J Hum Genet (2006) 51:523–532
DOI 10.1007/s10038-006-0393-6

ORIGINAL ARTICLE

Shu-Hui Wen · Jung-Ying Tzeng · Jau-Tsuen Kao
Chuhsing Kate Hsiao

A two-stage design for multiple testing in large-scale association studies

Received: 9 November 2005 / Accepted: 14 February 2006 / Published online: 12 May 2006

Abstract Modern association studies often involve a large number of markers and hence may encounter the problem of testing multiple hypotheses. Traditional procedures are usually over-conservative and with low power to detect mild genetic effects. From the design perspective, we propose a two-stage selection procedure to address this concern. Our main principle is to reduce the total number of tests by removing clearly unassociated markers in the first-stage test. Next, conditional on the findings of the first stage, which uses a less stringent nominal level, a more conservative test is conducted in the second stage using the augmented data and the data from the first stage. Previous studies have suggested using independent samples to avoid inflated errors. However, we found that, after accounting for the dependence between these two samples, the true discovery rate increases substantially. In addition, the cost of genotyping can be greatly reduced via this approach. Results from a study of hypertriglyceridemia and simulations suggest the two-stage method has a higher overall true positive rate (TPR) with a controlled overall false positive rate (FPR) when compared with single-stage approaches. We also report the analytical form of its overall FPR, which may be useful in guiding study design to achieve a high TPR while retaining the desired FPR.

Keywords Association studies · Cost-effectiveness · False positive rate · Multiple testing · Two-stage design

S.-H. Wen
Department of Public Health, College of Medicine,
Tzu-Chi University, Hua-Lien, 97004, Taiwan

J.-Y. Tzeng
Department of Statistics and Bioinformatics Research Center,
North Carolina State University, Raleigh, NC 27606, USA

J.-T. Kao
Department of Clinical Laboratory Sciences
and Medical Biotechnology, College of Medicine,
National Taiwan University, Taipei, 100, Taiwan

C.K. Hsiao (✉)
Division of Biostatistics, Institute of Epidemiology,
National Taiwan University, Taipei, 100, Taiwan
E-mail: ckhsiao@ha.mc.ntu.edu.tw
Tel.: +886-2-33228032
Fax: +886-2-23418562

Introduction

Genetic association analysis has been used widely in the search for genes contributing to complex diseases. One common study design is to collect genomic data from affected and unaffected unrelated individuals, and contrast their genetic features at the level of gene expression (e.g., Hedenfalk et al. 2001; Haga et al. 2002; Ozaki et al. 2002) or marker distribution frequencies (e.g., Risch and Merikangas 1996; Collins et al. 1999). Such studies tend to encounter the problem of multiple testing due to the use of densely spaced markers (Ohashi and Tokunaga 2001). In that case, the Bonferroni procedure, which aims at preventing the occurrence of a single false positive, is frequently adopted. However, the traditional Bonferroni procedure is known to be an over-conservative test because its critical value becomes extremely large when many hypotheses are examined simultaneously.

Several approaches have been proposed with which to pursue higher power in a multiple-testing setting. One direction is to modify the testing procedures, and focus on constructing more powerful tests (see Dudoit et al. 2003 for review). Improved procedures have been obtained either by considering measures different from overall type I error rate, such as false discovery rate (FDR; e.g., Benjamini and Hochberg 1995; Storey and Tibshirani 2003; Tsai et al. 2003), or by empirically obtaining the significance threshold of multiple tests (e.g., Ge et al. 2003; Becker and Knapp 2004). Another direction adopts the design point of view, and accounts for both power and cost. Our proposed approach is motivated by the design perspective.

Given that the cost of a genetic study is largely determined by the number of individuals recruited and the number of markers genotyped, sequential designs to enhance study efficiency are becoming increasingly popular (e.g., Böddeker and Ziegler 2001; Saito and Kamatani 2002; van den Oord and Sullivan 2003a, b; Thomas et al. 2004; Hirschhorn and Daly 2005). Either samples or marker density or both may be increased sequentially. Such strategies have also been proposed to deal with low power due to the Bonferroni correction. For example, for array experiments, Miller et al. (2001) suggested first selecting a smaller set of microarray sample and proceeding to the second stage of tests with another data set using only genes found significant in the first stage. Later, Allison and Coffey (2002) pointed out that this two-stage method can be even more conservative if the two significance levels are not chosen carefully. Several other papers have proposed different two-stage methods from different perspectives. Hoh et al. (2000) and Ott and Hoh (2001) proposed first selecting a smaller subset of single nucleotide polymorphisms (SNPs) and then using them in the modeling stage to reduce the number of coefficients to be estimated. Elston et al. (1996) and Guo and Elston (2000) proposed genotyping at two different spacings to save cost. Several authors (see, for instance, Saito and Kamatani 2002; Satagopan and Elston 2003; van den Oord and Sullivan 2003a, b) advocated the use of two-stage (or multi-stage) procedures that inflate the probability of false positives in the first stage, and control this probability in later stages. Others have focused on genotyping cost, allocation of sample sizes, and selection of significance levels in two-stage designs. Satagopan et al. (2002, 2004) considered a two-stage design under the constraint of a fixed total number of genotypings. Saito and Kamatani (2002) performed extensive simulation studies based on various combinations of type I error in the first stage, sample sizes at two stages, and genotype relative risks, in the search for an optimal design. Their two tests conducted at two stages use different, and hence independent, data. Thomas et al. (2004) used the likelihood approach to select tagging SNPs in the second stage using a larger sample that combines previous data.

