BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 6 2005, pages 781–787
doi:10.1093/bioinformatics/bti053

Data and text mining

An efficient Monte Carlo approach to assessing statistical significance in genomic studies

D. Y. Lin

Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420, USA

Received on July 22, 2004; revised on August 30, 2004; accepted on August 31, 2004
Advance Access publication September 28, 2004

ABSTRACT

Motivation: Multiple hypothesis testing is a common problem in genome research, particularly in microarray experiments and genomewide association studies. Failure to account for the effects of multiple comparisons would result in an abundance of false positive results. The Bonferroni correction and Holm's step-down procedure are overly conservative, whereas the permutation test is time-consuming and is restricted to simple problems.

Results: We developed an efficient Monte Carlo approach to approximating the joint distribution of the test statistics along the genome. We then used the Monte Carlo distribution to evaluate the commonly used criteria for error control, such as familywise error rates and positive false discovery rates. This approach is applicable to any data structures and test statistics. Applications to simulated and real data demonstrate that the proposed approach provides accurate error control, and can be substantially more powerful than the Bonferroni and Holm methods, especially when the test statistics are highly correlated.

Contact: lin@bios.unc.edu

1 INTRODUCTION

In genome research, it is common to examine a large number of features. For example, a microarray experiment involves the expression levels of thousands of genes. One may be interested in detecting genes that show differential expressions across two or more biological conditions or in relating gene expression levels to clinical outcomes. Spurred by the sequencing of the human genome and the advances in molecular technology, there is now a proliferation of genomewide association studies for complex diseases, which involve hundreds or thousands of single nucleotide polymorphisms (SNPs). It is of great interest to determine which SNPs or SNP-based haplotypes are associated with disease phenotypes. In these studies, a large number of hypotheses are tested simultaneously. Even a study with a limited number of candidate genes will involve several hypotheses.

When testing multiple hypotheses, one must guard against an abundance of false positive results. The traditional criterion for error control is the familywise error rate (FWER), which is the probability of rejecting one or more true null hypotheses. The most familiar method for controlling FWER is the Bonferroni correction. It is widely recognized that the Bonferroni method is overly conservative. A more liberal method is the step-down procedure proposed by Holm (1979). However, when the number of hypotheses is large there is little difference between the single-step and step-down procedures. These methods are designed to control FWER for all possible data structures and can be very conservative for the specific data at hand.

Several authors, including Westfall and Young (1993) and Ge *et al*. (2003), suggested the permutation resampling approach. This approach shuffles the phenotype values among the study subjects a number of times so as to create permuted datasets that have only random genotype–phenotype associations. The empirical joint distribution of the test statistics over the permuted datasets then serves as the reference distribution for determining the threshold levels. This approach incorporates the actual data structures into the calculations and thus tends to be less conservative than the aforementioned analytical methods.

The permutation resampling approach has its own limitations. First, this approach is computationally demanding since the analysis needs to be repeated for each permuted dataset. The computation can be prohibitive if the number of hypotheses is large and the calculation of each test statistic is time-consuming. More importantly, this approach requires complete exchangeability under the null hypothesis and thus may not be applicable when there are covariates or nuisance parameters. In particular, the permutation distribution may not be appropriate when the analysis involves covariates (e.g. disease stage) that are correlated with both the genotype and phenotype, as will be demonstrated in the sequel.

An alternative criterion for error control is the false discovery rate (FDR), which is the expected proportion of falsely rejected hypotheses. This error rate is equal to FWER when all null hypotheses are true but is smaller otherwise. Benjamini and Hochberg (1995) proposed a step-down procedure to control FDR for independent test statistics. Benjamini and Yekutieli (2001) showed that the Benjamini–Hochberg procedure controls FDR for certain dependence structures. They proposed a simple, but highly conservative modification to control FDR under arbitrary dependence. Storey (2002) and Storey and Tibshirani (2003) argued that it is more appropriate to consider the positive FDR (pFDR), which is the conditional expectation of the proportion of falsely rejected hypotheses given that at least one hypothesis is rejected. These authors showed how to directly calculate FDR and pFDR for independent test statistics. Storey and Tibshirani (2001) and Ge *et al*. (2003) used the permutation resampling approach to calculate FDR and pFDR for potentially dependent statistics. As mentioned above, the permutation approach has its important limitations.

In this paper, we develop a Monte Carlo procedure to approximate the joint distribution of the test statistics and then use the Monte Carlo distribution to evaluate the error rates, including FWER and

pFDR. Since the Monte Carlo procedure incorporates the actual joint distribution of the test statistics into the calculations, this approach provides an accurate error control. This approach removes the aforementioned drawbacks in the permutation approach. First, it does not involve repeated analyses of simulated datasets and is thus computationally less demanding. Second, it does not require complete exchangeability and is thus widely applicable.

