)\,,$$

and a χ^2_2 distribution under the null. This result is derived in Marchini *et al.* (2007) and, with the notation above, in Vukcevic (2009). Vukcevic (2009) includes a discussion of the properties of this test as compared to the Cochran-Armitage test. This test is less frequently used in GWAS than the trend test, but it will be useful in Chapter 6 for computing power to infer non-additive disease models.

1.4.4 Estimates of effect sizes

We will first discuss effect sizes in the setting of the additive model, taking up the subject of effect sizes in the more general disease model setting in Chapter 6. There are two main measures of effect size: the odds ratio and the relative risk. In the additive model, the odds ratio of allele A is defined as,

$$\begin{aligned}\mathrm{OR}_A &= \frac{\mathrm{Odds}(\mathrm{Disease}|G = Aa)}{\mathrm{Odds}(\mathrm{Disease}|G = aa)} = \frac{\mathrm{Odds}(\mathrm{Disease}|G = AA)}{\mathrm{Odds}(\mathrm{Disease}|G = Aa)} \\ &= \frac{\Pr(\mathrm{Disease}|G = Aa)\Pr(\mathrm{Healthy}|G = aa)}{\Pr(\mathrm{Disease}|G = aa)\Pr(\mathrm{Healthy}|G = aA))}\,,\end{aligned}$$

which is equivalent to e^β in the model defined by Equation 1.1. Note that we could have also compared the odds of disease at genotypes AA and Aa, since these will be equal to the above comparison under the additive disease risk model. The relative risk is defined as,

$$\mathrm{RR}_A = \frac{\Pr(\mathrm{Disease}|G = Aa)}{\Pr(\mathrm{Disease}|G = aa)}\,.$$

Each measure has its merits. The odds ratio (OR), while less intuitive, is directly tied to the logistic regression model and can be directly estimated from case control data. The relative risk (RR), on the other hand, is easier to interpret but requires knowledge of the prevalence of the disease in a case control study. Specifying an additive risk model for one measure does not give an additive risk model for the other measure. When the disease is rare, as is the case for many of the diseases considered here, OR is approximately equal to RR. In many GWAS studies, controls are selected not on the basis of health, but rather are drawn from the general population. These "population controls" are not phenotyped, so there is some probability of a control actually being a case. In this setting, using the logistic regression model estimates the RR and not the OR (Schouten *et al.*, 1993). In our simulations, we use population controls, and thus estimate the RR with the logistic regression model.

1.4.5 The winner's curse

Effect sizes estimated from a discovery sample are subject to upward bias, often called the "winner's curse" (Kraft, 2008). The term was first introduced in the setting of auctions, in which the value of a commodity is likely to be overestimated by analyzing winning bids (Capen *et al.*, 1971). The winner's curse arises as a result of imposing a significance threshold on observed data. Conditional on a result passing the significance threshold, it is more likely to have been inflated by chance. Thus if we estimate model parameters on the basis of significant data we are likely to overestimate as a result of this bias.

The corollary in association studies is that those SNPs selected as top hits are likely to have higher test statistics and lower p values on average, conditional on the fact that they generated the strongest association signal. One negative aspect of the winner's curse for GWAS is that it makes it difficult to design sufficiently-powered replication studies (Zollner & Pritchard, 2007). However, if the power of the study is high, ei-

ther because of the large number of cases and controls sampled, or because of high penetrance, this effect is small. Reviews of early association studies (prior to GWAS) suggested that the winner's curse caused consistent overestimation of discovered genetic effects (Lohmueller *et al.*, 2003). As a result, current GWAS now include a replication step rather than leaving this to future studies, and SNPs are not declared associated unless they show strong evidence for association in the original GWAS and compelling evidence in replication. There are no standard thresholds used, but, for example, many studies would adopt a p value threshold of $< 10^{-6}$ in discovery, < 0.05 in replication, and a combined p value of less than 10^{-8}.

A number of methods have been suggested to correct for the winner's curse (Xiao & Boehnke, 2011; Zhong & Prentice, 2010; Zollner & Pritchard, 2007). However, estimating odds ratios from the data in a replication study (disregarding the data in which the association was initially discovered) removes the effect of the winner's curse entirely. In order to avoid the impact of the winner's curse on our results, we chose to estimate effect sizes from the replication sample.

1.4.6 Priors on effect size

In Chapter 3, we will use a Bayesian approach to infer the true distribution of effect sizes arising from GWAS. To do this, we will assume a joint distribution of the true risk allele frequencies and effect sizes, which is of course unknown. A number of theoretical analyses (Pritchard, 2001; Wakefield, 2008; Wang *et al.*, 2005; Zondervan & Cardon, 2004) have argued for a relationship between effect size, disease model, and MAF. As there is no consensus on the exact form and extent of the relationship, we do not rely on them explicitly here. Instead we define two distributions which represent two different sets of assumptions about these unknowns, with the subsequent results representing a range of possibilities. Our first set of assumptions posits that the distribution of effect sizes is the same for all putative causal variants, regardless of their allele frequency,

and that effect sizes are close to those observed in GWAS studies. We call this the *conservative* prior. The second set of assumptions explicitly assumes that there might be larger effects at variants with smaller minor allele frequency (MAF). We refer to the second distribution as the *MAF-dependent* prior.[1]

We do not explicitly model the potential impact of selection on the effect size distribution. However, because the effect size and selective coefficient are likely to be related, classical population genetics provides insight into the motivation for our two priors. Sawyer & Hartl (1992) give a formula expressing the expected derived allele frequency distribution of variants as a function of their selective coefficient (defined as the proportional decrease in offspring attributable to each copy of the allele). For a population at equilibrium, the expected distribution of alleles with selective coefficient s can be expressed,

$$h(f,s) \propto \frac{e^{-4Ns(1-f)}-1}{f(1-f)(e^{-4Ns}-1)},$$

where f is the derived allele frequency and N is the long-term effective population size. To understand this formula, it is helpful to consider a few examples. If a common disease has population prevalence μ and causes a proportional reduction γ in fitness, an allele that increases disease risk by a factor $1+\delta$ will have selection coefficient $s = \mu\gamma\delta$. For instance, consider a disease with prevalence $\mu = 0.05$ that decreases fitness by $\gamma = 0.01$. Alleles that increase risk by less than 40% would have $s \leq 2 \times 10^{-4}$ and would be dominated by common variants.

If effect size and selection coefficient are indeed related in this way, the MAF-dependent prior represents a set of assumptions consistent with the presence of selection. The conservative prior supposes that no selection is acting, and simply imposes an effect size distribution consistent with current observations. Since we have little

[1] Although both priors are distributed around $\beta = 0$, we note that for the purposes of simulation, alleles with $\beta < 0$ were modelled by the other allele at the SNP, for which β is positive. As an example, if allele a has $e^{\beta} = 0.8$, then allele A has $e^{\beta} = 1.25$, and we chose to simulate a as protective by simulating A as risk-conferring. For this reason we will show distributions for relative risks greater than 1.

information about the true value of the selective coefficients for most common diseases, we chose not to define our priors in terms of selection coefficients.

1.4.6.1 Conservative prior

More precisely, the conservative prior posits that if e^{β} is the effect size at a causal variant, then β is normally distributed with mean 0 and standard deviation 0.2, independent of risk allele frequency. Its name derives from the observation that it places little weight on relative risks greater than 1.5 (see Figure 1.5). This distribution assumes that 81% of true effect sizes are less than 1.3 with 96% less than 1.5 and 99.9% less than 2. For a justification of this choice of standard deviation, see Wellcome Trust Case Control Consortium (2007).

1.4.6.2 MAF-dependent prior

The MAF-dependent prior again assumes a normal distribution for β with mean 0, but here the standard deviation, σ, is allowed to depend on the risk allele frequency in the following way:

$$\log(\sigma) = \log(0.2) + (2 - 8f(1-f))\,.$$

This equation is suggested in Spencer *et al.* (2011a) and is similar to the prior in Wakefield (2008), to which we refer the reader for further discussion. For f near 0.5 (a common SNP) this prior is approximately the same as the conservative prior. However, as f approaches 0 or 1 (corresponding to SNPs with rarer alleles), then the prior puts considerably more weight on larger ORs. For example, when the MAF is 5%, the MAF-dependent prior gives an approximately 5% chance that the risk associated with each copy of the causal allele is larger than 2.5. Figure 1.5 shows the two prior distributions side by side.

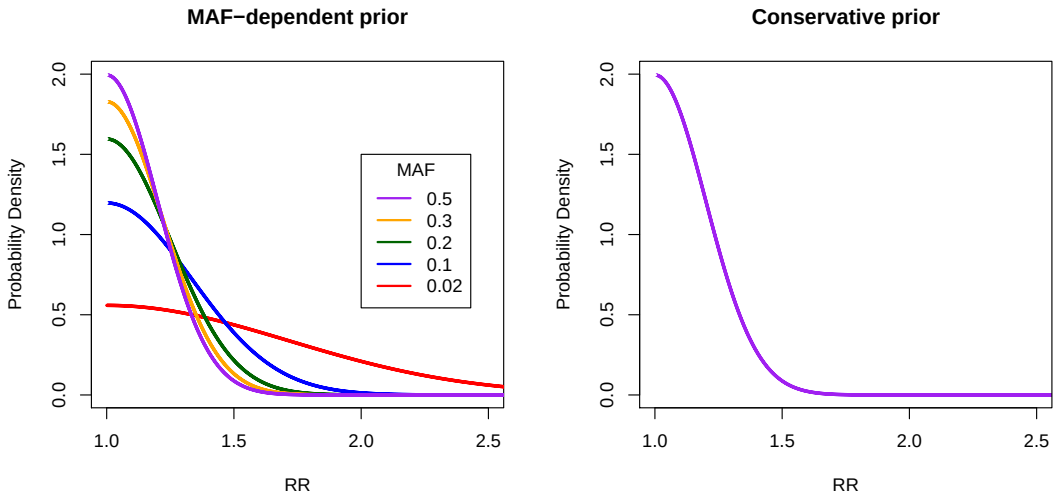


Figure 1.5: Priors on the effect size - Probability density curves showing the relationship between minor allele frequency and effect size for the conservative and MAF-dependent priors. Each curve in the MAF-dependent prior represents the probability density for one minor allele frequency. Note that as the minor allele frequency decreases, greater probability is placed on higher relative risks. The conservative prior is independent of causal allele frequency, and corresponds to the MAF-dependent prior for a causal allele frequency of 50% (shown in purple).

1.4.7 Imputation

Genotype imputation refers to a method, widely applied in GWAS, which predicts genotypes at untyped loci. We will briefly review the main ideas behind genotype imputation in samples of unrelated individuals, but for a full description of the method, as well as a comparison between existing imputation softwares and examples of GWAS that have utilized imputation, see the reviews by Li *et al.* (2009) and Marchini & Howie (2010). The performance of imputation using HapMap II reference panels is discussed in Huang *et al.* (2009).

A number of different software packages exist to perform imputation (see Table 1.1). All suggest different recipes applied to a common set of ingredients: (i) observed genotype data, often from a genotyping chip, in a sample of individuals, and (ii) a panel of reference haplotypes, often drawn from HapMap, and more recently from the 1000 Genomes project (1000 Genomes Project Consortium, 2010). The first step in the recipe is to separate out the genotypes into maternal and paternal haplotypes (called "phasing"). The next step is to compare these separated haplotypes to those in the reference set, which includes many more SNPs than are represented on the genotyping chip. Locally, the separated haplotypes will look like particular reference haplotypes, since they share a common ancestor. These stretches of identity may be small, depending on the recombination landscape and the number of generations since convergence in the unobserved genealogical tree. Imputation algorithms model the separated haplotypes as a mosaic of the reference haplotypes, and fill in genotypes at untyped loci on the basis of the reference haplotypes. Although these two steps are common to the different methods, each method uses a different underlying model to infer haplotype phase and relate the two sets of haplotypes to perform imputation. Most imputation methods will give a sense of the confidence of the imputation. For many SNPs, imputation is remarkably accurate ($> 95\%$, see Marchini & Howie (2010)).

Imputation has a variety of applications in GWAS, including combining studies

Table 1.1: Imputation Methods

Software	Reference
IMPUTE v1	Marchini *et al.* (2007)
IMPUTE v2	Howie *et al.* (2009)
fastPHASE	Scheet & Stephens (2006)
MACH	Scott *et al.* (2007)
BEAGLE	Browning & Browning (2007)

conducted with different genotyping chips in meta-analysis, fine-mapping, imputation of untyped variation (as we do in Chapter 4), imputation of non-SNP variation, and missing data imputation and correction of genotyping errors (reviewed for example in Marchini & Howie (2010)).

1.4.8 Estimating heritability

Ultimately we would like to be able to explain all of the genetic contribution to the correlation of trait variation among relatives. Knowing, for a given disease, how close we have come to accounting for this correlation is useful both in predictive terms, in order to implement the vision of "personalized medicine" as described in Hamburg & Collins (2010), and also as a benchmark against which to measure how much genetic contribution has been explained. Much attention has recently focused on evaluating GWAS in light of the heritability explained by its variants (Eichler *et al.*, 2010; Maher, 2008; Manolio *et al.*, 2009; Slatkin, 2009; Yang *et al.*, 2010a). Our focus is on understanding how different assumptions about underlying variants affect estimates of heritability.

The most commonly used measure of heritability is λ_s, the sibling recurrence risk ratio, the properties of which are discussed at length in Risch (1990). Let $\Pr(Y_i = 1)$ be the probability that individual i is affected and write the prevalence of the disease as $\Pr(Y = 1)$. The sibling recurrence risk ratio λ_s expresses how much more likely

individual j is to have the disease, conditional on j's sibling k being affected, than a member of the population at large (Thomas, 2004):

$$\lambda_s = \Pr(Y_j = 1 | Y_k = 1) / \Pr(Y = 1) \,.$$

We first show how to compute λ_s under an additive disease model at a single locus as a function of risk allele frequency and relative risk. The sibling recurrence risk λ_s calculated at a locus can be thought of as the value of λ_s if that locus were the only genetic factor in developing the disease. If a disease in fact has n loci contributing to risk, it turns out that λ_s for the disease is given by a product of the $\lambda_{s_1} ... \lambda_{s_n}$ at the individual loci, assuming that genetic effects of the loci are independent and there are no interactions between the loci (Risch, 1990).

Let G_i denote the genotype of the father (f), mother (m), child (j), and affected sibling (k) respectively, and let Q_g denote the genotype probabilities under Hardy-Weinberg equilibrium, which we assume to hold at the loci in question. Let q be the frequency of the A allele and $p = 1 - q$ the frequency of the a allele, so $Q_{AA} = q^2$, $Q_{aA} = 2pq$, and $Q_{aa} = p^2$. Finally, we need transition probabilities $T_{g_j g_m g_f} = \Pr(G_j = g_j | G_m = g_m, G_f = g_f)$. We can compute the sibling recurrence risk as,

$$\Pr(Y_j = 1 | Y_k = 1) / \Pr(Y_j = 1) = \frac{\Pr(Y_j = 1, Y_k = 1)}{\Pr(Y_k = 1) \Pr(Y_j = 1)} \,.$$

Assuming that the two sibs' phenotypes are conditionally independent given their genotypes, this expression can be computed (see Thomas (2004) pp 106-7) as,

$$\frac{\sum_{g_m} \sum_{g_f} \sum_{g_j} \sum_{g_k} RR_{g_j} RR_{g_k} T_{g_j g_m g_f} T_{g_k g_m g_f} Q_{g_m} Q_{g_f}}{(\sum_{g_j} RR_{g_j} Q_{g_j})^2} \tag{1.3}$$

To give a sense of the magnitude of the sibling recurrence risk that can be attained by a single locus, consider the contribution of the TCF7L2 variant to risk of type 2

diabetes. This locus has a risk allele frequency of 40%, a relative risk of 1.37, and contributes $\lambda_s = 1.025$ to the overall sibling recurrence risk of 3. In order to account for the observed risk to siblings of those with type 2 diabetes, 44 such loci would be needed.

As noted in Clayton (2009), the empirical value of λ_s is likely to be overestimated by family studies, partly because of shared environment, and partly because of ascertainment effects. Wallace & Clayton (2003) discuss the issue of shared environment further in the context of leprosy, another disease in which both the environment and genetics are expected to play a role. Thus, the "benchmark" set by sibling studies may be lower than sometimes assumed.

1.5 Summary

In this chapter we introduced the motivation behind GWAS and some particulars of the method. The outcome of a GWAS is shown in Figure 1.6, which contains signal plots from the Crohn's disease analysis in Wellcome Trust Case Control Consortium (WTCCC) study. Each plot illustrates an association between a region of the genome and risk of Crohn's disease.

A comparison between the two associations in Figure 1.6 reveals the importance of LD in the outcome of a GWAS. The hit region on chromosome 2 (left panel, Figure 1.6) is relatively narrow, and contains only one gene, ATG16L1, which has subsequently been shown to play a role in Crohn's disease via the autophagy pathway (Cadwell *et al.*, 2008). In contrast, the hit region on chromosome 3 contains 18 genes and displays little fine-scale recombination (right panel, Figure 1.6). The reported relative risks for the two associations were 1.19 and 1.09, respectively (Wellcome Trust Case Control Consortium, 2007). In each case, supposing a causal SNP exists in the regions shown, one can visually see how LD causes this signal to be carried to, and sometimes diluted

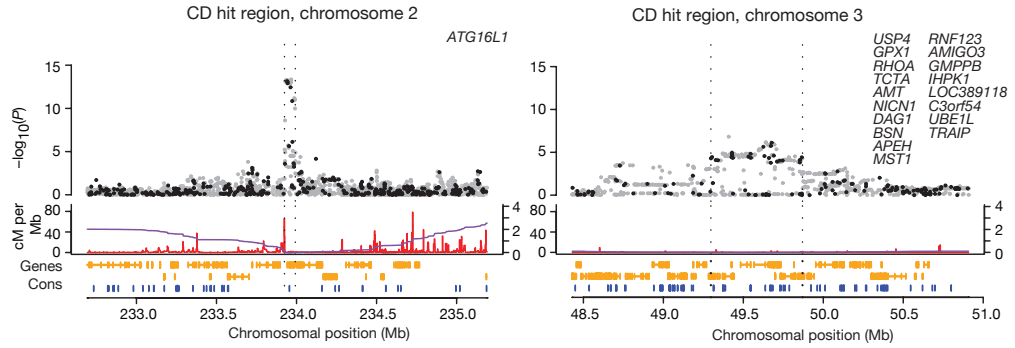


Figure 1.6: Examples of associations from the WTCCC - Figure taken from Wellcome Trust Case Control Consortium (2007). Each plot shows the signal in genomic regions 1.25 Mb to either side of the best-associated SNPs (called 'hit' SNPs) in the study. Vertical dotted lines indicate locations where test statistics returned to background levels. The upper panels show $-\log_{10}$ p values for the trend test at each SNP. Black points represent SNPs typed in the study, and grey points represent imputed SNPs. Middle panels show the fine-scale recombination rate in cM/Mb estimated from Phase II HapMap. The purple line shows the cumulative genetic distance in cM from the hit SNP. The lower panels show known genes and sequence conservation in 17 vertebrates. Known genes (in orange) in the hit region are listed in the upper right of each plot in chromosomal order.

at, other nearby SNPs. One might ask questions such as: what is the likely effect size at the causal variant? Might the hit SNP in fact *be* the causal SNP? What is the disease model at the causal SNP, and what disease model is observed at the hit SNP?

In this thesis, we examine the effects of imperfect LD between the causal SNP and the hit SNP on the outcomes of an association study. We design and implement a simulation framework described in Chapter 2 that allows us to ask, for a given causal variant, what hit SNP will arise in a simulated GWAS. In Chapter 3, we use this simulation framework, together with the priors introduced earlier, to estimate the effect sizes of causal variants from the observed hit SNPs arising from GWAS, and discuss the implications of the results both for individual risk prediction and for explaining the heritability of common diseases. Chapter 4 modifies the simulation framework to examine a variety of GWAS contexts. We show that, in three different populations, four different genotyping chips, and with and without imputation, LD between the causal SNP and the hit SNP drives effect size estimates. In Chapter 5, we ask how often the hit and causal SNPs are expected to coincide. Finally, in Chapter 6, we show how departures from the simplest disease risk model at the causal variant are distorted by LD, and to what extent such departures can be detected at the hit SNP.

2

A GWAS simulation framework

A central aim of this thesis is to understand properties of the causal variants underlying GWAS signals. This chapter develops a method by which we can address the forward problem: given a causal variant with a specified effect size, what signal would it generate in a typical GWAS? In the next chapter we will examine the results of this method applied to thousands of hypothetical causal variants, and go on to pose the inverse problem: given an observed GWAS signal, what are the likely properties of the causal variant?

GWAS exploit the correlation in genetic diversity along chromosomes in order to detect effects on disease risk without having to type the causal locus directly. As others have noted (Iles, 2008; Wang *et al.*, 2010), this means that most risk variants identified by GWAS will be tags for as-yet-unknown causal variants. Using simulations, where we know the true risk model at the causal locus, we can ask a number of questions of interest:

1. How often is the estimated risk associated with carrying the predisposing allele at the tag SNP smaller than the risk associated with carrying the predisposing allele at the causal SNP?

2. How does varying the SNP chip or the study population impact the risk estimates at the tag SNP?

3. How often do the causal and tag SNPs coincide?

4. How does the correlation between the tag SNP and the causal SNP affect the inferred disease model?

These questions will be addressed in later chapters. In this chapter, we describe a flexible simulation framework in which we can mimic a GWAS by simulation. By flexible, we mean that we can easily modify the genotyping chip used in the study, the population from which we draw cases and controls, the sample sizes, and the disease risk model at the causal SNP. For now, we consider an additive disease risk model. In Chapter 6 we will investigate further disease models.

Patterns of LD in human populations are complicated, and preclude analytical results, so we adopted a simulation approach. We describe this approach informally before going into detail. First, we chose each allele at each SNP in the HapMap ENCODE regions in turn, assuming it to be causative with a given effect size. We then used a previously reported simulation scheme (HAPGEN, (Marchini *et al.*, 2007; Spencer *et al.*, 2009)) to simulate a large population of chromosomes, whose patterns of LD match those in the HapMap analysis panel. A case-control sample was drawn from this population. The controls were sampled randomly from the population and the cases were chosen by oversampling chromosomes carrying the causal allele in the appropriate way, given its frequency and assumed effect size. To simulate a GWAS on a particular commercial chip, we examined data at only those SNPs on the chip in question and checked to see whether any of these SNPs showed a p-value for association of $< 10^{-6}$. If this occurred we then modelled a replication study. We took the best SNP from the simulated GWAS and examined it in the simulated replication sample to check whether it had a p-value of < 0.01 in this replication sample. We considered

only those simulations where the best SNP on the genotyping chip met both of these criteria, as these model the ascertainment implicit in reported GWAS associations. For these simulations, we estimated the effect size at this best-associated SNP, which we call the *hit* SNP.

2.1 Choice of genomic regions

In order to model the signal of association generated by disease-causing mutations, we chose to simulate data exploiting empirical surveys of human diversity. For this purpose we used the 10 ENCODE regions (ENCODE Project, 2004) within the CEU analysis panel of HapMap II (International HapMap Consortium, 2007), which have undergone SNP ascertainment by resequencing 48 individuals of diverse ancestry. These regions therefore show a fuller spectrum of SNPs than are represented in the HapMap data at large, and haplotypes are expected to be accurate due to the trio design of the HapMap panel (International HapMap Consortium, 2003). The regions over which we simulate data are centred on each of the 10 ENCODE regions (listed in Table 2.1) and include 500kb of flanking HapMap variation at the boundaries of each region.

2.1.1 Properties of the ENCODE regions

Table 2.1 summarizes the coordinates of the ENCODE regions and the number of SNPs in each region for each of the populations we considered in the simulation study. Note that the Yoruban panel (YRI) tends to contain more SNPs in each region than either the European (CEU) or Japanese and Chinese combined panels (JPT+CHB). This is not surprising, since genetic and archeological evidence suggests that modern humans originated in Africa (Barbujani & Goldstein, 2004); greater genetic diversity is expected as a consequence of the longer history of the African population during which more mutations can accrue. The ENCODE regions with an 'r' in their region name were

Table 2.1: ENCODE Regions: SNP Summary

Region	Chr	Genomic interval	CEU	JPT+CHB	YRI
ENm010	Chr 7	26924045 - 27424045	731	710	873
ENm013	Chr 7	89621624 - 90121624	1042	800	1185
ENm014	Chr 7	126368183 - 126865324	1116	953	1248
ENr112	Chr 2	51512208 - 52012208	1255	1173	1889
ENr113	Chr 4	118466103 - 118966103	1366	1109	1525
ENr123	Chr 12	38626477 - 39126476	1259	1570	1208
ENr131	Chr 2	234156563 - 234656627	1307	1154	1589
ENr213	Chr 18	23719231 - 24219231	863	776	1291
ENr232	Chr 9	130725122 - 131225122	718	741	1053
ENr321	Chr 8	118882220 - 119382220	837	836	1281

chosen in an automated random way in keeping with the overall aim that this set of regions be representative of the genome in some sense, while the regions with an 'm' were chosen manually to include extensively characterized genes and/or functional elements and also in parts of the genome with a substantial amount of comparative sequence data (ENCODE Project, 2004). Figure 2.1 shows the allele frequency distribution in the CEU reference panel.

Table 2.2 compares densities of the chips in the ENCODE regions with the genome-wide averages. Although there is a 2-2.5 fold enrichment, visual inspection of the MAF distributions of SNPs on the chips in the ENCODE regions versus the genome suggests that this enrichment is not particularly biased towards lower-allele frequency SNPs. One implication of this enrichment is that the tagging properties of the Affymetrix 500k chip in the ENCODE regions are likely to mimic those of the Affymetrix 6.0 chip in the genome at large. Further consequences for the results of our simulations will be discussed in Chapters 3, 4, and 5.

Table 2.2: ENCODE bias summary. Each line of the table shows, for a given chip, the genomic average marker density and the ENCODE marker density.

