On genetic variants underlying common disease

Eliana Hechter

Statistics Department

University of Oxford

A thesis submitted for the degree of

DPhil

2010

Contents

1

Introduction

Trait variation in human populations, including disease risk, shows correlation between relatives, suggesting an underlying genetic contribution. The number and frequency of alleles contributing to such 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's, 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 crucial importance to the development of new treatments (Hamburg & Collins, 2010). Diseases in which the pathophysiology is well-understood, such as inborn errors of metabolism, may be suited to candidate gene studies, in which particular genes of interest can be sequenced in order to determine which variants give rise to illness (Cichon *et al.*, 2009). However, this approach is fundamentally limited by the necessity for a strong hypothesis.

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

relate polymorphic loci with regions of the genome that are linked in this way. Linkage studies examine affected families, and have been very successful in the identification of Mendelian disease genes (Hirschhorn & Daly, 2005)[1], with over 2800 disorders tied to a specific locus (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 to detect many loci associated with complex diseases (Altmüller *et al.*, 2001). Recently, another type of genetic mapping, called a genome-wide association study (GWAS), has proved successful at identifying common variants that contribute to complex diseases. 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. Such modest 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 seems plausible. Consistent with this expectation, within the last few years, 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 around 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 neighbors. 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

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

between cases and controls that do not owe 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, considering the entire human population as a large pedigree. 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 they appear together in the population more often than expected under independence.

Linkage disequilibrium arises as a result of a process during meiosis called recombination in which parental chromosomes align and exchange segments so that the single chromosome passed on to one of the four gametes produced will not simply be a copy of one of the original chromosomes, but a mosaic of the two (Ewens, 2004). Recombination is neither randomly nor uniformly distributed across the genome but occurs especially often at sites termed 'recombination hotspots' (Myers *et al.*, 2005). The result is that chromosomal segments of varying lengths separated by recombination hotspots are in linkage disequilibrium.

A single nucleotide polymorphism (SNP) is a site in the genome at which individuals in a population differ by a single nucleotide base. SNPs are defined as "common" if 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). 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 (Frazer *et al.*, 2007). The HapMap project catalogued these correlations systematically (International HapMap Consortium, 2005).

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). Each of the 96 chromosomes in HapMap was resequenced 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 many consecutive common variants are highly correlated, separated by recombination hotspots (Hirschhorn & Daly, 2005).

A commonly used measure of the LD between two nearby SNPs is r^2. If we think of SNPs as 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 standard correlation coefficient between these random variables (Chen *et al.*, 2006). Suppose two SNPs A and B have minor allele frequencies p and q. Then r^2 is defined as,

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

where $x = pq$ if A and B are independent and $x = \Pr(a, b)$, the probability of the haplotype (a, b), otherwise. Note that the earlier choice of the minor allele was arbitrary. For the purposes of this thesis we will assume that all SNPs are biallelic.

Other measures of LD exist and each has different advantages. We choose to describe our results with the r^2 statistic because it is most commonly used in the association study context. For further discussion of LD measures, see for example Pritchard & Przeworski (2001b).

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 artifacts and point to references for methods developed to circumvent them.

1.2.1 Avoiding artifacts

Some of the potential sources of false positives due to systematic differences include:

- Population stratification

- Misclassification bias
- Differences in genotyping between cases and controls

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 (Wellcome Trust Case Control Consortium, 2007). Freedman *et al.* (2004) show empirically how such signals can arise by analyzing 11 case-control and case-cohort association studies to look for signals of stratification, and demonstrate that modest amounts of stratification are likely to exist even when samples are carefully chosen. Several methods, including the model-based clustering program STRUCTURE (Pritchard *et al.*, 2000), were developed to detect population stratification on the basis of markers spread throughout the genome and suggestions have been forwarded 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).

