2
Single Locus Association Analysis

2.1 Introduction

The simplest methods of genetic association analysis originate from case-control study designs used by epidemiologists to identify variables associated with a disease state. The basic premise is that DNA polymorphisms influencing individual disease susceptibility will display different frequencies in cases versus controls. From this perspective, genes are simply another potential disease covariate. This chapter considers a variety of such case-control methods that have been employed in genetic association studies. The methods make few explicit assumptions regarding the phenotypic effects of genes and instead rely on simple comparisons of proportions of particular genetic polymorphisms in cases versus controls, and so on, to detect differences that are expected to occur under virtually any models in which a locus is involved in disease susceptibility. Implicit assumptions of the methods typically include random sampling of individuals, random mating (Hardy-Weinberg equilibrium of genotypes), etc. Although the type I error of such tests usually does not depend on a model of gene effects, the power (type II error) typically does. Models and simulation methods for predicting the power of association studies will be discussed in Chapter 6.

The techniques presented here fall into the statistical realms of categorical data analysis (in the case of discrete disease states) discussed in the first portion of the chapter, and linear analysis (in the case of continuous disease phenotypes) discussed in the second portion. The statistical tests themselves fall into a natural hierarchy beginning with those based on the simplest models (with few parameters) and extending to those involving progressively more complex models (with many more parameters). Typically, a trade-off exists between the enhanced power of tests based on models with fewer parameters (e.g., those with fewer degrees of freedom) and those examining more complex, but potentially more realistic, models. A disadvantage of simple (parameter poor) models is that they may be more sensitive to so-called "spurious associations." Namely, those associations due not to an effect of the gene on a disease outcome but to the effects of other mutual covariates. More complex (parameter rich) models may account for such potentially confounding covariates, but at the cost of additional modeling assumptions.

In this chapter, we generally ignore the complexities arising from linkage disequilibrium (see Chapter 3) and proceed as if the locus under investigation were the actual causal locus influencing disease risk. Most of the methods can be expected to work reasonably well even when the association is due to indirect linkage to an undiscovered disease polymorphism. The first part of the chapter presents several simple tests for Hardy-Weinberg equilibrium (HWE) that may be applied to samples of cases or controls. Such tests should be carried out as an initial phase of exploratory data analysis. Departures from HWE may reveal excess genotyping errors at a locus, cryptic population substructure, etc. We reserve the discussion of disease association tests based on departures from HWE in cases versus controls until Chapter 4 as linkage disequilibrium (introduced in Chapter 3) is an important consideration in such tests that cannot be ignored.

2.2 Discrete disease states

Here, we consider tests of genetic association for a situation in which disease states are discrete (binary). In particular, we focus on case-control studies in which a disease is classified as either present (case) or absent (control). We assume that individuals are sampled at random from a defined population and consider a SNP locus with two alleles labeled 1 and 2. In most studies of single nucleotide polymorphisms (SNPs) in humans only two alleles (nucleotides) are found to be present at each polymorphic SNP site (presumably due to a recent demographic expansion of human populations and low rates of mutation in nuclear DNA). However, if needed the extension to multiple alleles is usually straightforward. Following the notation of Sasieni (1997), let r_i and s_i be the sample counts of genotype i in cases and controls, respectively, where $i \in \{0,1,2\}$ denotes the number of copies of allele 1 present in the genotype, and let $b_i = r_i + s_i$. Let R and S be the total number of cases and controls, respectively, and let $N = R + S$. The relationship of these variables is depicted in Table 2.1.

Number of Alleles	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	b_0	b_1	b_2	N

Table 2.1 *SNP genotype counts for a case-control association study*

2.2.1 Testing Hardy-Weinberg equilibrium

As discussed in Chapter 1, one possible indicator of some form of association between a marker and disease (whether due to a causal relationship or a spurious correlate such as cryptic population structure) is the observed departure of genotype frequencies from Hardy-Weinberg equilibrium (HWE) proportions. In particular, genetic association can cause a deviation from HWE among cases. Deviations from HWE may also be symptomatic of other problems with the data, however, such as genotyping errors (Xu et al., 2002). One should be particularly cautious when such deviations occur in the sample of controls.

Likelihood ratio test of HWE

Here, we describe a likelihood ratio test (LRT) of HWE for a sample of cases. The application to controls involves a straightforward substitution of variables (s_i for r_i, S for R, etc). Under the model of HWE, with SNP allele 1 having frequency p, the probability of the sample of genotypes in cases is

$$L_0 = \binom{R!}{r_0!r_1!r_2!} \left[p^2\right]^{r_0} [2p(1-p)]^{r_1} \left[(1-p)^2\right]^{r_2}. \qquad (2.1)$$

Under the alternative hypothesis, the genotype proportions g_0, g_1 and g_2, where the subscript denotes the number of copies of allele 1, are not constrained by HWE and the sample probability is

$$L_1 = \binom{R!}{r_0!r_1!r_2!} g_0^{r_0} g_1^{r_1} g_2^{r_2}, \qquad (2.2)$$

where $g_2 = 1 - g_0 - g_1$. There is a single free parameter, p, to be estimated under the null hypothesis H_0 while there are two free parameters, g_0 and g_1 to be estimated under the alternative hypothesis H_1. The likelihood ratio test statistic Λ is defined as

$$\Lambda = \frac{L_0}{L_1} = \frac{\left[\hat{p}^2\right]^{r_0}\left[2\hat{p}(1-\hat{p})\right]^{r_1}\left[(1-\hat{p})^2\right]^{r_2}}{\hat{g}_0^{r_0}\hat{g}_1^{r_1}(1-\hat{g}_0-\hat{g}_1)^{r_2}},$$

where

$$\hat{p} = \frac{2r_0 + r_1}{2R} \text{ and } \hat{g}_i = \frac{r_i}{R},$$

are maximum likelihood estimates of the parameters under each model. In this case, model 1 has one more free parameter than model 0 and so the hypothesis test has one degree of freedom. According to standard likelihood theory (see e.g., Rice, 1995), the LRT statistic $-2\log\Lambda$ has a sampling distribution that is asymptotically χ^2 with 1 df under the null hypothesis.

As an example, suppose that $R = 10$ cases are sampled in a case-control study. Let the observed sample configuration be $(r_0 = 1, r_1 = 9, r_2 = 0)$. The value of $-2\log\Lambda$ calculated from the above formula is 8.547. Under a χ^2 distribution with 1 df this value has a tail probability (significance) of $\alpha = 0.0035$.

χ^2 test of HWE

A χ^2 test of HWE can be formulated using the test statistic,

$$\chi^2 = \frac{[r_0 - p^2 \times R]^2}{p^2 \times R} + \frac{[r_1 - 2p(1-p) \times R]^2}{2p(1-p) \times R} + \frac{[r_2 - (1-p)^2 \times R]^2}{(1-p)^2 \times R},$$

where $p = (2r_0 + r_1)/(2R)$. There are two free observations (e.g., r_0 and r_1) and one estimated parameter, p, so the test has $2 - 1 = 1$ df. Analyzing the genotype data from the previous example, with observed sample configuration $(r_0 = 1, r_1 = 9, r_2 = 0)$, the value of the χ^2 test statistic is 6.694. Under a χ^2 distribution with 1 df this value has a tail probability (significance) of $\alpha = 0.0097$.

Exact test of HWE

The LRT and chi-square tests presented above both rely on asymptotic (large sample) theory to obtain the sampling distribution of the test statistic and determine significance. These methods may not provide accurate results for small samples and/or alleles that are in low frequency. Two possible solutions to this problem are to use an exact test of HWE (Levene, 1949; Haldane, 1954) or to use Monte Carlo simulation to generate the null distribution of the test statistic (see section 2.2.1 below). Here, we describe an exact test (Levene, 1949) of HWE for a locus with two alleles. See Louis and Dempster (1987) for an extension to multiple alleles.

We again consider a test of HWE in the sample of cases (the application to controls requires only a change of symbols). The basic objective of Fisher's exact test is to calculate the probability of each possible sample configuration under the null hypothesis of independence among elements of a contingency table with the marginal frequencies constrained to equal those observed in the sample. The possible samples are ranked according to their probabilities and the probabilities of the observed sample configuration, and all those samples less probable than the observed sample, are summed to predict the probability of observing a sample with a probability as small, or smaller, than that of the observed sample under the null hypothesis. If this tail probability is small the null hypothesis is rejected.