2 METHODS

2.1 Familywise error rates

Suppose that we are interested in testing m hypotheses $H_1, \ldots, H_m$. We denote the corresponding p-values by $p_1, \ldots, p_m$. FWER is the probability of rejecting at least one true hypothesis:

$$\text{FWER} = \Pr(\text{rejecting at least one } H_j, \\ j = j_1, \ldots, j_t | H_{j_1}, \ldots, H_{j_t} \text{ are true}).$$

A simultaneous test procedure is said to control the FWER at α if FWER $\leq \alpha$ regardless of which subset $\{j_1, \ldots, j_t\}$ of hypotheses is true.

The simplest approach is the single-step Bonferroni procedure, which rejects hypothesis H_j if the p-value p_j is less than α/m. Since the probability of rejecting at least one hypothesis is less than the sum of the probabilities of rejecting m hypotheses, the Bonferroni correction is conservative.

Some improvements can be made by employing a step-down procedure. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered p-values, and $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. We first test $H_{(1)}$ using the Bonferroni threshold α/m. Once $H_{(1)}$ is rejected, we should believe that $H_{(1)}$ is false. Then there are only $(m-1)$ hypotheses which may be true, implying the threshold $\alpha/(m-1)$ for $H_{(2)}$. If $H_{(1)}$ and $H_{(2)}$ are rejected, we use $\alpha/(m-2)$ for $H_{(3)}$ and so on. In general, we reject $H_{(j)}$, $j = 1, 2, \ldots,$ if $p_{(j)} \leq \alpha/(m-j+1)$ provided that $H_{(1)}, \ldots, H_{(j-1)}$ have been tested and rejected. Holm (1979) proved that this sequential rejective algorithm indeed controls the FWER at α. Since it is based on the Bonferroni probability inequality, this step-down procedure remains conservative, and is in fact nearly as conservative as the single-step Bonferroni procedure when m is large.

The overall probability of rejection depends on the joint distribution of the test statistics. In the extreme case where the m test statistics are perfectly correlated, no adjustment should be made for multiple testing. Thus, the aforementioned analytical methods, which makes no use of the joint distribution of the test statistics, are inevitably inaccurate, and can be very conservative when the test statistics are highly correlated. We describe below a Monte Carlo approach that provides an accurate control of FWER by incorporating the actual joint distribution of the test statistics into the calculations.

As shown in the Appendix section, all the commonly used statistics can be written in the following form or can be approximated by the statistics of the following form: for $j = 1, \ldots, m$,

$$T_j = U_j^T V_j^{-1} U_j, \tag{1}$$

where

$$U_j = \sum_{i=1}^{n} U_{ji},$$

n is the sample size, U_{ji} involves only the data from the i-th subject and

$$V_j = \sum_{i=1}^{n} U_{ji} U_{ji}^T.$$

When hypothesis H_j holds, U_j is approximately normal with mean zero and covariance matrix V_j in large samples, so that T_j has approximately a χ^2 distribution with r_j degrees of freedom, where r_j is the dimension of U_j. In general, the U_j are correlated, and so are the T_j. Suppose that $H_{j_1}, \ldots, H_{j_t}$ are the true hypotheses. Then for large samples, $(U_{j_1}, \ldots, U_{j_t})$ is approximately multivariate normal with mean zero and with covariance matrix

$$V_{jk} = \sum_{i=1}^{n} U_{ji} U_{ki}^T$$

between U_j and U_k, $j, k = j_1, \ldots, j_t$.

We define

$$\widetilde{U}_j = \sum_{i=1}^{n} U_{ji} G_i,$$

where $G_1, \ldots, G_n$ are independent standard normal random variables that are independent of the data. Also, define

$$\widetilde{T}_j = \widetilde{U}_j^T V_j^{-1} \widetilde{U}_j. \tag{2}$$

Conditional on the data, each $\widetilde{U}_j$ is a weighted sum of independent standard normal random variables, so that $(\widetilde{U}_{j_1}, \ldots, \widetilde{U}_{j_t})$ is a multivariate normal with mean zero and with covariance matrix V_{jk} between $\widetilde{U}_j$ and $\widetilde{U}_k$, $j, k = j_1, \ldots, j_t$. It follows that the conditional joint distribution of $(\widetilde{T}_{j_1}, \ldots, \widetilde{T}_{j_t})$ given the data is approximately the same as the unconditional joint distribution of $(T_{j_1}, \ldots, T_{j_t})$. Thus, we can use the former distribution to approximate the latter distribution. We obtain realizations from the distribution of $(\widetilde{T}_{j_1}, \ldots, \widetilde{T}_{j_t})$ by repeatedly generating the normal random samples $G_1, \ldots, G_n$ while holding the data at their observed values. In calculating the T_j and $\widetilde{T}_j$, we replace the unknown parameters in the U_{ji} with their sample estimators.