Here we introduce a design for a two-stage procedure for multiple testing that includes data from the first stage. We conduct a formal test in the second stage, conditional on the findings of the first, with both the original and the augmented data. In the first stage, the objective is to eliminate those markers that are very unlikely to be associated with the disease of interest among the total number (M) of markers. A large significance level is used, and markers with "large" P values are excluded. In other words, at this stage we intend to include as many true positives as possible by tolerating a larger-than-usual amount of false positives. This follows the same idea as that proposed by van den Oord and Sullivan (2003a, b), i.e., relaxing the false positive probability in the first stage. After obtaining a smaller and more promising set of markers, we conduct statistical tests with a stringently controlled overall type I error using a combination of data from the first-stage sample and the newly genotyped data. van den Oord and Sullivan (2003a) and others (Miller et al. 2001; Saito and Kamatani 2002; Satagopan et al. 2002) considered only new data in their second-stage test to avoid complexity in test statistics due to interdependence between the first and second samples. As argued by van den Oord and Sullivan (2003a), inclusion of first-stage data may elevate false positive discoveries, but can reduce the genotyping burden. Here, we recommend incorporating both samples and we show that the increase in false discovery can be reduced by use of a larger sample size and a more stringent significance level. A test with combined data will save on overall genotyping costs, as long as the dependence and its effect can be handled carefully. This point is similar to that mentioned in Satagopan and Elston (2003) although they did not pursue it further. The proposed procedure is designed to enlarge the power of each test in the first stage and reduce the type I error in the second stage. Overall, the procedure is more powerful at a controlled false positive rate (FPR), and avoids waste of resources in genotyping markers with no association.

In this paper, we use the term true positive rate (TPR) for the probability of rejection of truly associated markers, and FPR for the probability of rejection of non-associated markers. The FPR can be considered as a measure of the overall type I error rate. However, we use TPR and FPR when referring to the overall performance of M multiple tests, and retain type I and II errors when considering each single test. In the following section, we explain the rationale of the design, discuss its implementation, and derive theoretical success and failure rates. More technical details can be found in Appendices 1 and 2. Simulation studies and a real example are then considered. The performance of the method, in terms of TPR and cost saving, is evaluated and compared with other Bonferroni type procedures.

Materials and methods

Two-stage selection procedure

In the first stage, we consider a large significance level, α_1, for each test to ensure that even markers with mild effects will be detected. The aim of obtaining a large TPR at this stage will then be guaranteed. For instance, in population-based case-control studies, a chi-square test for contingency tables or z-test for comparing two proportions can be considered. Taking $\alpha_1 = 0.05$ results in a large power in the first stage for M SNPs (or markers), each with N_1 allele data value. For SNPs, N_1 is the number of total allele counts from the case group and $N_1/2$ is the number of cases. The resulting significant R markers will proceed to the second testing procedure. In this stage, the sample size for each of the significant R markers will be enlarged by the N_2 allele data for each marker. That is, only the R markers of the additional

individuals will be genotyped. Therefore, the total number of genotypings is MN_1 in the first stage and RN_2 in the second. The total cost ($MN_1 + RN_2$) is much smaller than that ($MN_1 + MN_2$) in a single-stage design. When testing the association of the R markers in the second stage, we adopt a smaller significance level, α_2, for each test such that the overall FPR decreases. Some power may be sacrificed but the inclusion of the additional sample will compensate for the loss. The final significant X markers can now be used for further studies such as fine mapping.

In the first stage, our proposal categorizes M markers into two groups: those with large P values (i.e., obviously removable) and the rest (possibly associated). Any induced error due to loose separation can be corrected in the second stage. For instance, one may apply the traditional stringent Bonferroni correction to test the R markers at this stage or use Benjamini and Hochberg's procedure (1995) to control the FDR. This combination ensures a high overall success rate (the TPR) and a low error probability (the FPR).

Notation and implementation

Among the total M markers, let w be the fraction of markers that are truly non-associated with the disease, and $M(1-w)$ be the number of associated markers that we want to identify. The proportion w is usually close to 1 in a large-scale association study (Ozaki et al. 2002). Let N_1 be the sample size of allele data considered in the first stage. Without loss of generality, here we assume the sample size of the case and of the control group to be the same, with N_1 allele data from cases and N_1 from controls. For the purposes of illustration, we take individual biallelic markers as the basic unit for association testing. The results can be generalized to other association tests at haplotypic or genotypic level.

For each marker, let $\alpha_1 = 0.05$ be the significance level in the first stage, and $1-\beta_1$ the corresponding power. Suppose that after M tests in the first stage, R ($R = R_0 + R_a$) markers result in significance, where R_0 are from the original Mw markers of no association and R_a from the $M(1-w)$ markers with association. In the second stage, we increase the sample size by N_2 for each case and control group. Next, we genotype only the R markers of the additional N_2 subjects and perform a second test on these markers of $N_1 + N_2$ individuals based on significance level α_2 where $\alpha_2 < \alpha_1$. In terms of genotyping, as suggested by Satagopan et al. (2002), the total number of gene evaluations up to this point is $MN_1 + RN_2$. The final total of significant markers, after the second stage, is denoted as X ($X = X_0 + X_a$, where X_0 results from R_0 and X_a from R_a).

True positive and false positive rates and numbers

An intuitive measure for a successful testing procedure is the TPR, which can be considered as the overall power. The TPR can be denoted as a product of two conditional probabilities, U_1 and U_2 (see Appendix 1), for success in detection at each stage, respectively. These two conditional probabilities depend on the testing procedures and significance levels employed. In the same manner, we express the overall FPR as a product of the two conditional probabilities Q_1 and Q_2 for incorrect rejection at each stage. Their values depend on both α_1 and α_2. The α_1 can be fixed at a large value, say 0.05, to ensure a high probability of true significance in the first stage. The α_2, however, will be made smaller to control the overall false positive results. For instance, we recommend α_1/R for α_2, where R is determined after the test in the first stage is complete. From Appendix 1, expectations of TPR and FPR can be approximated by $(1-\beta_2)$ and α_2, respectively, under conditions discussed later.

The sample sizes N_1 and $N_1 + N_2$ also affect FPR. Based on the results in Appendix 1, by setting the proportion of N_1 to $N_1 + N_2$ larger than $(z_{1-\alpha_1/2}/z_{1-\alpha_2/2})^2$, the overall FPR will be controlled at the level of α_2. A planned design with properly chosen significance levels ($\alpha_1 > \alpha_2$) and sample sizes ($N_1 > N_2$) will ensure a high overall TPR. Alternatively, the approximated TPR and FPR can be fixed together to determine the sample sizes.

