father transmitted the A_1 allele to his offspring. In fact, if the process generating these genotyping errors is random, then the end result is an apparent increase in transmission of the common allele purely through chance.

The actual increase in the type-I error rate depends upon a number of factors including: the number and frequency of alleles, the number of individuals in the study, as well as the process responsible for the genotyping error. In general, type-I error rate will increase as minor allele frequency decreases, sample size increases and the amount of genotyping error increases (Mitchell *et al.*, 2003). How much genotyping error is sufficient to increase the type-I error rate? Mitchell *et al.*'s simulations suggested that inflation is severe when the error rate is >5%, but even rates as low as 0.5% are sufficient to produce distortion when minor allele frequency is low.

To ensure that positive TDT results are not merely an artifact of genotyping error, investigators should exercise caution when interpreting significant results where the common allele is transmitted preferentially, particularly if the disease under study is likely to have selection pressure against it (Mitchell *et al.*, 2003). In contrast, if the rare allele is over-transmitted then one can be more confident in the results. It might also be informative to examine the pattern of linkage disequilibrium amongst adjacent markers, retype the significant marker (on a different platform if possible) or perform haplotype analysis. Alternatively, one could use a likelihood-based TDT which explicitly models the effect of genotyping error (Gordon *et al.*, 2001).

Phenotype misclassification

Genetic case-control studies are typically performed assuming that individuals are classified correctly into affected and unaffected groups. However, in practice phenotype misclassification can occur as a result of data entry or processing errors as well as errors in the diagnosis of individuals. These errors are particularly relevant in psychiatric genetics where making the correct diagnosis is often difficult. For example, it has been estimated that >15% of Alzheimer's and Parkinson's disease patients are initially misdiagnosed in the clinic, and it is only subsequently upon autopsy that a definitive diagnosis can be made unequivocally (Lansbury, 2004).

Although random errors in phenotype classification should not affect the type-I error rate of genetic case-control studies (Bross, 1954; Edwards *et al.*, 2005), any misclassification will reduce real differences between the groups and hence the power to detect genetic association (Mote and Anderson, 1965). Zheng and Tian (2005) examined the impact that misclassifying individuals had on the power to detect association using Armitage's (1955) trend test. For prevalent diseases (i.e., $K = 0.1$), when individuals were diagnosed with 95% sensitivity (i.e., the probability that an individual who is truly affected tests positive) and 95% specificity (i.e., the probability that a person without disease tests negative), the required sample size almost doubled as compared to the situation where cases and controls were correctly identified. For rarer diseases ($K = 0.01$), the increase in sample size needed to maintain power was even more alarming, with sample sizes needing to be over thirty times as large under some scenarios. The important point is that even small amounts of phenotype misclassification require large increases in the number of individuals required to maintain statistical power (Zheng and Tian, 2005).

The effect of misclassifying an affected individual as unaffected on the statistical power of the test is not the same as the reverse error, and the impact of each depends upon the prevalence of disease (Edwards *et al.*, 2005). For example, as prevalence of diseases approaches zero, the cost of misclassifying a control as a case becomes infinitely large, whereas penalty for misclassifying an affected individual as unaffected approaches zero (Edwards *et al.*, 2005). However, for most diseases where prevalence is equal or less than about 0.1, it is more important to ensure that cases are truly affected rather than ensuring that controls are really unaffected (Edwards *et al.*, 2005).

21.5 Genome-wide Association

Recent advances in high throughput technology and decreased genotyping costs (Matsuzaki *et al.*, 2004), as well as the publication of the International Haplotype Map of the human genome (Altshuler *et al.*, 2005) have meant that it is now possible to search for complex disease genes by screening thousands of individuals on hundreds of thousands of polymorphisms across the genome. Already, the first genome-wide association studies are beginning to appear in the literature with encouraging results (Cheung *et al.*, 2005; Klein *et al.*, 2005; Maraganore *et al.*, 2005; Ozaki *et al.*, 2002). Whilst these designs hold great promise for dissecting the genetic basis of complex traits and diseases, there are a number of statistical issues particular to this sort of study design. Two issues that we examine here are multiple testing and two-stage strategies.