Misclassification bias can stem from poorly phenotyped cases or the decision to use un-phenotyped controls (often called population controls), any number of whom may have or may develop the disease of interest. Unless the disease is quite common, this has a relatively modest effect on power (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 avoid spurious associations. In less than ideal conditions, cluster plots showing the fluorescent signal data points for the two alleles at a SNP can be inspected visually when the three genotypes do not appear distinct, a method adopted in 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 1.1 shows the distribution of effect sizes from the NHGRI GWAS Catalog. 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).

As an example of the contrast between Mendelian disease gene effects and those discovered by GWAS, compare the frequencies and effect sizes of rare mutations known to confer susceptibility to breast cancer with GWAS variants. The BRCA1 and BRCA2 genes that predispose to a Mendelian breast cancer subtype have frequencies of less than 1% and those who have the mutation are approximately 8 times more likely than someone without the mutation to develop the disease in their lifetimes (Easton, 1999). In contrast, 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 common and rare variation contribute about equally to the explained familial inheritance, rare variation contributes incrementally 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 views ranging from dubious about the role of common variants in disease risk (Goldstein, 2009; McClellan & King, 2010) to celebratory of the number and variety of new biological clues for disease mechanism (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.

1.4 Statistical methods

We now briefly review some of the statistical methods used in GWAS. This is by no means a comprehensive overview, since it covers only those methods which we employ

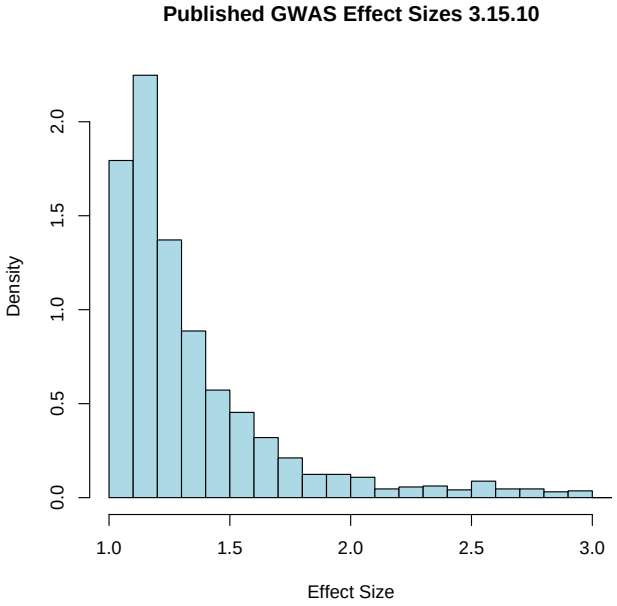

Figure 1.1: Effects discovered by GWAS - Distribution of effects

later to mimic a standard GWAS design. For further information on Bayesian methods for association testing, we refer the reader to Marchini *et al.* (2007); Wakefield (2008); Wellcome Trust Case Control Consortium (2007). In addition to the tests we will discuss below, Purcell *et al.* (2007) discusses 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. One common parameterization is,

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

which writes the log-odds of disease as a linear function of the number of allele copies. Under this model, e^{β} is a measure of the effect size.

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 and 0 for homozygotes. In GWAS it is standard to use (un-phenotyped) 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, as we do in Chapter 6.

Figure 1.2 helps guide an intuitive understanding of the parameterisation, which is similar to that of Balding (2006). The γ parameter measures the deviation from the simplest multiplicative model, represented as linear on the log risk scale of the figure. As shown in the figure, a recessive model, in which both allele copies are necessary to influence disease risk, can be achieved by setting $\gamma = -\beta$ and a dominant model, in which one allele copy is sufficient, is satisfied when $\gamma = \beta$. These disease models will be explored further in Chapter 6.