Given M and w, R_0 is binomially distributed with the number of trials Mw and probability Q_1, assuming independence between M tests. Similarly, the number of true positives, R_a, is binomially distributed with size $M(1-w)$ and probability U_1. When the significance level of the first stage is fixed at α_1 for each test, the overall Q_1 becomes α_1, and U_1 is equivalent to $(1-\beta_1)$. Similarly, when α_1 and α_2 are fixed, the conditional probability of incorrect rejection (Q_2) can be approximated by $1/[E(R_0) + E(R_a)]$, and U_2 by the expected value, $E(1-\beta_2)/(1-\beta_1)$ (details in Appendix 1), where the expectations, $E(R_0) = MwQ_1$ and $E(R_a) = M(1-w)U_1$, are taken with respect to the binomial distributions, respectively, at given values of M, w, Q_1, and U_1.

Therefore, the expected number of correctly identified markers (true positives), $E(X_a)$, can be approximated by $M(1-w) \times E(1-\beta_2)$ where $E(X_a) = R_a U_2$ if R_a and U_2 are given. Similarly, following the argument above, the number of false positives, X_0, is binomially distributed as $\mathrm{Bin}(R_0, Q_2)$. The probabilities of obtaining zero or one non-associated marker can be estimated. For instance, the sum is approximately 0.92 when $M = 500$, $w = 0.95$, and $\alpha_1 = 0.05$, which implies a less-than-one false alarm.

Results

Example: association with hypertriglyceridemia

The two-stage procedure was applied to a small data set containing only 15 SNPs as markers. These markers locate in the exons and introns of the four genes, lipoprotein lipase (*LPL*), apolipoprotein A1 (*APOA1*), apolipoprotein C3 (*APOC3*), and apolipoprotein A5 (*APOA5*). There are two, four, four, and five markers

contained in each gene, respectively. The last five markers on the *APOA5* gene have been previously tested for association with hypertriglyceridemia (Kao et al. 2003). The latter study revealed that, when all samples (290 hypertriglyceridemia and 303 controls) are considered, three markers (c.553G > T, c.1259T > C, and IVS3 + 476G > A) are statistically significantly associated with hypertriglyceridemia after Bonferroni's correction.

For the purposes of illustration, we selected 198 individuals randomly from each case and control group in the first stage, and conduct 15 tests each with $\alpha_1 = 0.05$. The significant markers (R markers) then enter the next stage with the data increased by adding the genotypes on each of the R markers of the remaining 197 ($197 = 593-198-198$) individuals. Each marker is tested with the significance level $\alpha_2 = 0.05/R$. The resulting significant markers are considered as showing evidence of association. This procedure is replicated 100 times to assess its performance and to account for sampling variation.

In each replication, we also compute the number of total genotypings, and divide it by 15 to derive the corresponding sample size $K (K = (15 \times 396 + R \times 197)/15)$ for a single-stage design. A total of K individuals are then sampled and their 15 markers are tested for association using Bonferroni correction. The final results from Bonferroni correction with K individuals [$B(K)$], with all individuals [B(all)], with 396 individuals [$B(N_1)$], and from the two-stage method are compared.

Table 1 lists the percentages of significance of each marker over 100 replications. For the purposes of comparison, we consider the results from Bonferroni correction with all data as the hypothetical standard. In other words, markers numbered 7, 9, 10, 12, 14, and 15 with a larger difference in minor allele frequency are considered associated with hypertriglyceridemia. When markers have a strong effect (numbers 12, 14, and 15 in *APOA5*), it is easy to distinguish between case and control groups. All methods were shown to be consistent. For markers of lesser strength (numbers 7, 9, and 10 in *APOC3*), the two-stage method reaches the same conclusion as the Bonferroni procedure with all data [B(all)]. However, the average number of genotypings for the two-stage method is only 7,313.09 ($7{,}313.09 = 15\times396 + 6.97\times197$ where 6.97 is the average R), i.e., 82% [$82\% = 7{,}313.09 / (15 \times 593)$] of the number required under B(all).

When compared with Bonferroni procedure with the same number of genotypings [$B(K)$], the two-stage method correctly identifyies the association with a much higher percentage (≥94%), while $B(K)$ performs poorly under the same number of genotypings.

In another analysis with $N_1 = 298$ (149 cases) and $N_2 = 295$, the two-stage selection procedure still identifies the same signals as Bonferroni's method (results not shown), and the average relative cost in genotypings is only 72%.

Simulation studies

In this section, we describe simulation studies to evaluate the success and error rates of the two-stage selection method. The final numbers of truly significant markers X_a and TPR $X_a/(M(1-w))$ are used to assess success, while the numbers of falsely significant markers X_0 and FPR $X_0/(Mw)$ are adopted to quantify error. The results are compared with those of single-stage methods such as Bonferroni's procedures. As stated in Reich et al. (2003), the minimal allele frequency of most discovered SNPs is greater than 10%. Therefore, we assume in the following that the allele frequencies range between 10% and 90%. That is, the allele frequencies p_c and p_n for the case and control groups, and their weighted average $\bar{p}$ all fall within this interval (10–90%). The range of their absolute difference, δ ($\delta = |p_c - p_n|$), can consequently be derived (Appendix 2). The ranges of the two quantities are shown in Fig. 1. If the genotyping can be made almost free of error so that the allele frequency takes values from 0 to 1, then the range of the two quantities would be larger, as indicated by the dashed lines in Fig. 1.