Multiple testing in genome-wide association

By its very nature, genome-wide association involves performing thousands of statistical tests. Evaluating each test against an uncorrected threshold would produce an excess of loci declared significant purely by chance. It is therefore necessary to control for multiple testing so that false positives are minimized and valuable resources are not wasted following up spurious associations. A commonly employed approach is to use a Bonferroni correction which limits the probability of making at least one false rejection of the null hypothesis to $\alpha_{\text{genome-wide}} \leq 0.05$. Specifically, each comparison is evaluated against a point-wise threshold of $\alpha_{\text{nominal}}/L$ where L is the number of tests of association which are performed. However, the Bonferroni correction is conservative when statistical tests are not independent (e.g., when linkage disequilibrium exists between markers), and also means that very low p-values are required in order to achieve significance when the number of comparisons is large as in genome-wide association. This means that very large sample sizes will be required in order to detect genes of small to moderate effect with appreciable power (see e.g., *Tables 21.1–21.3*).

Two less-conservative procedures which correct for the effect of multiple testing are permutation testing and controlling the false discovery rate (FDR). Permutation testing derives the distribution of the test statistic under the null hypothesis by shuffling case-control status (or the phenotypic values in the case of a quantitative trait) relative to the genotypes and performing the same analysis multiple times. It is then possible to compare the actual significance values with those obtained under permutation to obtain an empirical level of significance (Churchill and Doerge, 1994). Permutation testing also has an advantage that it is less likely to produce spurious results in small datasets where the influence of outliers or batch effects may invalidate the assumptions underlying traditional asymptotic tests of significance (Cheung *et al.*, 2005; Evans and Cardon, in press; Stranger *et al.*, 2005).

Whereas traditional multiple testing procedures such as the Bonferroni correction limit the overall probability of making a type-I error anywhere in the genome scan, controlling the FDR aims to limit the expected proportion of errors among those loci identified as significant (Benjamini and Hochberg, 1995; Efron and Tibshirani, 2002; Storey and Tibshirani, 2003). The basic idea behind the approach is that under the null hypothesis the observed p-values are expected to be distributed uniformly. However, under the alternative hypothesis, more of the significance values should be distributed close to zero. In other words, the observed distribution of p-values in a genome-wide scan should be a mixture of these two distributions. The FDR method finds a cutoff value so that results with smaller p-values are likely to be true positives from the alternative distribution. The FDR is equal to the genome-wide error rate when all hypotheses are true but is smaller otherwise.

The corollary is that by employing a FDR there is scope for increasing power of the genome-wide analysis compared to that obtained via traditional Bonferroni correction.

Benjamini and Hochberg (1995) describe a simple procedure for controlling the FDR. Let the ordered set of p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ correspond to tests of the m hypotheses $H_1, H_2, \dots, H_m$. Let k be the largest i for which $p_{(i)} \leq \frac{i}{m} q^*$; then reject all H_i for $i = 1, 2, \dots, k$. This procedure will control the FDR at q^* (Benjamini and Hochberg, 1995). In order to make this clear, we apply this procedure to the fictional dataset in *Table 21.6*. Seven different hypotheses were tested for association. We first rank the p-values from smallest to largest and then calculate $\frac{i}{m} q*$ in order to maintain the FDR at $q^* = 0.05$. Using this criterion, we find that the first three hypotheses are rejected. In contrast, if we had performed the more conservative Bonferonni correction (i.e., $\alpha = 0.05/7 = 0.007$), only the first two hypotheses would have been rejected.

Two-stage studies in genome-wide association

In order to decrease the cost of genotyping thousands of subjects with hundreds of thousands of markers, a more cost-effective strategy might be to genotype subjects in stages. That is, in an initial stage, a proportion of individuals ($p_{individuals}$) are genotyped on all markers ($N_{markers}$), and subsequently, a proportion of markers displaying the most promising results ($\pi_{markers}$) are typed in the remaining individuals (Sobell *et al.*, 1993). Compared to one-stage approaches where all markers in all individuals are genotyped, two-stage designs can lead to appreciable savings in genotyping costs whilst maintaining appreciable power (Maraganore *et al.*, 2005; Satagopan and Elston, 2003; Satagopan *et al.*, 2002, 2004; Skol *et al.*, 2006; Thomas *et al.*, 2005).

Table 21.6 Calculating the False Discovery Rate.