Chip	# on chip	Average Density	ENCODE Density
Illumina 1.2M	1,238,764	0.12	0.24
Affymetrix 6.0	934,986	0.09	0.15
Illumina 550	561,497	0.06	0.15
Affymetrix 500k	500,569	0.05	0.09

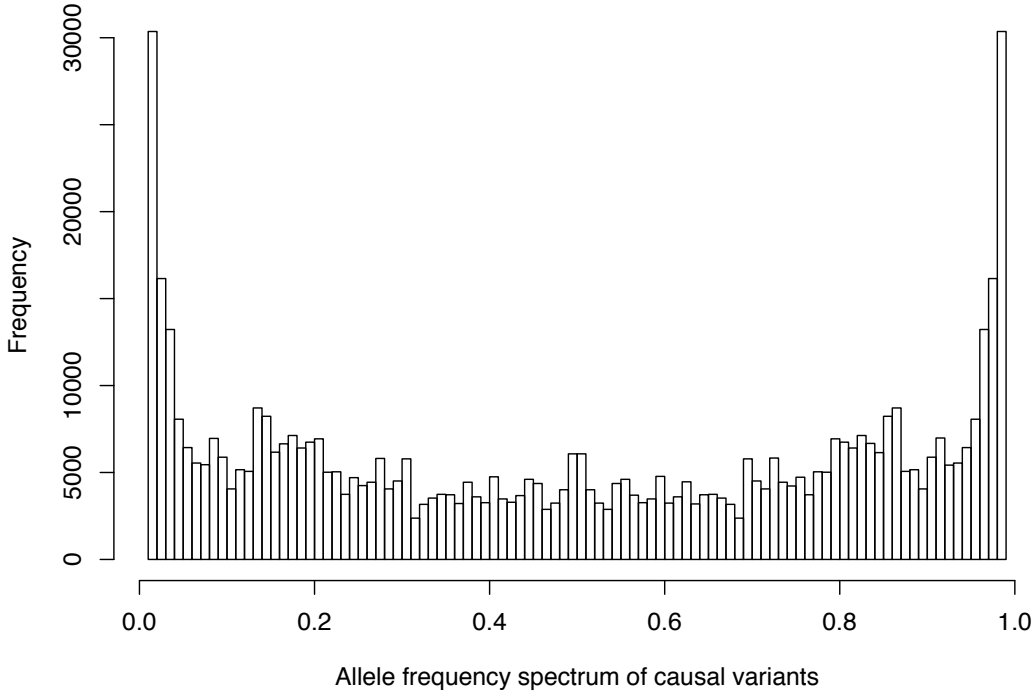


Figure 2.1: Allele frequency spectrum of the reference panel - Histogram of the allele frequencies of causal SNPs in the reference panel for the CEU population. Note that in simulations, we test both alleles at each SNP. The number of SNPs in a given allele frequency interval is shown on the y-axis. Note that only SNPs with minor allele frequency $> 1\%$ are shown, since we do not simulate causal SNPs at lower frequencies.

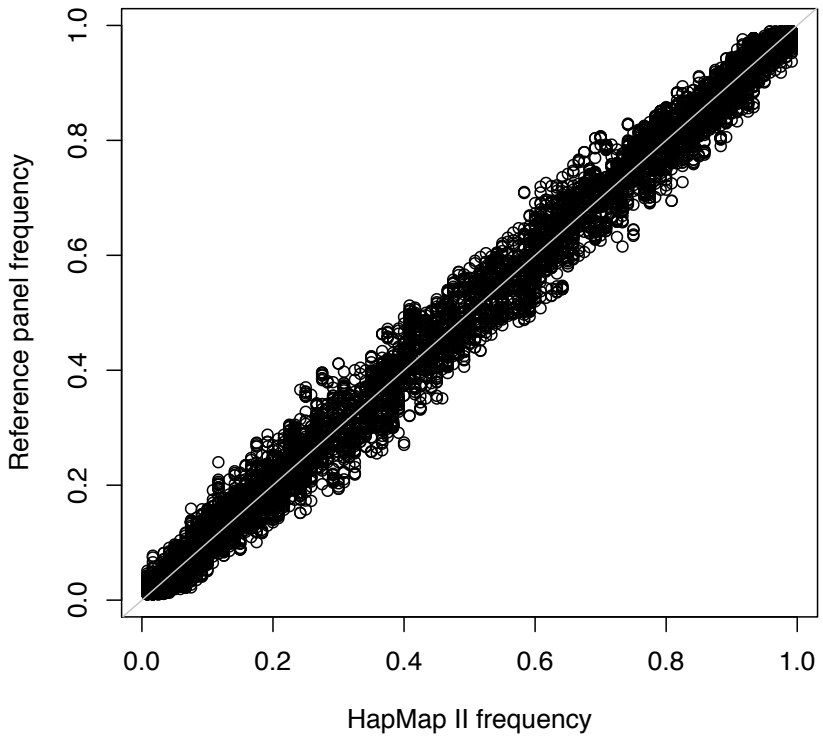


Figure 2.2: Comparison of the allele frequency distribution in the HapMap haplotypes and HAPGEN simulation output - Each point on the scatterplot represents a SNP. Its value on the x-axis is its frequency in HapMap II, while its value on the y-axis is its frequency in the reference panel produced by HAPGEN. The grey line shows $y = x$.

2.2 Simulation of population data

As the typical sample size of most GWAS is much larger than the number of CEU HapMap individuals, we simulated 100,000 chromosomes using the HAPGEN software package. We will refer to these 100,000 haplotypes as the *reference panel*. Figure 2.2 compares the reference panel allele frequency spectrum with that of the original HapMap II haplotypes from which it is simulated. Sampling variation causes small deviations from the line $y = x$. GWAS case and control samples were then subsampled from the reference panel, as described below.

2.2.1 A brief description of HAPGEN

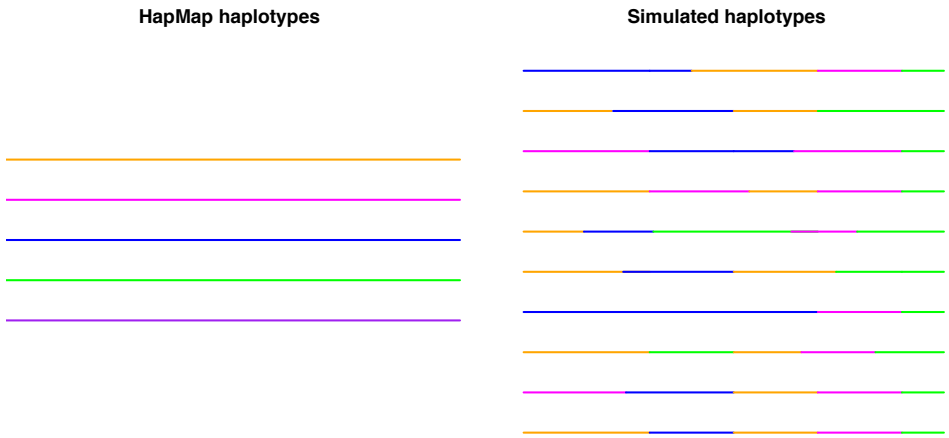


Figure 2.3: Cartoon of the HAPGEN program - The set of known haplotypes is extended by creating additional haplotypes that are a mosaic of known haplotypes, in a process described in §2.2.1. Note that this is a greatly simplified representation.

HAPGEN uses a population genetic model that incorporates the processes of mutation and fine-scale recombination to generate simulated haplotypes from an existing set of known haplotypes (here, the HapMap haplotypes). An intuitive, informal way of thinking about how HAPGEN works is to imagine the HapMap haplotypes as colored lines on a piece of paper. We begin with the set of known haplotypes, as shown on the

left side of Figure 2.3. To generate the next haplotype we place a special probabilistic ant down on one of the lines (the ant is special in that it obeys the rules of the Li and Stephens Hidden Markov Model (HMM), described in detail in Li & Stephens (2003)). The ant walks along the line and then, according to the HMM, jumps to another line and continues walking. SNPs mark various places along the lines. When the ant passes a SNP position a nucleotide flashes on a nearby screen, and is mutated, or not, again according to the HMM. At the end of the walk the ant's trajectory becomes another line at the bottom of the page and the ant begins again.

HAPGEN adds a new simulated haplotype to its list of reference haplotypes after each iteration. Thus the chance of a recombination event becomes very small as the size of the simulated population increases. Su *et al.* (2011) compared the empirical and simulated LD patterns between the HapMap reference haplotypes and 4,000 simulated individuals (2,000 cases and 2,000 controls). Figure 2.4 shows this comparison graphically. Although our simulated population is much bigger than 4,000, we expect that that the patterns of LD among SNPs with minor allele frequencies greater than 1% are comparable.

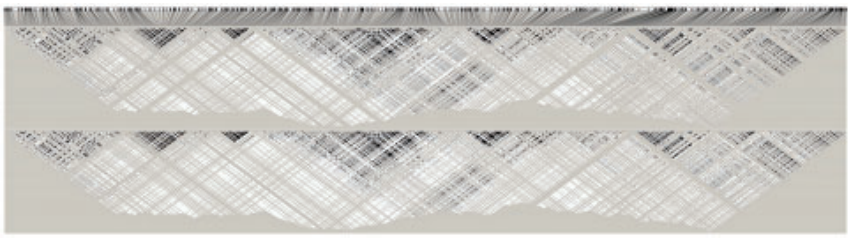


Figure 2.4: HAPGEN performance - The figure, taken from Su *et al.* (2011), shows LD patterns, in terms of r^2, in the HapMap haplotypes (top) and simulated haplotypes (bottom).

2.2.2 Inputs to HAPGEN

We ran HAPGEN with an effective population size of 11418 for the CEU population, 14269 for the JPT+CHB, and 17469 for the YRI, as recommended at `http://www.`

stats.ox.ac.uk/~marchini/software/gwas/hapgen.html, a population scaled mutation rate of 1 per SNP, population scaled recombination rate estimates as described in Myers *et al.* (2005), and with a known set of haplotypes taken from the analysis panels of HapMap II as described above.

2.3 Generating a case-control sample

For SNPs greater than 1% frequency in the ENCODE regions we performed two hypothetical GWAS by letting each of the two alleles be causal in turn. We denote the causal allele by A and the protective allele by a. To generate the control sample we drew the required number of haplotypes, without replacement, from the reference panel and combined these in pairs to form diploid individuals. This mimics the common use of population controls, rather than controls explicitly chosen for not having the disease under study. For the case sample, we drew pairs of haplotypes from the reference panel according to the genotype frequencies at the causal SNP dictated by the assumed disease model. If δ is the risk of the AA genotype, and α is the risk of the Aa genotype, both relative to the aa genotype, then we sample case individuals (without replacement) on the basis of their genotypes at the SNP assumed to be causal with success probabilities proportional to:

$$\Pr(AA) \propto \delta f^2\,, \tag{2.1}$$

$$\Pr(Aa) \propto 2\alpha f(1-f)\,, \tag{2.2}$$

$$\Pr(aa) \propto (1-f)^2\,, \tag{2.3}$$

where f is the frequency of the risk allele A in the reference panel. We adopted an additive disease model for disease risk defined by $\delta = \alpha^2$ for the material in Chapters 3-5; in Chapter 6 we will consider recessive and dominant models of the form $\alpha = 1$ and $\delta = \alpha$ respectively. Discussion of these alternative models will be postponed until

Chapter 6. Until then, we will refer to α as the relative risk (RR) or effect size associated with the causal variant. To approximate a GWAS, we thinned the generated data set to include only those SNPs present on the commercial chip under consideration which had a minor allele frequency in sampled controls of greater than 1%. This set may or may not include the assumed causal SNP.

For analyses involving only simulated data, we sampled 2,000 cases and 2,000 controls from the reference panel to emulate a typical GWAS. For the subsequent analyses of heritability and individual risk profiling for type 2 diabetes, breast cancer, and Crohn's disease (discussed below and in Chapter 3), we simulated 5,000 cases and 5,000 controls to obtain results more comparable to the size of study from which the associations were ascertained. We simulated under a range of relative risks at 24 grid points from 1.05 to 6. In attempting to simulate the signal of disease at rarer alleles (1% to 5%) in a GWAS of 5000 cases and controls there were a small number of simulations in which there were insufficient haplotypes in our reference panel to generate the required number of genotypes at the causal SNP for large effect sizes. These simulations were discarded, but as the numbers were small (3% when RR=4 and 11% when RR=6 for the CEU population in an Affymetrix 500k chip study) we do not believe this greatly affects the results presented in subsequent chapters.

Although we simulate causal variants at both alleles at each SNP, some disease models might favour causal variants that are rarer, derived mutations. The motivation for these models grows out of observations of Mendelian diseases. Mendelian diseases are generally subject to purifying selection. Purifying selection has the effect of removing deleterious variants as mutation creates them over time in a population. Therefore any observed mutations associated with Mendelian disease risk are likely to be newer. There is some evidence that this framework is applicable to common diseases, but there is also evidence that ancestral-susceptibility alleles act in common disease (Rienzo & Hudson, 2005). It is difficult to estimate selection coefficients for common diseases

because there is little power to detect the effects of rare alleles. In the absence of such knowledge we opted to simulate under no particular model of selection, and simulated causal variants at both ancestral and derived alleles.

2.4 Testing for association

Following common practice, for each simulated case control sample, we tested for association between genotype and case control status using the Cochran Armitage trend test (Armitage, 1955), introduced in §1.4.2, at each SNP with frequency greater than 1% in the simulated panel of chromosomes. We calculated the p-value of this test statistic which is χ^2 distributed with 1 degree of freedom under the null hypothesis of no association. If any test across the region obtained a p-value $< 10^{-6}$ the location of the most significant SNP (termed the *hit* SNP) was recorded and we simulated this SNP in an independent replication sample.

2.5 Simulating the replication process

We simulated the replication experiment in three stages. First we simulated the frequency of the causal allele in cases and controls in the replication population. We then simulated the frequency of the hit SNP conditional on the frequency of the causal allele. Finally, we simulated the genotype counts for a sample of cases and controls in this replication population.

We motivated sampling of the frequency of the causal allele in controls in the replication population by thinking of the replication sample as an additional sample from the same population as the original GWAS sample. (Other assumptions are possible here, but seem unlikely to affect the main conclusions.) Specifically, we placed a uniform prior distribution on the unobserved population frequency and sampled a value, f', from the posterior distribution of this frequency given the data in the reference

panel. (Given the large size of the reference panel, the frequency in the replication sample will be very close to that in the reference panel.) Conditional on f', the population replication frequency in cases was calculated from equations 2.1 - 2.3. To obtain the replication population frequencies at the hit SNP we estimated the conditional distribution in the reference sample of alleles at the hit SNP in cases and then controls, given those at the causal SNP, and used these for the replication sample. This corresponds to assuming that the LD between the causal and hit SNP in the replication sample will be the same as that in the reference sample. Finally, conditional on the population replication frequencies in cases and controls, we take multinomial samples of the required size to mimic the replication case and control samples. A test of association using the trend test was performed at the hit SNP on the simulated replication samples and deemed a significant replication if the p-value was less than 10^{-2}. The replication study sample sizes were 2,000 cases and 2,000 controls for simulations of the same discovery sample size, and 10,000 cases and 10,000 controls for simulations in which the discovery sample size was 5,000 cases and 5,000 controls.

2.6 Estimating effect sizes

We estimate the effect size, or relative risk, α, at the hit SNP by maximum likelihood under the model described by equations 2.1 - 2.3. For studies with population controls this can be achieved in practice by fitting a logistic regression model for case status, as discussed in §1.4.4. We fit the logistic regression model with case and control counts from the replication sample in order to mimic the estimates reported in GWAS studies, which are often taken from the replication stage of the study. Estimating the effect size from the replication sample has the advantage that it mitigates the winner's curse, whereby relative risks estimated in the discovery sample are biased upwards, conditional on having been identified (see §1.4.5).

2.7 Discussion

The approach taken in this chapter has been in close collaboration with Chris Spencer, whose previous work served as a foundation for the method utilized herein (Spencer *et al.*, 2009). One of his many contributions to this software was to write the code that iteratively samples from the reference panel and performs simulated GWAS and replication studies. I used this template in the extensions of the original simulations.

Among the many design decisions involved in creating the software, three limitations should be noted. Firstly, the ENCODE data is restricted in its coverage of low-frequency (1-5%) and rare (< 1%) variants because the sample size of the study was limiting and some rare variants were missed. As a result, most of our analyses excluded causal variants at frequencies of less than 5%. The role of these low-frequency variants in both GWAS results and disease more generally is of interest, but our approach is fundamentally limited by the original data from which we simulate our population data. Lower-frequency variants might be better represented if we were to simulate from a set of haplotypes from the 1000 Genomes dataset (1000 Genomes Project Consortium, 2010), for example, which was not available at the time we performed the study.

Secondly, the decision to use HAPGEN was motivated by previous work, although we recognize that other choices were possible, for instance simulating from a population genetic model (Hernandez, 2008). Although population data generated from such a model would include a wider variety of low-frequency variants, we felt that this advantage was outweighed by the advantages inherent in simulating from real data, with real patterns of linkage disequilibrium.

Thirdly, we did not test whether the direction of the effect estimated in the replication sample was consistent with the direction of the effect in the discovery sample. This is because it is unlikely that a SNP would be significantly disease-predisposing in the discovery sample and then would then replicate as a protective allele (or vice

versa).

The simulation framework described in this chapter provides the methodological foundation for the analysis of effect size estimation in GWAS, which is the topic of the next three chapters. Chapter 3 focuses on estimating true effect sizes from observed effect sizes in the CEU population with a genotyping chip used in the WTCCC study (Wellcome Trust Case Control Consortium, 2007) and applies the resulting analysis to three diseases interrogated by GWAS. Chapter 4 addresses how the genotyping chip and the underlying population affect observed relative risks, and shows that imputation can be used to combat effect size underestimation. Chapter 5 asks how often the causal and hit SNPs are expected to coincide. Finally, Chapter 6 looks at how variations in the disease risk model influence effect size estimates and whether these variations can be detected by statistical tests. In each case the simulation framework was adjusted to ask how GWAS results are affected by properties of the causal variants.

3

Effect size estimation in GWAS

3.1 Introduction

Virtually all associated variants identified from GWAS to date have relatively small effects: each additional copy of the risk allele typically increases disease risk by 10%-30% (see for example Manolio *et al.* (2008)). It has become clear that the variants discovered thus far account for only a small proportion of the genetic basis of each of the diseases under study, and there has been considerable speculation about where the "missing" heritability might lie (Eichler *et al.*, 2010).

One of several important factors in the success of GWAS design has been detailed knowledge about patterns of linkage disequilibrium in human populations. As discussed in the Introduction, the strong correlations between nearby SNPs mean that surrogates on genotyping chips can capture much of the common variation in the human genome. Because genotypes at the causative loci will often be correlated with those at SNPs that are typed on the genotyping chip, it is typically not necessary to assay the true causative variant directly in order to detect a genetic association with disease.

While LD is extremely helpful for GWAS discovery, the downside is that in most reported regions of association, the true causal variant or variants remain unknown.

Therefore it is possible that many of the associated SNPs are only surrogates for the true causal variant(s). When it comes to quantifying the genetic effect, the genotype at the reported SNP acts as a noisy measurement of the genotype at the causal variant. The noise can dilute the apparent strength of the effect, and obscure the true relationship between genotype and phenotype. As we progress towards the identification of causal variants, estimates of effect sizes for associated loci will thus tend to increase. In turn, the proportion of disease susceptibility explained by GWAS loci will also increase. Thus, in addition to other plausible sources, such as secondary signals in GWAS loci, rare variants ($< 1\%$ frequency), copy number polymorphisms, and epigenetic effects, some of the missing heritability is actually contained in loci already identified by GWAS, and is driven by common variation ($> 1\%$ frequency).

In this chapter we present results from an extensive simulation study, the details of which are described in Chapter 2, in order to investigate and quantify the phenomenon of effect size underestimation. We use the Affymetrix 500k chip in the CEU population, and will explore the effect of changing these parameters (both chip and population) on our conclusions in later chapters. We show that estimates of the size of the genetic effect based on the best SNP from the GWAS genotyping chip can often closely approximate the effect size at the true causal SNP. In some cases the causal SNP has a large effect size and is poorly tagged, leading to substantial underestimation of true effect size. We investigate how much of the 'missing' heritability could thus be hidden in reported GWAS loci, under several sets of assumptions about the nature of the effects at true causal SNPs. Our results also inform the design and value of fine mapping experiments of GWAS loci.

The results in this chapter, together with the methodology described in the previous chapter, have been published (Spencer *et al.*, 2011b). The paper, from which this chapter is drawn almost verbatim, was a highly collaborative work between myself and Chris Spencer under the direction of Peter Donnelly. Damjan Vukcevic provided

valuable help, especially on the sections that involve the two prior distributions we considered. I was largely responsible for the analyses presented here, with the exception of the individual risk estimates, for which Chris Spencer took the lead role.

3.2 Effect size estimates

As discussed earlier, patterns of LD in human populations are too complicated to draw analytical conclusions, so we adopted the simulation approach described in detail in Chapter 2. Here we discuss the outcomes of a simulation study in the CEU population with an Affymetrix 500k chip. Recall that the Affymetrix 500k chip is enriched in the ENCODE regions (see Table 2.2), so that the density of markers in the regions we simulated closely matches that of the Affymetrix 6.0 chip in the genome at large. Results in this section can therefore be interpreted to represent the outcomes of a GWAS study conducted in the Affymetrix 6.0 chip.

To begin, we compare the estimated effect size at the replicated SNP with the true effect size at the causal SNP in the simulation. Note that in the following figures we show only results for causal SNPs with minor allele frequencies >5%. Figure 3.1 illustrates this comparison for three different values of the true effect size. For each we see a peak of estimates around the true effect size assumed at the causal SNP. But note also that there is often underestimation of the true effect size, and that this underestimation can be substantial, especially when the true effect is large. For example, when the true relative risk is 4, the estimated effect size was less than 2 in 11.5% of the simulations in which GWAS successfully discovered the effect.

In Figure 3.2 we plot the relative under- (or over-) estimation of the effect size (estimated effect size divided by the true effect size) as a function of the correlation between the hit SNP and the true causal variant. The underestimation is seen to be due to imperfect tagging: when the true causal variant is not well tagged by SNPs on the

genotyping chip (the correlation is weak), the estimated effect at the hit SNP is often much lower than the true effect. Conversely, when the causal SNP is well tagged by a SNP on the chip, the estimated effects cluster around the true effect size. Note that while underestimation decreases as the correlation between the hit SNP and the causal SNP increases, there remains systematic underestimation even when the hit SNP has $r^2 \approx 0.8$ with the causative SNP. For example, in 30% of simulations when the true effect is 2, the estimated effect will be under 1.8. Note also that when the true effect size is large, significant and replicable associations can be detected when the best tag SNP has only $r^2 \approx 0.2$ with the causal variant (Figure 3.2, relative risk = 4).

Imperfect tagging and an ascertainment effect also explain the feature of the plots whereby the underestimation is much less for smaller true effect sizes. If the true effect is small and the true causal variant is not well-tagged on the genotyping chip, there will not be enough power for the GWAS and subsequent replication to reach significance (Iles, 2008), with the result that the corresponding simulation will not contribute to the plot. But if the true effect is large there may still be power to see a significant result when the true variant is not well tagged, so the simulation contributes to the plot and shows the underestimation. Put another way, if the true effect is small, it will only be detected in an association study if the causal SNP is well tagged, and in this case the effect size will be reasonably well estimated. This second ascertainment effect explains the lack of underestimation at hit SNPs not strongly correlated to the causal SNP in the leftmost panel of Figure 3.2.

3.3 What true effects might underlie the effects estimated from GWAS?

The results in the previous section describe the distribution of estimated effect sizes as a function of known true effect sizes. In practice we are interested in the reverse question,

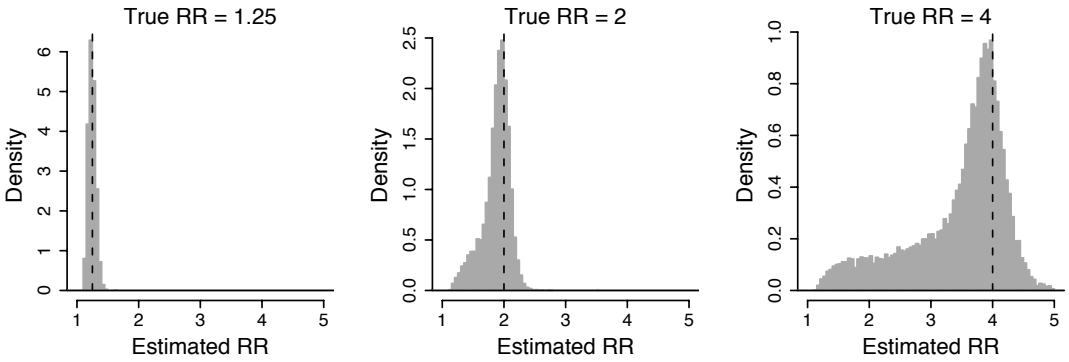


Figure 3.1: Distribution of estimated effect sizes - Histograms of estimated relative risks (RR), for three different true relative risks indicated by a vertical dashed line in each plot. Histograms include all simulations where the most associated (hit) SNP was significant in both the initial study and the replication study.

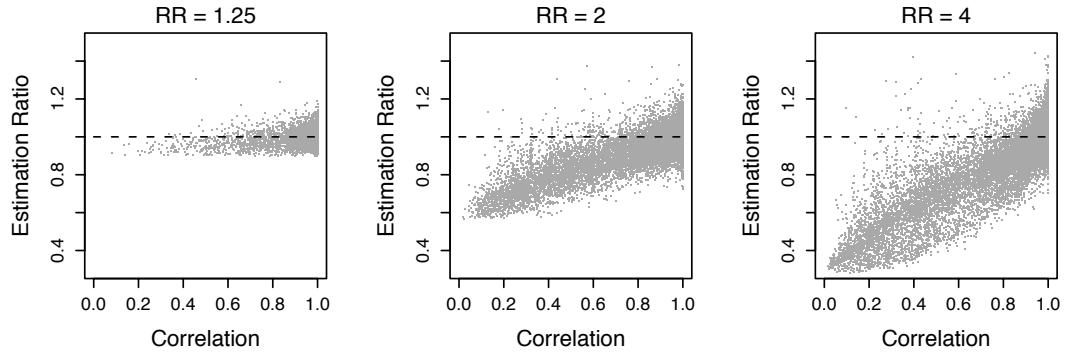


Figure 3.2: Relationship between underestimation and correlation - Scatter plots of the ratio of estimated to true relative risk against the correlation (r^2) between the disease SNP and the hit SNP, for three different true relative risks. The horizontal dashed line indicates a ratio of 1. Points below the line are underestimated, and above the line, overestimated. (Note that in the case where the hit SNP is also the causal SNP the correlation is 1.)

namely what true effect sizes are plausible in light of the effect size actually estimated from a GWAS and follow up study? We will see that this requires assumptions about the true distribution of effect sizes. Indeed, writing RR for the relative risk, and RAF (risk allele frequency) for the allele frequency at the risk allele, application of Bayes' theorem gives,

$$\begin{aligned} &\Pr(RR_{\text{true}}, RAF_{\text{true}}) \mid RR_{\text{observed}}, RAF_{\text{observed}}) \propto \\ &\Pr(RR_{\text{observed}}, RAF_{\text{observed}} \mid RR_{\text{true}}, RAF_{\text{true}}) \times \Pr(RR_{\text{true}}, RAF_{\text{true}}) \end{aligned} \tag{3.1}$$

where RR_{true} refers to the relative risk at the causal SNP and RR_{observed} refers to the relative risk at the hit SNP, and similarly for the risk allele frequency. Our simulation study allows us to estimate the first factor on the right hand side of Equation 3.1, and we do so by discretising both the observed and true RR and RAF and creating a matrix of counts based on our simulations over the ENCODE regions.