Table 1 Numbers in the last four columns are percentages of significance in 100 replications. B(all) denotes the Bonferroni procedure with all data, $B(K)$ with K individuals, $B(N_1)$ with only N_1 ($N_1 = 396$) individuals. The absolute difference in minor allele is denoted as δ

Gene	Number	δ	B(all)	Two-stage	$B(K)$	$B(N_1)$
LPL	1	0.030	0	0	0	0
	2	0.012	0	0	0	0
APOA1	3	0.007	0	0	0	0
	4	0.003	0	0	0	0
	5	0.070	0	0	9	5
	6	0.029	0	0	1	7
APOC3	7	0.090	100	94	51	30
	8	0.014	0	0	0	0
	9	0.107	100	98	94	78
	10	0.083	100	98	84	70
APOA5	11	0.009	0	0	0	0
	12	0.228	100	100	100	100
	13	0.010	0	0	0	0
	14	0.156	100	100	100	100
	15	0.165	100	100	100	100

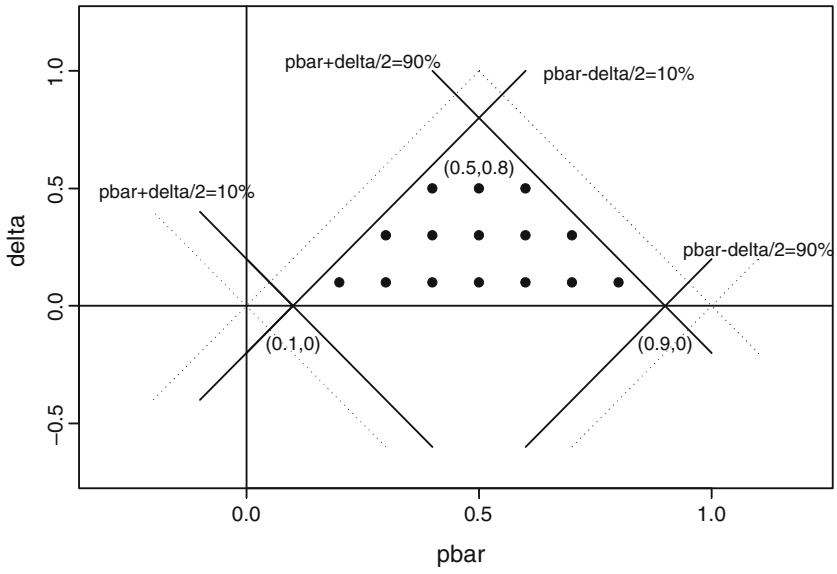

Fig. 1 The range for $\bar{p}$ and δ. *Solid lines* Range of possible values when frequencies are within 10 –90%, *dotted lines* range when frequencies are between 0 and 1

We have fixed M at 500 (Table 2) or 500,000 (Table 3), $w = 0.95$, and $N_1 = 2{,}000$. Numbers of allele counts for Mw non-associated SNPs were simulated under the null H_0:$\delta = 0$ with given $\bar{p}$, while other $M(1-w)$ associated SNPs were with a given $\delta > 0$. We then tested each marker with $\alpha_1 = 0.05$. If any significance was found, an additional sample of size $N_2 = 1{,}000$ would be generated and tested with $\alpha_2 = 0.05/R$. For the purposes of comparison, we also considered two Bonferroni's procedures. One uses only the first N_1 data [$B(N_1)$], while the other [B(all)] considers the data $N_1 + N_2$ in all M markers. The number of replications is 1,000 under each condition. Tables 2 and 3 list the theoretical values (numbers in parentheses) of TPR and FPR to allow comparison with the simulation results. It is evident that the theoretical values are very close to the simulation results.

Table 2 The false positive rate (FPR) and true positive rate (TPR) based on simulations (numbers in parentheses are theoretical values) for the proposed two-stage method and two Bonferroni procedures. The sample sizes are $N_1 = 2{,}000$ and $N_2 = 1{,}000$. The total number of markers is assumed to be $M = 500$ and the proportion of non-associated markers is $w = 0.95$. *FDR* False discovery rate

δ	TPR			FPR $\times 10^2$			FDR	
	$B(N_1)$	B(all)	Two-stage	$B(N_1)$	B(all)	Two-stage	B(all)	Two-stage
$\bar{p} = 0.15$								
0.01	0.001	0.002	0.021 (0.021)	0.009	0.008	0.158 (0.183)	0.431	0.623
0.03	0.108	0.263	0.489 (0.503)	0.007	0.011	0.110 (0.117)	0.007	0.042
0.05	0.708	0.939	0.981 (0.980)	0.010	0.010	0.104 (0.103)	0.002	0.019
0.07	0.990	1.000	1.000 (1.00)	0.008	0.013	0.093 (0.103)	0.002	0.019
0.09	1.000	1.000	1.000 (1.00)	0.011	0.012	0.099 (0.103)	0.002	0.019
$\bar{p} = 0.3$								
0.01	0.001	0.001	0.011 (0.012)	0.009	0.009	0.163 (0.190)	0.621	0.752
0.03	0.035	0.089	0.239 (0.251)	0.010	0.012	0.113 (0.134)	0.021	0.092
0.05	0.331	0.629	0.818 (0.830)	0.008	0.010	0.091 (0.106)	0.003	0.024
0.07	0.824	0.979	0.995 (0.996)	0.010	0.011	0.096 (0.103)	0.002	0.019
0.09	0.990	1.000	1.000 (1.00)	0.010	0.010	0.092 (0.103)	0.002	0.019
0.10	0.999	1.000	1.000 (1.00)	0.011	0.010	0.090 (0.103)	0.002	0.019
0.30	1.000	1.000	1.000 (1.00)	0.009	0.009	0.092 (0.103)	0.002	0.019
0.35	1.000	1.000	1.000 (1.00)	0.010	0.010	0.090 (0.103)	0.002	0.019
$\bar{p} = 0.5$								
0.01	0.001	0.001	0.010 (0.010)	0.014	0.012	0.169 (0.192)	0.675	0.786
0.03	0.023	0.060	0.183 (0.192)	0.008	0.010	0.128 (0.140)	0.031	0.122
0.05	0.230	0.499	0.715 (0.728)	0.008	0.010	0.089 (0.109)	0.004	0.028
0.07	0.699	0.938	0.981 (0.984)	0.009	0.012	0.096 (0.103)	0.002	0.019
0.09	0.964	0.999	1.000 (1.00)	0.010	0.012	0.103 (0.103)	0.002	0.019
0.10	0.993	1.000	1.000 (1.00)	0.011	0.013	0.094 (0.103)	0.002	0.019
0.30	1.000	1.000	1.000 (1.00)	0.007	0.011	0.099 (0.103)	0.002	0.019
0.50	1.000	1.000	1.000 (1.00)	0.012	0.011	0.104 (0.103)	0.002	0.019
0.70	1.000	1.000	1.000 (1.00)	0.011	0.011	0.096 (0.103)	0.002	0.019
0.75	1.000	1.000	1.000 (1.00)	0.009	0.010	0.096 (0.103)	0.002	0.019

