Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 28

A Comparison of Methods to Control Type I Errors in Microarray Studies

Jinsong Chen[*] Mark J. van der Laan[†]
Martyn T. Smith[‡] Alan E. Hubbard[**]

[*] Lawrence Berkeley National Laboratory, jchen@lbl.gov
[†] University of California, Berkeley, laan@berkeley.edu
[‡] University of California, Berkeley, martynts@berkeley.edu
[**] University of California, Berkeley, hubbard@stat.berkeley.edu

A Comparison of Methods to Control Type I Errors in Microarray Studies

Jinsong Chen, Mark J. van der Laan, Martyn T. Smith, and Alan E. Hubbard

Abstract

Microarray studies often need to simultaneously examine thousands of genes to determine which are differentially expressed. One main challenge in those studies is to find suitable multiple testing procedures that provide accurate control of the error rates of interest and meanwhile are most powerful, that is, they return the longest list of truly interesting genes among competitors. Many multiple testing methods have been developed recently for microarray data analysis, especially resampling based methods, such as permutation methods, the null-centered and scaled bootstrap (NCSB) method, and the quantile-transformed-bootstrap-distribution (QTBD) method. Each of these methods has its own merits and limitations. Theoretically permutation methods can fail to provide accurate control of Type I errors when the so-called subset pivotality condition is violated. The NCSB method does not suffer from that limitation, but an impractical number of bootstrap samples are often needed to get proper control of Type I errors. The newly developed QTBD method has the virtues of providing accurate control of Type I errors under few restrictions. However, the relative practical performance of the above three types of multiple testing methods remains unresolved. This paper compares the above three resampling based methods according to the control of family wise error rates (FWER) through data simulations. Results show that among the three resampling based methods, the QTBD method provides relatively accurate and powerful control in more general circumstances.

KEYWORDS: microarrays, experiment-wise error rates, permutation methods, bootstrap, adjusted p-values

1 Introduction

Microarray data analysis typically starts from testing thousands of hypotheses in order to identify differentially expressed genes among biological samples that differ by some trait or treatment. As microarray technologies have become more popular, many methods have been developed to evaluate and to find patterns from microarray data sets, such as algorithms referred to as multiple testing procedures (MTP), which first evaluate the association between gene expressions and variables of interest on a gene by gene basis and then derive larger experiment-wise inference over the entire set of genes or tests. To account for dependence in expression among different genes, resampling based methods have been used, such as permutation methods (Westfall and Young (1993)), the null-centered and scaled bootstrap (NCSB) method (Pollard and van der Laan (2003), Dudoit et al. (2004)) and the quantile-transformed-bootstrap-distribution (QTBD) method (van der Laan and Hubbard (2006)). Theoretically, permutation methods will fail asymptotically when certain conditions, which do not affect the asymptotic validity of the bootstrap-based methods, are unsatisfied. However, for a finite number of resamples, the NCSB method often has poor performance due to the difficulty of using the empirical distribution to accurately estimate the far right tail of a sampling distribution.

This paper offers some practical guidance for researchers to choose suitable multiple testing methods for microarray data analysis and to provide some evidence of the superiority of particular methods under particular scenarios through data simulations. The simulation studies are based on two-sample problems, which are common in microarray data analysis. Although our conclusions are obtained according to the study of two-sample problems, they can be applied more broadly to MTP issues in microarray data analysis.

The paper is organized as follows. Section 2 briefly describes two-sample problems and three resampling based methods. Section 3 shows a series of simulation studies used to evaluate the performance of these estimators. Section 4 describes an application of the competing methods to microarray data, where the goal was finding genetic markers of exposure to the cancer-causing chemical benzene in an occupational setting in China. A short discussion section is provided in section 5.