As noted above, the probability of the observed genotype counts under the null hypothesis is multinomial and can be rearranged to have the form,

$$\Pr(r_0, r_1, r_2) = \frac{R!}{r_0! r_1! r_2!} 2_1^r p^{(2r_0+r_1)} (1-p)^{(2r_2+r_1)}. \tag{2.3}$$

Let the total allele counts be $n_1 = 2r_0 + r_1$ and $n_2 = 2r_2 + r_1$. Assuming HWE, the marginal probability distribution of the allele counts is

$$\Pr(n_1, n_2) = \binom{2R}{n_1} p^{n_1} (1-p)^{n_2}. \tag{2.4}$$

We now consider the probability of the observed sample configuration conditioned on the fixed allele counts. The goal is to compare probabilities of all possible sample configurations with the same marginal probabilities of allele counts (and therefore the same population allele frequencies). The conditional distribution is obtained by dividing the joint probability of the sample (and allele counts) from equation 2.3 by the marginal probability of the allele counts from equation 2.4,

$$\begin{aligned}\Pr(r_0, r_1, r_2 | n_1, n_2) &= \frac{\Pr(r_0, r_1, r_2, n_1, n_2)}{\Pr(n_1, n_2)}, \\ &= \frac{R!/(r_0! r_1! r_2!) 2^{r_1} p^{n_1} (1-p)^{n_2}}{2n!/(n_1! n_2!) p^{n_1} (1-p)^{n_2}},\end{aligned}$$

$$= \frac{R!(2r_0 + r_1)!(2r_2 + r_1)!2^{r_1}}{r_0!r_1!r_2!(2R)!}.$$

This conditional probability no longer depends on the unknown population allele frequency, p. The computational challenge in applying this method is the neccessity to calculate the probability of every possible sample configuration with marginal allele counts n_1 and n_2. A useful algorithm for enumerating all possible sample configurations is given by Louis and Dempster (1987). Define the sample configuration to be (s_0, s_1, s_2). Construct the first sample of the set as either $(s_0 = 0, s_1 = n_1, s_2 = (n_2 - n_1)/2)$ if $n_1 \leq n_2$ or $(s_0 = (n_1 - n_2)/2, s_1 = n_2, s_2 = 0)$ if $n_2 \leq n_1$. The next sample is constructed as $(s_0 + 1, s_1 - 2, s_2 + 1)$. This operation is applied iteratively until either $s_1 = 0$ or $s_1 = 1$. For $n_i \leq n_j$ there will be $(n_i - 1)/2 + 1$ or $n_i/2 + 1$ distinct sample configurations if n_i is odd, or even, respectively. Thus, for a sample of R genotypes from cases there are at most $R/2$ distinct sample configurations and the tail probabilities for the exact test can be rapidly calculated on a modern computer.

Configuration	Probability	Rank
$(1, 9, 0)$	0.0305	4
$(2, 7, 1)$	0.2744	2
$(3, 5, 2)$	0.4801	1
$(4, 3, 3)$	0.2000	3
$(5, 1, 4)$	0.0150	5

Table 2.2 *Table of all possible sample configurations and exact test probabilities (and relative rank) for fixed marginal allele counts of* $n_1 = 11$ *and* $n_2 = 9$.

Consider the example given previously with sample configuration $(r_0 = 1, r_1 = 9, r_2 = 0)$. The marginal allele counts are $n_1 = 2(1) + 9 = 11$ and $n_2 = 2(0) + 9 = 9$. The possible sample configurations, conditional on the marginal allele counts, are shown in Table 2.2, along with the exact probability of each sample configuration. In this example, the observed sample configuration is ranked 4th in terms of its exact probability. Thus, the probability of seeing a sample with a probability this small (or smaller) under the null hypothesis of HWE proportions is $0.0305 + 0.0150 = 0.0455$. Thus the null hypothesis of HWE can be rejected at the $\alpha = 0.05$ level for these data.

Null distributions via Monte Carlo

The exact test of HWE outlined above requires that one enumerate all possible sample configurations, conditional on the marginal allele counts. For a locus with two alleles this exhaustive enumeration is feasible (even for large samples) since the number of possible configurations is never more than $R/2$. If more than about 5 alleles exist at a locus exhaustive enumeration is no longer feasible and a Monte Carlo simulation method must be used instead. Here, we consider a Monte Carlo method proposed by Guo and Thompson (1992) for m alleles. For simplicity, we describe the algorithm with only 2 alleles here; the extension to m alleles is straightforward (see Guo and Thompson, 1992). Usually, this algorithm is not needed for SNP loci and is most often applied to highly polymorphic markers such as microsatellites. However, if one is interested in testing for HWE among haplotypes (with distinct haplotypes equivalent to alleles) a Monte Carlo method may be needed when many haplotypes are present in a population. As shown in the previous section, the probability of the sample of genotypes, conditional on the marginal allele counts is,

$$\Pr(r_0, r_1, r_2) = \frac{R!(2r_0 + r_1)!(2r_2 + r_1)!2^{r_1}}{r_0!r_1!r_2!(2R)!}$$

The probability for the exact test of HWE is given by

$$P = \sum_s \Pr(s),$$

where $s = (s_0, s_1, s_2)$ is the set of all possible sample configurations with the same marginal allele counts as the observed sample $r = (r_0, r_1, r_2)$ and for which $\Pr(s) \leq \Pr(r)$. A Monte Carlo estimator of P is obtained as follows. First, create a vector v of length $X = n_1 + n_2$, with the first n_1 elements set to be allele 1 and the remaining n_2 elements set to be allele 2. Initialize variables $K = 0$ and $i = 0$. Define Z to be the number of iterations for the Monte Carlo analysis. Apply the following algorithm:

1. Randomly permute the elements of vector v so that the probability that an element occupies any particular position in the vector after the permutation is $1/X$.
2. Create a genotype vector of length $X/2$ by splitting the elements of vector v into pairs (genotypes) such that

$$w_k^* = \{(v_{2k-1}, v_{2k}), k = 1, 2, \ldots, X/2\}.$$

3. Let s^* be the genotype counts for vector w^*. If
$$\Pr(s^*) \leq \Pr(r),$$
then $K = K + 1$, otherwise $K = K$.
4. If $i < Z$ then $i = i + 1$, return to step 1. Otherwise $P = K/Z$, exit algorithm.

To illustrate this algorithm, we apply it to the example given in Section 2.2.1. In this example, $r = (r_0 = 1, r_1 = 9, r_2 = 0)$ so that $n_1 = 11$ and $n_2 = 9$. Five random permutations and the resulting sample configurations for data with these marginal counts are shown in Table 2.3. Three separate Monte Carlo analyses were done for the sample $r = (1, 9, 0)$, each with $Z = 10,000$ replicates. The estimates of the probability for the test were $P_1 = 0.0449$, $P_2 = 0.0441$, and $P_3 = 0.0471$. The exact value is $P = 0.0455$ and so the Monte Carlo estimates are accurate to the second decimal place. The accuracy of the Monte Carlo estimate can be increased by increasing the number of replicates. For example, with $Z = 5 \times 10^5$ we obtain $P_4 = 0.045558$ which is accurate to the 4th decimal place. Note that a similar Monte Carlo methodology could be used to generate the sampling distribution of the likelihood ratio and χ^2 tests described previously, potentially improving their statistical performance for small samples. Methods for predicting the error (and confidence intervals) of Monte Carlo estimates are discussed in Chapter 7.

Random Permutation (v)	$s^* = (s_0, s_1, s_2)$	$\Pr(s^*)$
$(1,1,1,2,1,1,1,1,1,1,2,2,2,2,1,2,1,2,2,2)$	$(3,3,4)$	0.200
$(1,1,2,1,1,1,2,2,2,2,2,1,1,2,1,1,2,1,1,2)$	$(3,5,2)$	0.480
$(1,2,1,1,2,1,1,1,2,1,2,1,2,1,1,2,1,2,2,2)$	$(2,7,1)$	0.274
$(2,1,2,2,2,1,1,1,2,2,2,1,2,1,1,2,1,1,1,1)$	$(3,5,2)$	0.480
$(1,2,1,1,1,1,1,2,1,2,2,2,2,1,2,2,1,1,1,2)$	$(3,5,2)$	0.480

Table 2.3 *Table of sample configurations, s^*, and exact test probabilities, $\Pr(s^*)$, for each of 5 random permutations, v, with fixed marginal allele counts of $n_1 = 11$ and $n_2 = 9$.*

Distributions of test statistics using discrete data

A well-known feature of statistical tests based on discrete data is that the true sampling distribution of the test statistic is also discrete. The

χ^2 test of HWE described above uses a continuous distribution (the chi-square distribution) as an approximation to the discrete sampling distribution. The accuracy of this approximation depends on factors such minor allele frequency (MAF) and sample size (Rohlfs and Weir, 2008). The exact test also has a discrete sampling distribution for the test statistic. An improvement of the χ^2 test statistic was suggested by (Yates, 1934). The Yates correction for continuity improves the performance of the χ^2 test when the counts in one or more elements of the contingency table are small (this could be due to low MAF, small sample size, extreme departures from HWE, etc), making the test more conservative. The Yates correction subtracts $1/2$ from each of the squared differences between expected and observed values,

$$\chi^2 = \frac{\sum(|Obs - Exp| - 0.5)^2}{Exp}.$$

One important consequence of the discreteness of the sampling distribution is that the distribution of p-values is not uniform. This is evident by examining the distribution of the p-values for all possible sample configurations enumerated under the exact test (see table 2.2). This can lead to anomalous outcomes. For example, if the sample configuration with the smallest p-value has $p = 0.10$ and a significance threshold of $p \leq 0.05$ is chosen the test will never reject (has power 0) and the type I error will be too large, whereas if $p \leq 0.10$ is chosen as the significance threshold the test will have the correct type I error and non-zero power. One proposed solution to this problem is to use a stochastic decision process. In the above example, this would involve randomly accepting half the outcomes with $p \leq 0.10$ to achieve a significance of $p \leq 0.05$. However, as noted by Rohlfs and Weir (2008) this could lead to different outcomes for parallel analyses of the same data set, which is not very appealing to scientists.

2.2.2 Allele frequencies in cases versus controls

A first step in a GWAS is often a locus-by-locus analysis for associations between SNP allele frequency at each locus and disease status. An important distinction between methods for testing allelic association is whether or not the genotypes are assumed to be in Hardy-Weinberg equilibrium. We begin by considering simple tests that assume HWE and then describe several additional tests that do not.