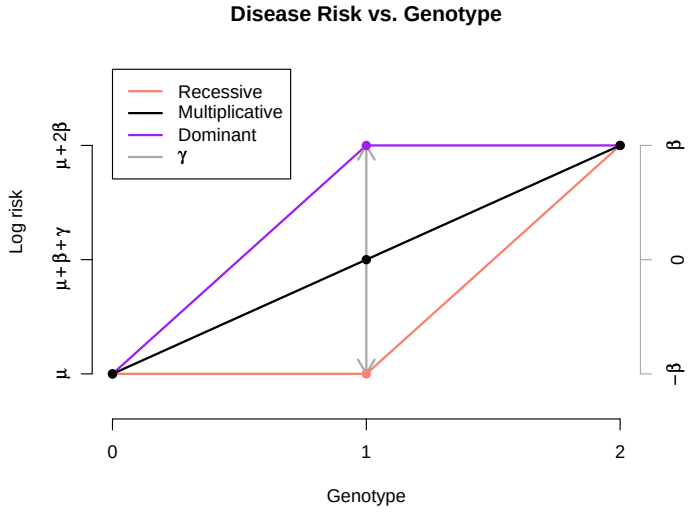

Figure 1.2: Representation of three models in the log risk regression - The grey arrows show how the γ parameter acts to perturb the model at the heterozygote away from its value under a multiplicative model.

Although we are mainly concerned with the setting of case-control studies, note that there is a correspondence between trait variation models and disease risk models

if one assumes that the disease risk can be modeled as a sum of small environmental and genetic contributions that are distributed normally, and only after some threshold risk is overcome does the individual develop the disease. This is referred to as a 'liability threshold' model and is described in Dempster & Lerner (1950). Yang *et al.* (2010b) give an analytic formula for the fraction of variance explained on the liability scale in terms of the frequency and effect size of an allele, which allows for conversion between the binary and the liability scales.

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 greater power under a multiplicative 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,

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

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 $S = \sum_i s_i$. Stated this way it is 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.

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 will inevitably arise by chance when performing so many association tests. 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,

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

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-multiplicative disease models.

1.4.4 Estimates of effect sizes

We will discuss effect sizes in the setting of the multiplicative 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. The odds ratio of allele A is defined as,

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

which is equivalent to e^{β} in the model defined by Equation 1.1. For clarity, we have written the OR in terms of the probabilities of health and disease, but in practice these are estimated on the basis of probabilities of being a control or a case. 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 multiplicative disease risk model. The relative risk is defined as,

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

Each measure has its merits. The odds ratio, while less intuitive, is directly tied to the logistic regression model and can be directly estimated from case control data. The relative risk, on the other hand, is easier to interpret but requires knowledge of the prevalence of the disease in a case control study. Specifying a multiplicative risk model for one measure does not give an multiplicative risk model for the other measure, although there are two special cases in which the measures coincide. Luckily for our purposes, one of these situations is in a GWAS with population-based controls, so we use the two measures interchangeably throughout the thesis depending on the context. We now review the special situations.

First, when the disease is rare, OR $\approx$ RR. Second is the setting in which 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 estimates the RR and not the OR (Schouten *et al.*, 1993) and the prevalence is needed to estimate the OR. Since GWAS often use population controls to increase sample size or to avoid the costs of phenotyping, we will assume this approximation holds.

1.4.5 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 need to assume a joint distribution of the true risk allele frequencies and effect sizes, which is of course unknown. Here we define two distributions which represent two different sets of assumptions about these unknowns. 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.

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, and instead our approach aims to capture two different perspectives on unknown effect sizes.

1.4.5.1 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 1.3). To get a sense of this distribution, it 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).

1.4.5.2 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:

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

Again, this form follows that in the Methods section of Wellcome Trust Case Control Consortium (2007). 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 rarer SNPs), then the prior puts considerably more weight on larger RRs. 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 1.3 shows the two prior distributions side by side.