In order for this approach to be maximally informative, the observed relative risks should be taken from a study that matches as closely as possible the particular features of the simulation design. As discussed in the previous section, and in Chapter 2, the Affymetrix 500k chip is enriched in the ENCODE regions. Therefore, the simulation results should ideally be used to estimate the posterior distribution of effect sizes for a GWAS study conducted with the Affymetrix 6.0 chip (or one with a similar genome-wide marker density to that of the Affymetrix 500k chip in the ENCODE regions).

The second factor on the right hand side of Equation 3.1 is the assumed joint distribution of true risk allele frequencies and effect sizes, which is of course unknown. We proceed by making two different sets of assumptions about these unknowns. In each case we assume that the distribution of risk allele frequencies is given by the empirical distribution of allele frequencies in the ENCODE regions. In effect this assumes that any SNP variant is, *a priori*, equally likely to affect disease status. What differs between the sets of assumptions is the assumed effect size of a particular variant. Our first set

of assumptions posits that the distribution of effect sizes is the same for all putative causal variants, regardless of their allele frequency, and that effect sizes are close to those observed in GWAS studies. The second set of assumptions explicitly assumes that there might be larger effects at variants with smaller minor allele frequency. These priors are described in detail in the Introduction.

Under either set of assumptions, we can use our simulation study, and Equation 3.1 to estimate the conditional distribution of true effect size and risk allele frequency in light of the observed data at the GWAS hit SNP. Figure 3.3 illustrates this, showing estimates of the posterior distribution of the true effect size conditional on observing risk estimates in different ranges for different observed risk allele frequencies, and under the two different prior assumptions on effect size distributions.

There is a subtle relationship between power and effect size underestimation, as our results in Figures 3.1 and 3.2 show. For causal variants of the same minor allele frequency, the non-centrality parameter of the chi-squared test is proportional to $N\beta^2$, where N is the sample size and e^{β} is the relative risk (Vukcevic *et al.*, 2011). Thus a variant with a relative risk of 1.3 in a study with a sample size of 2000 cases and controls has roughly the same non-centrality parameter as a variant with a relative risk of 1.1 in a study with a sample size of approximately 2800 individuals when the variants' minor allele frequencies are equal. We recognize that the relative risk ranges considered in Figure 3.3 may not be representative of the majority of GWAS hits, but qualitatively similar conclusions apply to lower relative risk ranges with the same observed risk allele frequencies, so long as the sample sizes are appropriately adjusted in the comparison.

A common feature of the histograms in Figure 3.3 is that the mode of the posterior distribution on the true effect size is on, or very close to, the upper bound of the observed estimate. That is, current estimates from GWAS studies of effect sizes from a common SNP in the ranges shown are most likely to be close to the truth. It is expected that estimated effects within any given range are more likely to be at the upper bound

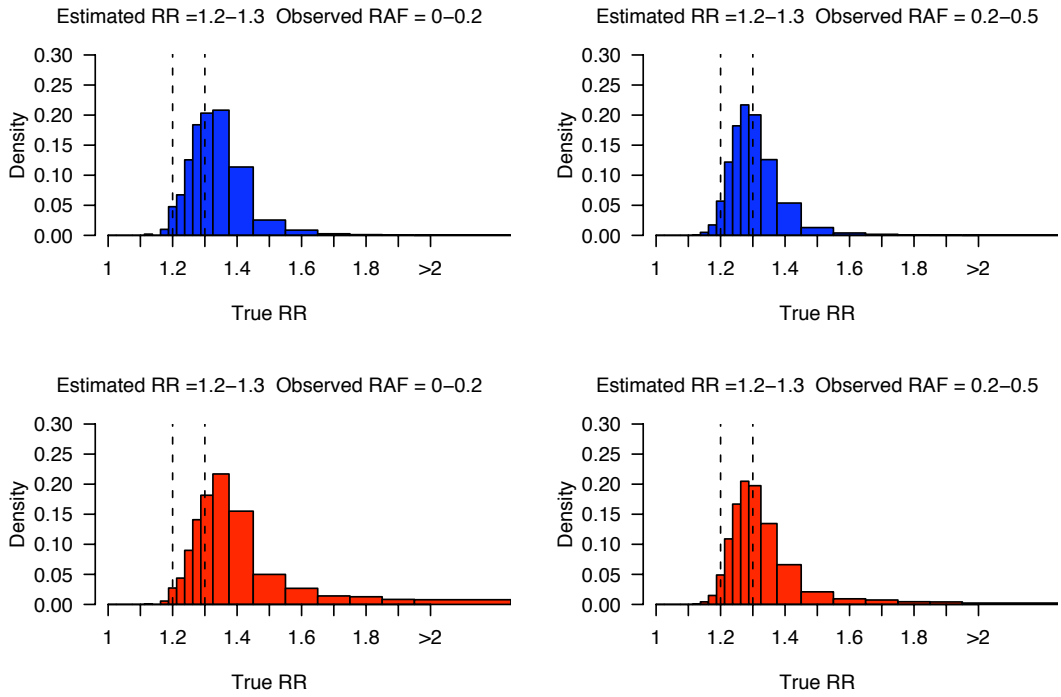


Figure 3.3: Posterior distribution on true relative risk - Histograms showing the posterior distribution of the true relative risk (RR) conditional on observing an estimated relative risk in the range 1.2-1.3 at the hit SNP (vertical dashed lines). Right hand plots condition on the observed risk allele frequency (RAF) being between 20 and 50%, while the left hand plots condition on a RAF less than 20%. Results are shown using two different priors on RR and RAF: the blue histograms are the posterior distribution obtained using a *conservative* prior, and the red histograms are the posterior distribution obtained using the *MAF-dependent* prior.

rather than the lower bound, because larger effects are more likely to generate a signal of association strong enough to pass the p-value thresholds commonly implemented in GWAS. This explains the left hand tail of the distributions represented in Figure 3.3.

Figure 3.3 also shows that there is some probability that the effect size at the causal variant is greater than that estimated from the most associated SNP. Interestingly, the observed risk allele frequency impacts our posterior belief about the true effect size, under either set of prior assumptions, with underestimation being more marked when the risk allele at the hit SNP is rarer. Under the conservative prior, when the risk allele at the hit SNP has less than 20% frequency in the control population, the probability that the relative risk is above 1.3 is 55%, compared to 39% when the risk allele frequency is between 20-50%. The corresponding numbers for the MAF-dependent prior are 67% and 44%. There are several different phenomena at work here. If the hit SNP is the causal SNP then, assuming that the association is strong enough to be detected and replicated in the GWAS, there is no systematic underestimation (and very little overestimation as we assume the effect size is estimated from the replication sample). However, conditional on the hit SNP not being causal, the distribution of LD with the true causal SNP, and therefore the propensity for underestimation, depends on its allele frequency. The posterior distribution on the true effect size given the observed frequency and effect of the hit SNP can be viewed as a mixture of these two scenarios, weighted by their conditional probabilities. Rarer SNPs are less likely to be tagged well by single markers (International HapMap Consortium, 2005), and, as noted above, poor tagging leads to underestimation of effect sizes. In contrast, for a common SNP, the associated allele is more likely to be well correlated with the causal allele, so there is relatively less underestimation. Under the MAF-dependent prior, when the associated allele is low-frequency, the causative allele will tend to be low-frequency as well, and so potentially of larger effect. In the scenario where we believe in larger effects at rare causal alleles and have observed a SNP with low RAF and estimated relative risk

Table 3.1: Replicated SNPs for Type 2 diabetes (Zeggini *et al.*, 2008).

Gene	SNP	RAF	OR
KCNJ11	rs5219	0.4	1.15
PPARG	rs1801282	0.85	1.23
TCF7L2	rs7901695	0.4	1.37
SLC30A8	rs13266634	0.72	1.12
WFS1	rs10010131	0.6	1.11
HHEX/IDE	rs5015480	0.63	1.13
CDKAL1	rs10946398	0.36	1.16
CDKN2A/B	rs10811661	0.86	1.19
FTO	rs8050136	0.45	1.23
IGF2BP2	rs4402960	0.35	1.11
TCF2	rs757210	0.43	1.08
JAZF1	rs864745	0.501	1.1
CDC123/CAMK1D	rs12779790	0.18	1.11
ADAMTS9	rs4607103	0.76	1.09
THADA	rs7578597	0.9	1.15
NOTCH2	rs10923931	0.106	1.13
TSPAN8/LGR5	rs79611581	0.269	1.09
MC4R	rs17782313	0.24	1.09

between 1.2 and 1.3, there is a 2% chance that the source of the signal is a variant which actually doubles, or more than doubles, the risk conferred by each copy of the causal allele.

3.3.1 Consequences for individual disease risk

One consequence of the potential underestimation of effect sizes from GWAS findings is that, as we move to better identification of the actual causal variants through fine mapping and/or functional studies of associated regions, our estimates of their effect sizes might well increase. Assuming a multiplicative model of risk across loci, these small expected changes could combine to increase the relative risk of disease in those individuals with highest genetic risk of disease.

To investigate this, we simulated genotypes at known associated loci in a population of individuals (assuming Hardy-Weinberg equilibrium and no linkage disequilibrium

Table 3.2: Replicated SNPs for Crohn's disease (Barrett *et al.*, 2008).

Gene	SNP	RAF	OR
IL23R	rs11465804	0.932	2.52
ATG16L1	rs3828309	0.53	1.25
MST1	rs9858542	0.567	1.17
PTGER4	rs4613763	0.125	1.35
—	rs2188962	0.412	1.36
IRGM	rs11747270	0.091	1.35
TNFSF15	rs4263839	0.676	1.2
ZNF365	rs10995271	0.393	1.26
NKX2-3	rs11190140	0.477	1.2
NOD2	rs2066847	0.018	4.01
PTPN2	rs2542151	0.152	1.32
PTPN22	rs2476601	0.899	1.33
—	rs9286879	0.244	1.18
—	rs11584383	0.674	1.2
IL12B	rs10045431	0.705	1.11
CDKAL1	rs6908425	0.78	1.21
—	rs7746082	0.288	1.18
CCR6	rs2301436	0.448	1.42
—	rs1456893	0.679	1.18
—	rs921720	0.606	1.17
JAK2	rs7849191	0.602	1.12
—	rs17582416	0.345	1.18
C11orf30	rs7927894	0.376	1.23
LRRK2, MUC19	rs11175593	0.017	1.63
—	rs3764147	0.221	1.25
ORMDL3	rs2872507	0.473	1.13
STAT3	rs744166	0.564	1.18
—	rs4807569	0.209	1.16
—	rs1736135	0.567	1.16
ICOSLG	rs762421	0.388	1.09

Table 3.3: Replicated SNPs for breast cancer (Pharoah *et al.*, 2008).

Gene	SNP	RAF	OR
FGFR2	rs2981582	0.38	1.26
TNRC9	rs3803662	0.25	1.2
MAP3K1	rs889312	0.28	1.13
LSP1	rs3817198	0.3	1.07
—	rs13281615	0.4	1.08
—	rs13387042	0.5	1.2
CASP8	rs1053485	0.86	1.13

across loci) for each of breast cancer, type 2 diabetes, and Crohns disease, based on the reported risk allele frequencies given in Tables 3.1 - 3.3.

Given a set of causal loci, we can stratify individuals in the population on the basis of their genotypes. At the lowest end of the risk spectrum are individuals with all protective alleles; at the highest end of the risk spectrum are individuals with all risk-conferring alleles. We then calculate the average risk of developing disease for individuals in a given risk bracket compared to the mean risk in the population. We make three sets of assumptions about the set of causal loci and calculate individual risks by simulating genotypes under each set of assumptions. First we suppose that the causal loci are in fact those reported in Tables 3.1 - 3.3. These are the unadjusted results. Second, we allow for uncertainty in the estimation of true effect sizes, under the conservative prior, by averaging over the posterior distributions on RAF and RR given the GWAS findings. Third, we we allow for uncertainty in the estimation of true effect sizes, under the MAF-dependent prior, by averaging over the posterior distributions on RAF and RR given the GWAS findings.

We assumed that risks combined multiplicatively across loci. For NOD2 and IL23R in Crohns disease, where the causal haplotypes are thought to be known, here and below we used the effect sizes for the known variants, and did not average over uncertainty. Because all three diseases have been extensively studied, we approximated the GWAS discovery process as corresponding to a GWAS discovery sample of 5,000 cases and 5,000 controls, and a replication sample of 10,000 cases and 10,000 controls. The actual discovery process for each of the diseases is complicated, often involving meta-analysis and/or multistage discovery, and not straightforward to model accurately, but the approach we use should capture the fact that GWAS-discovery were ascertained through study of large numbers of samples.

The results of the three simulations are given in Table 3.4. The unadjusted simulations give estimates of how much more at risk individuals with the greatest genetic

Table 3.4: Adjusted estimates of individual risks

Type II Diabetes							
x%	50	25	10	5	1	0.5	0.1
Unadjusted	1.4	1.7	2.0	2.2	2.8	3.0	3.6
Conservative	1.6	2.0	2.5	2.9	3.8	4.3	5.3
MAF-dependent	1.9	2.7	3.9	5.0	8.5	10.4	16.2
Crohn's Disease							
x%	50	25	10	5	1	0.5	0.1
Unadjusted	2.1	3.0	4.4	5.6	9.4	11.5	17.6
Conservative	2.4	3.4	5.1	6.6	11.5	14.1	22.1
MAF-dependent	2.9	4.4	7.0	9.7	18.5	23.9	41.2
Breast Cancer							
x%	50	25	10	5	1	0.5	0.1
Unadjusted	1.3	1.5	1.7	1.8	2.0	2.2	2.4
Conservative	1.4	1.6	1.9	2.1	2.5	2.7	3.0
MAF-dependent	1.5	1.9	2.4	2.8	3.8	4.3	5.5

propensity to disease are, based only on GWAS loci, relative to the average person in the population. As expected, the fold change in risk of individuals carrying a large fraction of risk variants is dependent on the number and magnitude of known loci. For example, individuals in the top 0.1% of risk for Crohns disease are 18 times more likely than the average person to develop the condition, whereas for breast cancer, where the number of common loci and associated relative risks is typically smaller, the equivalent number is just over two-fold.

The second and third simulations attempt to average over the possible outcomes of our future efforts to map causal mutations, in order to reveal the likely gains in our ability to stratify individuals on the basis of risk. These use the methodology above, under both prior distributions, to average over the posterior distribution of the allele frequency and effect size at the causal SNPs underlying reported GWAS loci for the three diseases. These adjusted estimates are also shown in Table 3.4. Across diseases we see that there is a significant increase in the risk associated with carrying multiple risk variants. In particular we see that the biggest differences in risk are for those

individuals in the extreme tail. It is these individuals who carry the stronger, likely rarer, risk alleles which are currently insufficiently characterised by the most significant signal of association in some regions identified to be important in disease. For example, the risk of an individual in the top 0.1% of the population for genetic risk typed at the causal loci underlying currently known GWAS loci will be increased by a factor of 3–5.5, 5.3–16.2, and 22–41.2 compared to an average individual, for breast cancer, type 2 diabetes and Crohns disease. These are notably greater increases in risk than predictions based on the hit SNPs from GWAS loci given in Tables 3.1 - 3.3, which would be 2.4, 3.6 and 17.6 respectively.

At the beginning of this section, we noted that the efficacy of our approach relies on the ability of our simulations to mimic the GWAS study from which the risk loci and their effect sizes derive. Particularly important, the pairwise correlations between potential causal SNPs and their markers on the chip should be recapitulated, so that the correction for effect size underestimation is properly calibrated. On the one hand, the type 2 diabetes loci were identified by a meta-analysis that tested over 2.1 million genotyped and imputed SNPs (Zeggini *et al.*, 2008), the Crohn's loci were identified by a meta-analysis that tested 635,547 genotyped and imputed SNPs (Barrett *et al.*, 2008), and the breast cancer loci were identified by two studies, one which tested 266,732 SNPs genotyped with a Perlegen microarray, and one which tested nine candidate SNPs (Pharoah *et al.*, 2008). On the other hand, our simulations report underestimation in the setting of a GWAS conducted with a genotyping chip with approximately 900,000 SNPs. Clearly these studies do not perfectly match the parameters of our simulations, and reflect the limitations of data available at the time. The interpretation of our results should be adjusted accordingly. In particular, we would expect the correction for the type 2 diabetes loci to be overly liberal, the correction for the Crohn's disease loci to be likely too conservative, and the correction for the breast cancer loci to be far too conservative. The same limitations apply in the next section, where we apply the

same methodology to the question of missing heritability.

3.3.2 Missing heritability

We have shown above that, as we move to identification of the true causal variants underlying GWAS associations, through fine mapping and functional studies, their effect sizes will tend to increase – in a minority of cases, substantially – compared to current estimates from GWAS. This will, in turn, increase the amount of heritability explained by these diseases. We can use the approach developed here to try to quantify this effect.

We investigated this question in the context of the three diseases just described, namely breast cancer, type 2 diabetes, and Crohns disease. For each disease we took the set of hit SNPs from published associated loci listed in Tables 3.1 - 3.3 and, for our two prior distributions on effect sizes, we estimated the posterior distribution of both the effect size and the allele frequency for the causal SNP at each locus, as described in the previous section. One commonly used measure of heritability is sibling recurrence risk ratio, often denoted by λ_s: the relative increase in risk to an individual if their sibling has the disease compared to the baseline risk in the population as a whole (Risch, 1990). Assuming, as is usual for heritability calculations (Thomas, 2004), that there is no interaction between loci, λ_s can be calculated as a function of the risk allele frequency and effect size for each causal variant (see Equation 1.3). In order to allow for uncertainty in the allele frequency and likely underestimation of the effect size at the causal variants underlying GWAS associations, we averaged this expression over the posterior distribution of these quantities, given the GWAS findings. Specifically, we calculated sibling relative risk as given by Equation 1.3 using estimates of RR and RAF of replicated loci. We then simulated 100,000 times from the posterior distribution of true RR and RAF of each locus, conditional on the reported RR and RAF, using the simulation approach and the two different priors as described in Chapter 2. For

each set of simulations, we recalculated λ_s at each locus and summed over loci, giving a sample from the posterior distribution of sibling risk.

The results are shown in Figure 3.4. For each disease they show that the heritability due to already identified GWAS loci will be higher than current estimates, under either set of assumptions about true effect sizes, but particularly under the MAF-dependent prior. Whereas the current estimates of the contribution to λ_s from GWAS loci are 1.05, 1.08, and 1.49 for breast cancer, type 2 diabetes, and Crohns disease, these may well be 1.08, 1.14, and 1.60 (the means under the conservative prior) and they could plausibly be as high as 1.71, 1.80 and 3.54 (the posterior 95th percentiles under the MAF-dependent prior). As before, these estimates should be interpreted with the understanding that the ascertainment method is not perfectly matched by the simulation study (particularly with respect to marker density), so that the true explained heritability captured by GWAS loci in type 2 diabetes is likely to be smaller than the results suggested by our method, while the true explained heritability in Crohn's disease and breast cancer are likely to be higher, and possibly much higher, respectively. Whilst some of the 'missing' heritability is thus disguised rather than missing, we note that this effect does not account for the extent of the gap between estimates of sibling relative risk – 1.8, 2, and 10, respectively, from family studies (Pharoah *et al.*, 1997; Tysk *et al.*, 1988; Weijnen *et al.*, 2002) – and those explained by currently known loci in Crohn's disease. We return below to a discussion of the discrepancy.

3.4 Discussion

The correlation between alleles along the human genome has allowed GWAS to look for regions associated with disease without having to either genotype all known genetic variation or guess *a priori* which regions of the genome may be important. Although this approach has been a significant success, there is a predictable downside of using

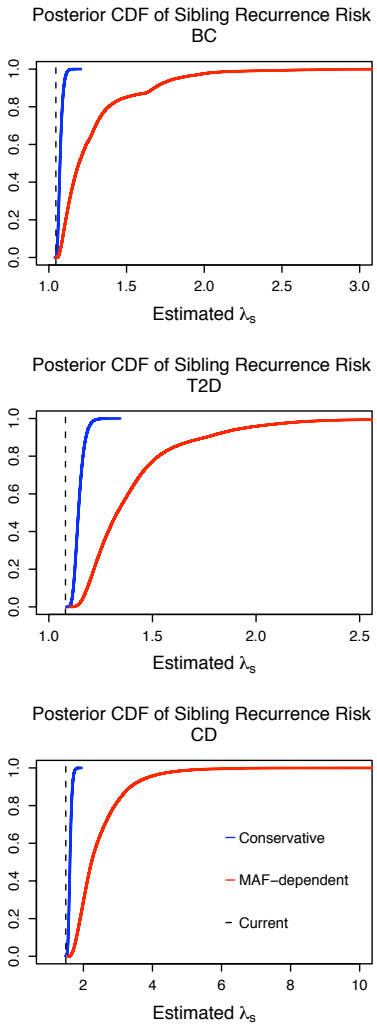


Figure 3.4: Adjusted estimates of explained heritability - Cumulative density functions for explained sibling risk (estimated λ_s) in breast cancer (BC), Type 2 Diabetes (T2D) and Crohn's disease (CD) under the conservative (blue) and MAF-dependent (red) priors. The dotted line indicates the value of λ_s based on replicated loci in each of the three diseases. The quoted values of sibling risk (lower right hand corner) are from references given in the text. Posterior distributions were calculated assuming a GWAS of 5,000 cases and controls and a replication study of 10,000 cases and controls.

a subset of variation to tag, or predict, untyped diversity: for the vast majority of the SNPs identified as mediating disease risk, we are left uncertain as to whether they are causally involved in the pathway from genotype to phenotype, or, much more plausibly, are just surrogates for the causal variation.

GWAS associations will thus typically relate to a noisy measurement of the causal variant. One consequence of this is that the size of the genetic effect associated with GWAS loci may be underestimated. We quantified this through an extensive simulation study designed to mimic patterns of linkage disequilibrium in European Caucasian populations. We draw two broad conclusions from these analyses. Firstly, a significant proportion of estimated relative risks will be biased downwards because the hit SNP is a powerful, but imperfect, tag for the true causal variation. Secondly, in most cases this effect will be relatively minor, but in some instances, the hit SNP may actually be a poor predictor of a putatively rarer SNP with a much larger effect, in which case the effect size estimated from the GWAS finding will substantially underestimate the true effect size. As denser chips are deployed in GWAS, we would expect that an increased number of SNPs represented on the chip will result in better tagging, and correspondingly less effect size underestimation. The number of cases in which the underestimation will be substantial because the hit SNP is a poor tag for the causal SNP will be reduced to the extent that denser chips better represent rarer variation.

The exact proportion of reported associations which fall into these two categories – minor effect size underestimation due to good tagging, or substantial effect size underestimation due to poor tagging – depends on properties of the design of the study from which the SNP was identified, and on one's belief about how likely common ($> 1\%$) variants of large effect are to cause common diseases. The statistical power afforded by any particular association strategy sets a lower limit on the size of effect that can be underestimated because an imperfect tag of an allele with a small effect size will simply fail to achieve genome-wide significance. Other properties of GWAS strategy, such as

sample ancestry and the number of markers typed, also change our interpretation of observed effect sizes because they influence the distribution of linkage disequilibrium between putative hit SNPs and causal variants.

Our findings show that, at any particular locus, especially if the associated SNP has a low MAF, the true effect could be quite large. But we would not expect this to be widespread. Were many true effects this large it would be extremely surprising for so few of them to have been observed: although any one such causal SNP may not be well tagged on the genotyping chips used for GWAS, some of them will happen to be at least moderately well tagged, and their detection would lead to much larger estimates than have been seen from current studies. This supports the conclusions of Anderson *et al.* (2011) in the setting of single rare variant synthetic associations: our results suggest that if the majority of underlying causal variants were rare, then their tags would tend to be rarer, on average, than the allele frequencies of the majority of SNPs arising from GWA studies.

One way of viewing the posterior distribution on the true effects shown in Figure 3 is as a probability distribution on the outcome of efforts to fine map current regions of association. Although the quantitative nature of the posterior depends on the prior, we have shown that the qualitative conclusions are the same for two prior distributions that represent different beliefs about the true effect size distribution. In this light, our results inform questions of the design and value of fine mapping experiments. First, simulations similar to those described above (assuming causal variation to be distributed like SNPs in ENCODE regions) suggest that less than 8% of the time will the hit SNP actually be the causal SNP. This will be discussed further in Chapter 5. On the one hand, we note that there may be more reward in terms of gains in predictive ability and increases in effect size from fine mapping SNPs with lower minor allele frequency because they are, on average, more likely to be in poor LD with an unobserved causal variant. On the other hand, our simulations show that although they are unlikely to be causal, most

common hit SNPs are likely to be very good surrogate markers for their causal variant. Indeed, in 25% of cases, the hit SNP will be a near-perfect surrogate (i.e. $r^2 > 0.99$) for the causal variant. Should this be the case, further genotyping will not reveal other SNPs with stronger associations unless sample sizes are extremely large.

Here we have quantified the increased spread of genetic risk with genotypes just at known loci, and only considering a multiplicative disease model. But even in this restricted setting, there will be substantial differences in risk between high- and low-risk groups based on these genetic factors. For example, the propensity of individuals in the top 0.1% of the population distribution of genetic risk of type 2 diabetes will be increased by a factor of 5-16, compared to the average. For breast cancer, in the analogous top-risk group this risk will be increased by a factor of 3-5 (on the basis of common variation). As discussed in §3.3.1, the increased risk in the high-risk groups for Crohn's and breast cancer are likely even higher, while the increased risk in the high-risk groups for type 2 diabetes is potentially overestimated by our method. Importantly, with the growth of GWAS findings, both in terms of numbers of diseases and numbers of loci for particular diseases, more and more of the population will be in this most-at-risk category for at least one disease: assuming 100 independent diseases, nearly 10% of the population will be in the top 0.1% of risk of at least one disease. Knowing which individuals these are, and what diseases they are most at risk of, is therefore potentially useful information, both to the individual and at the population level. The issues involved in utilising such information in screening programmes (discussed for example in Pharoah *et al.* (2008)) are complicated, but our results strengthen the arguments for consideration of this possibility.

We have shown that some of the missing heritability for common disease actually resides in known GWAS loci and have estimated this deficit for three particular diseases. While rather more heritability is likely to be explained by known GWAS loci than has been reported, this effect alone falls short of explaining all the missing heritability,

particularly in Crohn's disease. Note, however, that there are other reasons why existing loci may explain more heritability than currently thought. One reason is that power to detect small effects is typically low. Recent papers suggest that correcting for limited power can contribute substantially to the heritability explained by GWAS loci (Park *et al.*, 2010; Yang *et al.*, 2010a). In addition, current calculations (by others, and above) focus on a single causal variant in each associated region: more variants within regions will explain more heritability. They also ignore possible non-additive disease effects, and interactions between variants at different loci, topics which will be touched on in Chapter 6. Power to detect either is low (Wellcome Trust Case Control Consortium, 2007), so it is misleading to put much weight on the failure of existing designs to find such effects. As others have noted (Maher, 2008), parts of the missing heritability could be due to multiple rare variants of large effect, associations with other forms of genetic variation such as copy number polymorphisms, and epigenetic effects. Indeed it would be surprising if each did not play some role. Another possibility is that estimates of the "genetic" component of disease susceptibility, from epidemiological studies, confound shared environment with shared DNA, and so inflate heritability estimates (Clayton, 2009; Wallace & Clayton, 2003).