Let $t_{(1)}, \ldots, t_{(m)}$ be the observed values of the test statistics associated with $H_{(1)}, \ldots, H_{(m)}$. Our step-down procedure works as follows: starting with hypothesis $H_{(1)}$, we reject $H_{(j)}$, $j = 1, 2, \ldots,$ if

$$\Pr\left(\max_{j \leq k \leq m} \widetilde{T}_k \geq t_{(j)}\right) \leq \alpha,$$

provided that $H_{(1)}, \ldots, H_{(j-1)}$ have been tested and rejected. The probability calculations are based on a large number, e.g. 10 000, realizations of the $\widetilde{T}_j$. When m is small (<8), the numerical integration can be used instead. By the closure principle (Marcus *et al.*, 1976), the FWER of this procedure is approximately α in large samples.

The p-value is the level of the test at which the null hypothesis would just be rejected. Extending this concept to the multiple testing situation leads to the definition of adjusted p-values. The adjusted p-value for hypothesis H_j pertains to the smallest significance level at which H_j would be rejected by the multiple testing procedure (Westfall and Young, 1993, p. 11). Specifically, the FWER adjusted p-value for hypothesis H_j is

$$\widetilde{p}_j = \min\{\alpha\colon H_j \text{ is rejected at FWER} = \alpha\}.$$

We propose to estimate this probability by $\Pr(\max_{j \leq k \leq m} \widetilde{T}_k \geq t_{(j)})$, which is again obtained using our Monte Carlo method. By contrast Holm's (1979) adjusted p-value for H_j is $\min\{(m-j+1)p_j, 1\}$. The adjusted p-values are constrained to be monotone increasing.

Unlike permutation and other resampling methods, the proposed Monte Carlo procedure involves the simulation of normal random variables rather than the genotype or phenotype data and does not require repeated analyses of simulated datasets. The quantities involving the observed data, i.e. the U_{ji} and V_j, are calculated only once, and the evaluation of the $\widetilde{T}_j$ given these quantities is trivial. Thus, the proposed approach is much less time-consuming than permutation and other resampling methods. Significantly, this approach does not involve shuffling of data and can thus be applied to any data structures and test statistics.

2.2 False discovery rates

We reproduced the Table 1 of Benjamini and Hochberg (1995) below. It is natural to define the FDR by $E(R_0/R)$, the expected proportion of falsely rejected hypotheses among all rejected hypotheses. Different ways of

Table 1. Frequency distribution for the hypotheses

	Not rejected	Rejected	Total
True hypotheses	W_0	R_0	m_0
False hypotheses	W_1	R_1	m_1
Total	W	R	m

handling the case of $R = 0$ result in different definitions. Setting $R_0/R = 0$ when $R = 0$ yields the definition of Benjamini and Hochberg (1995):

$$\text{FDR} = E\left\{\frac{R_0}{R} I(R > 0)\right\},$$

where $I(\mathcal{A})$ indicates, by the values 1 versus 0, whether the event $\mathcal{A}$ occurs or not. The pFDR (Storey, 2002) is defined as the conditional expectation of the proportion of falsely rejected hypotheses among all rejected ones given that at least one hypothesis is rejected

$$\text{pFDR} = E\left\{\frac{R_0}{R}\middle| R > 0\right\}.$$

Clearly, $\text{pFDR} = \text{FDR}/\Pr(R > 0)$, so that the two measures will be similar if $\Pr(R > 0)$ is close to 1. When m is small or dependence exists, $\Pr(R > 0)$ can be less than one, resulting in different values of FDR and pFDR.

Benjamini and Hochberg (1995) defined the following Bonferroni-type multiple test procedure: if k is the largest j for which $p_{(j)} \leq (j/m)q^*$, then reject all $H_{(j)}$, $j = 1, \ldots, k$. This procedure controls the FDR at q^* if the test statistics are independent or have the so-called positive regression dependence (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Benjamini and Yekutieli (2001) developed a highly conservative procedure to control the FDR for arbitrarily dependent statistics.

Storey (2002) and Storey and Tibshirani (2001, 2003) proposed a direct approach to evaluate FDRs. Suppose that we reject those hypotheses whose p-values are less than p, then

$$R = \sum_{j=1}^{m} I(p_j \leq p).$$

Let π_0 be the proportion of true hypotheses, i.e. $\pi_0 = m_0/m$. Storey and Tibshirani (2001) derived the following formula:

$$E\left(\frac{R_0}{R}\right) \approx \pi_0 \frac{pm}{R}. \tag{3}$$

They also suggested to estimate π_0 from the observed data. Let p_0 be a number between 0 and 1 such that the p-values greater than p_0 correspond mostly to the null hypotheses. A conservative estimator of π_0 is

$$\widehat{\pi}_0 = \frac{W(p_0)}{(1 - p_0)m},$$

where $W(p_0)$ is the number of hypotheses not rejected at level p_0, i.e.

$$W(p_0) = \sum_{j=1}^{m} I(p_j > p_0).$$

It then follows from formula (3) that FDR and pFDR can be estimated conservatively by

$$\widehat{\text{FDR}}_{p_0}(p) = \frac{W(p_0)p}{(1 - p_0)\max(R, 1)}, \tag{4}$$

$$\widehat{\text{pFDR}}_{p_0}(p) = \frac{\widehat{\text{FDR}}_{p_0}(p)}{\Pr(R > 0)}. \tag{5}$$