$P_{(i)}$	FDR = q*	Rank = i/m	$\frac{i}{m}q^*$
0.001	0.05	1/7	0.007
0.006	0.05	2/7	0.014
0.010	0.05	3/7	0.021
0.050	0.05	4/7	0.029
0.200	0.05	5/7	0.036
0.500	0.05	6/7	0.043
0.800	0.05	1	0.050

Recently, Skol *et al.* (2006) examined the performance of two different approaches to analyzing the data from two-stage genome-wide association scans. In the 'replication-based' strategy, association was only tested on markers and individuals genotyped in the second phase of the procedure. In order to maintain the genome-wide error rate below $\alpha_{genome} = 0.05$, a Bonferonni-corrected significance level of $\alpha_{genome}/(\mathrm{N}_{markers} \times \pi_{markers})$ was employed. In the 'joint analysis' strategy, test statistics in stages one and two were combined and compared against an approximate significance level of $\alpha_{genome}/\mathrm{N}_{markers}$. Skol *et al.* found that despite the more stringent significance level, jointly analyzing the data from both stages almost always resulted in increased power to detect association as compared to a replication-based strategy; and with appropriately chosen thresholds, power comparable to the one-stage design. The power of the joint analysis decreased as the proportion of samples typed in stage one decreased, presumably because variants which dispose to disease were less likely to be selected for genotyping in stage two. Similarly, taking too few markers through to stage two also decreased the power of the joint analysis because of the decreased probability of taking a true risk variant through to the next stage. In contrast, power increased when fewer markers were selected for follow-up in the replication-based strategy because less statistical tests were performed and hence there was a lower penalty due to multiple testing. In fact the only situation in which the joint analysis was not more powerful than a replication-based strategy was when the strength of association was far greater in stage two than one. The authors therefore recommend that two-stage genome-wide association scans are analyzed using a joint analysis strategy, that a large proportion of samples are genotyped in stage one (i.e., $\pi_{individuals} > 30\%$), and that a relatively large proportion of markers are selected for follow-up in stage two ($\pi_{markers} > 1\%$). The authors provide a useful Web-based tool which may be used by researchers to calculate the power to detect association in two-stage designs at http://csg.sph.umich.edu.

21.6 Calculating Power to Detect Association

In this section, we describe the noncentrality parameters (NCP) for the case-control and transmission-distortion tests of association which are implemented in the GPC.

Case-control studies

The allelic test of association exploits the fact that the penetrance of the heterozygote is often intermediate between the two homozygotes, and collapses across genotypes to form a 2×2 contingency table (*Table 21.7*). Because each individual contributes two (alleles) counts rather than just a single genotype, this is a powerful test of association when alleles contribute to disease risk in a multiplicative manner. However, the approach assumes that alleles pair randomly as in random mating and thus is only inappropriate when genotypes are in Hardy-Weinberg equilibrium (Sasieni, 1997). Consider the 2×2 contingency table displayed in *Table 21.7*. The Pearson chi-square test for allelic independence is calculated as:

Table 21.7 Cell counts for the allelic test of association

	Cases	Controls	Total
A_1	n_{1A}	n_{1U}	$n_{1.}$
A_2	n_{2A}	n_{2U}	$n_{2.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

$$\chi^2 = \sum_{i=1,2} \sum_{j=A,U} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$$

where $E[n_{ij}] = \frac{n_{i.} n_{.j}}{n_{..}}$

The resulting test statistic is distributed as a chi-square distribution with 1 df. The noncentrality parameter for the allelic test of association is given by Mitra (1958):

$$\lambda = 2N_C N_U (p_C - p_U)^2 \frac{N_C + N_U}{(N_C p_C + N_U p_U)(N_C + N_U - N_C p_C - N_U p_U)}, \qquad (21.2)$$

where N_C denotes the total number of cases, N_U the total number of controls, and p_C and p_U are the expected allele frequencies in the case and control samples respectively. Thus, in order to calculate the NCP, we require the size of case and control samples (which we can assume is given), as well as the expected allele frequencies in the case and control samples. In order to calculate p_C and p_U we need to calculate the expected genotype frequencies given an individual is a case or control individual, that is $P(G|D)$. This quantity is calculated via Bayes' theorem:

$$P(G|D) = \frac{P(D|G)P(G)}{\sum_G P(D|G)P(G)}$$

where $P(G)$ is the probability of the genotype (i.e., given by Hardy-Weinberg), and $P(D|G)$ is the probability that an individual is a case (or a control) given their genotype. $P(D|G)$ for the 'A_2A_2' genotype