4

The influence of different SNP chips, different populations, and imputation on effect size estimates

Two primary features of GWAS design are the population in which the study is conducted and the genotyping chip used for sequencing. As discussed previously, an important factor in the success of a GWAS design is the pattern of linkage disequilibrium, and the consequent ability of genotyping chips to provide surrogates for common variation. Genotyping chips capture common variation differently, depending on the way in which SNPs on the chip were selected, and likewise different populations have different patterns of linkage disequilibrium. Thus the choice of population together with the choice of chip influences the outcome of a GWAS, even if the causal variants have the same properties; in practice, the causal variants may in fact have different allele frequencies or effect sizes in different populations.

In Chapter 3 we focussed on one population (CEU) and one genotyping chip (Affymetrix 500k). In this chapter we ask how varying the genotyping chip and the population influences the effect size underestimation observed in the previous chapter. We show that estimates of the size of the genetic effect based on the best SNP from the genotyping chip can often closely approximate the effect size at the true causal SNP, as before. However, in those cases in which the causal SNP has a large effect size and is poorly tagged, the choice of genotyping chip and population can influence the extent of the underestimation of the true effect size. Others have examined the ability of different genotyping chips to detect causal loci (see Eberle *et al.* (2007) and Spencer *et al.* (2009)) but have not addressed the issue of effect size underestimation.

Within the constraints of a given population and chip, a GWAS design can potentially be enhanced by the use of imputation. Genotype imputation has become a widely used tool in GWAS studies to boost power, fine-map associations, and combine results across different genotyping chips for the purpose of meta-analysis (Marchini & Howie, 2010). By probabilistically "filling in" data at ungenotyped SNPs, imputation offers a way to test more markers in a GWAS, and we hypothesized that this might mitigate effect size underestimation. We conducted a modified simulation study in which we used imputation to obtain genotypes at all of the SNPs in the ENCODE regions and show that the use of imputation improves effect size estimation modestly in the CEU population with the Affymetrix 6.0 chip and substantially in the YRI population with the Illumina 550 chip.

Spencer *et al.* (2009) use HAPGEN, in a similar experimental design to the one we employ here, to compare the performance of a variety of chips with and without imputation in detecting variants of different effect sizes and allele frequencies. There are two differences to the analysis discussed in this chapter, aside from the diverse objectives of the studies discussed above. The first difference concerns the regions in which the simulation study was conducted: whereas Spencer *et al.* (2009) use randomly

drawn regions of the genome, we specifically use the ENCODE regions (see Chapter 2). Because the ENCODE regions are more densely genotyped than regions of the genome at large, this allows us to better assess the potential contribution of causative alleles in the lower frequency end of the allele spectrum. The second difference is the variety of genotyping chips we choose to explore. Although the specific chips we chose in our study differ from those in Spencer *et al.* (2009), the same trends (chip density, and selection of markers based on LD or not) affect our results.

4.1 Different SNP chips

As GWAS have developed, so too have the genotyping chips with which they are conducted. Some genotyping chips (in particular, those produced by Illumina) are designed with particular patterns of LD in mind; others (specifically some produced by Affymetrix) are selected by a process that turns out to be nearly random (Spencer *et al.*, 2009). Consequently, different chips each have their own effects on the observed GWAS results. Here we compare effect size estimates across multiple SNP chips for simulations conducted in the CEU population.

Figure 4.1 shows the distribution of estimated effect sizes at the replicated hit SNP for three different values of the true effect size and four different genotyping chips. All of the distributions have peaks around the true effect size assumed at the causal SNP. However, as observed in the previous chapter, there is often underestimation of the true effect size, and this underestimation can be substantial, especially when the true effect is large. As seen in Figure 4.1, the pattern of underestimation depends on the genotyping chip. Because our simulations are conducted in the ENCODE regions, the comparisons between genotyping chips should be considered with attention to their densities in the ENCODE regions and not genome-wide. For instance, the Affymetrix 6.0 and Illumina 550 chips have similar numbers of SNPs in the ENCODE regions

(see Table 2.2), and correspondingly similar distributions of estimated relative risk, as evidenced in Figure 4.1. Genome-wide, we would expect the distribution of estimated relative risks for the Affymetrix 6.0 chip to closely resemble that of the Affymetrix 500k in Figure 4.1, while the distribution for the Illumina 550 would show even more underestimation.

When the true relative risk is 4, the estimated effect size was less than 2 in 11% of the simulations in which GWAS successfully discovered the effect with the Affymetrix 500k chip; this percentage was 6% for the Affymetrix 6.0 chip, 5% for the Illumina 550, and 3% in simulations with the Illumina 1.2M chip. As we saw in Chapter 3, the underestimation is due to imperfect tagging, but different chips show various patterns of tagging. In Figure 4.2 we plot the relative under- (or over-) estimation of the effect size (estimated effect size divided by the true effect size) as a function of the correlation between the hit SNP and the true causal variant. Substantial underestimation can occur in cases when significant and replicable associations are detected at tag SNPs with $r^2 \approx 0.2$ with the causal variant (Figure 4.2, RR=4, Illumina 1.2M chip).

The similar pattern of the plots shown in the columns of Figure 4.2 reveals an important feature of the results, namely, that effect size underestimation is driven by LD. For a given value of the correlation between the hit and the causal SNP, the effect size underestimation is the same across genotyping chips. The differences in observed effect size distributions shown in Figure 4.1 are due not to different underlying mechanisms, but to variations in the properties of the marker SNPs on the particular chips. For example, if we choose an arbitrary causal SNP, the closest proxy will be farther away if we conduct the GWAS with the Affymetrix 500k chip rather than the Affymetrix 6.0 chip, and so there is more underestimation, evidenced in Figure 4.1. But the distributions in Figure 4.2 for the Affymetrix 500k and Affymetrix 6.0 chip are remarkably similar in their overall patterns. The Affymetrix 500k plots have more points at lower values of the correlation between the hit and causal SNP than the

Affymetrix 6.0 plots because, for a given causal SNP, there is more likely to be a better proxy in the Affymetrix 6.0 chip than the Affymetrix 500k chip.

4.2 Different populations

GWAS have, as of this writing, predominantly focussed on populations of European ancestry, with populations of Asian and African ancestry generally underrepresented (Yang *et al.*, 2010b). In addition to the chip used in a GWAS, another factor that influences GWAS results is the population in which the study is conducted. Since patterns of linkage disequilibrium and minor allele frequency distributions differ among populations (International HapMap Consortium, 2005), in addition to the clinical manifestations and/or the prevalence of many disease phenotypes, we consider the impact of the underlying population in observed effect size estimates with the Affymetrix 6.0 chip. We chose the Affymetrix chip for the cross-population comparison because SNPs on this chip were not as biased towards the CEU population as the Illumina chips considered in the previous section (Spencer *et al.*, 2009).

In §4.1, the underlying population, and thus the number of potential causal variants, was fixed. By varying the population, we are now considering two factors that influence effect size underestimation: (1) the LD patterns between the causal variants and their markers on the genotyping chip, as before, and (2) the underlying distribution of causal variants. Figure 4.3 shows the allele frequency distributions in each of the three HapMap populations considered in our simulations. Note that there are a greater number of lower-frequency variants (<5%) in the YRI population (International HapMap Consortium, 2005). We expected that effect size underestimation would be greater in the YRI population because lower-frequency variants tend to be more poorly tagged by the genotyping chip (Spencer *et al.*, 2009).

This expectation is borne out in Figure 4.4, which shows the distribution of esti-

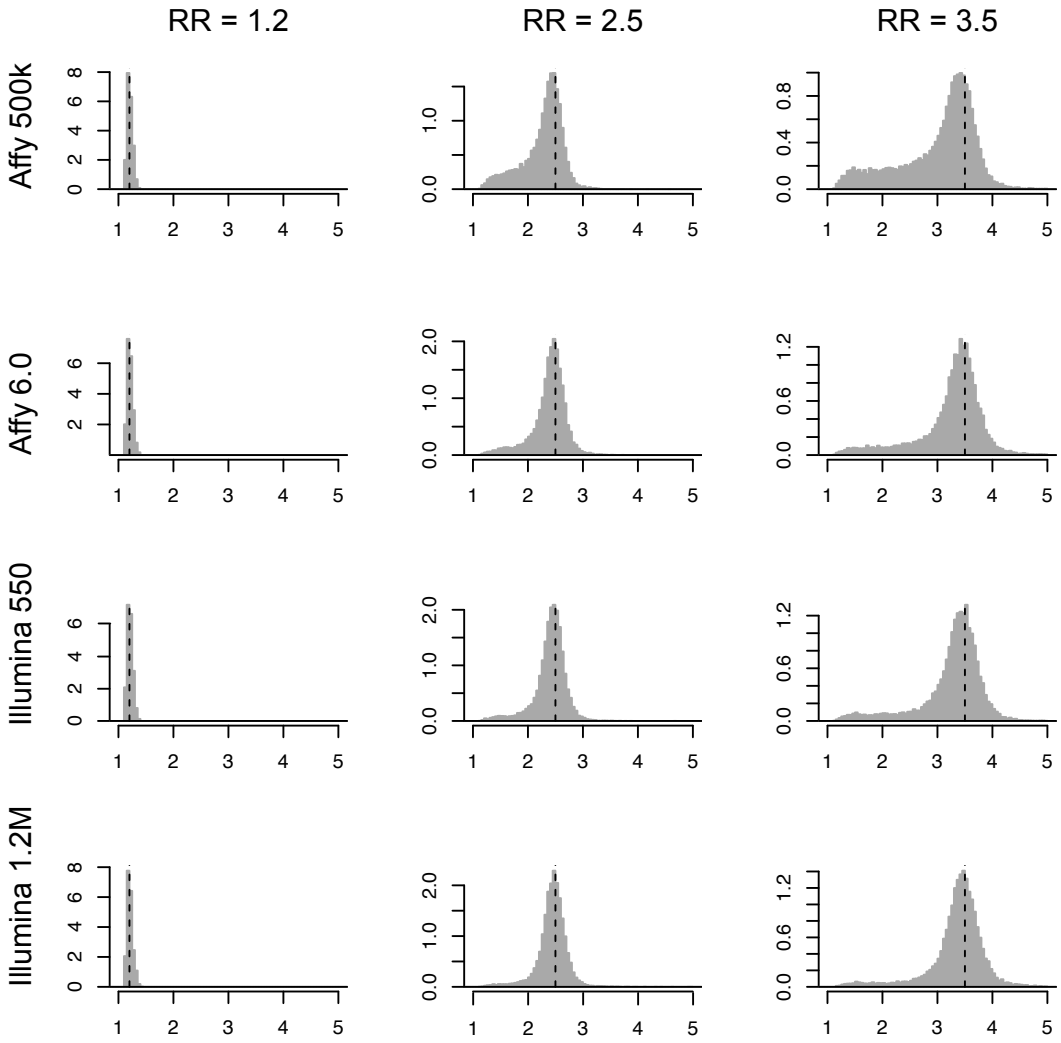


Figure 4.1: Distribution of estimated effect sizes for different genotyping chips - Histograms of estimated relative risks (RR), for three different true relative risks indicated by a vertical dashed line in each plot and four different genotyping chips. Histograms include all simulations where the most associated (hit) SNP was significant in both the initial and the replication study. Density is represented on the y-axis.

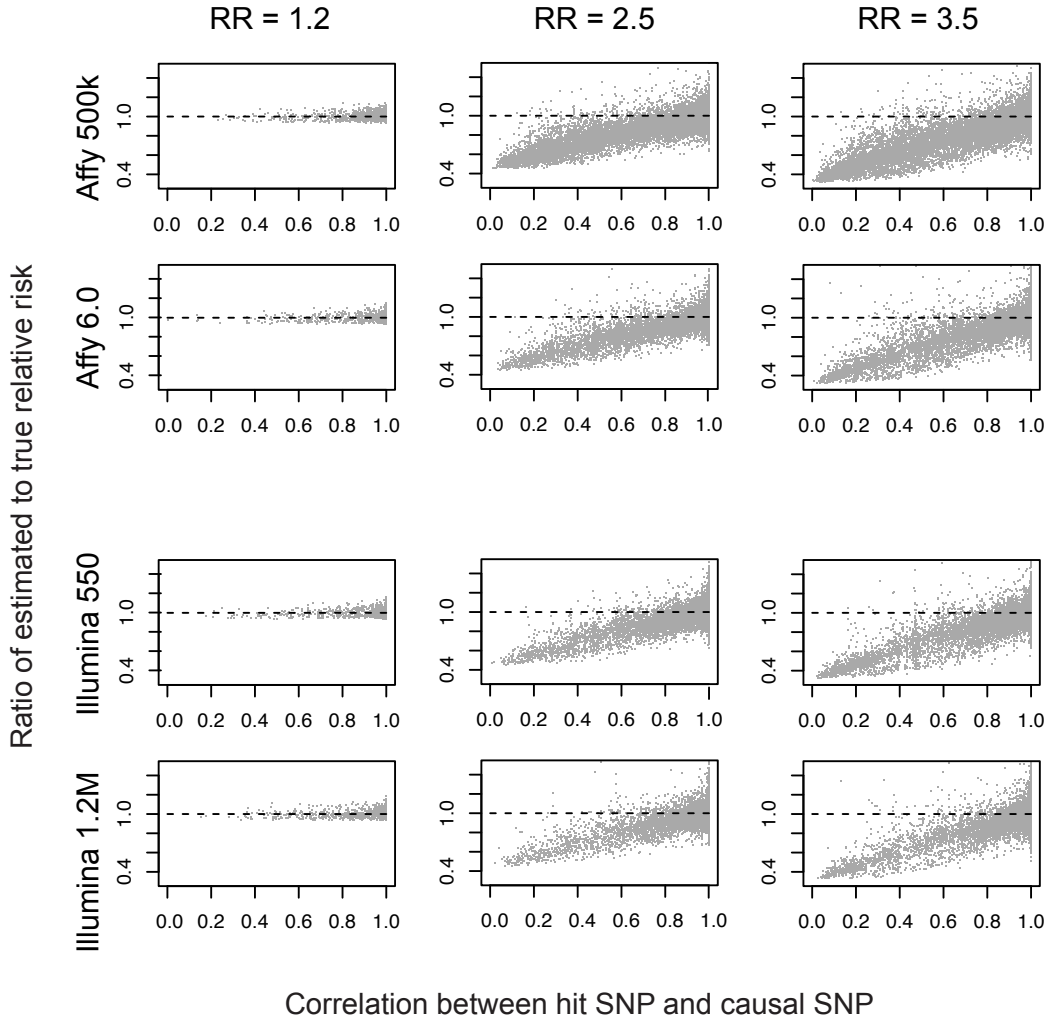


Figure 4.2: Relationship between underestimation and correlation for different genotyping chips - Scatter plots of the ratio of estimated to true relative risk against the correlation (r^2) between the disease SNP and the hit SNP, for three different relative risks and four different genotyping chips. The horizontal dashed line indicates a ratio of 1. Points below the line are underestimated, and above the line, overestimated. (Note that in the case where the hit SNP is also the causal SNP the correlation is 1.)

mated effect sizes at the replicated hit SNP for three different values of the true effect size and the three HapMap populations. Particularly in the YRI population, the underestimation can be substantial (see Figure 4.4). For example, when the true relative risk is 4, the estimated effect size was less than 2 in 6% of the simulations in the CEU and JPT+CHB panels, compared with 14% in the YRI panel. There is also notable underestimation in the YRI population when the true relative risk is 2: the estimated effect size is less than 1.5 12% of the time, compared with 4% of the time in the CEU and JPT+CHB panels. Therefore we expect that GWAS conducted in populations with African ancestry will show greater effect size underestimation, while GWAS conducted in populations with Asian ancestry will show comparable effect size underestimation to that in CEU.

Figure 4.5 shows the relative under- (or over-) estimation of the effect size (estimated effect size divided by the true effect size) as a function of the correlation between the hit SNP and the true causal variant for the same three different causal relative risks in the three populations. Despite the additional complication of comparing distributions that arise from different underlying sets of causal variants, Figure 4.5 has the same property as Figure 4.2 in §4.1: the columns of scatter plots all show a similar pattern. As before, for a given level of correlation between the causal and hit SNPs, the underestimation is the same across populations, suggesting that the effect size underestimation observed in Figure 4.4 largely owes to LD.

4.3 Imputation

Imputation has become a standard tool in GWAS, both for refining association signals and for meta-analysis (Marchini & Howie, 2010). Chapter 1 discusses the role of imputation in GWAS. In this section, we consider the potential of imputation to correct for effect size underestimation.

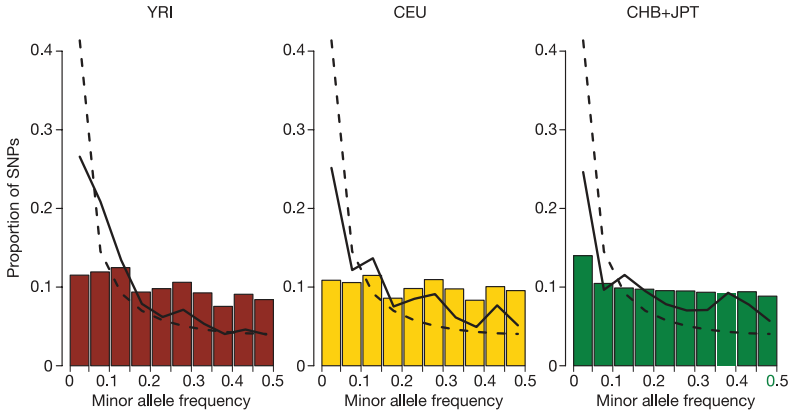


Figure 4.3: HapMap MAF distributions in three populations - Density of variants in the three HapMap populations we consider. The density in CEU (yellow), JPT+CHB (green), and YRI (maroon) is calculated by considering the number of variants in windows of width 5%. The solid line shows the MAF distribution for the ENCODE SNPs, and the dashed line shows the theoretical MAF distribution expected for the standard neutral population model with constant population size and random mating. The figure is taken from International HapMap Consortium (2005).

In practice, the implementation of imputation in a GWAS explicitly involves a phasing step. Popular imputation algorithms incorporate the phasing step into their software packages, often performing phasing and imputation together. So as not to confound the effects of phasing and imputation, we imputed into phased haplotypes rather than individuals. We therefore begin with phased haplotype data at SNPs represented on a particular genotyping chip, say the Affymetrix 6.0 chip. For concreteness, let's focus on a given ENCODE region, say ENm010. As we have seen, the Affymetrix chip contains only a subset of the full set of SNPs known in ENm010. We might wish to make an educated guess about the alleles at the unobserved SNPs. Suppose, furthermore, that a haplotype reference panel exists. In our example, this haplotype reference panel is the same panel from which we simulate our population data, namely the analysis panel from HapMap II (International HapMap Consortium, 2007) (see discussion below). As discussed in Chapter 2, the ENCODE regions show a fuller spectrum of

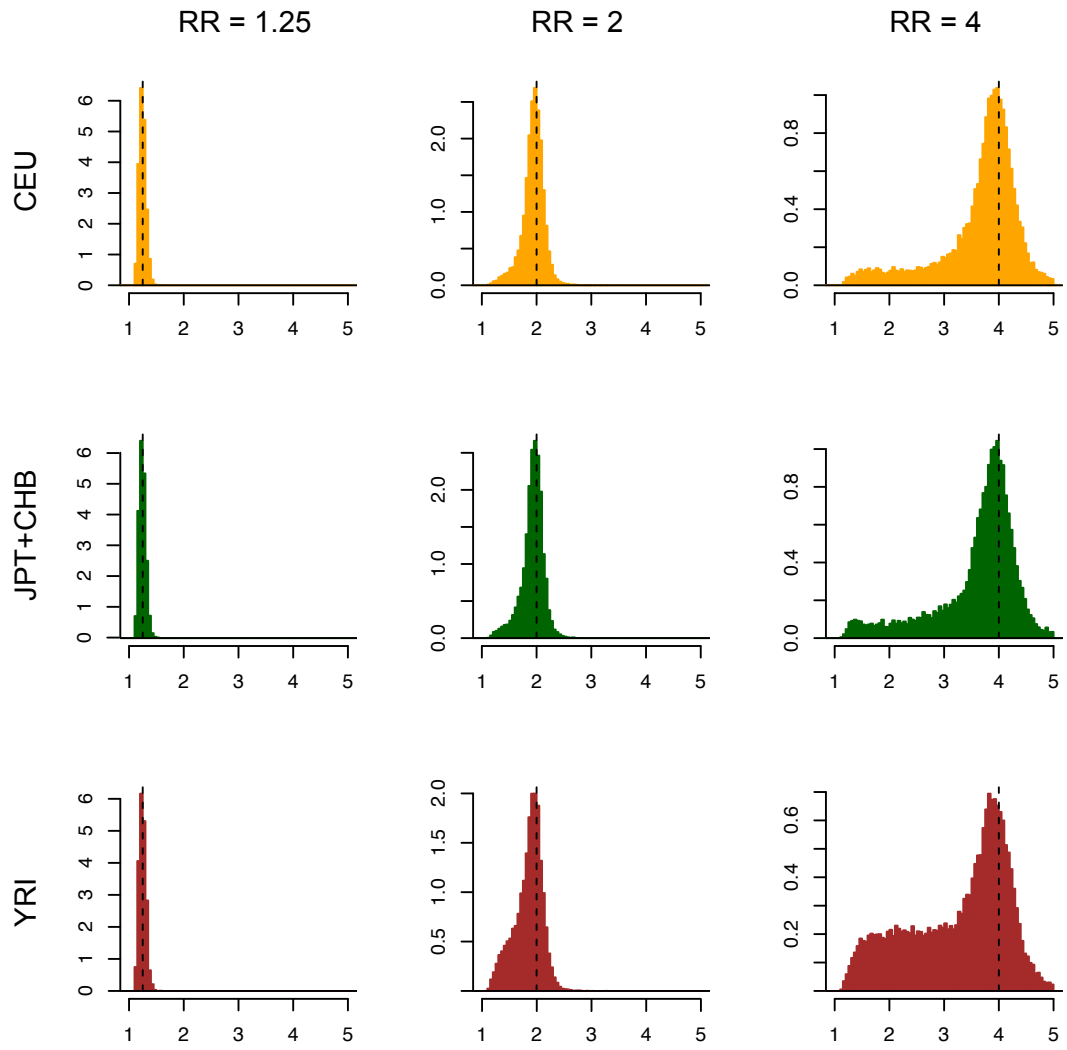


Figure 4.4: Distribution of effect sizes for different populations - Histograms of estimated relative risks (RR), for three different true relative risks indicated by a vertical dashed line in each plot and three different populations. Histograms include all simulations where the most associated (hit) SNP was significant in both the initial and the replication study. Density is represented on the y-axis.

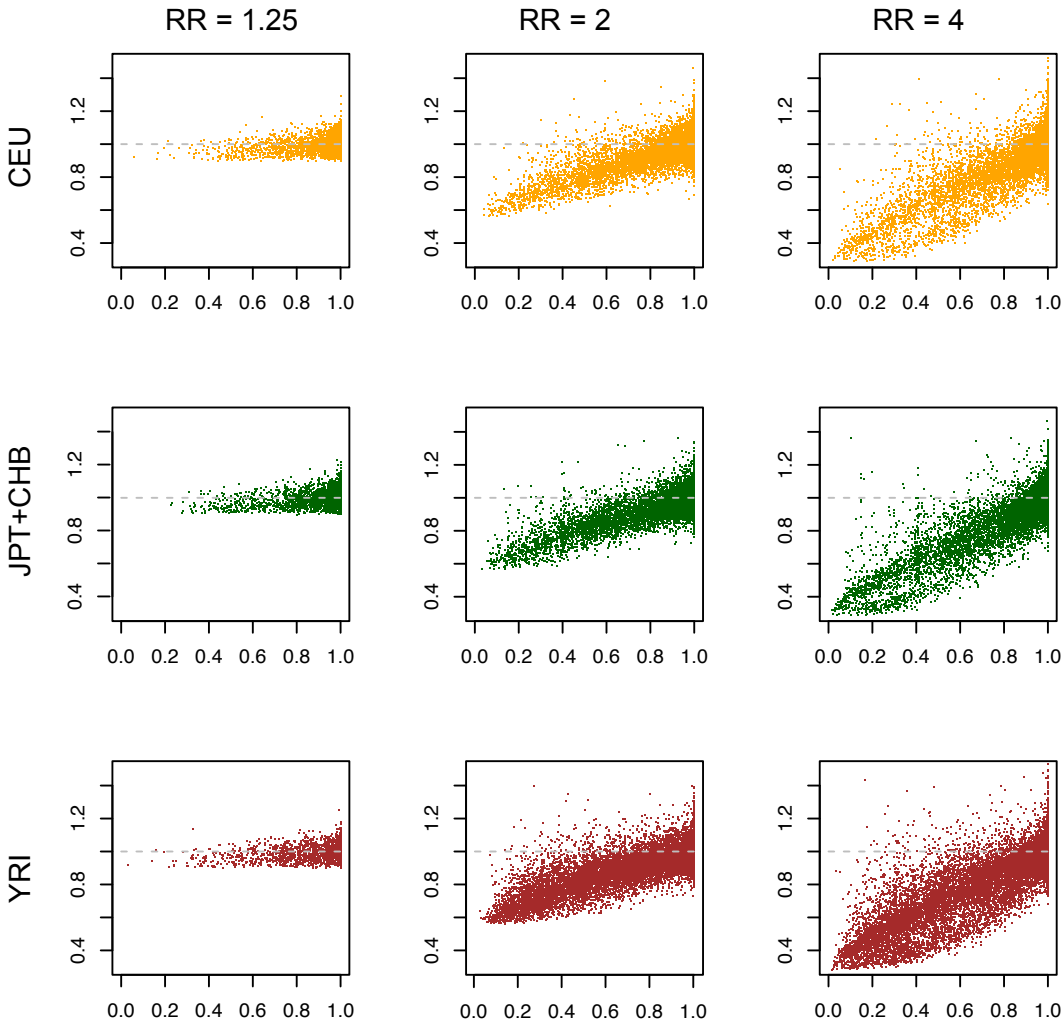


Figure 4.5: Distribution of LD between the causal and hit SNPs for different populations - Scatter plots of the ratio of estimated to true relative risk against the correlation (r^2) between the disease SNP and the hit SNP, for three different relative risks and four different genotyping chips. The horizontal dashed line indicates a ratio of 1. Points below the line are underestimated, and above the line, overestimated. (Note that in the case where the hit SNP is also the causal SNP the correlation is 1.)