If the test statistics are independent, then $\Pr(R > 0) = 1 - (1 - p)^m$. For potentially dependent test statistics, Storey and Tibshirani (2001) and Ge *et al.* (2003) suggested to estimate this probability by permutation resampling. As mentioned previously, permutation resampling has important limitations. We recommend to estimate this probability by our Monte Carlo approach. Specifically, we generate a large number of replicates of $\widetilde{T}_1, \ldots, \widetilde{T}_m$. The proportion of the replicates in which there is at least one $\widetilde{T}_j$ whose p-value is less than p provides an estimator of the desired probability.

The FDR adjusted p-value for hypothesis H_j is defined as:

$$\widetilde{p}_j^* = \min\{q^*\colon H_j \text{ is rejected at FDR} = q^*\},$$

which is estimated by $\min_{p \geq p_j} \widehat{\text{FDR}}_{p_0}(p)$. For pFDR, we estimate the analogous q-value (Storey, 2002) by $\min_{p \geq p_j} \widehat{\text{pFDR}}_{p_0}(p)$, which is again obtained using our Monte Carlo procedure.

3 RESULTS

3.1 Simulated microarray data

We simulated data from the following linear models with random effects:

$$Y_{ij} = \beta_0 + \beta_j X_i + \xi_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

where Y_{ij} represents the expression level of the j-th gene on the i-th subject, X_i indicates whether the i-th subject belong to group 1 (e.g. cancer patients) or group 0 (normal subjects), β_j is the group difference for the j-th gene, ξ_i is the random effect for the i-th subject and the ϵ_{ij} are the random errors. We let the ϵ_{ij} be independent zero-mean normal with variance σ_ϵ^2, and the ξ_i be independent zero-mean normal with variance σ_ξ^2, so that the correlation between any two expression levels of the same subject is $\sigma_\xi^2/(\sigma_\xi^2 + \sigma_\epsilon^2)$. The null hypotheses correspond to $H_j\colon \beta_j = 0$, $j = 1, \ldots, m$. We tested each hypothesis by the two-sample t-statistic.

For the results shown in Figure 1, we set $\sigma_\xi^2 + \sigma_\epsilon^2 = 1$, so that σ_ξ^2 becomes the intra-class correlation. In addition, we let $n = 100$ with $n/2$ subjects in each of the two groups, $m = 2000$, $\beta_0 = 0$ and

$$\beta_j = \begin{cases} 0 & \text{for } j = 1, \ldots, 1800; \\ 0.6(j - 1800)/200 & \text{for } j = 1801, \ldots, 2000. \end{cases}$$

Thus, 200 out of 2000 genes are differentially expressed, the differences ranging from 0.003 to 0.6 at the 0.003 increment. We set the nominal or target familywise type I error at $\alpha = 0.10$. The size pertains to the actual probability of rejecting at least one true hypothesis, and the power pertains to the actual probability of rejecting at least one false hypothesis. These probabilities were estimated from 10 000 simulated datasets. For each dataset, the proposed Monte Carlo method was based on 10 000 normal samples.

These results show that the proposed method has proper control of FWER, whereas the Holm method is conservative and thus less powerful, especially when the correlation is high. For the correlation of 0.5, the Holm method has a power of 50% whereas the proposed method has a power of 75%.

The permutation method is applicable to this two-sample problem. Figure 2 compares the proposed and permutation methods in the quantiles of the estimated null distribution of the supremum statistic $\max_{1 \leq j \leq m} T_j$ for the first dataset generated under $\sigma_\xi^2 = 0.5$. The two distributions agree well except at the extreme tails. The 90 and 95% quantiles are 12.8 and 14.6 under the proposed method, and are 12.7 and 14.4 under the permutation method. Thus, the proposed and permutation methods would have very similar power in this setting.

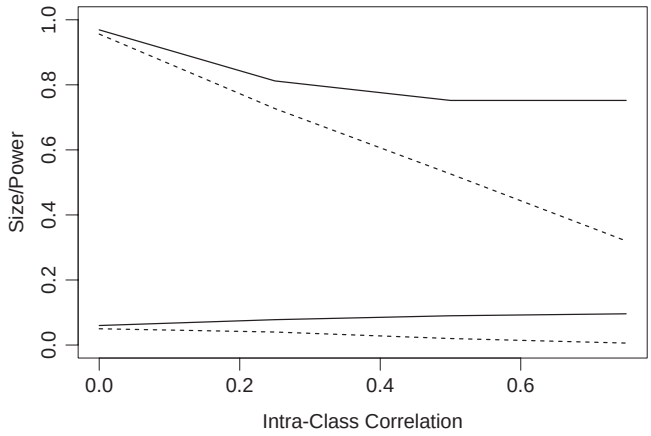

Fig. 1. Empirical size/power of multiple testing procedures at the target FWER of 0.10 for the simulated microarray experiments: the lower and upper solid curves pertain to the size and power of the proposed method; the lower and upper dashed curves pertain to the size and power of the Holm method.