Likelihood ratio test assuming HWE

We first consider a likelihood ratio test comparing allele frequencies at a given locus in cases versus controls. Let $a_1 = r_1 + 2r_2$ and $a_2 = r_1 + 2r_0$ be the number of copies of alleles 1 and 2, respectively, in cases and let $m_1 = s_1 + 2s_2$ and $m_2 = s_1 + 2s_0$ be the number of copies of alleles 1 and 2, respectively, in controls. Let p and q be the population frequencies of allele 1 in cases and controls, respectively. If we assume that the chromosomes represent a random sample from a large population (or that individuals are sampled at random and their genotypes are in Hardy-Weinberg equilibrium at the locus) then the sampling distribution of alleles follows a multinomial distribution. Under the null hypothesis that allele frequencies at this locus are identical in cases and controls $c = p = q$, there is one free parameter, c, and

$$\Pr(a_1, m_1, a_2, m_2) = \frac{(2R)!(2S)!}{a_1!a_2!m_1!m_2!} c^{(a_1+m_1)}(1-c)^{(a_2+m_2)}.$$

The maximum likelihood estimate of the population frequency of allele 1 among both cases and controls is

$$\hat{c} = \frac{a_1 + m_1}{2N}.$$

Under the alternative hypothesis that allele frequencies differ between cases and controls $p \neq q$ so that there are two free parameters, p and q,

$$\Pr(a_1, m_1, a_2, m_2) = \frac{(2R)!(2S)!}{a_1!a_2!m_1!m_2!} p^{a_1}(1-p)^{a_2} q^{m_1}(1-q)^{m_2}.$$

The maximum likelihood estimates of the population frequencies of allele 1 in cases and controls, respectively, are

$$\hat{p} = \frac{a_1}{2R} \text{ and } \hat{q} = \frac{m_1}{2S}.$$

The likelihood ratio test statistic Λ is constructed by taking a ratio of the probability of the data under the null hypothesis to that under the alternative hypothesis, with maximum likelihood estimates substituted for the unknown parameters p, q and c,

$$\begin{aligned}\Lambda &= \frac{\hat{c}^{(a_1+m_1)}(1-\hat{c})^{(a_2+m_2)}}{\hat{p}^{a_1}(1-\hat{p})^{a_2}\hat{q}^{m_1}(1-\hat{q})^{m_2}}, \\ &= \frac{(\frac{a_1+m_1}{2N})^{(a_1+m_1)}(1-[\frac{a_1+m_1}{2N}])^{(a_2+m_2)}}{(\frac{a_1}{2R})^{a_1}(1-[\frac{a_1}{2R}])^{a_2}(\frac{m_1}{2S})^{m_1}(1-[\frac{m_1}{2S}])^{m_2}}\end{aligned}$$

The asymptotic distribution of the test statistic $-2 \log \Lambda$ follows that of a χ^2 distribution with 1 degree of freedom (the difference in the number of free parameters under the null versus alternative hypotheses). To illustrate this test we apply it to a SNP locus, labelled rs11209026, from a genome-wide association study of a sample of patients with inflammatory bowel disease (IBD) and a sample of controls (Duerr et al., 2006). This SNP (A/G) polymorphism results in a change of the amino acid at position 381 (Arg/Gln) of the proinflammatory cytokine interleukin-23. The genotype counts in a non-Jewish case-control cohort are given in Table 2.4. In this example, the observed values are $2R = 1084$, $2S = 1082$, $a_1 = 21$, $a_2 = 1063$, $m_1 = 76$ and $m_2 = 1006$. The test statistic is $-2 \log \Lambda = 34.69$ and the probability of a value at least as great as this under the null hypothesis (the tail probability for a χ^2 distribution with 1 df) is 3.87×10^{-9}.

	Genotype			
Group	A/A	A/G	G/G	Total
IBD	0	21	521	542
Controls	3	70	468	541
Total	3	91	989	1083

Table 2.4 *IBD Data*

χ^2 test assuming HWE

A χ^2 test of the hypothesis that allele proportions are equal among cases and controls is asymptotically equivalent to the test outlined above. The χ^2 test statistic is

$$\chi^2 = \frac{(O_{cases} - E_{cases})^2}{E_{cases}} + \frac{(O_{controls} - E_{controls})^2}{E_{controls}},$$

where

$$E_{cases} = 2R\hat{c} = \frac{R}{N}(a_1 + m_1) \text{ and } O_{cases} = a_1$$

are the expected and observed counts of the allele among cases and

$$E_{controls} = 2S\hat{c} = \frac{S}{N}(a_1 + m_1) \text{ and } O_{controls} = m_1$$

are the expected and observed counts of the allele among controls. The test has 1 degree of freedom which is obtained as the difference between the number of free observations (2), given the total sample size, and the number of estimated proportions (1), so that $df = 2 - 1 = 1$. The value of the χ^2 test statistic for the IBD data is 31.28 which has tail probability 1.92×10^{-8}. Thus, there is a highly significant difference in allele frequency between IBD cases and controls at this locus. In this example, the observed number of copies of the allele in cases ($O_{cases} = 21$) is less than the expectation under the null hypothesis ($E_{cases} = 48.5$) and the authors conclude that the A allele of the IL23R gene may be protective against IBD.

Hardy-Weinberg disequilibrium

The tests outlined in the previous section assume that population genotype frequencies are in HWE proportions. If this assumption is violated the type-I error rate may be incorrect. Schaid and Jacobsen (1999) proposed an alternative test for differences in allele frequencies between cases and controls that allows a deviation of genotype frequencies from HWE. Let A and a be 2 alleles at a locus with population frequencies p and $1 - p$, respectively. Let x_i be the number of copies of allele A (either 0, 1, or 2) in the genotype of individual i. Assuming HWE, the probabilities of the 3 possible genotypes are

$$\begin{aligned}
\Pr(x_i = 0) &= (1-p)^2, \\
\Pr(x_i = 1) &= 2p(1-p), \\
\Pr(x_i = 2) &= p^2.
\end{aligned}$$

The expected value of x_i is

$$\mathbb{E}(x_i) = \sum_{x_i=0}^{2} x_i \Pr(x_i) = 0 \times (1-p)^2 + 1 \times 2p(1-p) + 2 \times p^2 = 2p,$$

and the variance is

$$\begin{aligned}
\mathrm{Var}(x_i) &= \mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2, \\
&= \sum_{x_i=0}^{2} x_i^2 \Pr(x_i) - (2p)^2, \\
&= 2p(1-p) + 4p^2 - (2p)^2, \\
&= 2p(1-p).
\end{aligned}$$

The maximum likelihood estimator of allele frequency is

$$\hat{p} = \frac{\sum_{i=1}^{N} x_i}{2N},$$

which has the expected value

$$\mathbb{E}(\hat{p}) = \frac{1}{2N}\sum_{i=0}^{N}\mathbb{E}(x_i) = \frac{N \times 2p}{2N} = p,$$

and the variance,

$$\begin{aligned}\mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{\sum_{i=0}^{N} x_i}{2N}\right), \\ &= \frac{1}{4N^2}\sum_{i=0}^{N}\mathrm{Var}(x_i), \\ &= \frac{1}{4N^2} \times N \times 2p(1-p), \\ &= \frac{p(1-p)}{2N}.\end{aligned}$$

Let p_d and p_c be the population frequencies of allele A in cases and controls, respectively, and let N_d and N_c be the sample sizes. For large samples the normal distribution approximation for the binomial distribution implies that the probability distribution of the sample allele frequency in cases and controls follows a normal distribution with means p_d and p_c, and variances $p_d(1-p_d)/(2N_d)$ and $p_c(1-p_c)/(2N_c)$, in cases and controls, respectively. The test statistic is

$$z = \frac{(\hat{p}_d - \hat{p}_c)}{\sqrt{V}} \tag{2.5}$$

Under the null hypothesis the frequency of A is identical in cases and controls so that $p_d = p_c = p$. The expectation of z is therefore

$$\mathbb{E}(\hat{p}_d - \hat{p}_c) = \mathbb{E}(\hat{p}_d) - \mathbb{E}(\hat{p}_c) = p - p = 0,$$

and the variance of z is 1 if we set

$$\begin{aligned}V = \mathrm{Var}(\hat{p}_d - \hat{p}_c) &= \mathrm{Var}(\hat{p}_d) + \mathrm{Var}(\hat{p}_c), \\ &= \frac{p(1-p)}{2N_d} + \frac{p(1-p)}{2N_c}, \\ &= p(1-p)\left(\frac{1}{2N_d} + \frac{1}{2N_c}\right).\end{aligned}$$

The test statistic in equation 2.5 thus follows a standard normal distribution under the null hypothesis (assuming HWE genotype proportions).

We now consider the situation when the population genotype frequencies are not in HWE proportions. Let F be an inbreeding coefficient that quantifies the degree of departure of genotype frequencies from HWE. The probabilities of the 3 possible genotypes under this model are

$$\begin{aligned}
\Pr(x_i = 0) &= (1-F)(1-p)^2 + (1-p)F, \\
\Pr(x_i = 1) &= 2p(1-p)(1-F), \\
\Pr(x_i = 2) &= (1-F)p^2 + pF.
\end{aligned}$$

The expected value of x_i is

$$\begin{aligned}
\mathbb{E}(x_i) &= x_i \Pr(x_i), \\
&= 0 \times [(1-p)^2(1-F) + (1-p)F] + \\
&\quad 1 \times [2p(1-p)(1-F)] + 2 \times [p^2(1-F) + pF] \\
&= 2p,
\end{aligned}$$

and the variance is

$$\begin{aligned}
\mathrm{Var}(x_i) &= \mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2, \\
&= \sum_{x_i=0}^{2} x_i^2 \Pr(x_i) - (2p)^2, \\
&= 0^2 \times [(1-p)^2(1-F) + (1-p)F] + \\
&\quad 1^2 \times [2p(1-p)(1-F)] + 2^2 \times [p^2(1-F) + pF] \\
&= 2p(1-p)(1+F).
\end{aligned}$$

The mean and variance of $\hat{p}$ for either cases or controls with a deviation from HWE are $\mathbb{E}(\hat{p}) = p$ and

$$\begin{aligned}
\mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{N} x_i}{2N} \right), \\
&= \frac{1}{4N^2} \sum_{i=1}^{N} \mathrm{Var}(x_i), \\
&= \frac{1}{4N^2} \times N \times 2p(1-p)(1+F), \\
&= \frac{1}{2N} p(1-p)(1+F).
\end{aligned}$$

Thus, the variance V for normalizing the test statistic under the null hypothesis when HWE is violated is

$$V_{nonHWE} = [p(1-p)(1+F)]\left(\frac{1}{2N_d} + \frac{1}{2N_c}\right). \tag{2.6}$$

Setting $f_{AA} = p^2(1-F) + pF$ and solving for F gives,

$$F = \frac{(f_{AA} - p^2)}{p(1-p)},$$

and substituting for F in equation 2.6 gives

$$V_{nonHWE} = [p(1-p) + (f_{AA} - p^2)]\left(\frac{1}{2N_d} + \frac{1}{2N_c}\right),$$

which matches the formula for the "pooled variance" model given in Schaid and Jacobsen (1999). Thus, the variance of the sampling distribution of the test statistic is increased when there are departures from HWE proportions so that using the variance derived under an assumption of HWE will increase the probability of rejecting the null hypothesis and inflate the type-I error over the nominal value.