SNPs than are represented in the HapMap data at large.

We conducted an experiment to mimic one type of imputation that might be performed in a typical GWAS with the 1000 Genomes data by imputing at all the markers typed in the ENCODE regions (1000 Genomes Project Consortium, 2010). There are a number of programs available for imputation, and comparisons of these methods have been described previously (Marchini & Howie, 2010; Pei *et al.*, 2008). In particular, we used the imputation software BEAGLE (Browning & Browning, 2007). We chose to use BEAGLE rather than IMPUTE, the companion software to HAPGEN, because of the concern that using the same underlying model for generating our simulation populations and imputing genotypes would bias the results.

Figure 4.6 shows the steps in the simulated imputation experiment. First the phased population panel generated by HAPGEN (see Chapter 2) was thinned to include only data at the SNPs on the genotyping chip. All of the SNPs in the ENCODE region were used for imputation. BEAGLE imputed the population data at the SNPs in the reference panel of ENCODE haplotypes. The phased output file gives the most likely haplotype pair for each individual, conditional on the genotypes from the population data together with BEAGLE's haplotype frequency model (Browning & Browning, 2007). The phased imputed haplotypes were then passed to the GWAS simulation software described in Chapter 2. Imputation quality of the causal SNP was not used as a criterion for discarding the simulation. Cases and controls were selected from the phased population panel, pre-imputation, and then the imputed genotypes were used for testing.

The way we implement imputation in the context of the simulation framework described in Chapter 2 doesn't exactly mirror how it is done in an actual GWAS experiment. We utilized imputation in the population panel, rather than imputing each case control sample individually, in order to avoid the computational intensiveness of running many parallel GWAS imputations.

The results of the imputation experiments should be considered as an upper bound on how well imputation can perform in improving effect size estimates. This is because we passed phased haplotypes, as well as a larger population, to the phasing software as described above. In practice, unphased genotypes and a smaller population will reduce the accuracy of the imputation experiment. Additionally, the Affymetrix 6.0 chip has enriched marker density in the ENCODE regions, so that an equivalent chip would have over 1.5 million SNPs genome-wide, similar to the latest chip releases.

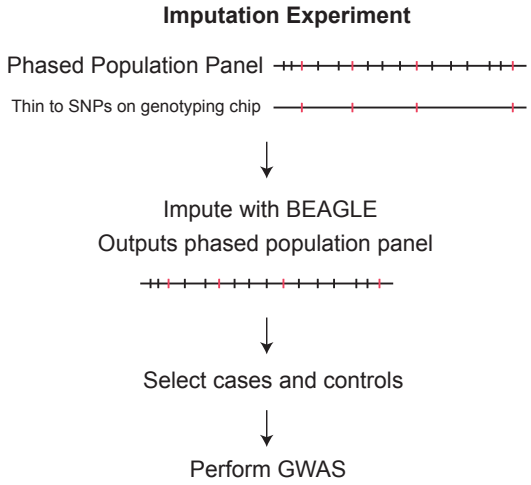


Figure 4.6: Cartoon of the imputation experiment - The figure shows the process whereby the population data, generated with HAPGEN, is imputed at the SNPs in the ENCODE regions before cases and controls are subsequently selected and a simulated GWAS is performed, as described in Chapter 2. SNPs on the genotyping chip are shown in red along the haplotype.

Figures 4.7 and 4.8 show the results from the imputation experiment with the Affymetrix 6.0 chip in the CEU population, contrasted with the results from simulations conducted without imputation. Figure 4.9 highlights estimates from simulations in which the causal and hit SNPs coincide. Note that, in this case, imperfect correlation between the causal and hit SNPs results from imputation inaccuracy. Imputation improves effect size estimation for higher causal relative risks. Table 4.1 gives the

relative number of SNPs passing the significance thresholds in the initial and replication experiments, as well as the proportion of SNPs underestimated in each experiment. In this setting, imputation modestly improves effect size estimation. Notably, Figure 4.8 reiterates the same conclusion we drew from Figures 4.2 and 4.5. Again, the similarity of the distributions for the same causal relative risks (down the columns in the figure) is striking, indicating that a given amount of correlation between the causal and hit SNP determines the degree of effect size underestimation.

The results of Figures 4.2, 4.5, and 4.8 together can be taken as evidence that the driving force behind effect size underestimation is the LD between the pool of potential candidate markers and a given causal SNP. If, for any given causal SNP, there is a pool of perfectly or nearly-perfectly correlated candidate makers, then any effect size underestimation will occur by chance. If, however, there are causal SNPs for which no good surrogates exist, then there will be effect size underestimation to the extent that diluted signals at the poor surrogates can pass initial and replication significance thresholds.

We can test this hypothesis by performing an experiment in which the population and the genotyping chip are explicitly mismatched, so that there is a significant pool of causal SNPs that are not well-tagged by SNPs on the chip. One such example is the Illumina HumanHap550 genotyping chip, which was designed with CEU patterns of variation in mind, together with the YRI population. Figure 4.10 shows the bimodal distribution of estimated relative risks that results, for high true relative risks, when the chip is designed to tag an LD distribution divergent from that of the population in the GWAS. The histogram shows a mixture of the two types of estimates that can result: (i) the distribution, similar to that observed with a chip with "random" SNP selection design, centered around the true estimate, and (ii) the distribution consistent with low LD between the hit and causal SNPs, which is enriched in YRI because of the number of low-frequency causal alleles. The second type of estimates can be removed

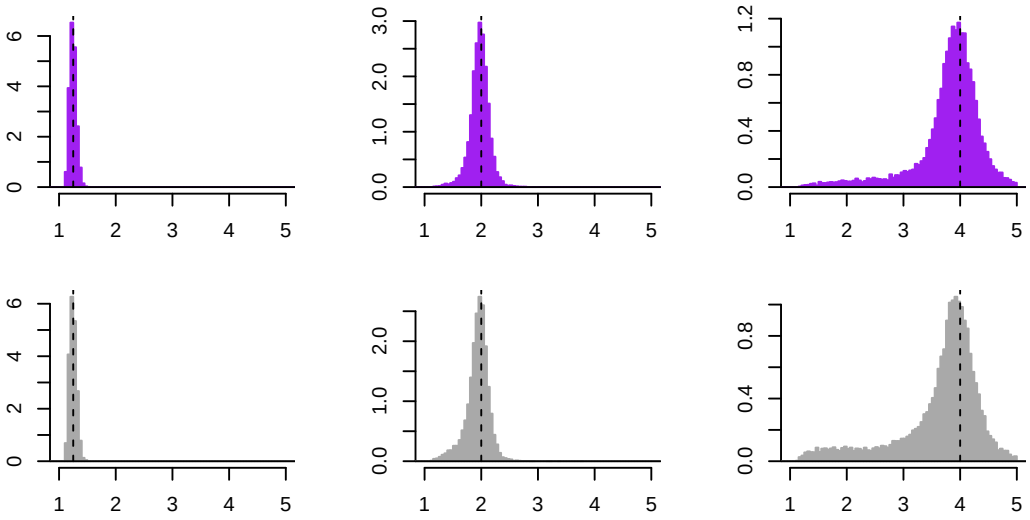


Figure 4.7: Imputation in CEU with the Affymetrix 6.0 chip - The figure shows histograms of effect sizes (RR) in the imputation experiment (purple) and without imputation (grey) for imputation with the Affymetrix 6.0 chip in the CEU population. The true relative risks (reading left to right) are 1.25, 2, and 4 respectively, as indicated by the dashed lines. As before, only simulations in which the hit SNP was both significant in the initial and the replication study are shown. Density is represented on the y-axis.

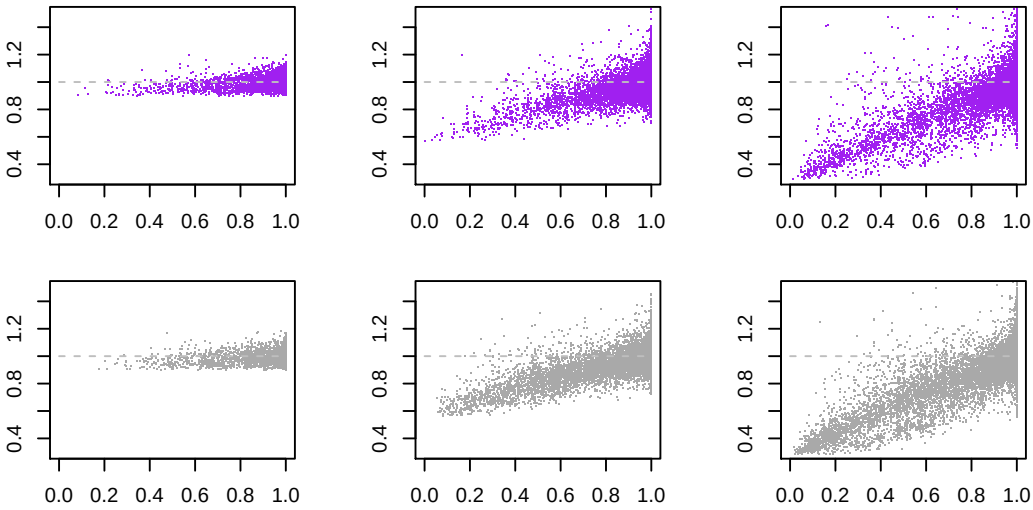


Figure 4.8: Imputation in CEU with the Affymetrix 6.0 chip - The figure shows scatters plot of LD (r^2) between the causal and hit SNPs plotted against the underestimation ratio (ratio of estimated to true relative risk) for the same relative risks and in the same settings as in Figure 4.7 above.

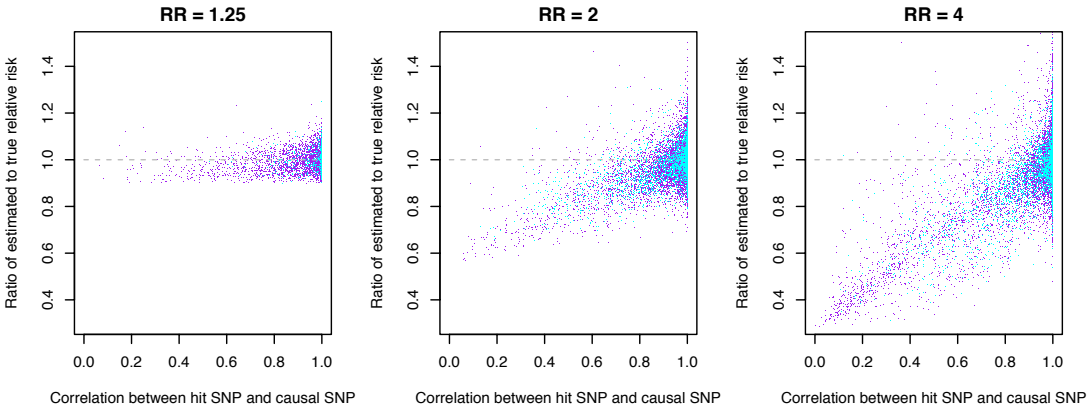


Figure 4.9: Effect size underestimation and correlation in the imputation experiment - The figure shows the same scatter plot of LD (r^2) and underestimation ratio as in Figure 4.8 for imputation in the CEU population genotyped with the Affymetrix 6.0 chip. In this figure, however, cyan points represent coincident causal and hit SNPs. The correlation between coincident causal and hit SNPs is not always 1 because the LD is measured between the causal SNP in the simulated population and the hit SNP in the imputed population.

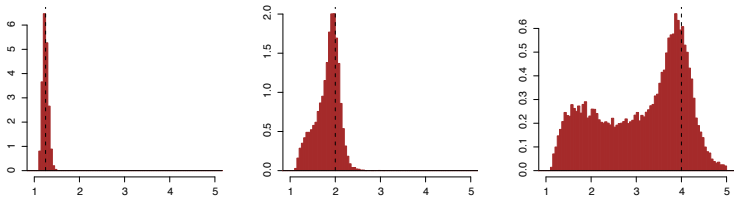


Figure 4.10: Histogram of effect sizes in YRI with the HumanHap550 chip - The figure shows histograms effect sizes (RR) for simulations with the HumanHap550 chip in the YRI population. The true relative risks (reading left to right) are 1.25, 2, and 4, as indicated by the dashed lines.

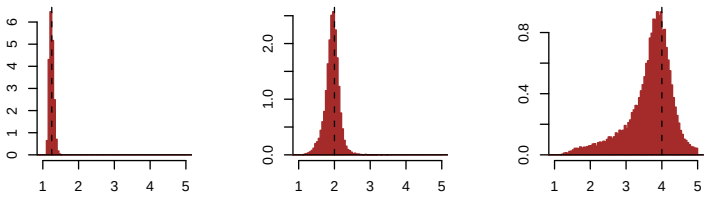


Figure 4.11: Imputation in YRI with the HumanHap550 chip - The figure shows histograms of effect sizes (RR) for simulations with the HumanHap550 in the YRI population with imputation. The true relative risks (reading left to right) are 1.25, 2, and 4, as above.

Table 4.1: Imputation Results Summary. For each of three true relative risks, the table shows the number of SNPs passing filters with and without imputation, and the proportion of hit SNPs where the relative risk estimated at the hit SNP was less than that at the causal SNP.

	RR	1.25	2	4
Affy 6.0 CEU population				
	# SNPs passing filter (imputation)	3478	16612	18704
	# SNPs passing filter (no imputation)	3239	16064	18152
	% Imputed SNPs underestimated	55%	57%	60%
	% Affy 6.0 SNPs underestimated	55%	62%	66%
HumanHap550 YRI population				
	# SNPs passing filter (imputation)	3069	18766	21966
	# SNPs passing filter (no imputation)	2232	15692	19748
	% Imputed SNPs underestimated	57%	64%	68%
	% HumanHap550 SNPs underestimated	55%	74%	80%

by imputation, as shown in Figure 4.11, which provides a significantly better pool of surrogates (namely, the imputed set of causal SNPs and those nearby SNPs in the ENCODE YRI panel) than those offered on the genotyping chip. The improvement in estimation is reflected in Table 4.1.

4.4 Conclusions

We began this chapter by asking how features of a GWAS design – the population in which the study is conducted, the genotyping chip used for sequencing, and the adoption of imputation – influence the effect size underestimation described in one particular GWAS design setting in Chapter 3. In Chapter 3 we showed that the linkage disequilibrium between the causal variant and the best-associated SNP on the genotyping chip determined the extent of underestimation for the CEU population with the Affymetrix 500k chip. In this chapter we extended that conclusion to three additional genotyping chips (Affymetrix 6.0, Illumina HumanHap550, and Illumina Human1.2M), to two additional populations (JPT+CHB, YRI), and in the setting of imputation. The

similarity between the results shown in Figures 4.2, 4.5, and 4.8 establishes that the mechanism behind effect size underestimation, in all of the different design contexts considered here, is the set of correlations between the pool of causal variants and the pool of markers tested for association.

This mechanism explains the observed differences in effect size distributions in Figures 4.1, 4.4, and 4.7. As we noted before, the correlation between alleles along the human genome differs among populations. Furthermore, genetic variation is tagged differently by various genotyping chips, each of which represents a different subset of variation that is used to predict untyped diversity. The main trends in Figures 4.1, 4.4, and 4.7 can be summarized as follows: (i) denser genotyping chips are less susceptible to effect size underestimation; (ii) GWAS conducted in populations of African ancestry are expected to exhibit greater effect size underestimation than GWAS conducted in either Caucasian or Asian populations; and (iii) imputation can improve effect size underestimation. Each of these trends follows from the LD properties of the pool of causal variants and the pool of markers, since (i) denser genotyping chips have the feature that, for a given causal variant, the best surrogate is closer (by genetic distance) to the causal variant than the best surrogate on a less dense chip; (ii) the greater number of lower-frequency variants in YRI increases the pool of poorly-tagged causal variants, making underestimation more likely; and (iii) imputed genotypes provide a pool of better surrogates (as measured by LD between the surrogate and causal SNPs) than the markers on a genotyping chip. Together these trends imply that there may be comparatively greater benefit to fine mapping signals discovered in GWAS with less dense genotyping chips or in African ancestry, and suggests that effect size estimates should be computed and reported from imputed genotypes in addition to the common practice of reporting effect sizes estimated from the replication sample.

Finally, we considered the ability of imputation to ameliorate effect size underestimation, and showed that its impact depends on the combination of chip density, design,

and population. Imputation has a modest impact for the CEU population. In YRI, imputation has a greater benefit, and can be particularly beneficial when the genotyping chip is designed for use in another population. Although we would hope studies in populations of African ancestry would avoid the use of such chips, imputation provides a way to "rescue" the significant effect size underestimation that results from poor tagging in this instance. In general, we demonstrated that if the SNP has surrogates, conditional on being discovered in a GWAS study, imputation goes some way to reversing the diminution of effect size observed at the hit SNP. The poorer the pool of available surrogates on the chip, the more dramatic is this impact.

In all, our findings are consistent with the broad conclusions reported in Spencer *et al.* (2009), while extending the approach to the question of effect size estimates.

Although obtaining accurate effect size estimates is an important goal, the ideal outcome of a GWAS is to identify the causal SNP directly. Unlike accurate effect size estimates, which can be obtained at any perfectly or nearly-perfectly correlated SNP, identification of a causal SNP by considering only the hit SNP in a GWAS requires the best-associated SNP to actually *be* the causal SNP. In the next chapter, we go on to ask how often we expect the best-associated SNP in a GWAS to be coincident with the causal SNP.

5

The relationship between the causal SNP and the hit SNP

The relationship between the causal SNP and the hit SNP – for instance, whether they coincide, and if they do not, their genetic distances from one another – is a key unknown in GWAS results. Ultimately, the goal of a GWAS is to point to variants of biological relevance in the disease or trait under study. As we discussed in Chapter 3, a predictable downside of exploiting patterns of linkage disequilibrium in the genome to identify regions associated with disease is that the associated SNPs reported in GWAS may not be causal variants themselves, but surrogates for the true causal variants.

In this chapter we quantify how often the causal and hit SNPs coincide in our simulations and discuss the implications of these results for the problem of fine mapping association signals. The analyses herein utilise the results of the simulation study described in Chapters 2 and 4. We show that, in GWAS conducted with a variety of chips in the CEU population, the causal variant is unlikely to be identified by association, although the hit SNP is often perfectly or nearly-perfectly correlated with the causal variant. Thus, although effect size estimates can be significantly improved by increasing the density of the genotyping chip or by a chip design that exploits LD patterns

in the CEU population, these changes do not dramatically alter the proportion of the time that the hit and causal SNPs will coincide.

5.1 Proximity between causal and hit SNPs

There are two main measures we will consider here to assess the proximity between the causal and hit SNPs. The first is LD, measured as before by r^2 between the causal and hit SNPs. We ask, for a variety of r^2 thresholds, in what proportion of simulations the LD between causal and hit SNPs exceeds a given threshold. The second measure is genetic distance, measured in centimorgans (cM). We ask, in analogy to our r^2 thresholds, in what proportion of simulations the causal and hit SNPs lie within a given genetic distance of one another.

5.1.1 Causal SNPs on the genotyping chip

To inform our intuition, we first examined simulation results in which the causal SNPs were represented on the genotyping chip. Figure 5.1 illustrates the possible scenarios that can arise when the causal SNP is on the genotyping chip. From the perspective of maximizing the number of coincident causal and hit SNPs, an ideal genotyping chip would have the property that, for a given causal SNP, there was only one perfect surrogate on the chip, and all other SNPs were imperfectly correlated. In practice, the genotyping chips do not have this property. For instance, of the SNPs on the Illumina 550 chip, less than one-third have only one perfect surrogate (as shown in Figure 5.1), and there are even fewer such SNPs on the Illumina 1.2M chip (27% in the Illumina 1.2M compared with 29% in the Illumina 550). This means that even when the causal SNP is on the genotyping chip, it is competing with other perfectly correlated SNPs for the lowest p-value in an association study, and as the number of perfect surrogates increases, the chance that the hit SNP and causal SNP will be coincident decreases.

(Note that we measure the amount of correlation in the simulated population panel.) Thus there will be slight differences among the chips in the percentages of coincident causal and hit SNPs, depending on the distribution of perfect surrogates. One can extend this argument when comparing the proportion of coincident causal and hit SNPs across populations with the same genotyping chip, since the different populations will have different numbers of perfectly correlated SNPs on the same chip. We note that in addition to perfectly correlated SNPs competing with the causal variant for association, imperfectly but very well-correlated SNPs can also compete, although they will have a smaller impact on the results.

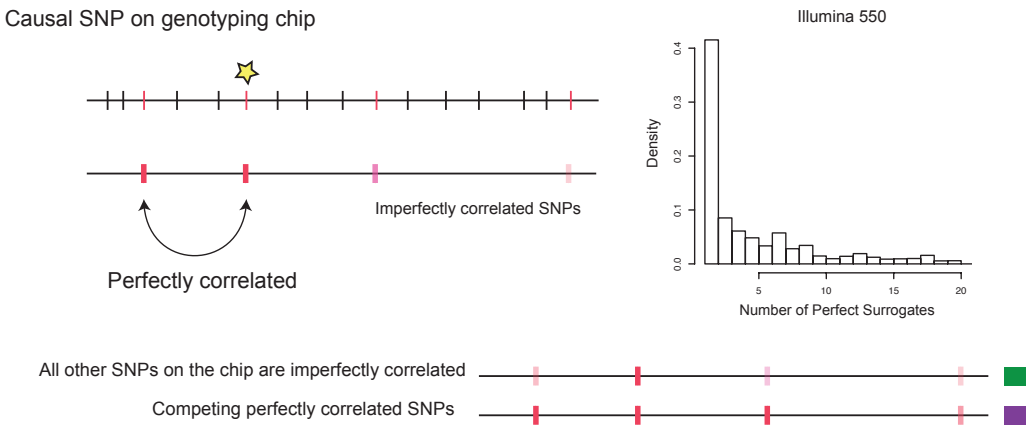


Figure 5.1: Cartoon of the possible relationships between the causal SNP and the hit SNP when the causal SNP is on the genotyping chip - In the cartoon, the causal SNP is represented by a yellow star above the haplotype on which it sits. The black hatch marks represent SNPs in the ENCODE regions, the red hatch marks SNPs on the genotyping chip in question. The histogram shows the distribution of perfect surrogates for the Illumina 550 genotyping chip. Below the histogram two scenarios are illustrated: one, in which there are no perfect surrogates on the chip (green) and another, in which there are two competing perfect surrogates (purple).

The simulation results, for four different genotyping chips and in three different populations, are shown in Figure 5.2. The non-centrality parameter of the trend test depends on the relative risk, so as the relative risk increases, the power to identify the causal SNP improves so long as the SNP is on the genotyping chip, as is the case here.

(The relationship between the power and the effect size parameter will be discussed further in Chapter 6.) This is illustrated by the fact that, for a given genotyping chip and population, increasing relative risk increases the proportion of simulations in which the causal and hit SNPs coincide (Figure 5.2).

As noted in Chapter 4, changing the population in which a GWAS is conducted alters both the LD patterns between the causal variants and their markers on the genotyping chip and also the underlying distribution of causal variants. The YRI population has a greater number of lower-frequency variants than the CEU and JPT+CHB populations (International HapMap Consortium, 2005), and these lower-frequency variants are more poorly tagged by the genotyping chip (Spencer *et al.*, 2009). Thus, on average, a causal variant in YRI on the Affymetrix 6.0 chip has fewer perfect surrogates on the chip, and correspondingly fewer SNPs competing with the causal variant for association. Table 5.3 illustrates that, as a consequence, the YRI population shows a greater proportion of simulations in which the causal and hit SNPs coincide when the causal SNP is on the genotyping chip than in the other two populations. Note that there is a lower bound on the MAF of the causal variants contributing to the percentages of coincident and perfectly-correlated hit SNPs. This is because, if the MAF of the causal variant is too low, there will be insufficient power to see an effect and hence no relevant simulations.

Because the selection of SNPs on the chips in the ENCODE regions differs from that in the genome at large (see Table 2.2), the variation represented in Figure 5.2 among chips should not be taken as representative. For example, in general, Illumina chips tend to select only one among many perfectly correlated SNPs. This results in higher percentages of coincident causal and hit SNPs among causal SNPs on the chip than the proportions observed here. Despite this limitation in our analysis, it is nonetheless useful to observe the trend noted earlier, namely that, if one picks a chip, the percentage of simulations in which the hit and causal SNPs coincide increases as

the relative risk increases.

For the remainder of the chapter we relax the restriction that the causal SNPs be on the genotyping chip and, as before, consider all SNPs in the ENCODE regions as potentially causal. The proportion of simulations in which the hit and causal SNP coincide is no longer necessarily linear with increasing relative risk, as it was when we only considered causal SNPs that were on the genotyping chip. This is because there is now a competing pool of causal SNPs *not* on the genotyping chip that nonetheless are detected by surrogates on the genotyping chip that pass significance thresholds. Depending on the specific properties of the causal SNP relative to the genotyping chip and the population, this pool can be of variable size, but as it grows, it will diminish the percentage of simulations in which the causal and hit SNPs coincide.

In the remainder of the chapter we ask how changing a number of factors – relative risk, genotyping chip, and population – affects the percentage of coincident and correlated hit and causal SNPs in our simulations, and compare the proximity of the hit and causal SNP in simulations with different genotyping chips. As in Chapters 3 and 4, the results should be interpreted in light of the marker density enrichment in the ENCODE regions (see Table 2.2). Corresponding GWAS results would arise from SNP chips with 2-2.5 times as many markers genome-wide as those we consider here. For example, simulations with the Affymetrix 500k chip could be interpreted to represent the outcomes of a GWAS study conducted in the Affymetrix 6.0 chip, since the density of markers on the Affymetrix 500k chip in the regions we simulated closely matches that of the Affymetrix 6.0 chip in the genome at large. Similar adjustments in interpretation apply to the other genotyping chips, and are discussed below. Finally, we discuss the implications of these results for fine mapping.