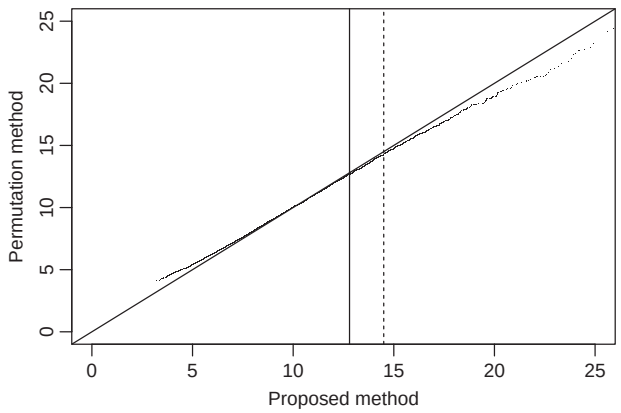

Fig. 2. The quantile–quantile plot of the estimated null distributions based on the proposed and permutation methods for a simulated microarray experiment: the vertical solid and dashed lines pertain to the 90 and 95% quantiles of the proposed method, respectively.

3.2 Simulated SNPs data

We considered a genomewide association study that scans 50 genome regions with 20 biallelic SNPs in each region. We assumed Hardy–Weinberg equilibrium and set the minor allele frequency for each SNP to be 0.3. There is a linkage equilibrium among the regions, and a linkage disequilibrium within each region. We assumed that there are two disease-predisposing SNPs located in the last two regions, which have dominant genetic effects and gene–gene interactions. Specifically, we generated disease incidences from the following logistic model:

$$\Pr(Y_i = 1) = \frac{\exp(-3 + X_{i,970} + X_{i,990} + X_{i,970} * X_{i,990})}{1 + \exp(-3 + X_{i,970} + X_{i,990} + X_{i,970} * X_{i,990})}, \quad i = 1, \ldots, n,$$

where Y_i indicates with the values 1 versus 0 whether the i-th subject is diseased or disease-free, $X_{i,j}$ takes the value 1 if the i-th subject has one or two minor alleles of the j-th SNP and the value 0 otherwise. The overall disease rate is ∼20%. We use the Pearson χ^2-statistics to test the null hypotheses that the SNPs are unrelated to the disease under the dominant genetic model. We let $\alpha = 0.10$.

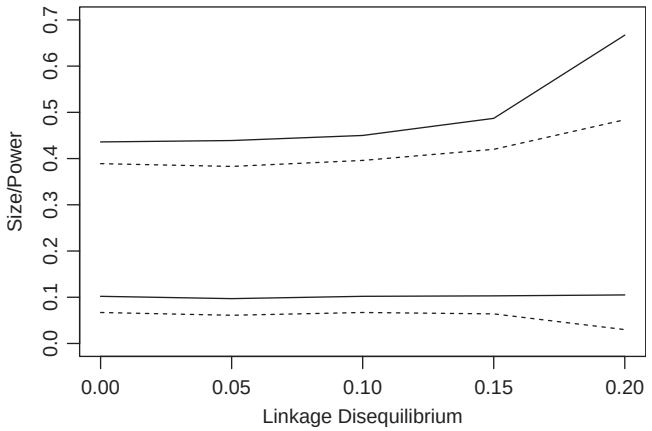

Fig. 3. Empirical size/power of multiple testing procedures at the target FWER of 0.10 for the simulated SNPs data: the lower and upper solid curves correspond to the size and power of the proposed method; the lower and upper dashed curves correspond to the size and power of the Holm method. The horizontal axis pertains to the linkage disequilibrium coefficient (Weir, 1996, p. 113) between two successive loci.

The results are shown in Figure 3. The size pertains to the actual probability of declaring a disease-predisposing SNP in any of the first 48 regions, and the power pertains to the actual probability of identifying any SNP in the last two regions. These probabilities were estimated from 10 000 simulated datasets, each with 100 subjects. For each dataset, the proposed Monte Carlo method was based on 10 000 normal samples.

These results show that the proposed method maintains its FWER near the nominal level and is more powerful than the Holm method, especially under strong linkage disequilibrium. For the pairwise linkage disequilibrium coefficient of 0.2, the power for the Holm method is 0.48 whereas that of the proposed method is 0.67.

3.3 Lung cancer studies

There is a growing interest in relating gene expression levels to survival and other clinical outcomes. Several such studies have been conducted in lung cancer. The objective of the CAMDA (Critical Assessment of Microarray Data Analysis) 2003 Conference was to discuss ways of pooling information across these studies so as to gain new biological insights. A paper by J. S. Morris and co-workers was voted by the attendees and the Scientific Committee as the best presentation in the conference. In their paper, the authors combined the data from the Harvard and Michigan studies (Bhattacharjee *et al*., 2001; Beer *et al*., 2002) and then assessed whether the gene expression levels provide predictive information on survival beyond clinical variables (Morris *et al*., 2004). Here, we apply the proposed method to the same data.