Schaid and Jacobsen (1999) suggested estimating the HWE deviation $\delta = f_{AA} - p^2$ from the pooled genotypes of both cases and controls and using the bias-corrected variance V_{nonHWE} in calculating the test statistic. They also suggested an alternative estimator of the bias-corrected variance obtained by separately calculating p and f_{AA} in cases and controls and then estimating the total variance as the sum

$$V_{NonHWE} = V_{NonHWE,d} + V_{NonHWE,c}.$$

Knapp (2001) suggested that this second "separate variance" formula for calculating V_{NonHWE} is preferable and will result in more powerful test with lower type-I error. He also pointed out that the square of the test statistic, z, calculated using the pooled variance estimator of V_{NonHWE}, is identical to the test statistic of Armitrage's trend test (see Section 2.2.4) so the two tests are equivalent.

2.2.3 Genotypes in cases versus controls

A genotype-based association test does not assume HWE of genotype frequencies and will be sensitive to differences of genotype proportions between cases and controls due to either deviations from HWE, differences of allele frequency, or both.

Likelihood ratio test of genotype association

Let g_i and h_i be the population frequencies of genotype i in cases and controls, respectively. Under the null hypothesis there is no difference of genotype frequencies between cases and controls so that $c_i = g_i = h_i$ for all $i = 0, 1, 2$ and there are 2 free parameters c_0 and c_1 because the genotype frequencies must sum to 1 and therefore $c_2 = 1 - c_0 - c_1$. The probability of the sampled genotypes under the null hypothesis is

$$\Pr(r_0, r_1, r_2, s_0, s_1, s_2) = \left(\frac{R!}{r_0!r_1!r_2!}\right)\left(\frac{S!}{s_0!s_1!s_2!}\right)\prod_{i=0}^{2} c_i^{n_i},$$

and the maximum likelihood estimator of parameter c_i is

$$\hat{c}_i = \frac{r_i + s_i}{N}.$$

Under the alternative hypothesis the genotype frequencies are different in cases versus controls so that $g_i \neq h_i$ for all $i = 0, 1, 2$ and there are 4 free parameters g_0, g_1, h_0 and h_1. The probability of the sampled genotypes under the alternative hypothesis is

$$\Pr(r_1, r_2, r_3, s_1, s_2, s_3) = \left(\frac{R!}{r_1!r_2!r_3!}\right)\left(\frac{S!}{s_1!s_2!s_3!}\right)\prod_{i=0}^{2} g_i^{r_i} h_i^{s_i},$$

and the maximum likelihood estimators of parameters g_i and h_i are

$$\hat{g}_i = \frac{r_i}{R} \text{ and } \hat{h}_i = \frac{s_i}{S}.$$

The likelihood ratio test statistic is

$$\Lambda = \prod_{i=0}^{2}\left(\frac{\hat{c}_i^{(r_i+s_i)}}{\hat{g}_i^{r_i}\hat{h}_i^{s_i}}\right),$$

and

$$-2\log\Lambda = -2\sum_{i=0}^{2}\left[(r_i + s_i)\log\left(\frac{r_i + s_i}{N}\right) - r_i\log\left(\frac{r_i}{R}\right) - s_i\log\left(\frac{s_i}{S}\right)\right] \tag{2.7}$$

which has an asymptotic distribution that is χ^2 with 2 degrees of freedom (the difference in the number of free parameters between the null and alternative hypotheses is $4 - 2 = 2$).

To illustrate this test we apply it to the data of Duerr et al. (2006) for SNP locus rs11209026 of the IBD case-control study considered previously. The genotype counts are presented in Table 2.4. Applying equation 2.7 to these data we obtain $-2\log\Lambda = 34.84$. The significance of this result (using the tail probability from a χ^2 distribution with 2 df) is 2.73×10^{-8}.

χ^2 test of genotype association

A chi-square test may be formulated that is asymptotically equivalent to the LRT. The chi-square test statistic is

$$\chi^2 = \sum_{i=0}^{2} \frac{\left(O^{(i)}_{cases} - E^{(i)}_{cases}\right)^2}{E^{(i)}_{cases}} + \frac{\left(O^{(i)}_{controls} - E^{(i)}_{controls}\right)^2}{E^{(i)}_{controls}},$$

where

$$E^{(i)}_{cases} = \frac{R}{N}(r_i + s_i) \text{ and } O^{(i)}_{cases} = r_i,$$

and,

$$E^{(i)}_{controls} = \frac{S}{N}(r_i + s_i) \text{ and } O^{(i)}_{controls} = s_i.$$

The chi-square test statistic has an asymptotic distribution that is χ^2 with 2 df (the difference between the number of free observations, 4, and the number of estimated proportions, 2). For the IBD data presented in Table 2.4 the test statistic is $\chi^2 = 32.22$ which has tail probability of 1.01×10^{-7}.

2.2.4 Score tests for disease trends

The tests for allele-disease, or genotype-disease, association described above do not specify any particular relationship between the number of allele copies present in a genotype and the risk of disease. It is reasonable to assume that the number of copies of the allele may be an important factor influencing disease risk and incorporating this assumption can lead to a more powerful test. The Cochrane-Armitage test (CATT) (Armitage, 1955; Cochran, 1954) can be used to introduce a score for each genotype that is a function of the number of allele copies. Furthermore, if the mode of disease inheritance is known (in the case of simple Mendelian disorders, for example), a score function can be chosen that provides optimal power for the CATT under a particular genetic model. The CATT essentially involves a regression of the frequency of cases in each genotype class (e.g., homozygous, heterozygous, etc) against the score for the class. A significant association indicates a relationship between the disease frequency and genotype score. Incorporating such relationships (when they exist) into the model used for a test of association increases the power of the test (Sasieni, 1997). The scoring function used in this test is subjective but can be chosen to mimic specific models of allelic effect.

Following (Sasieni, 1997) let the genotypes $g_0 = aa$, $g_1 = Aa$ and $g_2 = AA$ be represented by scores $(0, x, 1)$ for (g_0, g_1, g_2), where $0 \leq x \leq 1$. The values $x = 0$, $x = 1/2$ and $x = 1$ correspond to optimal choices of x for a recessive, additive (or multiplicative) and dominant model of allele effect on disease risk, respectively (Zheng et al., 2003). The CATT score statistic is

$$Z_x^2 = \frac{N\{\sum_{i=0}^{2} x_i(Sr_i - Rs_i)\}^2}{RS\{N\sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2\}}. \tag{2.8}$$

Under the null hypothesis, H_0, that no association exists between the genotype and disease risk the allele frequencies are equal in case and control samples, $q_i = p_i$ for $i = 0, 1, 2$, where p_i and q_i denote the population frequencies of genotype i in controls and cases, respectively. If H_0 is true then the sampling distribution of Z_x^2 is asymptotically a χ_1^2 distribution. One difficulty in applying this test to data from genome-wide association studies is that the model of allelic effect is unknown (e.g., x is unknown). If x is incorrectly specified the type I error will remain correct but the power may be reduced (e.g., the type II error rate may be inflated).

To illustrate the CATT we apply it to the IBD data of Duerr et al. (2006) in Table 2.4 under each of the three genetic models. Under the additive and recessive models, the score statistics are $Z_{1/2}^2 = 32.18$ and $Z_1^2 = 31.61$, respectively, which are both highly significant (e.g., a value of 3.84 or greater is significant at the $\alpha = 0.05$ level for a χ_1^2 distribution). Under the complete dominance model, the score statistic is $Z_0^2 = 3.014$, which is not significant at the $\alpha = 0.05$ level.

A CATT analysis using one particular score function (genetic model) is not robust when analyzing complex genetic diseases for which the underlying genetic model is unknown. Freidlin et al. (2002) studied the properties of several approaches to the development of robust tests when the true model is unknown. Zheng and Ng (2008) advocated a two-phase analysis strategy that uses the difference of Hardy-Weinberg disequilibrium coefficients (see Weir, 1996) between cases and controls to choose a genetic model followed by a CATT test of association using the optimal model chosen in the first phase of analysis.

2.3 Continuous disease states

Many complex traits that play a role in human disease arise from measurements on individuals (or functions of measurements), such as the body mass index (BMI) used as a phenotype measure in obesity research. The approaches described in section 2.2 are for use with discrete phenotype variables such as a binary variable describing disease outcome. It is possible to accommodate continuous measurement data using such methods by truncating the variable into 2 or more discrete classes. For example, physicians use thresholds for the BMI to categorize individuals as normal, overweight or obese. Information may be lost by carrying out such transformations, however, and it is frequently better to make explicit use of continuous variables in an analysis. Here we consider several methods for detecting associations between SNP alleles and continuous phenotype measures with the aim of determining whether a particular locus influences a phenotype.