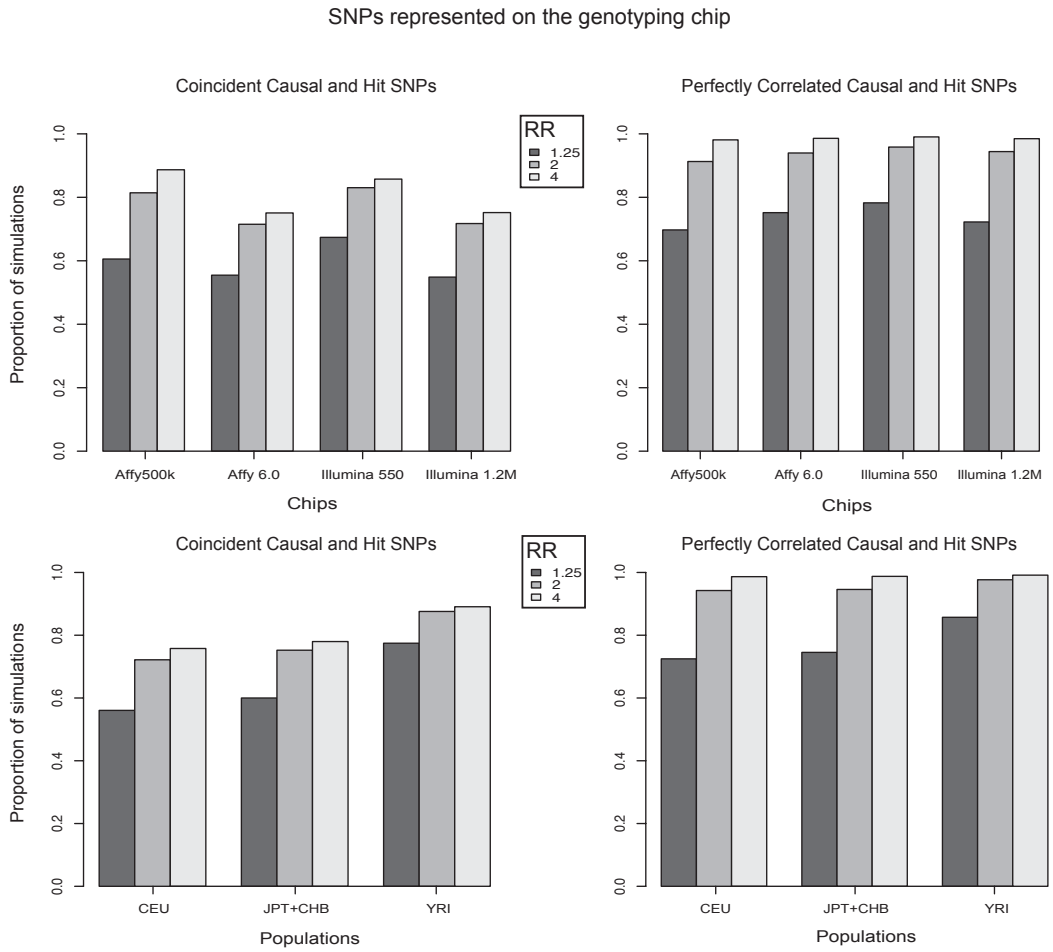


Figure 5.2: Proportion of simulations in which the hit and causal SNPs are coincident and highly correlated when the causal SNPs are on the genotyping chip - Each column in the bar chart shows the proportion of simulations in which the causal and hit SNPs are coincident or almost perfectly correlated ($r^2 > 0.99$) for relative risks of 1.25, 2, and 4 (see figure legends) in the four genotyping chips in CEU and in the three populations with the Affymetrix 6.0 chip. In each of the simulations, the causal SNP is on the genotyping chip.

5.1.2 The impact of relative risk and sample size

Next we examined simulations in the Affymetrix 500k chip to see to what extent the relative risk of the causal variant and the sample size of the GWAS influenced the percentage of hit SNPs within a given value of the correlation (measured by r^2) with the causal SNP. The results are shown in Table 5.1. In addition to hit SNPs that have $r^2 = 1$ with the causal SNP, we also looked at hit SNPs that have $r^2 > 0.99$ with the causal SNP. These second class of hit SNPs are effectively perfect surrogates, and represent either (i) r^2 measured with error, (ii) one or a small number of recombinants, or (iii) genotyping errors. (We measure r^2 between the causal and hit SNPs in the population cohort rather than in HapMap.) Notably, neither the relative risk of the causal variant nor the sample size of the GWAS have much influence on the percentage of coincident or perfectly-correlated hit SNPs in simulations with the Affymetrix 500k chip.

The percentage of coincident hit SNPs is consistently 7-8%, while the percentage of perfectly-correlated hit SNPs is consistently 24-30% across relative risks and sample sizes. Thus, while we would expect the optimum outcome of a GWAS, in which the hit SNP *is* the causal SNP, to be relatively infrequent with the Affymetrix 500k chip, the situation of nearly-perfect or perfect LD between the causal and hit SNPs is more likely.

In the next section, we consider how the choice of genotyping chip influences the percentage of hit SNPs within a given value of correlation of the causal SNP, as well as the genetic distances between the causal and hit SNPs. Because, as discussed above, the relative risk of the causal variant and the size of the GWAS have little impact on these measures (data not shown), for future comparisons we will consider simulations in which the causal variant has a relative risk of 1.25, 2, and 4, and for which the GWAS is conducted in an initial sample of 2000 cases and 2000 controls, and a replication study with 2000 cases and 2000 controls.

5.1.3 Comparisons in the CEU population between different genotyping chips

To compare the four genotyping chips, we calculated in what percentage of simulations the hit SNP and causal SNP are correlated above a given r^2 threshold in the CEU population. The results are shown in Table 5.2 and graphically in Figure 5.3. The table and figure reveal differences among the chips.

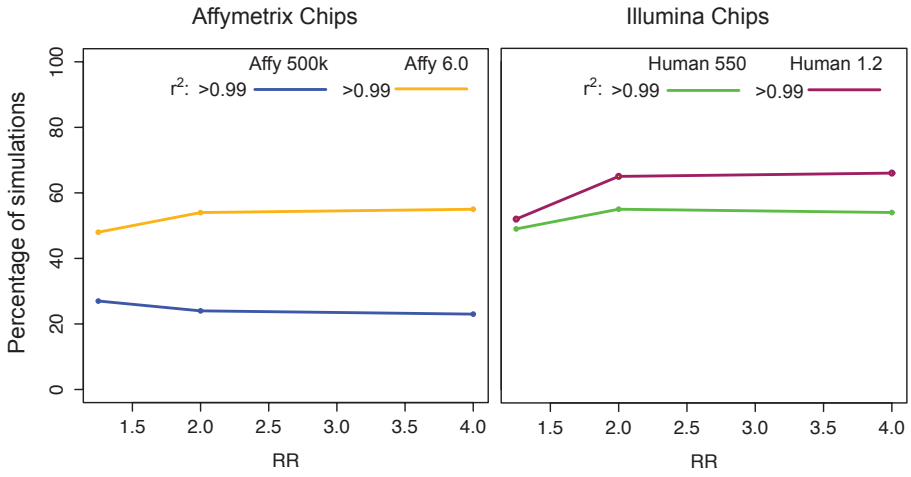


Figure 5.3: Percentage of hit and causal SNPs that are coincident and highly correlated in simulations with four genotyping chips - Each line shows the percentage of simulations in which the causal and hit SNPs are coincident or nearly-perfectly correlated ($r^2 > 0.99$) with the Affymetrix chips (left panel) or Illumina chips (right panel).

The Illumina chips consistently show a greater percentage of both coincident and perfectly-correlated hit SNPs. Less common causal SNPs (MAF $< 20\%$) are only slightly less likely to be coincident with the hit SNPs arising from GWAS. The extent to which this observation holds depends on the genotyping chip, and, to a lesser extent, on the relative risk of the causal variant. The Affymetrix 500k chip has the lowest percentage of coincident causal and hit SNPs, followed by the Affymetrix 6.0, the Illumina 550, and finally, the Illumina 1.2M. This trend is also observed in the percentage of hit SNPs that are perfectly correlated with the causal SNP. The Illumina

550 and Affymetrix 6.0 chips have similar numbers of SNPs in the ENCODE regions, but the Affymetrix 6.0 has a higher density genome-wide. Thus we would expect the Affymetrix 6.0 to perform better than the Illumina 550 in a real GWAS.

Increasing relative risk has a small impact on the percentages of coincident and nearly-perfectly correlated causal and hit SNPs, and the direction of the impact is not always consistent (Figure 5.3). This owes to the fact that, at higher relative risks, a greater number of less-well-correlated SNPs will pass significance thresholds. Depending on the particular chip and the minor allele frequency of the causal variant, this can either decrease the number of coincident and nearly-perfectly-correlated causal variants (Illumina 1.2M, low MAF; Affymetrix 500k, all MAF) or increase the number of such variants (Affymetrix 6.0), but in either case the effect is small.

Genetic distance is used to compare another aspect of the relative merits of the four genotyping chips. Comparisons of the proximity of the four chips measured by the proportion of hit and causal SNPs within a certain genetic distance of one another are shown in Figure 5.4. The 0.1cM threshold was chosen to coincide with the width of the fine mapping regions following the WTCCC study (Chris Spencer, personal communication). Note that, as shown in Figure 5.4 panel (a), there is a slight but consistent downward trend in the percentage of simulations for which the hit and causal SNP are within 0.1cM of one another as the relative risk increases, conditional on the hit SNP passing p-value thresholds in the initial GWAS and subsequent replication. Panel (b) shows that this trend is amplified in the Affy 6.0, Illumina 550, and Illumina 1.2M chips for causal SNPs with minor allele frequency $< 10\%$. Panel (c) shows the proportion of hit and causal SNPs within 0.01cM of one another. Panel (a) of Figure 5.4 demonstrates that in nearly all of the simulations passing significance thresholds the causal SNP will be within 0.1cM of the hit SNP.

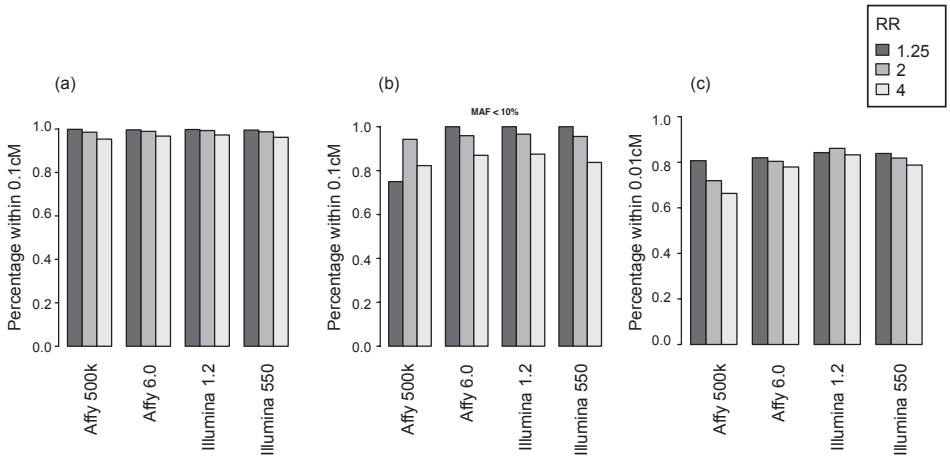


Figure 5.4: Bar chart of genetic distances between hit and causal SNPs in simulations of GWAS conducted with four chips - Panel (a) shows the percentage of simulations in which the hit and causal SNPs were within 0.1cM of one another for three relative risks and four genotyping chips. This distance was chosen to be representative of a potential fine-mapping region. Panel (b) shows the same analysis for causal variants with minor allele frequency less than 10%. Panel (c) shows the percentage of simulations in which the hit and causal SNPs were within 0.01 cM of one another.

5.1.4 Comparisons across populations

Table 5.3 summarizes the extent to which the Affymetrix 6.0 chip tags the causal variants in the three different populations. Note that the proportion of simulations in which the hit SNP perfectly tags the causal variant in YRI is only slightly less than in the other two populations (16% in YRI vs 18-20% in the other populations for a causal relative risk of 2), but that at relative risks of 2 and 4, the proportion of hit SNPs with $r^2 > 0.9$ with the causal SNP is substantially lower (52% in YRI vs 71-74% in the other populations for a causal relative risk of 2). Our results suggest that there may be utility in combining results across populations which have different patterns of LD, and thus tag the causal variant differently.

5.1.5 The effect of imputation

Table 5.4 summarizes the extent to which imputation impacts the proportion of hit SNPs that are expected to be causal and perfectly correlated with the causal SNP in a CEU population genotyped with the Affymetrix 6.0 chip. The imputation results are further partitioned into the percentages of coincident and perfectly correlated hit SNPs when the causal SNP is on the genotyping chip and when the causal SNP is imputed. Notably, imputation itself drops performance very little (with the understanding that, as discussed in §4.3, the imputation experiment we performed represents an upper bound). At a relative risk of 2, for instance, the causal and hit SNPs coincide 28% of the time when the causal SNP is on the chip and 26% of the time when the causal SNP is imputed.

As expected, imputation increases the percentage of coincident causal and hit SNPs, by 1.5 to 2-fold across relative risks compared to the simulations without imputation. That this holds true, even given the increased marker density of the Affymetrix 6.0 chip in the ENCODE regions, is a compelling reason to perform imputation with the high density genotyping chips currently available, and suggests that the results for lower density chips are likely to be more impressive. At relative risks of 2 and 4, imputation modestly improves the number of hit SNPs that are perfectly correlated with the causal variant, but actually does slightly worse at a relative risk of 1.25. These results foreshadow the likely outcomes of sequencing experiments, which are discussed below.

5.2 Implications for fine mapping

We now consider the implications of our results for the prospect of fine mapping a GWAS hit. This corresponds to the question of whether, by increasing the number of SNPs genotyped in a region showing association, an investigator would be likely

to identify the causal SNP rather than a surrogate, or, in other words, under what conditions one might expect a hit and a causal SNP to coincide.

As a hypothetical example, consider the following situation. Suppose we are fine mapping a GWAS signal in ENm010, and suppose further that the causal SNP happens to be one of the SNPs identified in the ENCODE project. How likely would we be to find the causal SNP if we genotyped our replication study at all the ENCODE SNPs in ENm010? One might, on first glance, think that the answer would be nearly 100% of the time. But consider the following. On average, in ENm010, there are 4 perfectly correlated SNPs (other than the SNP itself) for every SNP in the region in the CEU population. Thus we would expect that, even if we genotyped every SNP under consideration, in an association study the hit SNP and causal SNP would coincide only 20% of the time in this region. Among common SNPs in CEU in the ENCODE data across all regions, one in five SNPs has 20 or more perfect surrogates, three in five have five or more, and one in five has no perfect surrogates (International HapMap Consortium, 2005).

This is consistent with the observation that the percentage of exact simulations (where the hit SNP is coincident with the causal SNP) in Table 5.2 is bounded above by 20%, even as the chip density increases, while the percentage of hit SNPs that are perfectly correlated with the causal SNP rises above 50% (for example, with relative risk 2 in the Human 1.2M chip, as shown in Table 5.2). As expected, simulations using Illumina chips show a large number of simulations in which the causal and hit SNPs coincide. However, even when the SNPs on the chip are selected with LD in mind, one cannot hope to distinguish an association signal among four perfectly correlated SNPs. The situation is actually slightly more complicated, since SNPs that are merely very well-correlated ($r^2 > 0.99$ and $0.99 < r^2 < 0.95$) with the causal SNP are also strong candidates for the hit SNP. By chance, the strength of the association signal at one of these very good surrogates may surpass that at the causal SNP itself, or even

one of its perfect correlates. This is illustrated graphically in Figure 5.5, which shows pairwise correlations in a 20kb segment of ENm010. As demonstrated in the figure, the number of correlated nearby SNPs is not uniform, with some SNPs showing many good surrogates (in the figure shown with darker grey boxes) and some showing few. Thus the average of 4 perfectly correlated SNPs is not necessarily expected to hold for all or even most SNPs, but nonetheless it is a helpful simplification to explain why we might not expect to find the causal SNP even in a perfect genotyping experiment.

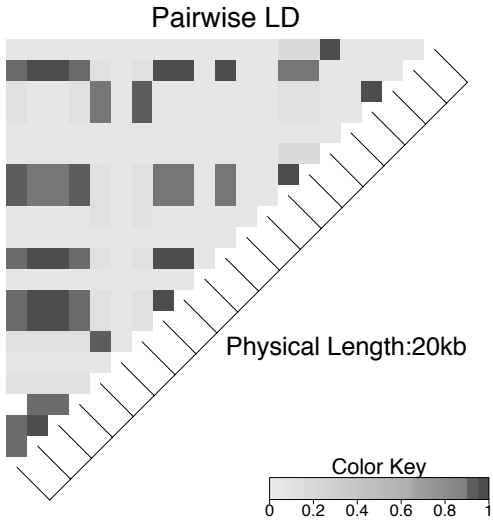


Figure 5.5: Pairwise correlations in a 20kb segment of ENm010 -

Table 5.1: Proximity of causal and hit SNP for a variety of true relative risks in the Affy 500k chip. The table shows the percentage of simulations that passed significance thresholds in which the LD between the causal and hit SNPs (measured by r^2) was greater than the stated threshold, except in the case of the $r^2 = 1$ column, in which case the number represents the percentage of simulations in which the causal and hit SNPs were in perfect LD. The "Exact" column gives the percentage of simulations in which the hit SNP and causal SNP were actually coincident. Two sample sizes, of 2000 cases and controls in the initial and replication studies and 5000 cases and controls in the initial and 10000 cases and controls in the replication study, are considered. Finally, the last column shows the total number of simulations passing significance thresholds. There were insufficient simulations with relative risk of 1.1 in the smaller sample size to give meaningful results.

		r^2 threshold					
RR	Sample Size	0.99	0.95	0.8	0.5	Exact	# Sims
1.1	5000/10000	28%	52%	78%	95%	8%	266
1.15	2000/2000	30%	59%	83%	94%	8%	115
1.15	5000/10000	28%	55%	81%	97%	8%	2755
1.2	2000/2000	25%	53%	81%	98%	8%	779
1.2	5000/10000	26%	51%	79%	95%	8%	6981
1.25	2000/2000	27%	53%	80%	96%	8%	2599
1.25	5000/10000	25%	48%	75%	93%	7%	10096
1.3	2000/2000	27%	53%	80%	96%	8%	5144
1.3	5000/10000	25%	47%	72%	90%	8%	11987
1.4	2000/2000	25%	49%	75%	93%	8%	9315
1.4	5000/10000	24%	45%	69%	87%	7%	13813
1.5	2000/2000	25%	47%	73%	91%	8%	11684
1.5	5000/10000	24%	44%	67%	84%	7%	14832
1.6	2000/2000	24%	46%	71%	89%	7%	12915
1.6	5000/10000	24%	43%	66%	83%	7%	15547
2	2000/2000	24%	43%	66%	84%	7%	15181
2	5000/10000	24%	41%	62%	79%	7%	17251

Table 5.2: Proximity of causal and hit SNP for a variety of true relative risks and four different genotyping chips in CEU. The table shows the percentage of simulations that passed significance thresholds in which the LD between the causal and hit SNPs (measured by r^2) was greater than the stated threshold, except in the case of the $r^2 = 1$ column, in which case the number represents the percentage of simulations in which the causal and hit SNPs were in perfect LD. The "Exact" column gives the percentage of simulations in which the hit SNP and causal SNP were actually coincident. As before, the "# Sims" column shows the total number of simulations passing significance thresholds. Finally, the last column gives the chip identifier. For each choice of chip and causal relative risk we consider both all causal SNPs and, separately, those with minor allele frequencies less than 20%.

		r^2 threshold						
RR	MAF	0.99	0.95	0.8	0.5	Exact	# Sims	Chip
1.25	< 0.2	19%	44%	75%	94%	6%	173	Affymetrix 500k
1.25	all	27%	53%	80%	96%	8%	2599	Affymetrix 500k
2	< 0.2	21%	37%	63%	82%	7%	3812	Affymetrix 500k
2	all	24%	43%	66%	84%	7%	15181	Affymetrix 500k
4	< 0.2	17%	28%	49%	63%	6%	5433	Affymetrix 500k
4	all	23%	39%	60%	76%	7%	17957	Affymetrix 500k
1.25	< 0.2	47%	59%	79%	99%	7%	135	Affymetrix 6.0
1.25	all	48%	64%	85%	98%	11%	3039	Affymetrix 6.0
2	< 0.2	54%	64%	79%	91%	11%	3612	Affymetrix 6.0
2	all	54%	66%	82%	94%	12%	14745	Affymetrix 6.0
4	< 0.2	50%	58%	69%	79%	9%	4727	Affymetrix 6.0
4	all	55%	65%	78%	89%	12%	16659	Affymetrix 6.0
1.25	< 0.2	45%	54%	79%	98%	11%	170	Illumina 550
1.25	all	49%	63%	85%	98%	15%	3202	Illumina 550
2	< 0.2	52%	62%	81%	93%	14%	3562	Illumina 550
2	all	55%	67%	87%	97%	15%	14797	Illumina 550
4	< 0.2	44%	52%	66%	76%	11%	4667	Illumina 550
4	all	54%	64%	81%	91%	15%	16500	Illumina 550
1.25	< 0.2	51%	61%	84%	98%	20%	173	Illumina 1.2M
1.25	all	52%	67%	87%	98%	17%	3314	Illumina 1.2M
2	< 0.2	63%	73%	86%	96%	18%	3763	Illumina 1.2M
2	all	65%	77%	91%	98%	20%	15118	Illumina 1.2M
4	< 0.2	60%	68%	78%	86%	16%	4825	Illumina 1.2M
4	all	66%	76%	88%	95%	19%	16936	Illumina 1.2M

Table 5.3: Population Tagging Summary. Each line of the table shows, for a given population, the percentage of hit SNPs with $r^2 > 0.9$ with the causal SNP (and, parenthetically, the percentage of hit SNPs in perfect LD with the causal SNP).

RR	1.25	2	4
CEU	75% (15%)	74% (20%)	72% (21%)
JPT+CHB	73% (15%)	71% (18%)	68% (19%)
YRI	67% (19%)	52% (16%)	46% (15%)

Table 5.4: Imputation Summary. Each line of the table shows, for simulations with the Affymetrix 6.0 chip, the percentage of hit SNPs that are coincident (exact) and those that have $r^2 > 0.99$ with the causal SNP (perfect LD), with and without imputation.

RR	1.25	2	4
CEU (exact)	11%	12%	12%
CEU (perfect LD)	48%	54%	55%
Imputed (exact)	17%	24%	25%
Imputed (perfect LD)	36%	57%	62%
Causal SNPs on chip (exact)	20%	28%	30%
Causal SNPs on chip (perfect LD)	38%	57%	63%
Imputed causal SNPs (exact)	17%	26%	27%
Imputed causal SNPs (perfect LD)	32%	53%	58%

6

Non-additive disease risk models

The statistical disease model that best describes the data between alleles and among loci is of interest in the analysis of GWAS signals. Most findings from GWAS to date are consistent with the simplest disease model at a SNP, in which each additional copy of the risk allele increases risk by the same multiplicative factor (Hindorff *et al.*, 2010, 2009). Biological interpretation of disease models is complex and outside the scope of this thesis (see Cordell (2002) for a discussion of some of the issues involved). However, purely from a statistical standpoint, the underlying disease model influences the power and the observed effect size. In this chapter, we explore whether LD plays a role in the inferred disease model by asking how different genotype-risk relationships impact our power to detect associations and departures from the simplest model.

We have assumed the simplest disease model in the preceding chapters, namely that the relationship between genotype and disease risk is log-linear. That is, each additional copy of the causal allele at the disease locus multiplies the disease risk by a factor: the relative risk. However, other assumptions are possible, including dominant and recessive models, in which either one copy (dominant) or both copies (recessive) of an allele are necessary to influence disease risk.

As discussed previously, LD plays a crucial role in the ability of the GWAS method

to detect associations with untyped SNPs. We have previously seen how, as the LD between the causal and tag SNPs decreases, the observed effect size gets smaller, and in particular, smaller than the true effect size. In the setting of a more general disease model with more than one parameter describing the relationship between genotype and disease risk, we can ask how the observed model compares to the true model based on the strength of the LD between the causal and marker SNPs.

In particular, we focus on two related questions:

- How do the disease model parameters ('effects') change as the LD between the causal and tag SNPs diminishes?
- How does the power to detect departure from the simplest model change?

The impact of imperfect LD has been described for additive models, both in terms of effect sizes and power (Chapman *et al.*, 2003; Pritchard & Przeworski, 2001; Zondervan & Cardon, 2004). Measuring LD using the squared correlation coefficient r^2, a well-known rule of thumb holds that a sample size of roughly N/r^2 is required at a tag SNP to achieve the same power to detect an association as a study with sample size N that types the causal SNP directly (Pritchard & Przeworski, 2001). An interesting question is whether there is a more general rule relating effects to power that does not assume an additive risk model. We describe a more general rule here, first in the context of recessive and dominant risk models, and then in a particular 2-locus interaction model.

We extend the framework introduced in Chapter 2 to simulate case-control studies when the causal SNP has a non-additive disease model to investigate the expected impact on effect size and inferred disease model with empirical patterns of LD. Power estimates are calculated from the simulation results.

In addition to studying disease effects, we investigate the impact of LD and model mis-specification on heritability estimates. We derive an expression for the relationship between λ_s and the penetrances and frequencies of each genotype. This gives an analytic

expression for the impact of LD on heritability estimates in both additive and non-additive models.

A number of papers have dealt with power in association studies, for instance Spencer *et al.* (2009) and Bhangale *et al.* (2008). These focus principally on how power to detect associations varies with different chips and different sample sizes. Although Bhangale *et al.* (2008) considered recessive and dominant models in GWAS simulations, as we do, they limited their inquiry to power to detect associations with a trend test. We have a slightly different focus.

First, we are interested in what can be thought of as "disease model distortion". That is, if we aim to infer the disease model on the basis of disease model parameters estimated at a tag SNP, how different do these parameters look than they would under an additive model? Second, we are interested not only in power to detect an association, but power to detect deviations from the additive model once the association is detected.

We note that Zheng *et al.* (2009) derived analytical results for non-additive models assuming the same allele frequency at the causal and marker SNPs, but we make no such simplifying assumptions. While we focus on the setting of case-control studies, we acknowledge related work has been published for studies of quantitative traits using variance components models. Sham *et al.* (2000) derived similar results for the impact of LD on QTL effects, and Hill *et al.* (2008) showed that additive variation (analagous to additive effects in case-control studies) will tend to dominate even when non-additive effects are present and the impact of LD is discounted.

A subset of results in this chapter has been published (Vukcevic *et al.*, 2011). Although the work presented was collaborative, this chapter focuses on those aspects of the work in which I took the lead role.

6.1 Recessive and dominant disease risk models

In earlier chapters we adopted the commonly used logistic regression framework described in §1.4.1. In this more general setting it turns out to be slightly easier to work with a log risk regression model. Let G denote the genotype as before. Here we model the penetrances $p(G) = \Pr(D|G)$ as

$$\log(p) = \mu + \beta G + \gamma \chi_1(G)\,, \tag{6.1}$$

where $\chi_1(G)$ is an indicator function that takes the value 1 for heterozygotes and 0 for homozygotes. Recall from §1.4.4 that in GWAS it is standard to use population controls in place of true controls but analyse the results as a typical case-control study using logistic regression. Schouten *et al.* (1993) show that this is equivalent to fitting the log risk regression model described above at 6.1. Then with a slight abuse of notation we use the same parameter names as before and refer to β as the additive parameter and γ as the dominance parameter.

We now briefly review two key results from Vukcevic *et al.* (2011).