The expression levels for 1036 probesets are available on 200 patients, 124 from the Harvard study and 76 from the Michigan study. Following Morris *et al*. (2004), we fit 1036 multivariable Cox (1972) proportional hazards models with age, stage, institution and the log-expression of each of the 1036 genes as predictors. We obtained the

Table 2. Top 15 genes in the lung cancer studies

Gene identity	Regression coefficient	Test statistic	Unadjusted p-value	FWER p-value		FDR p-value	pFDR q-value	
				Holm	New		Storey	New
ENO2; enolase 2	1.46	18.45	0.00002	0.018	0.014	0.016	0.156	0.247
RRM1; ribonucleotide reductase M1 polypeptide	1.81	15.49	0.00008	0.086	0.060	0.027	0.156	0.247
OST; oligosaccharyltransferase	−1.64	15.13	0.00010	0.103	0.070	0.027	0.156	0.247
DDX3; DEAD/H box polypeptide 3	−2.37	14.72	0.00012	0.129	0.084	0.027	0.156	0.247
FCGRT; Fc fragment of IgG receptor	−2.06	14.41	0.00015	0.151	0.097	0.027	0.156	0.247
Similar to phosphoglycerate mutase 1	1.92	13.76	0.00021	0.214	0.128	0.032	0.156	0.247
CPE; carboxypeptidase E	0.72	11.95	0.00055	0.563	0.268	0.072	0.156	0.253
TBCE; tubulin-specific chaperone e	−2.35	11.50	0.00070	0.716	0.315	0.080	0.156	0.253
STK25; serine/threonine kinase 25	2.29	10.05	0.00152	1.000	0.506	0.142	0.178	0.271
ATIC; IMP cyclohydrolase	1.80	10.03	0.00154	1.000	0.509	0.142	0.178	0.271
TPS1; tryptase, alpha	−0.64	9.40	0.00217	1.000	0.602	0.173	0.188	0.271
CLU; clusterin	−0.52	9.23	0.00238	1.000	0.628	0.173	0.188	0.271
FSCN1; fascin homolog 1, actin-bundling protein	0.66	9.18	0.00244	1.000	0.635	0.173	0.188	0.271
BZW1; basic leucine zipper and W2 domains 1	1.33	8.89	0.00286	1.000	0.678	0.188	0.198	0.273
PFN2; profilin 2	0.63	8.50	0.00355	1.000	0.736	0.218	0.223	0.273

The regression coefficient pertains to the log hazard ratio. A negative coefficient indicates that survival is improved with a larger expression level of the gene. The test statistic is based on the likelihood ratio. For the FWER adjusted p-values and the pFDR q-values, the results under the new method are based on the simulation of 100 000 normal samples.

Table 3. Numbers of genes significantly related to survival according to the Holm and proposed methods of controlling FWER

Method	FWER											
	0.05	0.1	0.15	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Holm	1	2	4	5	6	6	6	7	7	8	8	8
Proposed	1	5	6	6	7	8	8	10	14	18	27	1036

p-value for each gene by the likelihood ratio statistic. The results for the 15 most significant genes are shown in Table 2. Morris *et al*. (2004) provided a good description on these genes.

Morris *et al*. (2004) obtained somewhat different p-values by randomly permuting the gene expression values across the subjects while keeping the clinical variables fixed. This strategy is likely to inflate the type I error because the gene expression levels and stage are correlated. It would be even more problematic to permute the data without fixing the clinical variables because the clinical variables are related to survival. Another potential problem is that censoring may be related to clinical variables and possibly to gene expression levels. This example amplifies the point made earlier that it may not be possible to obtain a suitable permutation distribution when the analysis involves covariates or nuisance parameters.

The results for the FWER analysis are summarized in Tables 2 and 3. The adjusted p-values are considerably smaller under the proposed method than under the Holm method. One would declare more significant genes using the proposed method than by using the Holm method. At the target FWER of 10%, for instance, the proposed method would identify five genes, whereas the Holm method would only identify two.