Table 3 The FPR and TPR based on simulations (numbers in parentheses are theoretical values) for the proposed two-stage method and two Bonferroni procedures. The total number of markers is assumed to be $M = 500{,}000$

δ	TPR			FPR $\times 10^5$			FDR	
	$B(N_1)$	B(all)	Two-stage	$B(N_1)$	B(all)	Two-stage	B(all)	Two-stage
$\bar{p} = 0.15$								
0.01	< 0.001	< 0.001	< 0.001 (0.0001)	0.011	0.009	0.177 (0.183)	0.147	0.216
0.03	0.004	0.019	0.053 (0.068)	0.006	0.011	0.114 (0.117)	< 0.0001	< 0.0001
0.05	0.182	0.540	0.706 (0.752)	0.011	0.011	0.100 (0.103)	< 0.0001	< 0.0001
0.07	0.813	0.989	0.997 (0.998)	0.011	0.011	0.096 (0.103)	< 0.0001	< 0.0001
0.09	0.997	1.000	1.000 (1.00)	0.012	0.010	0.101 (0.103)	< 0.0001	< 0.0001
$\bar{p} = 0.3$								
0.01	< 0.001	< 0.001	< 0.001 (< 0.0001)	0.011	0.011	0.191 (0.190)	0.339	0.427
0.03	0.001	0.003	0.011 (0.014)	0.012	0.012	0.132 (0.134)	0.001	0.002
0.05	0.030	0.135	0.256 (0.302)	0.008	0.011	0.103 (0.106)	< 0.0001	< 0.0001
0.07	0.309	0.723	0.849 (0.880)	0.009	0.007	0.092 (0.103)	< 0.0001	< 0.0001
0.09	0.813	0.989	0.997 (0.998)	0.009	0.006	0.108 (0.103)	< 0.0001	< 0.0001
0.10	0.944	0.999	1.000 (1.00)	0.009	0.010	0.096 (0.103)	< 0.0001	< 0.0001
0.30	1.000	1.000	1.000 (1.00)	0.012	0.010	0.097 (0.103)	< 0.0001	< 0.0001
0.35	1.000	1.000	1.000 (1.00)	0.011	0.009	0.102 (0.103)	< 0.0001	< 0.0001
$\bar{p} = 0.5$								
0.01	< 0.001	< 0.001	< 0.001 (< 0.0001)	0.010	0.009	0.193 (0.192)	0.417	0.502
0.03	0.0003	0.0013	0.006 (0.008)	0.009	0.011	0.118 (0.140)	0.001	0.003
0.05	0.015	0.072	0.156 (0.192)	0.012	0.009	0.100 (0.109)	< 0.0001	< 0.0001
0.07	0.183	0.536	0.702 (0.752)	0.010	0.010	0.103 (0.103)	< 0.0001	< 0.0001
0.09	0.643	0.950	0.981 (0.987)	0.010	0.010	0.116 (0.103)	< 0.0001	< 0.0001
0.10	0.842	0.992	0.998 (0.999)	0.010	0.011	0.097 (0.103)	< 0.0001	< 0.0001
0.30	1.000	1.000	1.000 (1.00)	0.009	0.011	0.095 (0.103)	< 0.0001	< 0.0001
0.50	1.000	1.000	1.000 (1.00)	0.009	0.010	0.092 (0.103)	< 0.0001	< 0.0001
0.70	1.000	1.000	1.000 (1.00)	0.009	0.009	0.101 (0.103)	< 0.0001	< 0.0001
0.75	1.000	1.000	1.000 (1.00)	0.009	0.011	0.095 (0.103)	< 0.0001	< 0.0001

TPR and FPR

The average TPR for the two-stage method is substantially greater than that of either of the two Bonferroni procedures. Even when the difference in frequency is small ($\delta < 0.05$), the two-stage method is still superior. However, when the difference is extremely small ($\delta = 0.01$), all methods fail to perform satisfactorily. In fact, the required sample size in this case would be larger than 15,000. That is, more subjects are needed to achieve a reasonable power. The second factor of TPR is the average frequency $\bar{p}$. When $\bar{p}$ approaches 0.5, all TPRs decrease, but the two-stage method still outperforms the rest. Between the two Bonferroni procedures, the one with the greater sample size performs better than the other. It should be kept in mind, however, that the cost for B(all) method is larger than others and it requires more laboratory work.

When looking at FPR and FDR, the two-stage method is not as good as the other two. However, these numbers are indeed small. For example, the FPR of 0.158×10^{-2} (Table 2) implies less-than-one false alarm. The FPRs for both Bonferroni procedures are small, indicating a very slim chance of identifying an unassociated marker, which is as conservative as we had expected. The TPR and FPR from simulation studies also match the theoretical values quite well in both cases ($M = 500$ or 500,000), indicating that the equations derived in Appendix 1 are good approximations.

Figure 2 is an alternative presentation of the simulation results. Figure 2a, b compares the TPRs of the three methods under different δ when $\bar{p}$ is 0.5 or 0.15. The curve of the two-stage method is located well above the rest, indicating a larger power. In addition, these curves climb quickly up to 1 over a short range of δ. In contrast, Fig. 2c demonstrates the same TPRs with δ fixed at 0.04 (upper three lines), and 0.1 (lower three curves). The power is lower when $\bar{p}$ is around 0.5. Figure 2d shows the range of FPRs with different values of $\bar{p}$ and δ. The values all lie within a small range (1×10^{-4}, 20×10^{-4}). This finding is consistent with the theoretical derivations in the Methods section.

Figure 3 displays TPR, FPR, and FDR with respect to various values of M, and different settings of N_1 and N_2. First, it can be seen that the patterns do not change even if M is as large as 500,000. Second, in Fig. 3a and d, the TPR of the two-stage method with dependent samples outperforms both the two-stage method with independent samples and Bonferroni's method. Third, Fig. 3b,c,e,f show that the errors, in terms of FPR and FDR, of the two-stage method with dependent samples are larger than those seen with the other two methods; however, the difference is very small. We conclude that when correct detection is of concern, the test with dependent samples will be better. When reducing the possibility of false detection is the focus, the test with independent samples should be adopted.