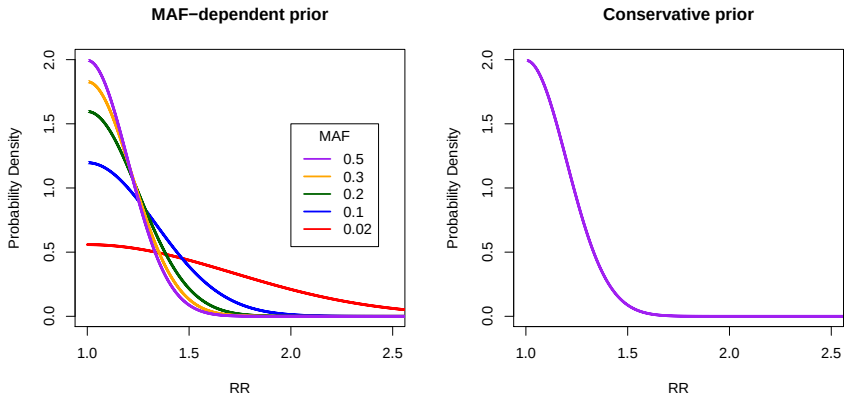

Figure 1.3: Priors on the effect size - Relationship between allele frequency and effect size for the conservative and MAF-dependent priors.

1.4.6 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 explaining 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 to serve as a benchmark by which we can measure how much further we have to go. 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.

There are two primary measures of heritability reported at the conclusion of a GWAS. The first is λ_s, the sibling recurrence risk ratio, the properties of which are discussed at length in Risch (1990). It is given by the formula,

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

where $\Pr(Y = 1)$ is the prevalence and k denotes the affected sib; here, $Y_j = 1$ indicates a that the proband has a disease phenotype. Sibling recurrence risk expresses how much more likely a sibling is to have a disease than a member of the population at large (Thomas, 2004). We compute the sibling relative recurrence risk of an individual allele under an additive disease model as a function of the effect size and allele frequency, and

then combine the sibling recurrence risks at all of the discovered loci for a particular disease by multiplication. For details on why multiplication across alleles estimates λ_s, see Risch (1990).

Let G_i denote the genotype of the f father, m mother, j proband, and k affected sibling respectively, and let Q_g denote the genotype probabilities under Hardy-Weinberg equilibrium. Let q be the frequency of the A allele and $p = 1 - q$ the frequency of the a allele. Let $f_g = Pr(Y = 1|G = g)$ be the penetrance for genotype g. Finally, we need transition probabilities $T_{g_j g_m g_f} = Pr(G_h = g_j|G_m = g_m, G_f = g_f)$. 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,

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

What remains is to convert relative risks or odds ratios into penetrances. To do this recall that,

$$\Pr(D) = \Pr(D|aa)Q_0 + \Pr(D|aA)Q_1 + \Pr(D|AA)Q_2\,.$$

Dividing through by $\Pr(D|AA)$ and rearranging we obtain,

$$\Pr(D|AA) = \frac{\Pr(D)}{\sum_g RR_g Q_g}\,,$$

so that $f_g = \mathrm{RR}_g \Pr(D|AA)$ give the penetrances.

As noted in Clayton (2009), λ_s is likely to be overestimated by sibling 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 or twin studies may be lower than sometimes assumed.

The second measure of heritability is the variance explained, which can be estimated directly from the data in a continuous trait but requires imposing a liability threshold model for a binary trait. Yang *et al.* (2010b) give a formula that allows us to convert from the binary to the liability scale. Let f be the frequency of the risk allele, μ the

disease prevalence, and OR its effect size. Then the fraction of variance explained is given by,

$$\frac{2\mu^2 f(1-f)(\mathrm{OR}-1)^2}{z^2(1-f+f\mathrm{OR})^2}$$

where z is the height of the normal distribution at the point y such that the area under the curve past y is equal to μ. Figure 1.4 shows the relationship between variance explained and λ_s for breast cancer and Crohn's disease, using loci discovered in Turnbull *et al.* (2010) and Barrett *et al.* (2008), respectively. The grey points denote as-yet-undiscovered loci using the simplest version of the power correction suggested by Park *et al.* (2010). From the figure, it is clear that these measures broadly agree for the two diseases shown. We use λ_s throughout the thesis since it does not rely on the assumption of an underlying liability threshold model.