Figure 4 shows the distribution of the 1036 p-values. The histogram looks fairly flat for p-values greater than 0.4, which indicates that there are mostly null p-values in this region. The estimates of π_0 are ∼0.9 based on $p_0 > 0.4$. Using the estimate of 0.9, we obtained

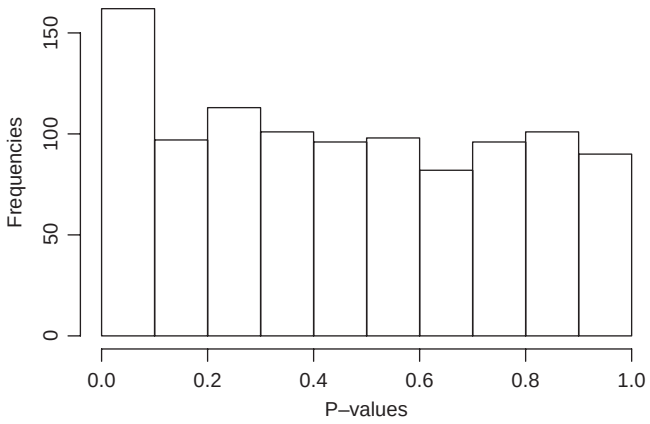

Fig. 4. Density histogram of the 1036 p-values from the lung cancer data.

the estimated FDR and pFDR shown in Figure 5. The corresponding FDR adjusted p-values and pFDR q-values are presented in Table 2. When R is small, the proposed method, which accounts for the dependence of the test statistics, provides considerably smaller estimates of $\Pr(R > 0)$ than what would be expected under independence, and thus yields appreciably higher estimates of pFDR. When R is large, $\Pr(R > 0)$ is close to 1, so that the estimates under dependence

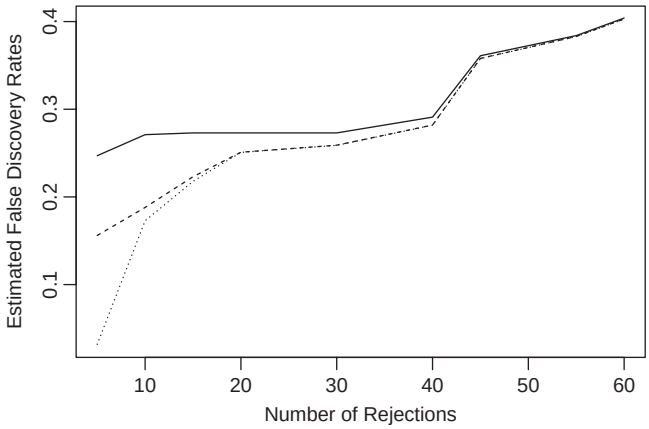

Fig. 5. Estimates of FDR and pFDR for the lung cancer studies: the dotted curve pertains to FDR, the dashed curves to pFDR based on Storey's method, and the solid curve to pFDR based on the proposed method.

and under independence are similar to each other and to the estimated FDR.

As shown in Table 2, the estimated FDR adjusted p-values are smaller than their FWER counterparts, although the proposed method yields a slightly smaller value for the first gene. One would declare 8 significant genes at the FDR of 0.1 and 14 significant genes at the FDR of 0.2. Using Storey's (2002) method, one would also identify the 14 genes at the pFDR of 0.2, but none at the pFDR of 0.15 or less. If the dependence of the test statistics is taken into account, then no gene would be declared significant at the pFDR of 0.2 or less, although six would be at the pFDR of 0.25.

4 DISCUSSION

We have developed an efficient Monte Carlo approach to evaluate error rates for arbitrary test statistics in genome studies. This approach is computationally less demanding than the permutation and other resampling methods and is applicable to more general data structures. Our approach requires a reasonably large sample size. We do not consider this as a serious limitation because properly powered association studies will enroll at least several hundred subjects and even the microarray experiments that are conducted nowadays tend to involve more than 100 subjects. If the sample size is indeed small, then it may be more appropriate to use the permutation test.

Our approach provides accurate control of FWER. It is difficult to accurately control FDR and pFDR for two reasons. First, there are sampling variations associated with the estimators of these error rates. (The Monte Carlo error can be made negligible by using a large number of replicates.) Second, formula (3) tends to be too conservative for dependent test statistics. Unfortunately, there does not exist a better approximation. When the vast majority of the null hypotheses are true, as would be the case in association studies, we recommend the use of FWER. It is particularly desirable to use FWER for candidate genes and other confirmatory studies. For studies involving a large number of false null hypotheses, it may be more appealing to use FDR and pFDR.

In association studies, it is useful to assess the effects of SNP-based haplotypes on disease phenotypes. When there are a large number of SNPs, one possible approach is to use the moving windows of 5–10 SNPs and test for the haplotype-disease association in each window. Since all but one SNPs are common between two adjacent windows, the test statistics tend to be highly correlated. In such situations, it would be wise to use the proposed approach rather than the Bonferroni-type correction since the latter would be extremely conservative.

The asymptotic theory presented in the Appendix section assumes that m is fixed and $n \to \infty$. Such an asymptotic theory may not work well when $m \gg n$. Our simulation studies show that the proposed asymptotic approximations yield proper control of FWER for commonly used statistics when $n > 100$ and m is a few hundreds to a few thousands. Further theoretical and numerical investigations are warranted.

We have focused on two-sided tests so far. It is trivial to modify the formulas to handle one-sided tests for scalar statistics (i.e. $r_j = 1$ for all j). If the U_js are multidimensional, then formulas (1) and (2) will need to changed considerably. Since this is not a common situation, we omit the details here.