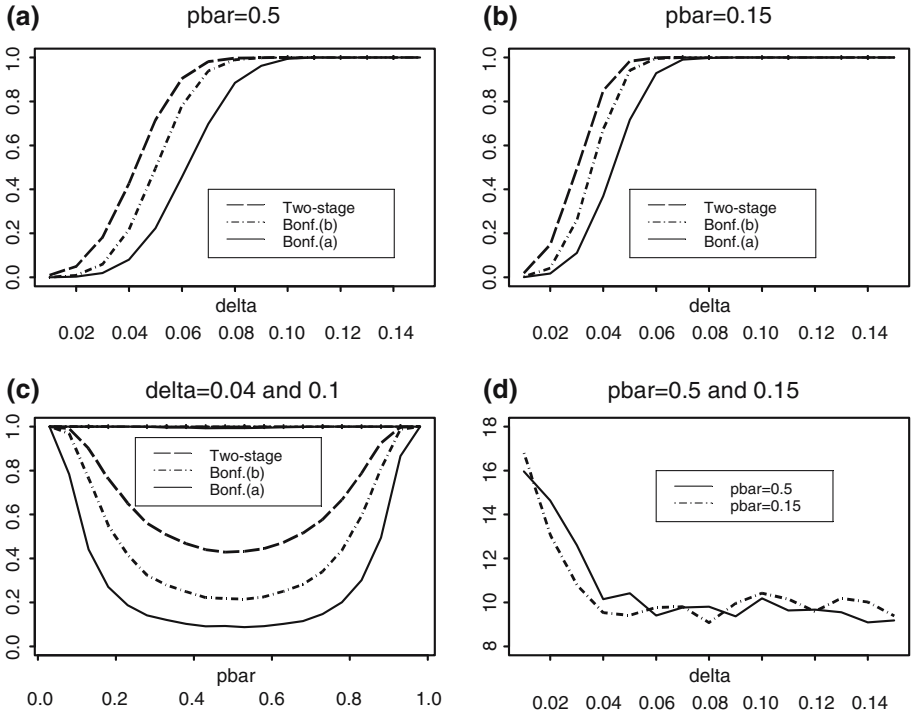

Fig. 2a–d The curves for true positive rates (*TPR*) and false positive rates (*FPR*) under various settings of $\bar{p}$ and δ. **a** TPR with $\bar{p} = 0.5$. **b** TPR with $\bar{p} = 0.15$. **c** TPR with δ fixed at either 0.04 (*upper three lines*) or 0.1 (*lower three lines*). **d** FPR ($\times 10^4$) with $\bar{p}$ fixed at 0.15 or 0.5

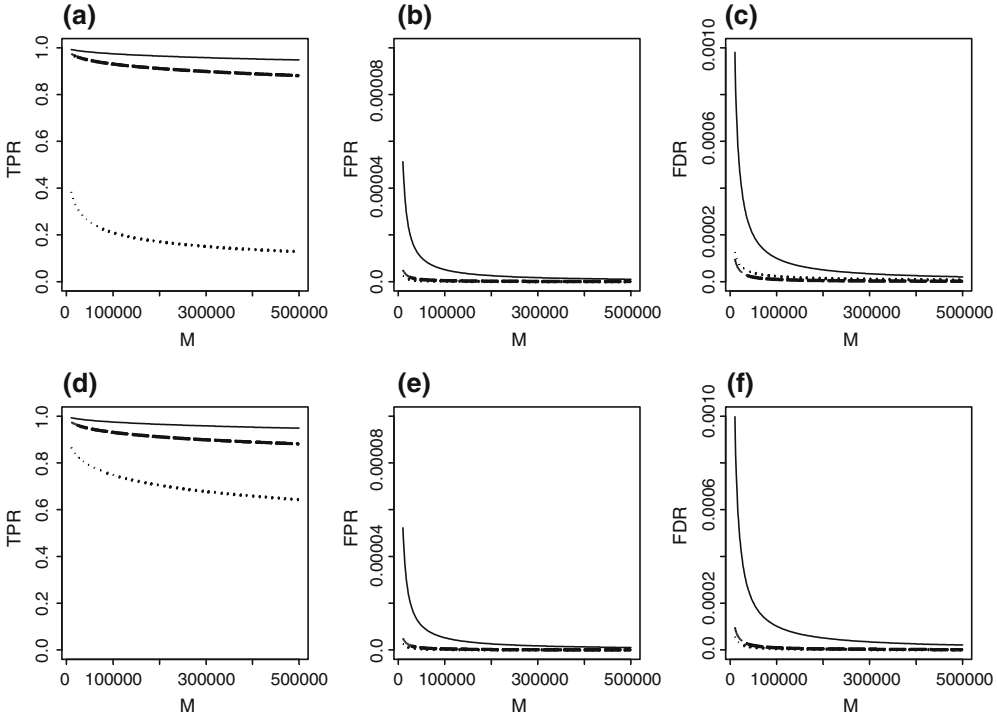

Fig. 3 TPR (**a**, **d**), FPR (**b**, **e**), and false discovery rate (*FDR*) (**c**, **f**) curves with respect to various M. In all figures, $\bar{p} = 0.15$, $w = 0.95$, and $\delta = 0.06$. In **a**–**c** $N_1 = 2{,}000$, $N_2 = 1{,}000$ and in **d**–**f** $N_1 = 1{,}000$, $N_2 = 2{,}000$. *Solid line* Two-stage method with dependent samples, *dashed line* Bonferroni method with all data, *dotted line* two-stage method with independent samples

Cost-effectiveness

When looking at the cost-effectiveness in terms of the number of genotypings and TPR, the two-stage procedure also outperforms. The total numbers of genotypings are MN_1 for $B(N_1)$, $MN_1 + MN_2$ for B(all), and $MN_1 + RN_2$ for the two-stage method, respectively. Taking the simulation for example, the increased cost for the two-stage method, using $B(N_1)$ as the reference, is only

$$\frac{RN_2}{MN_1} \approx \frac{(\hat{R}_0 + \hat{R}_a)N_2}{MN_1} \leq \frac{(50)(1,000)}{(500)(2,000)} = 5\%,$$

while the increase in TPR can be as high as 39% when $\bar{p} = 0.15$ and $\delta = 0.05$. In fact, the increase in TPR is quite dramatic and can be larger than 40%, particularly when $\delta \leq 0.05$. Meanwhile, the magnitude of FPR is still less than 0.001. A similar pattern can be found when comparing B(all) with the two-stage method. The increase in TPR is greater than 30% for $\delta \leq 0.05$ (except when $\delta = 0.05$ and $\bar{p} = 0.15$), and the two-stage procedure saves about 30% in costs compared with B(all).