The classical population genetics approach to model continuous traits, known as quantitative genetics (see section 1.6), assumes that continuous trait variation arises through the effects of many genes, each with relatively small effect, and of the environment. Let x_{ijk} be the allele present in individual i at SNP j that was inherited from the father ($k = 1$) or the mother ($k = 2$). Let P_i denote the phenotype of individual i, and let ϵ denote the effect of environment. Given the individual's genome and environment the phenotype is predicted by the equation,

$$P_i = \sum_{k=1}^{2} \sum_{j=1}^{L} x_{ijk} + D + I + \epsilon.$$

The first term of the equation models "additive effects" of alleles, while the terms D and I incorporate the non-additive gene effects (dominance and epistasis). The main effects are deterministic (non-random) while the environmental influences are assumed to be due to unknown random factors and are modeled such that ϵ follows a normal distribution with mean 0 and variance σ^2. This is the same basic model that is used in linear regression to predict the value of a random variable that is a deterministic linear function of a predictor variable plus a random error component (the errors are equivalent to environmental deviations under the genetic model and the predictor variable is the genotype).

2.3.1 Linear regression

Treating the effect of alleles at locus j as additive and subsuming the genetic effects at other (possibly unstudied) loci, as well as dominance and epistasis effects, into the random "environment" term, the standard regression equation can be used to predict individual phenotypes based on the genotype at locus j,

$$P_i = \beta_j y_{ij} + \epsilon_i,$$

where β_j is a parameter that summarizes the relative effect of each allele copy at locus j on the phenotype, $y_{ij} = \{0, 1, 2\}$ denotes the number of copies of the allele possessed by individual i, and ϵ_i is a normal random variable representing the "random" effects on phenotype of environment, additional causal loci, etc. The intercept term is dropped here because variables can always be transformed by subtracting a constant so that the intercept becomes zero. By applying standard linear regression with the genotype at locus j, $\mathbf{y}_j = \{y_{ij}\}$, as the predictor (independent) variable and the phenotype $\mathbf{P} = \{P_i\}$ as the dependent variable one can test whether, for example, $\beta_j > 0$, and one can use the squared correlation coefficient R^2 to predict the proportion of phenotypic variation that is attributable to the additive effects of allele copy number at locus j.

To illustrate this approach, we consider a recent study by van Vliet-Ostaptchouk et al. (2008) that examined the association between SNP polymorphisms in the TUB gene, measures of body composition (weight, BMI, etc), and eating behavior in middle-aged women. The TUB gene (and tubby protein that it encodes) is known to be expressed in regions of the hypothalamus involved in regulating appetite and satiety. A loss-of-function mutation in tubby results in late-onset diabetes, insulin resistance, and related phenotypes in mice. van Vliet-Ostaptchouk et al. (2008) genotyped three SNP loci in 1680 middle-aged Dutch women who were subjected to anthropometry and a macronutrient intake questionnaire. A linear regression analysis was performed regressing the anthropometrical characteristics of the women against the genotypes at each locus. Table 2.5 shows several of the results for locus rs1528133. Two significant associations were identified at the $\alpha = 0.05$ level. These were between either weight or BMI and the rs1528133 genotype. There is a positive association of both BMI and weight with the number of copies of the C allele at this locus; individuals that were either heterozygous A/C or homozygous C/C had a slightly elevated weight

Phenotype	Mean ± SD	β	95% CI	p-value
Weight (kg)	69.56 ± 0.30	1.88	$(0.27, 3.48)$	0.02
Waist (cm)	83.13 ± 0.26	1.23	$(-0.17, 2.64)$	0.09
Hip (cm)	105.13 ± 0.22	1.10	$(-0.07, 2.27)$	0.06
BMI (kg/m^2)	25.81 ± 0.11	0.56	$(0.00, 1.12)$	0.05

Table 2.5 *Summary of the results of a linear regression analysis of several body composition measures for middle-aged Dutch women against the genotypes at SNP locus rs1528133. The paramater β is the slope of the regression line.*

(an increase of +1.88 kg per copy of allele C). However, associations at the $\alpha = 0.05$ are marginal and the significant results in this study would disappear, for example, if one were to correct for multiple testing (the analysis of 3 SNP loci) using a Bonferroni correction (see section 2.5)

2.3.2 Multiple regression

The regression approach for analyzing continuous phenotypes can be extended to allow multiple predictor variables that are potential phenotypic covariates using standard multiple regression procedures (see e.g., Abraham and Ledolter, 2006). Let P_i be the phenotype measured for individual i and let $x_1(i), \ldots, x_p(i)$ be measurements of p traits, for individual i, suspected to influence the phenotype. Assuming a standard linear model, the phenotype of individual i is modeled as

$$P_i = \beta_1 x_1(i) + \beta_2 x_2(i) + \cdots + \beta_p(i) x_p(i) + \epsilon_i.$$

Note that one or more of the x's may be gene counts at SNP loci and other traits could include environmental factors, or other measurable factors of potential relevance. The coefficients $\beta_1, \ldots, \beta_p$ are estimated by finding the values that minimize the squared difference between observed and predicted phenotypic values. The squared difference can be expressed in matrix form as

$$S(\beta) = (P - X\beta)'(P - X\beta),$$

where

$$P' = \left[\begin{array}{cccc} P_1 & P_2 & \cdots & P_n \end{array}\right],$$
$$\beta' = \left[\begin{array}{cccc} \beta_1 & \beta_2 & \cdots & \beta_p \end{array}\right],$$

and

$$X = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_p(1) \\ \vdots & \vdots & & \vdots \\ x_1(n) & x_2(n) & \cdots & x_p(n) \end{bmatrix}$$

Univariate linear regression is a special case of this more general regression method (with $p = 1$). Standard methods can be used to test whether any particular regression coefficient β_i is significantly different from zero; this indicates whether predictor variable i influences phenotype. If a genetic variant is only indirectly associated with the phenotype (namely, it covaries with one of the factors and that factor is a predictor of phenotype) then including the factor in the multiple regression analysis can remove the effect of mutual covariance and reduce the apparent association between the genetic variant and the phenotype. This increases the likelihood that detected genetic associations are causal rather than spurious.

As an example, the population association study of Lanktree et al. (2009) replicated several previously identified associations between SNP polymorphisms in various genes (including *LPL* and *APOE*) and lipoprotein traits known to be associated with cardiovascular disease such as high-density lipoprotein (HDL), low-density lipoprotein (LDL) and triglycerides (TG). These authors used a multiple linear regression model that included age, sex, BMI and ethnicity as covariates in addition to the number of risk alleles at the SNP loci.

2.3.3 Analysis of variance

Analysis of variance methods can be used to partition the phenotypic variance in a regression analysis. The following relationship holds among the sums of squares:

$$SST = SSR + SSE,$$

where,

$$\begin{aligned} SST &= \sum_{i=1}^{n} (P_i - \overline{P})^2, \\ SSE &= S(\hat{B}). \end{aligned}$$

The total sum of squares (SST) sums the squared deviations of phenotypes about the population mean, $\overline{P} = \sum_{i=1}^{n} P_i$, and the error sum of squares (SSE) sums the deviations of phenotypes about the values predicted using the regression equation with coefficients $\hat{\beta}$ estimated from the data. The regression sum of squares (SSR) is the sum of squared deviations removed by the fitting of the regression equation and can be estimated as

$$SSR = SST - SSE = \sum_{i=1}^{n}(P_i - \overline{P})^2 - S(\hat{B}).$$

The proportion of phenotypic variance explained by the linear regression is referred to as the coefficient of determination R^2 and is estimated as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

With multiple predictor variables, the proportion of phenotypic variance explained by the ith predictor variable can be quantified as

$$\Delta R^2 = \frac{SSR_i - SSR_{i-1}}{SST},$$

where SSR_i and SSR_{i-1} are the regression sum of squares calculated with (and without) predictor variable i, respectively. The population association study of traits associated with cardiovascular disease (Lanktree et al., 2009) mentioned in the previous section also quantified the proportion of lipoprotein trait variation explained by age, sex, BMI and ethnicity, with and without a set of genetic markers included in the regression. Including all traits and genetic markers the regression model explained 25% of variation in TG, 34% of variation in HDL and 14% of variation in LDL. Eliminating the genetic markers from the model reduced the proportion of variance explained by the model by roughly 3% for LDL, 5% for HDL and 7% for TG. Multiple regression approaches for modeling the joint effects of multiple loci on phenotype will be discussed in more detail in Chapter 4.

2.4 Relative risk and the odds ratio

To investigate the causes of a disease phenotype, P, a case-control study design focusing on a potential genetic risk factor, G, empirically compares $f(G|P)$ versus $f(G| \neq P)$ in searching for genetic factors. This can be achieved using one or more of the association study approaches outlined previously. Once a risk factor has been identified, it is of interest to quantify the increase in risk for an individual exposed to the risk factor. In genetic studies the risk factors are genetic polymorphisms. The odds ratio and the relative risk are powerful methods for objectively quantifying the magnitude of disease risk conferred by a factor and are fundamental measures used by epidemiologists.

2.4.1 Odds ratio

The odds ratio has been independently discovered several times during the last century, attesting to its importance and generality. Fisher (1935) studied an odds ratio statistic in the context of proportions of criminality in monozygtic versus dizygotic twins, Berkson (1953) derived an odds ratio (logit) in the context of logistic regression methods for analyzing dose response curves, and Woolf (1955) proposed an odds ratio as a measure to quantify the disease risk conferred by blood group type, removing the effect of blood type population frequencies. Woolf's seminal study appears to be the first application of an odds ratio in human genetic association analysis.