It is customary to conduct genomewide linkage analysis, in which a large number of genetic markers are measured and in which possible genetic linkage is tested at all possible positions along the genome. The proposed approach can be applied to this setting, although 'subject' now corresponds to 'family' and the test statistics are typically one-sided (Lin and Zou, 2004).

Zaykin *et al*. (2002) developed the truncated product method that combines evidence from all the tests whose significance exceeds certain threshold. Dudbridge and Koeleman (2003) considered a complementary strategy by forming the product of the K most significant p-values and demonstrated its advantages in genomewide association scans. They suggested to use the permutation test to adjust for the dependence of the test statistics. We can use the proposed Monte Carlo approach to combine evidence from the correlated test statistics in an accurate and flexible manner.

ACKNOWLEDGEMENTS

The author is grateful to Dr Jeffery S. Morris and his colleagues for sharing their version of the CAMDA 2003 data, and to the reviwers for their helpful comments. This research was supported by the National Institutes of Health.

REFERENCES

Beer,D.G., Kardia,S.L., Huang,C.C. Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

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

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

Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Bickel,P.J., Klassen,C.A.J., Ritov,Y. and Wellner,J.A. (1993) *Efficient and Adaptive Estimation in Semiparametric Models.* Johns Hopkins University Press, Baltimore, MD.

Cox,D.R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. Ser.* B, **34**, 187–220.

Dudbridge,F. and Koeleman,P.C. (2003) Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360–366.

Ge,Y., Dudoit,S. and Speed,T.P. (2003) Resampling-based multiple testing for microarray data analysis (with discussion). *Test*, **12**, 1–77.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Lin,D.Y. and Wei,L.J. (1989) The robust inference for the Cox proportional hazards model. *J. Am. Stat. Assoc.*, **84**, 1074–1078.

Lin,D.Y. and Zou,F. (2004) Assessing genomewide statistical significance in linkage studies. *Genet. Epidemiol.* **27**, 202–214.

Marcus,R., Peritz,E. and Gabriel,K.R. (1976) On closed testing procedure with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Morris,J.S., Yin,G., Baggerly,K.A., Wu,C. and Zhang, L. (2004) Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer. In Shoemaker, J.S. and Lin, S.M. (eds) *Methods of Microarray Data Analysis III*. Springer, New York, pp. 51–66.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. Ser.* B, **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report 2001–28*, Department of Statistics, Stanford University, CA.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Wei,L.J. and Lachin,J.M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Am. Stat. Assoc.*, **79**, 653–661.

Weir,B.S. (1996) *Genetic Data Analysis II*. Sinauer Associates, Inc., Publishers, Sunderland, MA.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, NY.

Zaykin,D.V., Zhivotovsky,L.A., Westfall,P.H. and Weir,B.S. (2002) Truncated product method for combining *P*-values. *Genet. Epidemiol.*, **22**, 170–185.

APPENDIX: SOME THEORETICAL DETAILS

All the commonly used statistics are related to the score statistics under parametric or semiparametric regression models. For example, the two-sample t-statistic and the Pearson χ^2 statistic used in the simulation studies correspond to the score statistics under the normal linear model and logistic regression model with a binary predictor, while the likelihood ratio statistic used in the lung cancer studies is asymptotically equivalent to the (partial-likelihood) score statistic for testing one parameter in the presence of other (nuisance) parameters under the semiparametric proportional hazards model (Cox, 1972).

Let U_j be the efficient score function for β_j. In the presence of nuisance parameters, the efficient score function is the projection of the score function for β_j on the orthocomplement of the space of the score functions for the nuisance parameters (Bickel *et al*., 1993, p. 30). For a random sample of n subjects,

$$U_j = \sum_{i=1}^{n} U_{ji}, \tag{A1}$$

where U_{ji} involves the data from the i-th subject only. For parametric models, the expressions for U_{ji} can be found in mathematical statistics texts (Bickel *et al*., 1993, p. 28). For the proportional hazards model, the expressions are given by Lin and Wei (1989). For the Wilcoxon statistics with potentially censored outcomes, the expressions can be found from Wei and Lachin (1984).

We are interested in testing the hypotheses H_j: $\beta_j = \beta_{0j}$, $j = 1, \ldots, m$, where β_{0j} is zero or some other null value. Suppose that hypotheses $H_{j_1}, \ldots, H_{j_t}$ are true. In view of Equation (A1), the multivariate central limit theorem implies that the random vector $n^{-1/2}(U_{j_1}, \ldots, U_{j_t})$ is asymptotically multivariate normal with mean 0 and with the limit of $n^{-1}\sum_i U_{ji}U_{ki}^T$ as the covariance matrix between $n^{-1/2}U_j$ and $n^{-1/2}U_k$. In calculating the test statistics, we evaluate the U_{ji} at $\beta_j = \beta_{0j}$ and replace the unknown parameters with their sample estimators.