Discussion

In this paper, we have proposed a two-stage method for multiple hypotheses testing to avoid multiplicity effects. The selection procedure aims at retrieving markers with little or mild association, incorporating two types of errors simultaneously, thus saving on overall costs in genotyping. We also derived the theoretical boundaries of TPR, FPR, and expected counts of true and false significance. When compared with two traditional Bonferroni procedures, the two-stage method outperforms. However, when δ is extremely small, none of the methods tested provides satisfactory results unless a study of a much larger size can be conducted.

Although we have used SNPs as an illustration throughout this paper, the method is not restricted to this type of data. If genotypes are considered, say three genotypes per locus, chi-square tests can be conducted based on 3×2 contingency tables, and the two-stage method can be similarly implemented. When pedigree data are of interest, the unit of observation becomes the vector for each family per locus. The test statistic can be constructed based on multivariate data and the two-stage procedure can be applied. The method can also be used in microarray data provided the number of subjects is not too small. In any case, the scheme does not change with the statistics used.

Through use of the two-stage design, the original power of a test statistic is always increased. This design guarantees a greater overall power than that achievable in a single stage of testing. If dependence among data exists, tests incorporating such relationships may be considered (Benjamini and Yekutieli 2001; Dale 2004). When a multi-stage approach is employed (van den Oord and Sullivan 2003a; Hirschhorn and Daly 2005), this procedure can be generalized with care for the derivation of TPR and FPR. It is worth noting that pooling data from all previous stages will be more powerful, but the tests must be carefully conditioned on results from previous stages to handle dependence, as illustrated here.

Another issue concerns the relative magnitude of sample sizes. Once a statistic is chosen to test the significance between case and control groups, the standard formula for the sample size can be applied based on given α_1 and α_2. The magnitudes of N_1 and N_2 are then determined. Our findings also assure that, since power is the primary interest in the first stage, a larger N_1 will guarantee a greater overall TPR. If one is more concerned about error rates, then a larger N_2 should be considered (van den Oord and Sullivan 2003a). Further studies on an optimal choice of the relative magnitudes of N_1 and N_2, and an investigation into constraints on haplotype size (or the number of SNPs) will be considered in a future manuscript.

Appendix 1

TPR is defined as

$$\begin{aligned}\text{TPR} &\equiv \Pr(\text{reject null}|\text{truly associated})\\ &= \Pr(\text{significant in first and second stage}|\\ &\quad \text{truly associated})\\ &= \Pr(\text{significant in first}|\text{truly associated})\\ &\quad \times \Pr(\text{signigicant in second}|\text{significant in first},\\ &\quad \text{truly associated})\\ &= U_1 \times U_2.\end{aligned}$$

Similarly, by the same argument for conditional probability, the overall FPR is

$$\begin{aligned}\text{FPR} &\equiv \Pr(\text{significant}|\text{no association})\\ &= \Pr(\text{significant in first}|\text{no association})\\ &\quad \times \Pr(\text{significant in second}|\text{significant in first},\\ &\quad \text{no association})\\ &= Q_1 \times Q_2.\end{aligned}$$

Define $T(\mathbf{X})$ as the test statistic used in the first stage where $\mathbf{X}$ denotes the data from MN_1 genotypings, and $T(\mathbf{X}^*)$ for the second stage where $\mathbf{X}^*$ contains the information based on the $R(N_1 + N_2)$ genotypings. The test statistic can be the difference in allele frequency between the case and control groups, or a chi-square statistic derived from contingency tables. These two statistics are correlated because partial data are used in both $\mathbf{X}$ and $\mathbf{X}^*$. Without loss of generality, we assume here that both statistics converge asymptotically to normality, providing N_1 is large, and that the allele frequencies obtained from N_1 and $N_1 + N_2$ are roughly equal. Let ρ denote the correlation between $T(\mathbf{X})$ and $T(\mathbf{X}^*)$, the overall FPR at α_1 and α_2 is then:

$$\begin{aligned}&\Pr(T(\mathbf{X}) \geq a, T(\mathbf{X}^*) \geq b)\\ &\quad = \Pr\left(\Sigma^{-1/2} \cdot \begin{pmatrix} T(\mathbf{X}) \\ T(\mathbf{X}^*) \end{pmatrix} \geq \Sigma^{-1/2} \cdot \begin{pmatrix} a \\ b \end{pmatrix}\right)\\ &\quad = \Pr\left(\Sigma^{-1/2} \cdot \begin{pmatrix} T(\mathbf{X}) \\ T(\mathbf{X}^*) \end{pmatrix} \geq \begin{pmatrix} s \cdot a + t \cdot b \\ t \cdot a + s \cdot b \end{pmatrix}\right),\end{aligned}$$

where $\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\sum^{-1/2} = \begin{pmatrix} s & t \\ t & s \end{pmatrix}$ whose elements depend on ρ. Applying their asymptotic normality with zero mean vector and the covariance matrix $\sum$, the above becomes

$$\Pr\left(\Sigma^{-1/2}\cdot\begin{pmatrix}T(\mathbf{X})\\T(\mathbf{X}^*)\end{pmatrix}\geq\begin{pmatrix}s\cdot a+t\cdot b\\t\cdot a+s\cdot b\end{pmatrix}\right)$$
$$\approx\Pr\left(\begin{pmatrix}Z_1\\Z_2\end{pmatrix}\geq\begin{pmatrix}s\cdot a+t\cdot b\\t\cdot a+s\cdot b\end{pmatrix}\right)$$
$$=\Pr(Z_1\geq s\cdot a+t\cdot b)\cdot\Pr(Z_2\geq t\cdot a+s\cdot b),$$

where $a=z_{1-\alpha_1/2}$ and $b=z_{1-\alpha_2/2}=z_{1-\alpha_1/2R}$ are the percentiles of standard normal, and Z_1 and Z_2 are independent standard normal distributions. This FPR will be bounded closely from above by $\Pr(Z_2\geq t\cdot a+s\cdot b)$ if ρ is large. This boundary is determined by α_1, α_2, R, and ρ, and is smaller than 0.001 when $R\geq 10$; while a quick guess for R would be $M\ \alpha_1$. If $\rho=0$, this boundary becomes $\Pr(Z_2\geq b)$, which is simply α_2.