The odds ratio (OR) of disease is defined as

$$OR = \frac{P_1}{1-P_1} \bigg/ \frac{P_2}{1-P_2} = \frac{P_1(1-P_2)}{P_2(1-P_1)}, \tag{2.9}$$

where P_1 and P_2 are the proportions of individuals exposed to the risk factor among cases and controls, respectively, and $1 - P_1$ and $1 - P_2$ are the proportions not exposed to the risk factor. Often the natural logarithm of the odds (the log-odds) is used instead because a change in the labels of the risk factors only changes the sign of the log-odds (LOD) but changes the value of the OR. For example, let $OR = 2$ so that $LOD = 0.693$. If we relabel the risk factors symmetrically, then $OR = 1/2$ but $LOD = -0.693$. The log-odds (LOD), is defined as

$$LOD = \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_2}{1-P_2}\right).$$

A major advantage of using the odds ratio (or log-odds), rather than comparing disease incidence directly between risk-exposed and -unexposed groups, is that the odds ratio is independent of the population frequency of the risk factor. To see this, let p be the population frequency of the risk factor and let z_1 and z_2 be the conditional probabilities that an individual is a case given that they are (or are not) exposed to the risk factor, respectively. The expected population proportions are,

$$\begin{aligned} P_1 &= z_1 p, \\ 1 - P_1 &= z_2(1-p), \\ P_2 &= (1-z_1)p, \\ 1 - P_2 &= (1-z_2)(1-p). \end{aligned}$$

Substituting these values into equation 2.9 above, we obtain

$$\begin{aligned} OR &= \left[\frac{z_1 p}{z_2(1-p)}\right] \Big/ \left[\frac{(1-z_1)p}{(1-z_2)(1-p)}\right], \\ &= \left(\frac{z_1}{z_2}\right) \Big/ \left(\frac{1-z_1}{1-z_2}\right), \\ &= \left(\frac{z_1}{1-z_1}\right) \Big/ \left(\frac{z_2}{1-z_2}\right), \end{aligned}$$

which is simply the ratio of the odds of being a case given an exposure to the risk factor, $z_1/(1-z_1)$, versus the odds of being a case given no exposure, $z_2/(1-z_2)$. This effectively quantifies the increased risk due to exposure to the risk factor independent of the population frequency of the risk factor. Clearly, if the factor does not influence risk this ratio will be 1, otherwise it will be greater than 1.

2.4.2 Relative risk

The relative risk (RR) is defined as the probability that an individual exposed to the risk factor develops the disease (e.g., becomes a case) divided by the probability that an unexposed individual develops the disease,

$$RR = \frac{\Pr(\text{case}|\text{exposed})}{\Pr(\text{case}|\text{unexposed})} = \frac{z_1}{z_2}.$$

The relationship between the OR and RR is

$$OR = RR \times \left(\frac{1-z_2}{1-z_1}\right).$$

Therefore, $OR \approx RR$ in the case that a disease is rare, so that risks for both exposed and unexposed individuals are small (e.g., $z_i << 1$, $i = 1,2$). It is not always possible to estimate the RR directly in case-control studies. The relevant ratio of population parameters is

$$RR = \frac{\Pr(\text{exposed}|\text{case})}{\Pr(\text{unexposed}|\text{case})} \times \frac{\Pr(\text{unexposed})}{\Pr(\text{exposed})}$$

$$= \left(\frac{p_C}{1-p_C}\right) \times \left(\frac{1-p}{p}\right),$$

where p_C is the frequency of the risk factor among cases and p is the overall population frequency of the risk factor (usually unknown). In much of the human genetics literature the terms relative risk and odds ratio are used interchangeably to refer to what in this book we call the odds ratio.

2.4.3 Odds ratio estimators

A straightforward estimator of the OR (Woolf, 1955) uses the observed sample proportions to estimate the population proportions, applying equation 2.9 above. Table 2.6 presents a contingency table of the outcomes for a case-control association study of a binary disease trait. The estimator of Woolf (1955) is

	Risk Factor		
Disease	+	-	Total
+	a	b	n_+
-	c	d	n_-

Table 2.6 *Contingency table of possible outcomes for a case-control study of a binary risk factor. Plus and minus signs indicate the presence or absence of either the risk factor (row) or the disease (column).*

$$OR_W = \frac{a \times d}{b \times c}. \tag{2.10}$$

For small samples Haldane (1956) suggested the formula,

$$OR_H = \frac{(a+1/2)(d+1/2)}{(b+1/2)(c+1/2)}. \tag{2.11}$$

Haldane (1956) and Anscombe (1956) showed that $\log OR_H$ is an approximately unbiased estimator of $\log OR$. One approach for applying these formulae to biallelic SNPs is to partition genotypes into binary classes, for example if we label the alleles 1 and 2 we could consider 11 versus 12 or 22, and so on (e.g., Thomson, 1981). This quantity is often referred to as the genotype relative risk.

To illustrate, we apply equation 2.10 to the IBD data of Duerr et al. (2006) given in table 2.4. Treating genotypes A/A and A/G as having equivalent risk and G/G as the non-risk genotype gives,

$$OR_W = \frac{21 \times 468}{70 \times 521} = 0.269.$$

The log-odds is $LOD_W = -1.31$. Applying equation 2.11 to the IBD data gives,

$$OR_H = \frac{(21 + 1/2) \times (468 + 1/2)}{(70 + 1/2) \times (521 + 1/2)} = 0.274.$$

The log-odds is $LOD_H = -1.29$. In this example, the sample size is quite large and so the two methods produce very similar results. As noted previously, the A allele at this locus appears to be protective and the *LOD* score is therefore negative. The proportion of IBD cases among individuals with no copies of the A allele is about 3-fold higher than among individuals with either one or two copies of A.

2.4.4 Hypothesis tests and confidence intervals

Woolf (1955) developed an approximate method for inferring the standard deviation of the *LOD*,

$$\sigma_W = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

and Haldane (1956) suggested the approximation

$$\sigma_H = \sqrt{\frac{1}{a+1} + \frac{1}{b+1} + \frac{1}{c+1} + \frac{1}{d+1}},$$

which produces smaller values, particularly when one or more cell counts in the contingency table are low (typically the case for small samples or low frequency alleles). Using either of these equations to infer σ an approximate 95% confidence interval for the *OR* is (see Fleiss, 1979),

$$OR \pm 1.96 \times OR \times \sigma,$$

and an approximate 95% confidence interval for the *LOD* is

$$LOD \pm 1.96 \times \sigma,$$

where here OR and LOD indicate estimates of the odds ratio or log-odds obtained using either of the equations 2.10 or 2.11 above. Although the approximations presented above tend to be quite accurate for the sample sizes and allele frequencies found in many GWASs, if needed exact confidence intervals for OR and LOD can be calculated numerically based on the theory outlined in Cornfield (1956). An efficient algorithm is described by Thomas (1971). If one is willing to assume HWE in controls, even more precise estimates of the OR are possible (Lathrop, 1983).

One test of the hypothesis that an allele has no influence on disease risk is to examine whether the 95% confidence interval of OR (or LOD) includes 1 (or 0), if so we accept the null hypothesis of no effect on disease risk, otherwise we reject the null hypothesis. This is essentially a test of disease-locus association. To illustrate calculation of OR confidence intervals and hypothesis tests, we analyze the data of Duerr et al. (2006) considered in the previous section. This gives $OR_H = 0.274$, the standard error of the log-odds is estimated to be

$$\sigma_H = \sqrt{\frac{1}{21+1} + \frac{1}{70+1} + \frac{1}{521+1} + \frac{1}{468+1}} = 0.252,$$

and the approximate 95% CI of OR_H is,

$$\begin{aligned} OR_H &= 0.27 \pm (1.96 \times 0.274 \times 0.252) \\ &= 0.27 \pm 0.14. \end{aligned}$$

The approximate 95% CI for LOD is

$$\begin{aligned} LOD &= -1.29 \pm (1.96 \times 0.252) \\ &= -1.29 \pm 0.49. \end{aligned}$$

Thus we reject the null hypothesis of no effect on disease risk at the $\alpha = 0.05$ level.

Another approach to test the hypothesis that $OR = 1$ (or $LOD = 0$) makes use of the fact that under the null hypothesis the statistic

$$T = \frac{(\log OR_H)^2}{\sigma_H^2},$$

follows an asymptotic distribution that is chi-square with 1 degree of freedom. For the data of Duerr et al. (2006), the test statistic has the value

$$T = \frac{(-1.29)^2}{0.252^2} = 26.3,$$

which has significance (tail probability) $P = 2.9 \times 10^{-7}$ assuming a chi-square distribution with 1 df.

2.5 Multiple testing corrections

In a locus-by-locus GWAS many tests are performed and corrections for multiple testing are needed to prevent the number of significant associations from increasing as a function of the number of loci. However, the GWAS is not typical of situations in which multiple test corrections are applied. Most multiple testing corrections assume that a common hypothesis is being tested with each additional experiment. The set of experiments constitute a "family" of hypothesis tests and the family-wise significance is the probability of a type I error (rejection of a true null hypothesis) in any test. Thus, if multiple experiments are performed a rejection of the null hypothesis under any experiment is assumed to be incompatible with the null hypothesis.