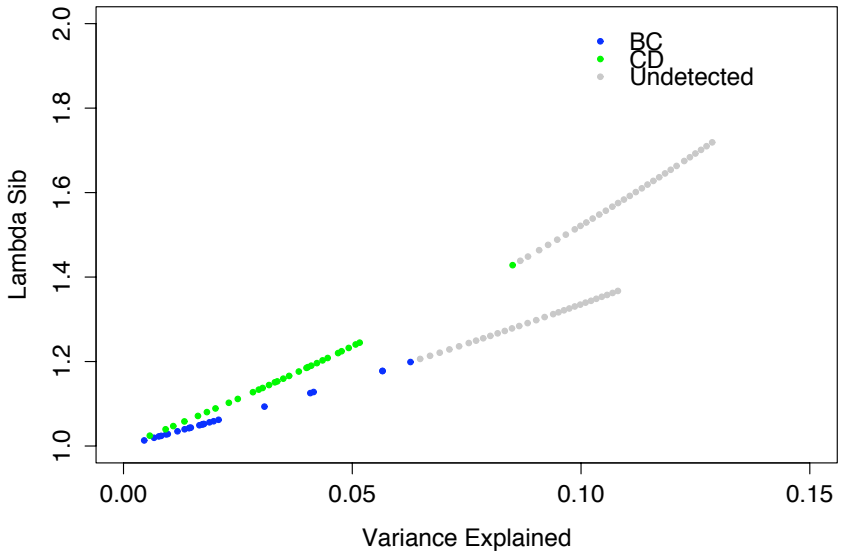

Figure 1.4: Relationship between λ_s and variance explained - λ_s and variance explained are plotted for loci discovered in breast cancer (BC) and Crohn's disease (CD). Note that the prevalence affects the slope of the correlation – Crohn's disease is much rarer than breast cancer (Steed *et al.*, 2010).

1.4.7 Imputation

Description of imputation with references to further reading (Marchini *et al.*, 2007).

2

A GWAS simulation framework

A central aim of this thesis is to try and understand properties of the causal variants underlying GWAS signals. This chapter develops a method by which we can solve 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 ask the reverse 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 a multiplicative 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)) to simulate a large population of chromosomes, whose patterns of LD match those in the HapMap analysis panel. From this population a case-control sample is drawn, with the controls sampled randomly from the population and the cases 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 only considered 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 (Frazer *et al.*, 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 (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 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 panel.

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*. GWAS case and control samples were then subsampled from the reference panel, as described below.

2.2.1 A brief description of HAPGEN

HAPGEN uses a population genetic model that incorporates the processes of mutation and fine-scale recombination to generate individuals 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

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

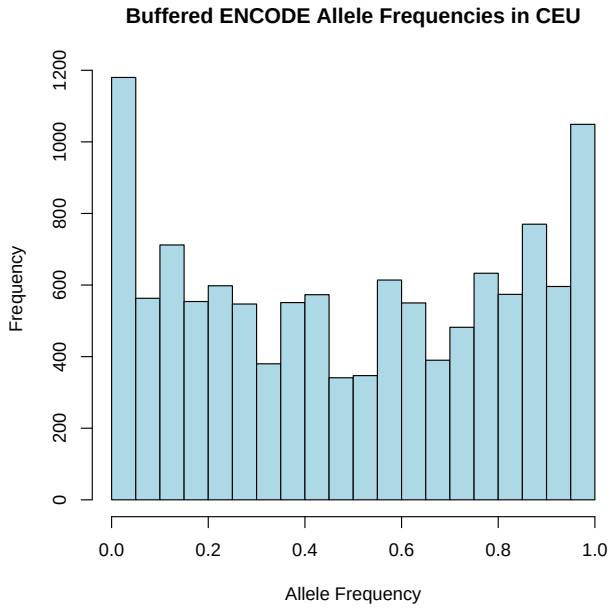

Figure 2.1: Allele frequency distribution in combined ENCODE regions in the CEU panel - Distribution of allele frequencies; note that in simulations, we test both alleles.

a piece of paper. We begin with the set of known haplotypes, as shown on the left side of Figure 2.2. 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.

Intuitive description of HAPGEN as a mosaic of HapMap haplotypes with references for further reading (Spencer *et al.*, 2009).