To investigate the effect of sample size, note that $T(\mathbf{X})$ is close to $\sqrt{N_1/(N_1+N_2)}\times T(\mathbf{X}^*)$ if the average allele frequencies in $\mathbf{X}$ and $\mathbf{X}^*$ are similar, therefore,

$$\mathrm{FPR}=\Pr\left(T(\mathbf{X}^*)\geq z_{1-\alpha_2/2},T(\mathbf{X})\geq z_{1-\alpha_1/2}|\text{no association}\right)$$
$$\cong\Pr\left(T(\mathbf{X}^*)\geq\max\right.$$
$$\left.\times\left(z_{1-\alpha_2/2},\sqrt{\frac{N_1+N_2}{N_1}}\times z_{1-\alpha_1/2}\right)|\text{no association}\right)$$
$$=\min\left(\alpha_2,2\times\left[1-\Phi\left(\sqrt{\frac{N}{N_1}}\times z_{1-\alpha_1/2}\right)\right]\right),$$

where Φ is the cumulative density function of the standard normal. If the proportion $N_1/(N_1+N_2)$ is larger than $(z_{1-\alpha_1/2}/z_{1-\alpha_2/2})^2$, the overall FPR is approximately α_2. Otherwise, it becomes $2\times\left[1-\Phi\left(\sqrt{N/N_1}\times z_{1-\alpha_1/2}\right)\right]$. Similarly, the TPR at fixed α_1 and α_2 is approximately $(1-\beta_2)$.

To derive the expected number of success and error counts, it suffices to calculate the path probability Q_2 in the second stage. Conditional on $\alpha_2=\alpha_1/R$, its upper bound becomes α_2/α_1, and is approximately $1/[E(R_0)+E(R_a)]$. Similarly, the conditional probability U_2, at fixed α_1 and α_2, is approximately $(1-\beta_2)/(1-\beta_1)$.

Appendix 2

Let p_c and p_n represent the allele frequencies for the case and control groups, respectively, and δ their absolute difference. Let $\bar{p}$ be their average, weighted by corresponding sample sizes N_c and N_n. Following the assumption that the allele frequency ranges from 10% to 90%, it can be derived that

$$\begin{cases}90\%\geq\max(p_c,p_n)=\bar{p}+\delta/2\geq 10\%,\\90\%\geq\min(p_c,p_n)=\bar{p}-\delta/2\geq 10\%,\\0.1\leq\bar{p}\leq 0.9,\\-0.8\leq p_c-p_n\leq 0.8,\\0\leq\delta\leq\min(0.8,2\bar{p}-20\%).\end{cases}$$

Therefore, the relationship between $\bar{p}$ and δ can be formulated. In the general form, when p_0 stands for the detectable minimal allele frequency, the above equation becomes

$$\begin{cases}\forall\bar{p}\in(p_0,0.5), & \delta\in(0,2\bar{p}-2\times p_0),\\\forall\bar{p}\in(0.5,1-p_0), & \delta\in(0,2\times(1-p_0)-2\bar{p}).\end{cases}$$

References

Allison DB, Coffey CS (2002) Two-stage testing in microarray analysis: what is gained? J Gerontol A Biol Sci Med Sci 57:B189–B192

Becker T, Knapp M (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27:21–32

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Böddeker IR, Ziegler A (2001) Sequential designs for genetic epidemiological linkage or association studies a review of the literature. Biomed J 43:501–525

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173–15177

Dale RN (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769

Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Stat Sci 18:71–103

Elston RC, Guo X, Williams LV (1996) Two-stage global search designs for linkage analysis using pairs of affected relatives. Genet Epidemiol 13:535–558

Ge Y, Dudoit D, Speed TP (2003) Resampling-based multiple testing for microarray data analysis. Test 12:1–77

Guo X, Elston RC (2000) Two-stage global search designs for linkage analysis. I. Use of the mean statistic for affected sib pairs. Genet Epidemiol 18:97–110

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese millennium genome project: identification of 190,562 genetic variations in the human genome. J Hum Genet 47:605–610

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Benjamin W, Borg A, Trend J (2001) Gene-expression profiles in hereditary breast cancer. N Engl J Med 344:539–548

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J (2000) Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet 64:413–417

Kao JT, Wen HC, Chien KL, Hsu HC, Lin SW (2003) A novel genetic variant in the apolipoprotein A5 gene is associated with hypertriglyceridemia. Hum Mol Genet 12:2533–2539

Miller RA, Galecki A, Shmookler-Reis RJ (2001) Interpretation, design, and analysis of gene array expression experiments. J Gerontol A Biol Sci Med Sci 56:B52–B57

Ohashi J, Tokunaga K (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46:478–482

Ott J, Hoh J (2001) Statistical multilocus methods for disequilibrium analysis in complex traits. Hum Mutat 17:285–288

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33:457–458

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Saito A, Kamatani N (2002) Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. J Hum Genet 47:360–365

Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene–disease association studies. Biometrics 58:163–170

Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene–disease association studies with sample size constraints. Biometrics 60:589–597

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445

Thomas D, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. Genet Epidemiol 27:401–414

Tsai CA, Hsueh H, Chen JJ (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. Biometrics 59:1071–1081

van den Oord EJCG, Sullivan PF (2003a) A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. Hum Hered 56:188–199

van den Oord EJCG, Sullivan PF (2003b) False discoveries and models for gene discovery. Trends Genet 19:537–542