The GWAS does not fit the standard multiple-testing paradigm because each experiment (locus) tests a different hypothesis (that a particular locus is associated with disease) and a rejection of the null hypothesis for one experiment (locus) normally does not invalidate the null hypothesis for other experiments. The usual motivation for controlling type I error rates in GWASs is thus to control the frequency of false positives, not to fix the global (or family-wise) type I error rate for the entire array of SNP loci.

In this section, we discuss several widely-used strategies for controlling type I error rates in GWASs. However, it is important to recognize that the GWAS hypothesis testing problem is complicated by factors other than simply the large number of hypothesis tests. The tests performed at individual loci have a complicated dependence structure, both because SNP alleles at closely-linked loci can be in linkage disequilibrium (see Chapter 3) or have epistatic interactions (and are therefore correlated) and because even when the loci are on different chromosomes and have independent effects the errors associated with the phenotype measures on individuals are shared across all loci tested for that individual.

To illustrate the dependence among tests, consider the quantitative genetic model of traits influenced by loci with additive effects presented in Section 1.6.2. If P_i is the phenotype measured for individual i and two genetic loci contribute to the trait, then the phenotype is given by the equation

$$P_i = a_1 x_1(i) + a_2 x_2(i) + \epsilon_i.$$

If locus-by-locus significance tests are carried out using genotypes $x_j \in (0, 1, 2)$, we ignore the fact that the same error distribution (terms ϵ_i) applies for all loci. We also ignore potential non-causal relationships between loci due, for example, to linkage disequilibrium (LD) between locus 1 and 2, with locus 2 having a causal influence on disease risk (e.g., $a_2 > 0$) but locus 1 only appearing to have a causal influence (e.g., $a_1 = 0$) through its correlation with locus 2 via LD. We have already considered multiple regression methods that can accommodate covariance among loci in their effects on a continuous phenotype. In Chapter 3 we consider methods that attempt to deal with these non-independence issues by simultaneously analyzing multiple SNP loci and explicitly incorporating such factors as linkage disequilibrium. Here, although we make corrections to deal with the multiplicity of tests, the results should be interpreted with caution.

2.5.1 Bonferroni correction

Suppose that m hypothesis tests are performed and the probability of a false positive (type I error) for any given test is α. In a GWAS the number of tests, m, equals the number of loci L multiplied by the number of traits measured, t, so that $m = L \times t$. For example, in a case-control association study there is a single binary trait (case/control) and $m = L$. If the tests are independent and the null hypothesis H_0 is true then the family-wise probability of at least one rejection is

$$\Pr(\text{at least on rejection}|H_0) = 1 - (1 - \alpha)^m.$$

For a fixed significance level α, this probability increases monotonically towards 1 with increasing m. The Bonferroni correction controls the family-wise rejection rate by instead using a test-wise rejection threshold for p-values that depends on the number of tests performed. The null hypothesis is rejected for test i if

$$P_i < \frac{\alpha}{m}$$

With this correction the probability of at least one rejection is

$$\Pr(\text{at least on rejection}|H_0) = 1 - (1 - \alpha/m)^m$$

$$
\begin{aligned}
&= \alpha - \frac{3\alpha^2}{m} + \cdots \\
&< \alpha. \qquad (2.12)
\end{aligned}
$$

As the number of tests increases the global rejection probability converges to,

$$\lim_{m \to \infty} 1 - \left(1 - \frac{\alpha}{m}\right)^m = 1 - e^{-\alpha} = \alpha + \frac{\alpha^2}{2!} - \frac{\alpha^3}{3!} + \cdots$$

Thus, the global significance is never greater than α and (for small α) it approaches the nominal level α with increasing numbers of loci. If tests are dependent, the Bonferroni correction will become increasingly conservative (e.g., the actual type I error will be much smaller than α). For example, if the tests performed at m loci are completely dependent then either all tests reject (with probability α/m) or all tests accept (with probability $1 - \alpha/m$) and so the actual type I error rate is α/m.

As an example, consider the p-values for association at 4 loci (experiments) presented in table 2.7. If the family-wise significance level is taken to be 10^{-3}, then the significance threshold for evaluating each p-value under a Bonferroni correction is $10^{-3}/4 = 2.5 \times 10^{-4}$. Thus, without a Bonferroni correction, both locus 1 and 3 would show a significant association, whereas with the correction only the association at locus 3 is considered significant.

Locus	p-value	Bonferroni	False Discovery
1	5×10^{-4}	NS	NS
2	1×10^{-2}	NS	NS
3	2×10^{-4}	S	S
4	0.21	NS	NS

Table 2.7 *Results obtained by applying Bonferroni and FDR corrections for multiple testing to analyze the significance of p-values obtained from 4 loci (experiments). The family-wise significance level was* $\alpha = 10^{-3}$ *in all cases.*

To further illustrate the outcome of a Bonferroni correction, we consider the GWAS of metabolic traits by Sabatti et al. (2009) who performed tests of association between $N = 329,091$ SNP loci and 9 metabolic traits including levels of triglycerides, lipoprotein, glucose, C-reactive protein, and so on, for a sample of individuals from Finland. The test-wise p-value rejection threshold, α^*, for this study, assuming a family-wise type I error rate of $\alpha = 0.05$, and only correcting for the number of loci (using a Bonferroni correction) is

$$\alpha^* = \frac{0.05}{329091} = 1.5 \times 10^{-7}.$$

Correcting for both the number of loci and the number of traits tested gives,

$$\alpha^* = \frac{0.05}{329091 \times 9} = 1.7 \times 10^{-8}.$$

For loci with small effects or low frequencies a very large sample of cases and controls is needed to achieve this level of significance.

2.5.2 False discovery rate correction

A different approach to controlling significance in multiple testing problems is to control the "false discovery rate" (FDR). Under this paradigm, rejecting a null hypothesis is a "statistical discovery" and rejecting a true null hypothesis is a "false discovery." The FDR is defined as the expected proportion of hypothesis rejections that are incorrect (e.g., those that reject a true null hypothesis and are therefore false discoveries). Table 2.8 shows the possible outcomes for m hypothesis tests, where m is fixed by the experiment and the outcomes are random variables. The proportion of erroneously rejected null hypotheses is

	Accepted H_0	Rejected H_0	Total
True H_0	U	V	m_0
False H_0	T	S	$m - m_0$
Total	$m -$ R	R	m

Table 2.8 *Variables summarizing counts of possible outcomes for m experiments that each test a null hypothesis H_0.*

$$Q = \frac{V}{V+S},$$

and the false discovery rate (FDR) is defined to be the expected value of this ratio of random variables,

$$FDR = \mathbb{E}(Q) = \mathbb{E}\left[\frac{V}{V+S}\right],$$

where the variables V and S are as defined in table 2.8.

Suppose that a total of m independent tests are performed and that m_0 of the data sets to which the tests are applied were generated under the null hypothesis (e.g., the null hypothesis is true for m_0 of the tests). The Benjamini-Hochberg method for adjusting p-values in multiple tests guarantees that,

$$FDR \le \frac{m_0}{m}\alpha \le \alpha,$$

There are 4 steps in the algorithm: (1) Rank order the p-values from each experiment so that

$$P_{(1)} < P_{(2)} < \cdots < P_{(m)}.$$

(2) Calculate a vector of values for the function l_i and find the position of the largest p-value $P_{(i)}$ that is less than l_i,

$$l_i = \frac{i\alpha}{m}, \quad R = max\{i : P_{(i)} < l_i\}.$$

(3) Set the test-wise p-value rejection threshold to be $\alpha^* = P_{(R)}$. (4) Reject all $H_{0(i)}$ for which $P_i \le T$.

As an example, we again consider the p-values for association at 4 loci (experiments) presented in table 2.7. If the family-wise significance level is taken to be 10^{-3}, then the rank-ordered p-values and values of function l are

$$\begin{aligned} P_{(1)} &= 2 \times 10^{-4}, \\ P_{(2)} &= 5 \times 10^{-4}, \\ P_{(3)} &= 1 \times 10^{-2}, \\ P_{(4)} &= 0.21, \end{aligned}$$

and

$$\begin{aligned} l_1 &= \frac{1 \times 10^{-3}}{4} = 2.5 \times 10^{-4}, \\ l_2 &= \frac{2 \times 10^{-3}}{4} = 5 \times 10^{-4}, \end{aligned}$$

$$l_3 = \frac{3 \times 10^{-3}}{4} = 7.5 \times 10^{-4},$$

$$l_4 = \frac{4 \times 10^{-3}}{4} = 10^{-3}.$$

Note that $P_{(1)} < l_1$ but $P_{(i)} \geq i_i$ for all $i > 1$ and therefore $\alpha^* = P_{(1)} = 2 \times 10^{-4}$ and only the association for locus 3 is considered significant.

Sabatti et al. (2009) used a false discovery rate of 0.05 in their GWAS of 9 metabolic traits described above. There were $N = 329,091$ SNP loci in that study and the test-wise p-value rejection threshold, obtained by applying the Benjamini-Hochberg algorithm, was determined to be $\alpha^* = 1.2 \times 10^{-6}$. Because the tests are not independent, Sabatti et al. (2009) chose to use the more conservative threshold of 1.5×10^{-7} for determining genome-wide significance of associations between SNPs and the 9 metabolic traits in their study.