Let A be a causal SNP with minor allele frequency f_A and B a tag SNP with minor allele frequency f_B. The first result expresses the β and γ parameters at the tag SNP as a function of the minor allele frequencies at A and B, the parameters at the causal SNP A, and the r^2 between A and B:

$$\beta_B \simeq \beta_A r \sqrt{\frac{f_A(1-f_A)}{f_B(1-f_B)}}\,,$$
$$\gamma_B \simeq \gamma_A r^2 \frac{f_A(1-f_A)}{f_B(1-f_B)}\,.$$

This equation shows that the dominance effect at the marker SNP decreases quadratically with r as the LD becomes weaker, in contrast to the the additive effect, which decreases linearly with r.

One can test the hypothesis that $\gamma = 0$ against the hypothesis that $\gamma \neq 0$ by a maximum likelihood ratio test which we will refer to as the *deviation* test. We refer the reader to Vukcevic *et al.* (2011) for further discussion of this test and its properties.

As a consequence of the first result, the non-centrality parameters at the causal and tag SNPs for the deviation test can be expressed as,

$$\begin{aligned}\eta_1 &\simeq 4N_A f_A^2(1-f_A)^2\phi(1-\phi)\gamma_A^2\,,\\ \eta_2 &\simeq 4N_B f_B^2(1-f_B)^2\phi(1-\phi)\gamma_B^2\,,\\ &\simeq 4N_B f_A^2(1-f_A)^2\phi(1-\phi)\gamma_A^2 r^4\,,\end{aligned}$$

where ϕ is the proportion of cases in the sample (Vukcevic *et al.*, 2011). Thus, a sample size of $N_B = N_A/r^4$ is required to achieve the same power to detect a deviation as typing the causal SNP directly.

6.1.1 Simulation results

We extended the framework developed in Chapter 2 to the setting of interaction effects between alleles at a single SNP to study how empirical patterns of LD are likely to affect effect size estimation and detection of non-additive disease models. In this section we will restrict ourselves to recessive and dominant disease models for a single causal SNP. We describe only the ways in which the procedure differs from the framework detailed in Chapter 2, to which we refer the reader for a fuller discussion on the specifics of the simulations. For definiteness, we will present results for the Affymetrix 6.0 chip.

6.1.2 Simulation sampling strategy

We sampled 2,000 diploid cases and 2,000 diploid controls from the population cohort produced by HAPGEN. For the controls, we sampled haplotypes uniformly from the population cohort without replacement and combined them in pairs, as before. We

sampled the cases according to the genotype frequencies at the causal SNP as dictated by Bayes' theorem. Let $(\alpha_0, \alpha_1, \alpha_2)$ be the relative risks of genotypes 0, 1, and 2 and let f be the frequency of allele 1 in the population cohort. Then the frequencies of the three genotypes in cases are given by,

$$\begin{aligned}
\Pr(G = 0 \mid \text{Disease}) &\propto \left(\frac{\Pr(\text{Disease} \mid G = 0)}{\Pr(\text{Disease})} \right) Pr(G = 0)\,, \\
&\propto \alpha_0 (1 - f)^2\,, \\
\Pr(G = 1 \mid \text{Disease}) &\propto \alpha_1 2f(1 - f)\,, \\
\Pr(G = 2 \mid \text{Disease}) &\propto \alpha_2 f^2\,.
\end{aligned}$$

We then sampled pairs of haplotypes without replacement uniformly from the panel such that they were consistent with the genotype frequencies when combined in pairs.

6.1.3 Power

In a typical GWAS, if testing for deviation from the additive disease model is done at all, it is performed after the initial association scan, a procedure to which we adhered in the simulations. We ascertained the hit SNPs on the basis of the p-values from the trend test in the initial and replication scan, and then applied a subsequent deviation test using genotype counts from the replication scan. The hit SNP was deemed "non-additive" if this p-value was less than 0.05. Each simulation has three possible outcomes: (1) no association is detected; (2) an association is detected but the model does not significantly deviate from the additive model; or (3) both association and deviation are detected. Table 6.1 shows power estimates of the two tests for recessive, additive, and dominant disease models across a variety of relative risks.

These results are presented visually in Figure 6.1, which shows the number and percentage of simulations that are undetected, detected without deviation, and detected with deviation for each of the three models for a homozygote relative risk of 1.4^2. Table

Table 6.1: Simulation results. Entries in the table give proportions of simulations achieving particular outcomes over a range of models and effect sizes. The three possible outcomes are: the hit SNP doesn't pass the scan and replication criteria ('Undetected'); that it passes these criteria but a subsequent deviation test is not significant ('Assoc. only'); or that this test is significant ('Assoc. + deviation'). The effect size is given by the homozygous RR (Hom. RR), which compares the risk of the two homozygotes.

		Outcome (%)		
Model	**Hom. RR**	**Undetected**	**Assoc. only**	**Assoc. + deviation**
Additive	1.1^2	100	0	0
Additive	1.2^2	94	6	0
Additive	1.4^2	47	50	3
Additive	2.0^2	19	77	4
Dominant	1.1^2	100	0	0
Dominant	1.2^2	84	4	12
Dominant	1.4^2	53	6	41
Dominant	2.0^2	38	6	56
Recessive	1.1^2	100	0	0
Recessive	1.2^2	83	4	13
Recessive	1.4^2	49	6	45
Recessive	2.0^2	27	8	65

6.1 suggests that most of the time when we detect association, the deviation test will give the correct outcome, even at smaller effect sizes. This is despite the fact that the deviation parameter diminishes more rapidly with LD than the additive parameter, as we showed in the previous section. Figure 6.1 provides insight as to why: the typical LD distribution for a set of simulations is enriched at both ends of the spectrum. Either the SNP will be nearly perfectly tagged, in which case there is little erosion of the deviation parameter, or it will be so poorly tagged that the trend test will not be able to detect it at all. As expected, as LD decreases, the relative number of simulations that pass the trend test but fail the deviation test increases.

6.1.4 Effect size estimates and parameter estimation

Figure 6.2 shows how the additive and interaction parameters measured at the hit SNP vary by LD compared to the true parameters for homozygous relative risks of 1.2^2, 1.4^2, and 2^2 under a dominant disease model. As predicted by theory, the interaction parameter decays more rapidly as LD diminishes than does the additive parameter.

6.1.5 Heritability estimates for non-additive models

Recall that a commonly used measure of genetic heritability is the sibling recurrence relative risk,

$$\lambda_s = \frac{\Pr(Y = 1 \mid Y_s = 1)}{\Pr(Y = 1)} .$$

This is a risk ratio; the numerator expresses the probability of disease given an affected sibling and the denominator is the population prevalence of the disease. Equation 1.3 expresses λ_s in terms of the relative risks and frequencies of the alleles at a locus,

$$\frac{\sum_{g_m} \sum_{g_f} \sum_{g_j} \sum_{g_k} RR_{g_j} RR_{g_k} T_{g_j g_m g_f} T_{g_k g_m g_f} Q_{g_m} Q_{g_f}}{(\sum_{g_j} RR_{g_j} Q_{g_j})^2}$$

Figure 6.1: Additive, dominant, and recessive disease model results stratified by LD between the causal and hit SNPs. The distribution of simulation outcomes is shown by LD intervals. Each LD interval spans 10% of the range and is represented by its upper bound on the y-axis label under the bar. Red indicates undetected simulations, yellow indicates association detected, but no deviation from an additive model, and blue indicates those associations detected with deviation. Plots on the right show relative proportions of the total number of SNPs in each category, while the plots on the left show absolute counts. In order from top to bottom, the simulations were performed under an additive, dominant, and recessive model with homozygous relative risk 1.4^2.

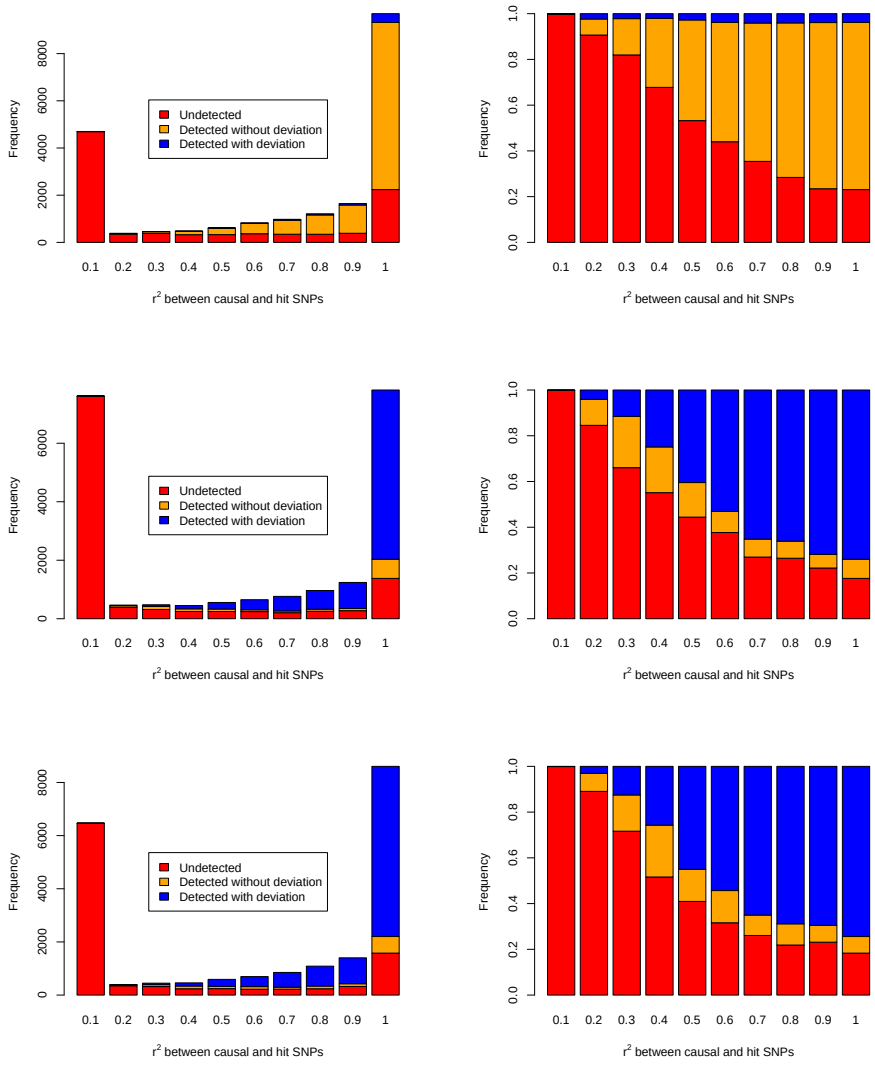
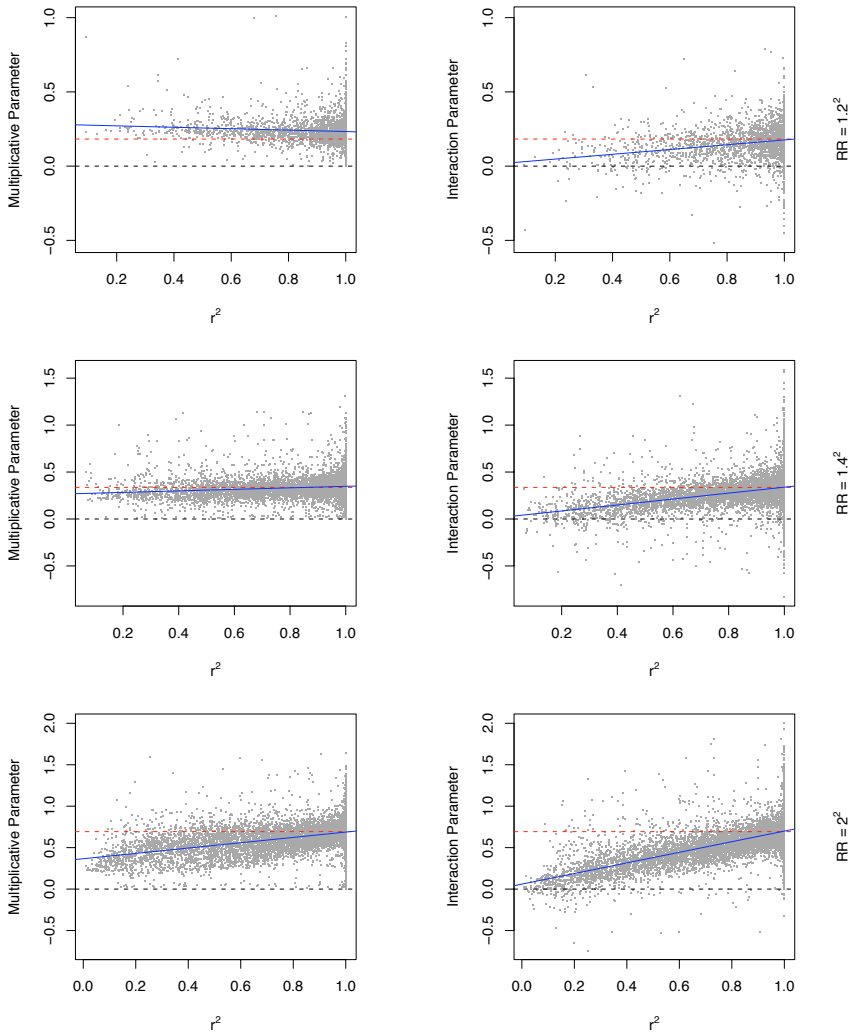


Figure 6.2: Estimated disease model parameters when the true disease model is dominant with a homozygous relative risk of 1.2^2, 1.4^2, and 2^2 respectively. The plots on the left show the distribution of the additive parameter, while plots on the right show the interaction parameter. The red dotted line indicates the true value of each parameter and the black dotted line, the null value. A line of best fit is shown in blue. Each grey dot represents the parameter values estimated in one simulation plotted against the r^2 between the causal and hit SNPs.



where the T_* are transition probabilities and the Q_* are genotype probabilities under Hardy-Weinberg equilibrium (see Section §1.4.8). This equation suggests two ways in which λ_s can be mis-estimated in the context of a GWAS. The first is model mis-specification, calculating λ_s using relative risks estimated assuming a additive model rather than the true model. The second is dilution due to LD, similar to that observed for parameter estimates.

Figure 6.3 shows both effects in a GWAS setting for three disease models. The red curves show the true heritability based on the disease model and allele frequency of the causal SNP. The blue curves show the mean heritability estimate if it were calculated at the causal SNP but assuming an additive model—this is computed by first calculating β and then using equation 1.3 with penetrances that follow an additive model with parameter β. The difference between the two lines shows the effect of mis-specification. Note that this results in underestimation for recessive models and overestimation for dominant models.

The points in the figure show λ_s estimates from the simulations, calculated at the hit SNP assuming an additive model. The difference between the points and the blue lines shows the effect of dilution due to LD. Note that the x-coordinate is always the allele frequency of the causal SNP, so that the corresponding y-coordinates are naturally comparable (i.e. the true value and estimate from the same simulation).

When the true model is in fact additive (Figure 6.3, middle column), then the only underestimation owes to LD. When the correlation between the causal SNP and the hit SNP is good (top row) then the estimates are centered around the true value of λ_s. Note that there would be much greater variability at causal allele frequencies near 0 and 1 if we were not excluding simulations that failed significance thresholds. In analogy to our earlier results on effect sizes, as the LD between the causal and hit SNPs is reduced (second and third rows of Figure 6.3), the estimates of λ_s decrease. The effect of model mis-specification (shown as the distance between the red and blue lines in

the figure) depends on the allele frequency and the disease model, but in each case, reducing pairwise LD between the causal and hit SNP reduces estimates of λ_s. Given our prediction that we will often detect model mis-specification when it occurs, and adjust our λ_s estimates accordingly, we do not expect this mis-estimation to contribute greatly to heritability-explained estimates in comparison with the impact of LD under the model considered here.

6.2 2-locus interactions

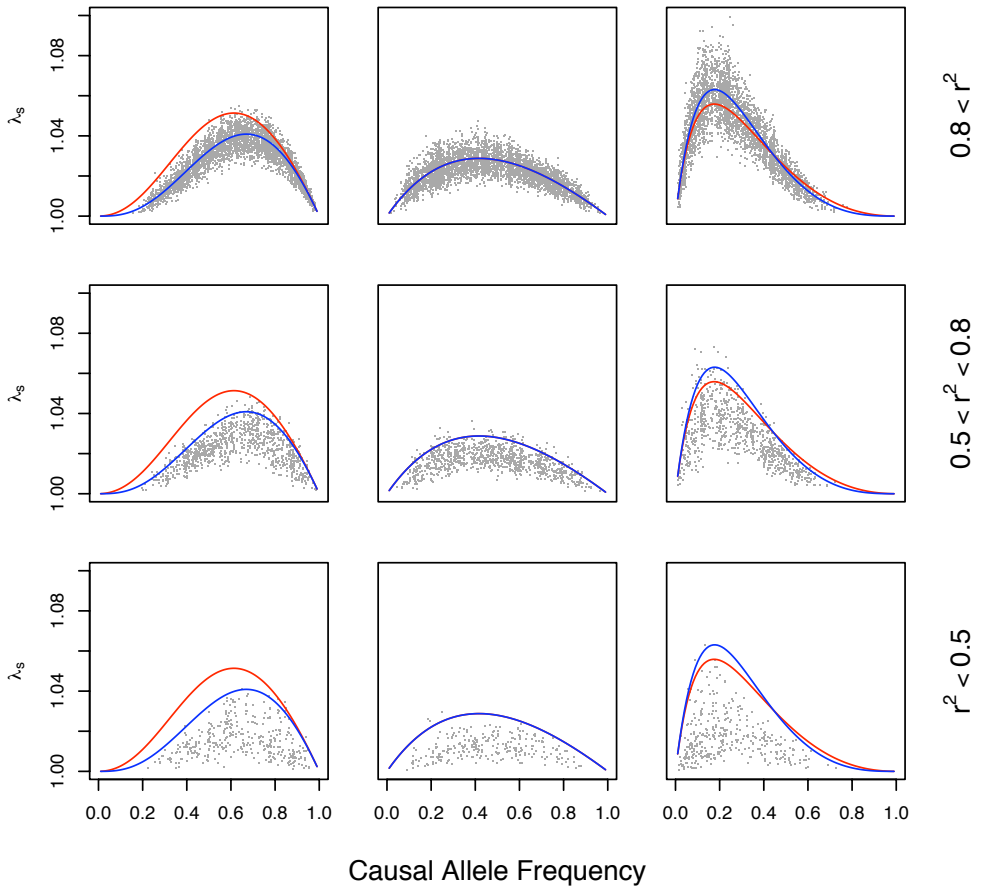
6.2.1 Empirical evidence for interactions in GWAS

Despite the observation in Marchini *et al.* (2005) that interactions between loci with non-marginal effects are plausible, and their subsequent demonstration that searching for interactions is computationally feasible and potentially more powerful than traditional single-locus tests at a genome-wide scale, there are very few examples of observed interactions in the context of GWAS. Various statistical methods for detecting interactions have been proposed; some of these are discussed in Thornton-Wells *et al.* (2004) and Cantor *et al.* (2010), to which we refer the reader for further background. Our focus here is not in developing new statistical technology for interrogating or identifying interactions, but rather in trying to understand the impact LD has on changing disease model parameters away from the causal variant. In discussing power we necessarily adopt a statistical test. We choose the two tests most commonly applied in the GWAS context, namely the trend test and the general test.

6.2.2 An interaction model

Interactions between alleles on two haplotypes and among two unlinked loci can be seen as special cases of a general interaction model which allows for non-additive disease risk. First we present theoretical results for interactions between unlinked loci and later

Figure 6.3: Heritability estimates under the three models. Estimates and true values of λ_s plotted against the allele frequency at the causal SNP. Results are shown for three different models, split into columns (recessive, additive, and dominant, from left to right), all with a homozygous RR of 1.4^2. The grey points show estimates from simulations, split into different rows depending on the r^2 between the causal and marker SNPs, as labelled. The red curves show the true heritability. The blue curves show the mean estimate if it were calculated at the causal SNP but assuming an additive model.



discuss the special case of interactions among alleles, where we re-derive the result in §6.1. The challenges of detecting dominance and interaction effects are somewhat different in practice. From an abstract mathematical standpoint, however, dominance and interaction can be thought of in the same mathematical language. We will introduce a model that facilitates this comparison.

There are many different ways of modeling interactions (Marchini *et al.*, 2005) and correspondingly many different parameterisations are possible. Let B and B' be two unlinked biallelic SNPs and let G_B denote the genotype at SNP B, likewise for SNP B'. Here we consider the simplest form of interaction from a statistical standpoint: a two-SNP model with a single additive interaction parameter,

$$\log(p) = \mu + \beta_B G_B + \beta_{B'} G_{B'} + \tau_{BB'} G_B G_{B'} \,, \tag{6.2}$$

where β_B and $\beta_{B'}$ are the additive parameters for each of the SNPs B and B' separately and $\tau_{BB'}$ is an additive interaction parameter. Again, we choose log risk regression for convenience, but we refer the reader back to the discussion in §6.1 on its connection with the log-odds model.

As discussed in previous chapters, we expect GWAS to highlight markers correlated with causal loci, rather than necessarily the loci themselves. We should imagine, then, that SNPs B and B' used to measure the model parameters β_B, β'_B and $\tau_{BB'}$ in an association study are actually markers for unobserved causal SNPs A and A'. We seek to elucidate the relationship between the parameters β_B, β'_B and $\tau_{BB'}$ and β_A, β'_A and $\tau_{AA'}$.

It is well known that, for a single SNP under the additive disease risk model in a diploid population, it is equivalent to regard individuals as haploid. It turns out that for estimating the parameters of interest, namely β and τ, the same is true for the 2-locus additive model with additive interactions defined at 6.2 (Vukcevic *et al.*, 2011).

For simplicity, we now turn to analysing the model in this haploid framework. We will slightly abuse terminology and refer to the type of an individual at two loci A and A' or B and B' as its haplotype even if the loci are on different chromosomes.

We would like to consider how the interaction parameters change as we move away from SNPs A and A'. Let SNPs B and B' be biallelic SNPs and code the alleles at each as 0 and 1. In the following discussion we choose the convention that the A and A' are causal SNPs, and B and B' are marker SNPs for A and A' respectively. Let $f_A = \Pr(A = 1)$ be the frequency of allele 1 in the population at SNP A, and define the frequencies of the other SNPs similarly.

First we consider the relationship between the causal SNPs and their tags. Define the following conditional probabilities,

$$q_0 = \Pr(A = 1 \mid B = 0)\,, q_0' = \Pr(A' = 1 \mid B' = 0)\,,$$
$$q_1 = \Pr(A = 1 \mid B = 1)\,, q_1' = \Pr(A' = 1 \mid B' = 1)\,.$$

As a matter of convenience, we will refer to the haplotype with $A = i$ and $B = j$ as ij. The conditional probabilities defined above allow us to describe the probability of each pair of haplotypes (for brevity, we will describe only the pairs of haplotypes for A and B since the situation for A' and B' is entirely analogous),

$$\Pr(00) = (1 - q_0)(1 - f_B)\,,$$
$$\Pr(01) = (1 - q_1) f_B\,,$$
$$\Pr(10) = q_0(1 - f_B)\,,$$
$$\Pr(11) = q_1 f_B\,,$$

and give the identity

$$f_A = q_0(1 - f_B) + q_1 f_B \,.$$

Manipulating the definition of r^2 in §1.1, we can express the correlation coefficient in terms of our parameters as,

$$r^2 = (q_1 - q_0)^2 \frac{f_B(1 - f_B)}{f_A(1 - f_A)} \,.$$

Let $\mathcal{A}$ and $\mathcal{B}$ denote the 2×2 matrix of penetrances with entries $\mathcal{A}_{ij} = \Pr(\text{Disease} \mid A = i, A' = j)$,

$$\mathcal{A} = \begin{bmatrix} \Pr(D \mid A = 0, A' = 0) & \Pr(D \mid A = 0, A' = 1) \\ \Pr(D \mid A = 1, A' = 0) & \Pr(D \mid A = 1, A' = 1) \end{bmatrix} ,$$

and similarly for $\mathcal{B}$:

$$\mathcal{B} = \begin{bmatrix} \Pr(D \mid B = 0, B' = 0) & \Pr(D \mid B = 0, B' = 1) \\ \Pr(D \mid A = 1, B' = 0) & \Pr(D \mid B = 1, B' = 1) \end{bmatrix} .$$

We seek to write $\mathcal{B}$ in terms of $\mathcal{A}$ and the conditional probabilities q_* that define the LD relationship between SNPs A and B.

For the purposes of exposition and concreteness, let us focus on $\mathcal{B}_{00}$. The derivation of the other entries is nearly identical: simply replace $B = 0$ with $B = i$ and $B' = 0$ with $B' = j$ in what follows to obtain the general result. There are four ways that disease can arise, corresponding to the four entries of $\mathcal{A}$. To calculate $\mathcal{B}_{00}$, we need to multiply each of these penetrances by the probability of the genotypes at A and A',

given the genotypes at B and B':

$$\begin{aligned}\Pr(D \mid B=0, B'=0) =& \Pr(D \mid A=0, A'=0)\Pr(A=0 \mid B=0)\Pr(A'=0 \mid B'=0) \\ &+ \Pr(D \mid A=1, A'=0)\Pr(A=1 \mid B=0)\Pr(A'=0 \mid B'=0) \\ &+ \Pr(D \mid A=0, A'=1)\Pr(A=0 \mid B=0)\Pr(A'=1 \mid B'=0) \\ &+ \Pr(D \mid A=1, A'=1)\Pr(A=1 \mid B=0)\Pr(A'=1 \mid B'=0) \\ =& \mathcal{A}_{00}(1-q_0)(1-q_0') + \mathcal{A}_{01}(q_0)(1-q_0') + \mathcal{A}_{10}(1-q_0)(q_0') + \mathcal{A}_{11}(q_0)(q_0')\end{aligned}$$

The result of the calculation suggests that matrix notation might be appropriate for these sums, and indeed it simplifies the expressions if we define a matrix of LD parameters,

$$\mathcal{L} = \begin{bmatrix} 1-q_0 & q_0 \\ 1-q_1 & q_1 \end{bmatrix}.$$

Thus we obtain the following identity,

$$\mathcal{B} = \mathcal{L}\mathcal{A}\mathcal{L}'^{\mathrm{T}}. \tag{6.3}$$

Recall that non-additive single-SNP models, in which the interaction parameter relates to alleles at a given SNP, are a special case of the more general interaction model. We now turn to these non-additive models before concluding the derivation in the general setting.

6.2.3 Recessive and dominant models as special cases

We can use the results obtained in the previous subsection and a different perspective to re-derive the results about model distortion or to provide a different derivation to that appearing in Vukcevic *et al.* (2011) and Vukcevic (2009). We assume in the following that the parental origin of the haplotype is immaterial in determining the

disease risk, so that what matters is the number of copies of the disease allele. In §6.2.2 we considered two interacting SNPs A and A' with markers B and B'. Now we consider two interacting *alleles* A and A' on haplotypes at a single locus tagged by alleles B and B' at a tag locus.