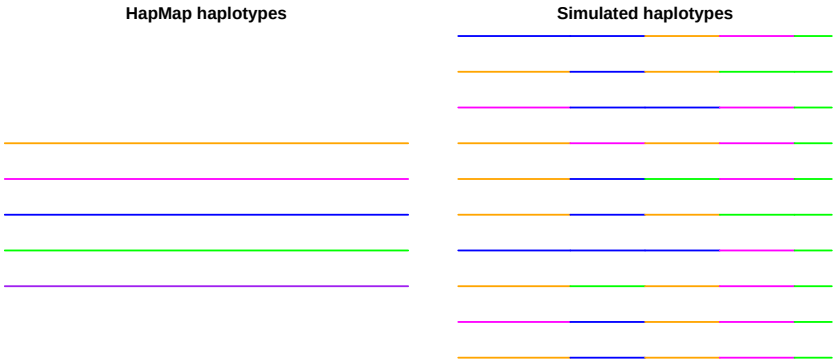

Figure 2.2: Cartoon of the HAPGEN program - The set of known haplotypes is extended by creating additional haplotypes that are a mosaic of the known haplotypes.

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 reigons 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 a multiplicative disease model for disease risk (additive on the log scale) 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 large 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 the 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.

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}.

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.

References

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

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386. 9, 23

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, **7**, 781–791, PMID: 16983374. 8, 9, 10

Barbujani, G. & Goldstein, D.B. (2004). AFRICANS AND ASIANS ABROAD: genetic diversity in europe. *Annual Review of Genomics and Human Genetics*, **5**, 119–150. 19

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. 15

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. 42

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. 1

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. 52

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. 42

Chen, Y., Lin, C. & Sabatti, C. (2006). Volume measures for linkage disequilibrium. *BMC Genetics*, **7**, 54. 4

Cichon, S., Craddock, N., Daly, M., Faraone, S.V., Gejman, P.V., Kelsoe, J., Lehner, T., Levinson, D.F., Moran, A., Sklar, P. & Sullivan, P.F. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American Journal of Psychiatry*, **166**, 540–556, PMID: 19339359. 1

Clayton, D.G. (2009). Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genet*, **5**, e1000540. 14

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. 5

Consortium, I.H. (2003). The international HapMap project. *Nature*, **426**, 789–796, PMID: 14685227. 19

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. 41

Dempster, E.R. & Lerner, I.M. (1950). Heritability of threshold characters. *Genetics*, **35**, 212–236, PMID: 17247344. 9

Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**, 997–1004, PMID: 11315092. 5

Easton, D.F. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Research*, **1**, 14–17, PMID: 11250676 PMCID: 138504. 6

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. 13

ENCODE Project (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640. 3, 18, 19

Ewens, W.J. (2004). *Mathematical Population Genetics*. Springer, 2nd edn. 3

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang,

REFERENCES

H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M.M., Tsui, S.K., Xue, H., Wong, J.T., Galver, L.M., Fan, J.B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M.S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.F., Phillips, M.S., Roumy, S., Sallée, C., Verner, A., Hudson, T.J., Kwok, P.Y., Cai, D., Koboldt, D.C., Miller, R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.C., Mak, W., Song, Y.Q., Tam, P.K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C.P., Delgado, M., Dermitzakis, E.T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B.E., Whittaker, P., Bentley, D.R., Daly, M.J., de Bakker, P.I., Barrett, J., Chretien, Y.R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D.J., Sabeti, P., Saxena, R., Schaffner, S.F., Sham, P.C., Varilly, P., Altshuler, D., Stein, L.D., Krishnan, L., Smith, A.V., Tello-Ruiz, M.K., Thorisson, G.A., Chakravarti, A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Lin, S., Abecasis, G.R., Guan, W., Li, Y., Munro, H.M., Qin, Z.S., Thomas, D.J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L.R., Clarke, G., Evans, D.M., Morris, A.P., Weir, B.S., Tsunoda, T., Mullikin, J.C., Sherry, S.T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Ajayi, I., Aniagwu, T., Marshall, P.A., Nkwodimmah, C., Royal, C.D., Leppert, M.F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I.F., Knoppers, B.M., Foster, M.W., Clayton, E.W., Watkin, J., Gibbs, R.A., Belmont, J.W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G.M., Wheeler, D.A., Yakub, I., Gabriel, S.B., Onofrio, R.C., Richter, D.J., Ziaugra, L., Birren, B.W., Daly, M.J., Altshuler, D., Wilson, R.K., Fulton, L.L., Rogers, J., Burton, J., Carter, N.P., Clee, C.M., Griffiths, M., Jones, M.C., McLay, K., Plumb, R.W., Ross, M.T., Sims, S.K., Willey, D.L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J.C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A.L., Brooks, L.D., McEwen, J.E., Guyer, M.S., Wang, V.O., Peterson, J.L., Shi, M., Spiegel, J., Sung, L.M., Zacharia, L.F., Collins, F.S., Kennedy, K., Jamieson, R. & Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861. 3, 18