2 Methods

2.1 Errors in Multiple Testing

In general, there are two relevant performance measures (types of errors) considered by multiple testing procedures. The first is the rate of false positives or Type I errors, caused by rejecting true null hypotheses; the second is the rate of false negatives or Type II errors caused by accepting false null hypotheses. Let H_{0j} be the null hypothesis for the $j-th$ gene, where $j = 1, 2, \cdots, p$ and p is the total number of genes under study. Let $R_n = \{j : H_{0j} \text{ is rejected}, j = 1, 2, \cdots, p\}$, $H_0 = \{j : H_{0j} \text{ is true}, j = 1, 2, \cdots, p\}$, $H_1 = \{j : H_{0j} \text{ is false}, j = 1, 2, \cdots, p\}$. Thus the number of Type I errors is given by $V_n = |R_n \cap H_0|$, and the number of Type II errors is given by $U = |R_n^c \cap H_1|$. Both V_n and U_n are random variables that depend on an unknown data-generating distribution. There are many different ways to quantify the Type I and Type II errors. This study focuses on the family-wise error rate (FWER) to measure the Type I errors, which are defined as the probability that at least one Type I error occurs, i.e. FWER $= Pr(V_n \geq 1)$. We use the average power to measure the Type II error, which is defined as the expected proportion of accepted false null hypothesis and is given by $E[|R_n \cap H_1|]/|H_1|$. Though there are other error rates, such as the less conservative false discovery rate, we concentrate on the FWER, for which re-sampling based methods have been more fully developed.

2.2 Two-sample Problems

Let $X_{ki}(j)$ be the expression value of the ith array (e.g., biological replicate) within group k for the jth gene. Suppose there are p genes (typically in the tens of thousands) and $n = n_1 + n_2$ arrays; thus, for gene j, the n_1 gene expressions are $X_{11}(j), X_{12}(j), \cdots, X_{1n_1}(j)$ from Population 1 and n_2 gene expressions $X_{21}(j), X_{22}(j), \cdots, X_{2n_2}(j)$ from Population 2. Let $\mu_1(j)$ and $\mu_2(j)$ denote the means of vectors of random variable $X_{k\cdot}(j) = (X_{k1}(j), X_{k2}(j), \ldots, X_{kn_k}(j))$ in Populations $k = 1$ and $k = 2$, respectively. A set of null hypotheses related to the mean expressions of the two groups is given as follows:

$$H_{0,j} : \mu_1(j) = \mu_2(j), j = 1, 2, \cdots, p. \tag{1}$$

A possible two-sample t-statistic for testing the above set of hypotheses is given by:

$$T_n(j) = \frac{\bar{X}_2(j) - \bar{X}_1(j)}{\sqrt{S_2^2(j)/n_2 + S_1^2(j)/n_1}}, \tag{2}$$

where $\bar{X}_1(j)$ and $\bar{X}_2(j)$ represent the averaged expression levels of gene j in Population 1 and 2, respectively, and $S_1^2(j)$ and $S_2^2(j)$ represent their corresponding sample variances. In the following studies, we will apply four different methods to control FWER for the set of p hypotheses given in Equation 1, which include three resampling based methods and one traditional marginal p-value based method (the Bonferroni procedure).

2.3 Permutation Method

Permutation methods, described in detail by Westfall and Young (1993), are an attempt to control Type I error rates while accounting for the dependence between the p test statistics in order to increase the power of the procedure (i.e. finding more true positives while still properly controlling the rate of false positives). As a contrast, the Bonferroni procedure for controlling the rate of false positives makes the most conservative assumption of independence of those test statistics. The procedure allows the multiple testing procedure to use only the marginal p-values from the test statistics (2), i.e. simply rejecting the null hypothesis if the p-value is less than α/p, where α is the desired Type I error rate. Such conservativeness of the Bonferroni method comes from two sources. First, it assumes that all the null hypotheses are true(step-down methods attempt to improve the power of such procedures by allowing for some false null hypotheses). Second, it ignores the potential correlation among test statistics calculated from microarray data.

The limitation of the independence assumption in the Bonferroni procedure can be seen intuitively from the following extreme case. Suppose, for every repeated random sample of the data from the target population, all the t-statistics have precisely equal p-values. In essence, there is only one effective test statistic, and thus the cut-off should be determined by dividing the desired FWER not by p, the number of tests, but by 1. As the dependence becomes less extreme, the principal still holds. If one could use the dependence structure, then one should be able to derive a multiple testing procedure that provides tighter control under more general situations. Because gene expressions from different genes on the same sample can be strongly correlated (e.g. co-regulated), deriving multiple testing procedures that can take advantage of the joint distribution (statistical dependence) of test statistics becomes very compelling.