With this new definition of A and B, the penetrance matrices $\mathcal{A}$ and $\mathcal{B}$ are symmetric, since $\mathcal{A}_{01} = \mathcal{A}_{10}$ and similarly for $\mathcal{B}$. In addition, it imposes the restriction $\beta_A = \beta_{A'}$. The log risk regression then looks like,

$$\begin{aligned}\log(p) &= \mu + \beta_A H_A + \beta_{A'} H_{A'} + \tau_{AA'} H_A H_{A'}\,, \\ &= \mu + \beta_A (H_A + H'_A) + \tau_{AA'} H_A H_{A'}\,, \\ &= \mu + \beta_A G_A + \tau_A \chi_2(G_A)\,,\end{aligned}$$

where χ is a standard indicator function which takes the value 1 when $G_A = 2$ and 0 elsewhere. For our purposes, it will be convenient to re-parameterise this as

$$\log(p) = \mu + \beta_A G_A + \gamma_A \chi_1(G_A)\,.$$

We now return to Equation 6.3, noting that $\mathcal{L}$ is now equal to $\mathcal{L}'$. Under an additive disease model, recall that $\mathcal{A}_{00} = 1$ and that $\mathcal{A}_{11} = \mathcal{A}_{10}^2 = \mathcal{A}_{01}^2$, so in particular $|\mathcal{A}| = 0$ when A follows an additive disease model. By Equation 6.3, $|\mathcal{B}|$ is also 0 when A follows an additive disease model, so that, even as the correlation between SNPs A and B diminishes, the observed disease model remains additive. The situation for non-additive models differs,

$$\begin{aligned}|\mathcal{B}| &= |\mathcal{L}\mathcal{A}\mathcal{L}^{\mathrm{T}}| \\ &= |\mathcal{A}|(q_1 - q_0)^2\,.\end{aligned}$$

Rewriting this in terms of the disease model parameters, allele frequencies, and r^2,

$$\frac{|\mathcal{B}|}{b_{01}^2} = \frac{|\mathcal{A}|}{a_{01}^2}\frac{a_{01}^2}{b_{01}^2}(q_1 - q_0)^2$$
$$e^{2\gamma_B} - 1 = (e^{2\gamma_A} - 1)\frac{e^{2(\mu_A+\beta_A)}}{e^{2(\mu_B+\beta_B)}}\frac{f_A(1-f_A)}{f_B(1-f_B)}r^2\,.$$

When the γ and β parameters are small, the approximation $e^x - 1 \simeq x$ holds, and also $\mu_A + \beta_A \simeq \mu_B + \beta_B$, giving us the simplified expression,

$$\gamma_B \simeq \gamma_A \frac{f_A(1-f_A)}{f_B(1-f_B)}r^2\,. \tag{6.4}$$

This result is derived in Vukcevic *et al.* (2011); our derivation considers this as a special case of the general interaction model. Equation 6.4 shows that the interaction parameter γ decays proportional to r^2 between the causal and tag SNPs.

6.2.4 Theoretical derivations for the interaction model

When $\mathcal{L} \neq \mathcal{L}'$ and the penetrance matrices $\mathcal{A}$ and $\mathcal{B}$ are no longer symmetric, we return to our earlier parameterisation of the interaction model and can follow the same procedure for describing the relationship between $\tau_{BB'}$ and $\tau_{AA'}$. Again by Equation 6.3,

$$|\mathcal{B}| = |\mathcal{L}\mathcal{A}\mathcal{L}'^{\mathrm{T}}|$$
$$= |\mathcal{A}|(q_1 - q_0)(q_1' - q_0')\,.$$

Dividing through by $b_{01}b_{10}$ we obtain,

$$\frac{|\mathcal{B}|}{b_{01}b_{10}} = \frac{|\mathcal{A}|}{a_{01}a_{10}}\frac{a_{01}a_{10}}{b_{01}b_{10}}(q_1 - q_0)(q_1' - q_0')$$
$$e^{\tau_{BB'}} - 1 = (e^{\tau_{AA'}} - 1)\frac{e^{(\mu_{AA'}+\beta_A)+(\mu_{AA'}+\beta_{A'})}}{e^{(\mu_{BB'}+\beta_B)+(\mu_{BB'}+\beta_{B'})}}\sqrt{\frac{f_A(1-f_A)f_{A'}(1-f_{A'})}{f_B(1-f_B)f_{B'}(1-f_{B'})}}(rr')\,.$$

When the interaction effect is small, this becomes

$$\tau_{BB'} \simeq \tau_{AA'} \sqrt{\frac{f_A(1-f_A)f_{A'}(1-f_{A'})}{f_B(1-f_B)f_{B'}(1-f_{B'})}}(rr') \tag{6.5}$$

Equation 6.5, analogous to Equation 6.4, shows a quadratic relationship between the LD between the causal and marker SNP and the interaction parameter measured at the marker SNP. Here, though, there are multiple sources of LD, and r and r' need not be equal. As discussed in Vukcevic *et al.* (2011), the comparison to keep in mind is that with the effect of the additive parameter, which decreases linearly with LD.

Since we showed theoretically that interaction effects between alleles and among loci of the type we considered have the same order decay with LD, we expect simulations of interactions among loci to show similar results to the between-allele interactions presented in §6.1.3-6.1.4. One direction for future work concerns the validity of this expectation, as well as how more complex models of interaction compare to the simplest parameterisation considered here.

6.2.5 Heritability under interactions

How might the presence of interactions distort estimates of λ_s? We can rewrite Equation 1.3 as,

$$\lambda_s = \frac{\sum_i \sum_j f_A^i f_{A'}^j p_{ij} \sum_k \sum_l T_{ik} T_{jl} p_{kl}}{(\sum_i \sum_j f_A^i f_{A'}^j p_{ij})^2}$$

where $f_A^i = \Pr(A = i)$, p_{ij} are the penetrances as defined in Equation 6.2, and T_{ik} is the probability that a sibling has genotype i given that the affected sibling has genotype k. If there is no interaction term in Equation 6.2 ($\tau = 0$), then this expression can be partitioned into a product of the λ_s values calculated at A and A' independently. Suppose, however, that τ is nonzero. For instance, we could imagine a model in which mutant forms at both loci A and A' are required for increased disease risk. Let $f_A^0 = 0.1$, $f_{A'}^0 = 0.3$, and suppose that the genotype pair $(A, A') = (0, 0)$ confers a relative risk of

2. In this setting, the true λ_s of the combined loci is equal to 1.0093, and calculating λ_s by multiplying the values of sibling recurrence risk calculated at A and A' independently gives 1.0048. Now suppose that the loci A and A' are not observed directly, but are tagged by loci B and B'. Let the r^2 between A and B, and between A' and B', be 0.8, and let B and B' have the same frequencies as A and A'. Then Equation 6.5 gives that the dilution of the interaction parameter results in the genotype pair $(B, B') = (0, 0)$ conferring a relative risk of 1.74, with a corresponding λ_s of 1.0051. If both types of underestimation – model mis-specification and dilution of effect size parameters due to imperfect correlation between the causal and marker loci – occur simultaneously, so that an additive disease model is assumed at B and B', then the λ_s attributable to these loci is estimated to be 1.0027.

Our example shows that the presence of unidentified interactions can cause underestimation of λ_s. Four loci with effects equivalent to those observed at B and B' would be required to recover the true λ_s explained by A and A'. This example is meant only as an illustration of the potential impact of interactions on heritability estimates. Although the relationship between the disease model and sibling recurrence risk has been studied (see for example Risch (1990) and Rybicki & Elston (2000)), our analysis shows that the current practice of assuming additivity when calculating the cumulative λ_s across a number of loci may, in some instances, be misleading.

6.3 Conclusions

The correlation structure in the human genome has enabled the GWAS method to find regions associated with disease without having to genotype all known genetic variants. Although this approach has been successful, it follows that observed GWAS associations often will be only surrogates for the causal variants and will typically represent a noisy measurement of them. One consequence of this is that the disease model inferred from

associated loci may be a distorted version of the true disease model. Through analytical derivations, we have characterised the relationship between disease model parameters and LD, and the resulting impact on power. These derivations show that dominance interaction effects tend to decay quickly, and that such distortions therefore tend to make the disease model look more like a additive model as the correlation between causal and hit SNPs decreases.

To quantify the effect of distortion on observed GWAS outcomes, we ran an extensive simulation study designed to mimic patterns of LD in European Caucasian populations. We considered recessive and dominant models, both representing natural extremes for deviation away from an additive model. We were specifically interested in the power of detecting such deviations, and also ran simulations under the additive model for comparison.

Our analyses showed that, if the true model is recessive or dominant but the locus is nonetheless detected by using the trend test, then a standard test will often also successfully detect deviation away from an additive model. Informally, for the relatively small effect sizes typical at GWAS loci, the effect is unlikely to be detected unless the causal variant is relatively common and well tagged by the SNPs on the chip. The high correlation between the causal and hit SNPs then means that there is reasonable power to detect deviation from the additive model, even under conditions of model distortion. Firstly, while these results are encouraging, we note: (1) that the dominant and recessive models are extreme, and (2) that power to detect non-additive models which are "closer" to the additive model will be lower. Secondly, as our simulations show, there will be settings where the model distortion is such that under the recessive and dominant models the locus is not detected at all using the trend test.

Relatively few GWAS associations are known to follow specific, non-additive models. It may be that testing for deviations is not done routinely, although even in studies where such investigations have been carried out, few SNPs have shown convincing

evidence of recessive or dominant effects (e.g. Wellcome Trust Case Control Consortium (2007)). Our simulations have shown that such effects will often be detectable, and therefore, it is worth explicitly testing associated loci for deviations. As noted above, real disease effects may not deviate as much as fully recessive and dominant effects, and small deviations from multiplicativity will be relatively hard to detect, and easily disguised with only a slight amount of distortion.

One consideration in the analysis of GWAS data is which statistical test or model to use for the initial genome-wide scan, as discussed in the Introduction. Since we expect to detect SNPs that are affected to a greater or lesser extent by distortion, a sensible default choice is the trend test, which is well-powered for additive effects. It also has the benefit of being more robust to genotyping error, than, for example, the general test (Ahn *et al.*, 2007), which performed comparably in our simulations. Others have made similar recommendations (Cantor *et al.*, 2010; Iles, 2008). Nevertheless, the trend test can be usefully complemented by the general test (Wellcome Trust Case Control Consortium, 2007), or other approaches for investigating nonadditive models, such as the deviation test.

6.3.1 Final thoughts on missing heritability

In the next few years, two types of new data are likely to become available that may affect the analyses presented in this thesis: genotyping data at larger chips which assay 2-5 million variants instead of the half million to million considered here, and the 1000 Genomes data set, which is predicted to extend the reach of imputation down to lower-frequency alleles (1000 Genomes Project Consortium, 2010). It is difficult to forecast the impact of these innovations on the types of questions considered here, since it depends on the extent to which the higher-density chips and the 1000 Genomes data accurately capture alleles at frequencies of 1-10%, and, further, how important this frequency class turns out to be as a contributor to common disease risk. To the

extent that imputation with the 1000 Genomes data and higher-density chips improves pairwise LD between potential causal SNPs and their surrogates, both of these advents are likely to reduce effect size underestimation and, as a corollary, improve heritability estimates.

Another trend worth mentioning is increasing study sample size, most commonly by meta-analysis. Our method of simulation is not designed to assess effect size underestimation in the context of meta-analysis. Among other complications, these studies often combine information across genotyping chips, each of which tags the underlying causal variants differently. However, these studies have been successful at increasing power, and have correspondingly detected more variants contributing to common diseases and traits, which in turn explain more of the heritability (Lander, 2011). Recently, statistical methods have been suggested that predict the likely impact of as-yet-undiscovered variants on heritability estimates (Park *et al.*, 2010; Yang *et al.*, 2010a). These methods point to limited sample size as a source of missing heritability.

The range of potential types of interactions is huge, and our analysis considered only recessive and dominant alleles at a single SNP and their theoretical analogs, two unlinked loci with a single additive interaction parameter. This is a small subset of even the space of possible two-locus interactions (Hallgrímsdóttir & Yuster, 2008), and we did not investigate the potential impact of multiple SNPs in LD, compound recessive alleles, or the rich variety of other models that could be considered. The more general question, of the potential impact of interactions among many loci on estimates of heritability, and our ability to detect these, remains open and seems likely to be a fruitful area of inquiry. It is simple to construct a model that explains any amount of heritability; the more interesting question is to find models that are consistent with observed data and to construct experiments that would distinguish these models from the simplest additive model which underlies most heritability calculations to date. Addressing this question would determine the extent to which heritability is a reliable measure of progress in

genetic mapping studies.

Finally, looking into the more distant future, it is likely that the cost of whole-genome sequencing will some day drop to the level at which it is feasible to conduct studies without the need for tag SNPs at all, thereby eliminating effect size underestimation entirely. Although this will ultimately allow us to gain direct access to the causal alleles, it is worth remembering that the same patterns of correlation that enabled association studies in the first place will limit our ability to distinguish these alleles from their many surrogates in the genome.

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073. 22, 41, 76, 123

Ahn, K., Haynes, C., Kim, W., Fleur, R.S., Gordon, D. & Finch, S.J. (2007). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Annals of Human Genetics*, **71**, 249–261, PMID: 17096677. 123

Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics*, **69**, 936–950, PMID: 11565063. 2

Altshuler, D., Daly, M.J. & Lander, E.S. (2008). Genetic mapping in human disease. *Science (New York, N.Y.)*, **322**, 881–888, PMID: 18988837. 1, 2, 3

Anderson, C.A., Soranzo, N., Zeggini, E. & Barrett, J.C. (2011). Synthetic associations are unlikely to account for many common disease Genome-Wide association signals. *PLoS Biol*, **9**, e1000580. 61

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386. 14, 39

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, **7**, 781–791, PMID: 16983374. 12, 13, 14, 15

Barbujani, G. & Goldstein, D.B. (2004). AFRICANS AND ASIANS ABROAD: genetic diversity in europe. *Annual Review of Genomics and Human Genetics*, **5**, 119–150. 31

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton, A., Dassopoulos, T., Datta, L.W., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R.J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.P., de Vos, M., Vermeire, S., Louis, E., Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C.G., Parkes, M., Georges, M. & Daly, M.J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962. 53, 56

Bhangale, T.R., Rieder, M.J. & Nickerson, D.A. (2008). Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.*, **40**, 841–843. 103

Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331, PMID: 6247908. 2

Browning, S.R. & Browning, B.L. (2007). Rapid and accurate haplotype phasing and Missing-Data inference for Whole-Genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**, 1084–1097. 23, 76

Cadwell, K., Liu, J.Y., Brown, S.L., Miyoshi, H., Loh, J., Lennerz, J.K., Kishi, C., Kc, W., Carrero, J.A., Hunt, S., Stone, C.D., Brunt, E.M., Xavier, R.J., Sleckman, B.P., Li, E., Mizushima, N., Stappenbeck, T.S. & IV, H.W.V. (2008). A key role for autophagy and the autophagy gene atg16l1 in mouse and human intestinal paneth cells. *Nature*, **456**, 259–263. 25

Cantor, R.M., Lange, K. & Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, **86**, 6–22, PMID: 20074509 PMCID: 2801749. 112, 123

Capen, E., Clapp, R. & Campbell, W. (1971). Competitive bidding in High-Risk situations. *Journal of Petroleum Technology*, **23**. 17

Chapman, J., Cooper, J., Todd, J. & Clayton, D. (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. 102

Chen, Y., Lin, C. & Sabatti, C. (2006). Volume measures for linkage disequilibrium. *BMC Genetics*, **7**, 54. 6

Clayton, D.G. (2009). Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genet*, **5**, e1000540. 25, 63

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., Nutland, S., Howson, J.M.M., Faham, M., Moorhead, M., Jones, H.B., Falkowski, M., Hardenbol, P., Willis, T.D. & Todd, J.A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, **37**, 1243–1246, PMID: 16228001. 9

Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468. 101

Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**, 997–1004, PMID: 11315092. 9

Donnelly, P. (2006). Modelling genes: mathematical and statistical challenges in genomics. *Proceedings of the International Congress of Mathematicians.*. 4

Easton, D.F. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Research*, **1**, 14–17, PMID: 11250676 PMCID: 138504. 10

Eberle, M.A., Ng, P.C., Kuhn, K., Zhou, L., Peiffer, D.A., Galver, L., Viaud-Martinez, K.A., Lawley, C.T., Gunderson, K.L., Shen, R. & Murray, S.S. (2007). Power to detect risk alleles using Genome-Wide tag SNP panels. *PLoS Genet*, **3**, e170. 66

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. & Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews. Genetics*, **11**, 446–450, PMID: 20479774. 23, 43

ENCODE Project (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640. 5, 31, 32

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, **42**, 558–560, PMID: 20581876. 10

Goldstein, D.B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1698. 10

REFERENCES

Hallgrímsdóttir, I.B. & Yuster, D.S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, **9**, 17–17, PMID: 18284682 PMCID: 2289835. 124

Hamburg, M.A. & Collins, F.S. (2010). The path to personalized medicine. *The New England Journal of Medicine*, **363**, 301–304, PMID: 20551152. 1, 23

Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)*, **24**, 2786–2787, PMID: 18842601. 41

Hill, W.G., Goddard, M.E. & Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, **4**, e1000008, PMID: 18454194. 103

Hindorff, L., Junkins, H., Mehta, J. & Manolio., T. (2010). A catalog of published genome-wide association studies. 3, 10, 11, 101

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362–9367, PMID: 19474294. 101

Hirschhorn, J.N. (2009). Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699–1701. 10

Hirschhorn, J.N. & Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, **6**, 95–108, PMID: 15716906. 2, 8

Howie, B.N., Donnelly, P. & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**, e1000529, PMID: 19543373. 23

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A. & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*, **84**, 235–250, PMID: 19215730. 22

Iles, M.M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.*, **4**, e33. 29, 46, 123

International HapMap Consortium (2003). The international HapMap project. *Nature*, **426**, 789–796, PMID: 14685227. 31

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320. 5, 6, 7, 51, 69, 73, 88, 96

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861. 5, 31, 73

Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A.D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E., Schwartz, S.M., Voight, B.F., Elosua, R., Salomaa, V., O'Donnell, C.J., Dallinga-Thie, G.M., Anand, S.S., Yusuf, S., Huff, M.W., Kathiresan, S. & Hegele, R.A. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, **42**, 684–687, PMID: 20657596. 10

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Jonasdottir, A., Frigge, M.L., Gylfason, A., Olason, P.I., Gudjonsson, S.A., Sverrisson, S., Stacey, S.N., Sigurgeirsson, B., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U. & Stefansson, K. (2009). Parental origin of sequence variants associated with complex diseases. *Nature*, **462**, 868–874, PMID: 20016592. 12

Kraft, P. (2008). Curses–winner's and otherwise–in genetic epidemiology. *Epidemiology (Cambridge, Mass.)*, **19**, 649–651; discussion 657–658, PMID: 18703928. 17

Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, **11**, 241–247. 2

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197. 124

Lewontin, R.C. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, **49**, 49–67, PMID: 17248194. 8

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233. 36

Li, Y., Willer, C., Sanna, S. & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, **10**, 387–406. 22

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, **33**, 177–182. 18

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21, PMID: 18987709. 23, 63

Manolio, T.A., Brooks, L.D. & Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590–1605. 3, 4, 5, 10, 43

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A. & Visscher, P.M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753. 23

Marchini, J. & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, **11**, 499–511. 22, 23, 66, 72, 76

Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512–517. 9

Marchini, J., Donnelly, P. & Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, **37**, 413–417, PMID: 15793588. 112, 114

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913. 16, 23, 30

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. & Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics*, **9**, 356–369, PMID: 18398418. 9

McClellan, J. & King, M. (2010). Genetic heterogeneity in human disease. *Cell*, **141**, 210–217, PMID: 20403315. 10

REFERENCES

McKusick-Nathans Institute of Genetic Medicine, M., Johns Hopkins University (Baltimore & National Center for Biotechnology Information, M., National Library of Medicine (Bethesda (????). Online mendelian inheritance in man. 2

Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, **310**, 321–324, PMID: 16224025. 4, 6, 37

Newton-Cheh, C. & Hirschhorn, J.N. (2005). Genetic association studies of complex traits: design and analysis issues. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **573**, 54–69. 9

Park, J., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J. & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, **42**, 570–575, PMID: 20562874. 63, 124

Pei, Y., Li, J., Zhang, L., Papasian, C.J. & Deng, H. (2008). Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE*, **3**, e3551. 76

Pharoah, P., Day, N., Duffy, S., Easton, D. & Ponder, B. (1997). Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int. J. Cancer*, **71**, 800–809. 58

Pharoah, P.D., Antoniou, A.C., Easton, D.F. & Ponder, B.A. (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.*, **358**, 2796–2803. 53, 56, 62

Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403 –411. 13

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909, PMID: 16862161. 9

Pritchard, J. & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14. 6, 8, 102

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, **69**, 124–137, PMID: 11404818. 3, 18

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959, PMID: 10835412. 9

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575, PMID: 17701901. 12

Reich, D.E. & Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends in Genetics*, **17**, 502–510. 1, 3

Rienzo, A.D. & Hudson, R.R. (2005). An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends in Genetics*, **21**, 596–601. 38

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.*, **46**, 222–228. 23, 24, 57, 121

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, **273**, 1516–1517, PMID: 8801636. 2

Rybicki, B.A. & Elston, R.C. (2000). The relationship between the sibling recurrence-risk ratio and genotype relative risk. *American Journal of Human Genetics*, **66**, 593–604, PMID: 10677319. 121

Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, **53**, 1253–1261, PMID: 9423247. 14, 15

Sawyer, S.A. & Hartl, D.L. (1992). Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176. 19

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. *American Journal of Human Genetics*, **78**, 629–644, PMID: 16532393. 23

Schouten, E.G., Dekker, J.M., Kok, F.J., Le Cessie, S., Van Houwelingen, H.C., Pool, J. & Vanderbroucke, J.P. (1993). Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. *Stat. Med.*, **12**, 1733–1745. 13, 17, 104

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S. & Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science (New York, N.Y.)*, **316**, 1341–1345, PMID: 17463248. 23

Sham, P.C., Cherny, S.S., Purcell, S. & Hewitt, J.K. (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. 103

Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics*, **182**, 845–850, PMID: 19416939. 23

Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. (2011a). Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genetics*. 20

Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. (2011b). Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genetics*, **7**, e1001337, PMID: 21437273. 44

Spencer, C.C.A., 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. 30, 41, 66, 67, 69, 83, 88, 103

Stephens, M. & Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews. Genetics*, **10**, 681–690, PMID: 19763151. 12

Su, Z., Marchini, J. & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics (Oxford, England)*, **27**, 2304–2305, PMID: 21653516. 36

Teare, M.D. & Barrett, J.H. (2005). Genetic linkage studies. *The Lancet*, **366**, 1036–1044. 2

Thomas, D.C. (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, New York. 24, 57

Thornton-Wells, T.A., Moore, J.H. & Haines, J.L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics*, **20**, 640–647. 112

REFERENCES

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S., Hughes, D., Warren-Perry, M., Tapper, W., Eccles, D., Evans, D.G., Hooning, M., Schutte, M., van den Ouweland, A., Houlston, R., Ross, G., Langford, C., Pharoah, P.D.P., Stratton, M.R., Dunning, A.M., Rahman, N. & Easton, D.F. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet*, **42**, 504–507. 10

Tysk, C., Lindberg, E., Järnerot, G. & Flodérus-Myrhed, B. (1988). Ulcerative colitis and crohn's disease in an unselected population of monozygotic and dizygotic twins. a study of heritability and the influence of smoking. *Gut*, **29**, 990–996, PMID: 3396969 PMCID: 1433769. 58

Vukcevic, D. (2009). *Bayesian and frequentist methods and analyses of genome-wide association studies*. Ph.D. thesis, University of Oxford. 16, 117

Vukcevic, D., Hechter, E., Spencer, C. & Donnelly, P. (2011). Disease model distortion in association studies. *Genetic Epidemiology*, PMID: 21416505. 49, 103, 104, 105, 114, 117, 119, 120

Wakefield, J. (2008). Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.*, **33**, 79–86. 18, 20

Walker, F.O. (2007). Huntington's disease. *Lancet*, **369**, 218–228, PMID: 17240289. 1

Wallace, C. & Clayton, D. (2003). Estimating the relative recurrence risk ratio using a global cross-ratio model. *Genetic Epidemiology*, **25**, 293–302, PMID: 14639699. 25, 63

Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B. & Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. *The American Journal of Human Genetics*. 29

Wang, W.Y.S., Barratt, B.J., Clayton, D.G. & Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics*, **6**, 109–118, PMID: 15716907. 3, 18

Weijnen, C.F., Rich, S.S., Meigs, J.B., Krolewski, A.S. & Warram, J.H. (2002). Risk of diabetes in siblings of index cases with type 2 diabetes: implications for genetic studies. *Diabetic Medicine: A Journal of the British Diabetic Association*, **19**, 41–50, PMID: 11869302. 58

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678. 9, 20, 25, 26, 42, 63, 123

Xiao, R. & Boehnke, M. (2011). Quantifying and correcting for the winner's curse in quantitative-trait association studies. *Genetic Epidemiology*, **35**, 133–138, PMID: 21284035. 18

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E. & Visscher, P.M. (2010a). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569, PMID: 20562875. 23, 63, 124

Yang, W., Shen, N., Ye, D., Liu, Q., Zhang, Y., Qian, X., Hirankarn, N., Ying, D., Pan, H., Mok, C.C., Chan, T.M., Wong, R.W.S., Lee, K.W., Mok, M.Y., Wong, S.N., Leung, A.M.H., Li, X., Avihingsanon, Y., Wong, C., Lee, T.L., Ho, M.H.K., Lee, P.P.W., Chang, Y.K., Li, P.H., Li, R., Zhang, L., Wong, W.H.S., Ng, I.O.L., Lau, C.S., Sham, P.C., Lau, Y.L. & (ALGC), A.L.G.C. (2010b). Genome-Wide association study in asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet*, **6**, e1000841. 69

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., BostrÖØm, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J., Deodhar, P., Ding, C.J., Doney, A.S., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., JÕØrgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N., Payne, F., Perry, J.R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbaek, A., Shields, B., SjÕØgren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J., Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., McCarthy, M.I., Boehnke, M. & Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645. 52, 56

Zheng, G., Joo, J., Zaykin, D., Wu, C. & Geller, N. (2009). Robust tests in Genome-Wide scans under incomplete linkage disequilibrium. *Statistical Science*, **24**, 503–516. 103

Zhong, H. & Prentice, R.L. (2010). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology*, **34**, 78–91, PMID: 19639606. 18

Zollner, S. & Pritchard, J.K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, **80**, 605–615, PMID: 17357068. 17, 18

Zondervan, K.T. & Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, **5**, 89–100. 18, 102