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N. & Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**, 388–393, PMID: 15052270. 5

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, **42**, 558–560, PMID: 20581876. 6

Goldstein, D.B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1698. 6

Hamburg, M.A. & Collins, F.S. (2010). The path to personalized medicine. *The New England Journal of Medicine*, **363**, 301–304, PMID: 20551152. 1, 13

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. 43

Hindorff, L., Junkins, H., Mehta, J. & Manolio., T. (2010). A catalog of published genome-wide association studies. 2, 6, 41

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. 41

Hirschhorn, J.N. (2009). Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699–1701. 6

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, 4

Iles, M.M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.*, **4**, e33. 17

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320. 3

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. 6

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. 7

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233. 21

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21, PMID: 18987709. 13

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. 2, 3, 6, 25

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. 13

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. 52

REFERENCES

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. 7, 10, 15, 18

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. 5

McClellan, J. & King, M. (2010). Genetic heterogeneity in human disease. *Cell*, **141**, 210–217, PMID: 20403315. 6

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. 3, 21

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. 5

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. 15

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. 5

Pritchard, J. & Przeworski, M. (2001a). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14. 42, 54

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, **69**, 124–137, PMID: 11404818. 12, 27

Pritchard, J.K. & Przeworski, M. (2001b). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**, 1–14, PMID: 11410837. 4

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959, PMID: 10835412. 5

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. 7

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

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.*, **46**, 222–228. 13, 14, 47

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

Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, **53**, 1253–1261, PMID: 9423247. 9

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. 8, 11, 44, 53

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. 43

Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics*, **182**, 845–850, PMID: 19416939. 13

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. 21, 42

Steed, H., Walsh, S. & Reynolds, N. (2010). Crohn's disease incidence in NHS tayside. *Scottish Medical Journal*, **55**, 22–25, PMID: 20795513. 15

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. 13, 14

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. 52

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. 6, 15

Vukcevic, D. (2009). *Bayesian and frequentist methods and analyses of genome-wide association studies*. Ph.D. thesis, University of Oxford. 10, 55

Vukcevic, D., Hechter, E., Spencer, C.C.A. & Donnelly, P. (2010). Disease model distortion in association studies. *In preparation*. 55, 56

Wakefield, J. (2008). Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.*, **33**, 79–86. 7, 12, 27

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

REFERENCES

Wallace, C. & Clayton, D. (2003). Estimating the relative recurrence risk ratio using a global cross-ratio model. *Genetic Epidemiology*, **25**, 293–302, PMID: 14639699. 14

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*. 17

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. 2, 12, 27

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. 5, 7, 12

Wray, N.R. & Goddard, M.E. (2010). Multi-locus models of genetic risk of disease. *Genome Medicine*, **2**, 10–10, PMID: 20181060 PMCID: 2847701. 47

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. 13

Yang, J., Wray, N.R. & Visscher, P.M. (2010b). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genetic Epidemiology*, **34**, 254–257, PMID: 19918758. 9, 14

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. 43

Zondervan, K.T. & Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, **5**, 89–100. 12, 27, 42