Typical multiple testing procedures for microarray data analysis entail two steps. First, genes are ranked by the raw p-values from the smallest to largest. Second, adjusted p-values are reported as opposed to a simple reject/accept indicator for each gene. The adjusted p-values can be thought of as the estimated error rates that results if one rejects the null hypotheses for test statistics that are bigger than the one on that row, hence all the genes above that one on the ordered list. Below is an algorithm for deriving adjusted p-values for permutation MTP methods (Dudoit et al. (2003)).

1. Compute t-statistics using (2) to get $T_n(j)$, where $j = 1, 2, \cdots, p$, and let $b = 1$.

2. Permute the n columns of the data matrix $X = \{X_{ki}(\cdot), k = 1, 2, i = 1, 2, \cdots, n_k\}$, where vector $X_{ki}(\cdot) = (X_{ki}(1), X_{ki}(2), \cdots, X_{ki}(p))^T$, by ignoring the group numbers.

3. Compute test statistics $T_n^b(j), j = 1, 2, \cdots, p$ from the permuted data matrix and let $b = b + 1$.

4. Repeat Steps 2 and 3 if $b < B$ and go to Step 5, otherwise. The above procedure will produce a $p \times B$ test statistics matrix $T_n^{\#} = \{T_n^b(j), j = 1, 2, \cdots, p, b = 1, 2, \cdots, B\}$.

5. Use the approach called maxT to control the FWER. For each column of the $T_n^{\#}$ matrix, calculate the maximum absolute value of the test statistics by $T_{max}^b = \max_j(|T_n^b(j)|)$, $j = 1, 2, \cdots, p$. Calculate adjusted p-values based on the original test statistic $T_n(j)$, $j = 1, 2, \cdots, p$, and the matrix $T_n^{\#}$ using the following formula: $Adjp_{value}(j) \equiv \hat{P}rob(T_{max}^b \geq |T_n(j)|)$, or simply the proportion of maximum t-statistics that are not less than the observed test statistic.

The above method for computing adjusted p-values has limitation because it requires the the so-called subset pivotality condition. By definition, this requires that the true covariance matrix of the test statistics for the true nulls is asymptotically equal to the covariance matrix of the test statistics implied by the chosen null data distribution (in Westfall and Young (1993), the permutation distribution). In the two-sample problem, this is satisfied when either the covariance matrices of the test statistics in the two populations are equal or the sample sizes in the two groups are equal. There are situations, such as all possible pair-wise correlations of genes, where the subset pivotality requirement can not be satisfied and this limitation motivated the development of new multiple testing approaches.

2.4 NCSB Method

A resampling based method, originally proposed in Pollard and van der Laan (2003) and further expanded in Dudoit et al. (2004), suggests using the bootstrap (i.e. randomly resampling arrays with replacement) as a way to create an appropriate joint null distribution of test statistics. In bootstrap methods, one approximates the unknown data-generating distribution by using the empirical distribution obtained from sampling with replacement from the original data. Since these methods preserve the dependence structure of original data without assumptions about the data-generating distribution, they require fewer assumptions than permutation methods. The way to re-draw samples depends on the specific experimental design. For typical microarray data analysis, there are a certain number of arrays, by design, in the control and experimental subgroups. To preserve such designs during the re-sampling procedure, arrays are sampled with replacement within each subgroup. The detailed procedures of obtaining test statistics matrices are given as follows:

1. Let $b = 1$.
2. Randomly draw n_1 columns from $p \times n_1$ matrix $X_1 = \{X_{1i}(\cdot), i = 1, 2, \cdots, n_1\}$ and n_2 columns from $p \times n_2$ matrix $X_2 = \{X_{2i}(\cdot), i = 1, 2, \cdots, n_2\}$ with replacement.
3. Compute test statistics $T_n^b(j), j = 1, 2, \cdots, p$ from the re-sampled data matrix and let $b = b + 1$.
4. Repeat Steps 2 and 3 if $b < B$ and otherwise, stop. The calculated test statistics form a $p \times B$ matrix $T_n^{\#} = \{T_n^b(j), j = 1, 2, \cdots, p, b = 1, 2, \cdots, B\}$.