2.5.3 Tail strength

Let m experiments be performed and a statistical hypothesis test applied to the data from each experiment. Taylor and Tibshirani (2006) proposed a test of whether the null hypothesis is true for all the experiments. Under this hypothesis, the m p-values, p_i for $i = 1, 2, \ldots, m$ are independent and identically distributed (i.i.d) uniform random variables on the interval $[0, 1]$. If the null hypothesis is false for some of the experiments there will be an excess of small p-values. To capture this effect, they proposed a "tail-strength" statistic that is a function of the p-values,

$$TS(p_1, \ldots, p_m) = \frac{1}{m} \sum_{k=1}^{m} \left(1 - p_{(k)} \frac{m+1}{k}\right), \tag{2.13}$$

where

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)},$$

are the rank-ordered p-values (referred to as the order statistics). This statistic takes positive values when p-values are small and negative values when they are large. Because the distribution of the p-values is uniform, the order statistics divide the interval into segments of equal size so that the expectation (mean) of the kth rank-ordered p-value under the null hypothesis is

$$\mathbb{E}[p_{(k)}] = \frac{k}{m+1},$$

and the expected value of the tail-strength statistic is zero, $\mathbb{E}[TS] = 0$. Furthermore, the asymptotic distribution of TS as $m \to \infty$ is a normal distribution with mean 0 and variance,

$$\sigma_{TS}^2 = \frac{1}{m}, \tag{2.14}$$

if the experiments are independent. Thus, under the null hypothesis the test statistic has 95% CI,

$$TS \pm 1.96 \times m^{-1/2},$$

If this interval excludes zero then the null hypothesis is rejected. If the experiments are not independent, equation 2.14 tends to underestimate the variance (and thus rejects too often). Taylor and Tibshirani (2006) proposed that a permutation procedure instead be applied to the original data to estimate $\sigma_{(TS)}$. Permutation procedures are discussed in the next section.

2.5.4 Permutation methods

A permutation test provides a non-parametric method for evaluating the hypothesis that two or more samples are from the same probability distribution (see Wasserman, 2004). Permutation is most useful in cases where sample sizes are small (and asymptotic theory therefore invalid) because it provides an exact test. Here we will be interested mainly in permutation procedures for obtaining family-wise significance values when multiple tests are performed. However, the method can also be used for estimating the significance of a single hypothesis test.

We first give a generic description of the permutation procedure for comparing two samples, then we consider the specific case of a test for genetic association. Let $x_1, \ldots, x_k$ and $y_1, \ldots, y_n$ be two independent samples of size k and n. Let $F_1(x_1, \ldots, x_k)$ and $F_2(y_1, \ldots, y_n)$ denote the cumulative distribution functions (CDFs) for each sample. Under the null hypothesis the samples have the same CDF so that $H_0 : F_1 = F_2$ while under the alternative hypothesis their CDFs are different so that $H_1 : F_1 \neq F_2$. Let $T(x_1, \ldots, x_k, y_1, \ldots, y_n)$ be a test statistic and let T_{obs} be its value calculated for the observed data. The permutation procedure is carried out as follows:

1. label each data point with an index from 1 to $n + k$.
2. enumerate all $(n + k)!$ permutations of the indexes.
3. order data points according to index for each permutation

4. for each permutation calculate T_i^* where $i = 1, 2, \ldots, (n+k)!$.

Under the null hypothesis, all sample permutations have the same joint probability distribution and each test statistic value calculated for the permuted data is equally likely. Thus the distribution of T^* for the permuted data provides the sampling distribution of the test statistic under the null hypothesis. The p-value for the observed data is calculated as

$$\mathrm{p-value} = \Pr(T^* > T_{obs}) = \frac{1}{(n+k)!} \sum_{i=1}^{(n+k)!} I(T_i^* > T_{obs}),$$

where $I(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise.

Permutation test of allelic association

We now consider a case-control study of association for a biallelic SNP locus with three genotype classes g_0, g_1 and g_2. Our test statistic compares genotype frequencies in cases versus controls. The expected frequency of genotype g_i among cases is

$$f(g_i|case) = \frac{f(case|g_i)f(g_i)}{f(case)} = f(case|g_i) \times \frac{f(g_i)}{f(case)}.$$

We now apply permutation to the binary case-control variable; this does not influence the marginal distributions $f(case)$ or $f(g_i)$ but alters the conditional distribution so that $f(case|g_i) \approx f(case)$ and $f(g_i|case) \approx f(g_i)$ for the permuted datasets. Calculating an association test statistic, T, from each of these permuted datasets will produce the null distribution for T against which the observed value can be compared to determine significance.

To illustrate the permutation procedure in the case of a single locus, consider the IBD data of Duerr et al. (2006) shown in table 2.4 We will use permutation to generate the sampling distribution of the genotype association likelihood ratio test statistic given by equation 2.7 Permuting case-control status among individuals in each genotype class, the probability of any particular partition with x, y and z cases in the g_0, g_1 and g_2 genotype classes, respectively, is

$$\Pr(x, y, x) = \binom{n_0}{x}\binom{n_1}{y}\binom{n_2}{z}\left(\frac{1}{2}\right)^{x+y+z},$$

where n_0, n_1 and n_2 are the total counts of cases and controls in each genotype class. It is efficient to sample directly from this distribution

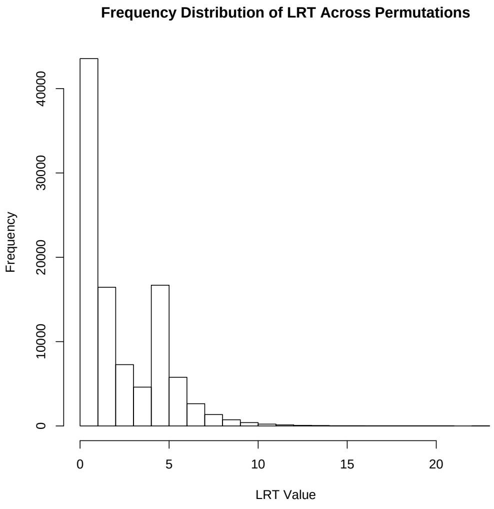

Figure 2.1 Frequency distribution of simulated likelihood ratio test (LRT) values for genotype association under null hypothesis obtained by permutation of case-control status.

to obtain permuted datasets, rather than permuting indexes of individuals, as described above. The two procedures are equivalent. Simulating samples from this distribution, calculating the LRT test statistic, and rank ordering the results we can predict the probability of a value at least as large as the observed value under the null hypothesis. The distribution of LRT values for 100,000 randomly permuted datasets is shown in Figure 2.1 The observed value of $-2 \log \lambda = 34.84$ is greater than any of the values observed in the simulation and therefore has a p-value no larger than 10^{-5}. The p-value calculated based on the asymptotic chi-square distribution is 2.73×10^{-8} suggesting that as many as 10^9 random permutations may be needed to accurately infer the p-value in this example.

Churchill and Doerge (1994) proposed the use of permutation to determine significance for quantitative trait loci mapped via breeding crosses in which many markers were studied and many tests carried out. Analogous procedures can be applied for determining significance in GWASs (e.g., Purcell et al., 2007). Churchill and Doerge (1994) permute the trait value across individuals in one of two ways: (1) comparisonwise: carrying out a separate permutation analysis across individuals for each locus; (2) experimentwise: carrying out a combined permutation analysis across individuals for all loci.

To determine the comparisonwise threshold, generate N permutations of the data by permuting phenotypes (e.g., trait measures or case-control status) across individuals. This could be done either by labeling each individual with an index (as described above) and permuting indexes across phenotypes, or by marginally permuting phenotypes across individuals at each locus. The second approach is preferable from a statistical perspective as otherwise the same permutations are shared across all loci, creating potential correlations in test results. However, computationally the first procedure is much more efficient (Churchill and Doerge, 1994) because it requires only N permutations versus $N \times L$ permutations in the second procedure. In either case, the threshold for significance at each locus is estimated by finding the $100(1-\alpha)$ percentile of the distribution of test statistic values calculated for the N permuted datasets. For any given locus, this is the same procedure as was presented in the previous example.

To determine the experimentwise threshold for significance (equivalent to what we earlier called the family-wise threshold) we generate N datasets by permuting individual indexes across phenotypes. A test statistic is calculated for each locus for the ith permuted dataset and we find the largest test statistic $\max T_i$ across loci. The distribution of $\max T$ is used to determine the experimentwise threshold by taking the $100(1-\alpha)$ percentile of the distribution of $\max T$ across permutations. The experimentwise critical value is used to detect an association at one or more loci in the genome, controlling the type I error rate.

2.6 Graphical representation of GWAS results

Many computer software packages are available for visualizing the results of a GWAS (e.g., Chen et al., 2008; Martin et al., 2009). Most allow the results of multiple studies (or high-dimensional phenotypes) to be displayed simultaneously. One of the simplest procedures for visualing locus-wise significance scores is the Manhattan plot. A Manhattan plot is a scatterplot that arrays the SNP loci according to physical position along the x-axis and plots the height of the absolute value of $\log_{10} p$, where p the probability score (obtained using one of the association methods described earlier) for a SNP on the y-axis. Increasing height on the y-axis indicates increasing significance. Figure 2.2 shows a Manhattan plot generated using the software package AssociationViewer (Martin et al., 2009). The figure is based on an analysis of an example data file included with the program. A quantile-quantile plot (see

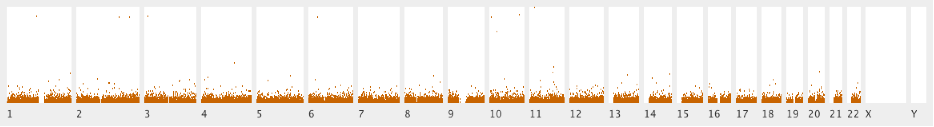

Figure 2.2 Manhattan plot of data from a GWAS. Each square represents a chromosome and SNPs are ordered across the chromosome according to physical position. The points are log-probability scores with increased height indicating an increased significance level.

Rice, 1995) can also be useful for identifying unusual distributions of p-values that might be due to population substructure, SNPs of unusually large effect, etc (Clayton et al., 2005). The use of Q-Q plots for identifying cryptic population substructure will be further discussed in Chapter 5.