Similar to permutation methods, the matrix $T_n^{\#}$ is an estimate of the dominating multivariate null distribution of the test statistics. As discussed in Pollard and van der Laan (2003), the empirical correlation of test statistics in the matrix provides an estimate of the dependence among them. However, the test statistics in matrix $T_n^{\#}$ need to be centered by the corresponding null mean value, and can be scaled by its corresponding null variance. Let $\lambda_0(j)$ and $\tau_0(j)$ be the known null mean and variance of the test statistics for gene j. The null-centered and scaled bootstrap null distribution is given by

$$Z_n^b(j) = (T_n^b(j) - E[T_n^b(j)])\sqrt{min\left(1.0, \frac{\tau_0(j)}{Var[T_n^b(j)]}\right)} + \lambda_0(j). \qquad (3)$$

For the two-sample problems, typically the null mean value $\lambda_0(j) = 0$ and the null variance (based on the t-distribution) is $\tau_0(j) = df/(df - 2)$, where df is the degrees of freedom, for all j. As discussed in Dudoit et al. (2004), the re-scaling adjustment to the original test statistics does not affect the asymptotic control of Type I errors but can increase the power. The main advantage of the NCSB method over permutation methods is they do not need restrictions on the joint distribution of the test statistics, such as the subset pivotality restriction. However, a limitation from which the NCSB method suffers relative to permutation methods is that for a large number of tests (i.e. big p), a very large number of bootstrap resamplings is typically is needed to get accurate adjusted p-values, that is, to estimate the distant tails of the maximum null-centered, scaled test statistics.

2.5 QTBD Method

The QTBD method, proposed by van der Laan and Hubbard (2006), is a modification of the original bootstrap procedure by using knowledge about the marginal null distribution (e.g., a t-distribution with df degrees of freedom for two-sample problems). The primary goal of the methods is to remedy practical performance problems of the NCSB method. The QTBD method shares the main idea with the NCSB, by trying to take advantage of the possible dependence among test statistics. However, instead of adjusting the null bootstrap distribution to have the correct marginal null mean and variance as in the NCSB method, the QTBD method transforms the bootstrap distribution to insure that each of the marginal test statistic distributions has the desired dominating null distribution.

Suppose T_n^b is the original bootstrap null distribution matrix and let Q_{0j}, $j = 1, 2, \cdots, p$, be the known marginal null distributions. In some cases, such as in this study, they are the same for all the test statistics, $j = 1, \ldots, p$. The QTBD method is given by:

$$\tilde{T}_n^b(j) = Q_{0j}^{-1}(Q_{nj}^b(T_n^b(j))),\ j = 1, 2, \cdots, p, \tag{4}$$

where $Q_{nj}^b(t)$ is the empirical bootstrap distribution for the jth test and $Q_{0j}^{-1}(x)$ is the inverse probability distribution for the desired marginal null distribution (for instance if $x = 0.5$, then $Q_{0j}^{-1}(x)$ is the median of the null distribution). The benefits of such a modification are that the procedure both takes advantage of the dependence among test statistics (in this case, the dependence is preserved solely through the ranks) and the marginal (row by row) distribution is the optimal chosen null distribution. This latter property should gain power relative to the original bootstrap method and still does not rely on the subset pivotality assumption of the permutation methods.

For implementation we used a slightly modified technique, which appears to work better in practice. This modification maps the bootstrap distribution into the empirical (as opposed to the actual) distribution of samples randomly drawn from the desired marginal null distribution. Specifically, suppose one has B numbers of bootstrap statistics $T_n(j)^b$, $b = 1, 2, \cdots, B$ for row (gene), j. First, randomly generate B samples $q_1, q_2, \cdots, q_B$, from the given marginal distribution function Q_{0j} and sort them in ascending order: $q_{(1)}, q_{(2)}, \cdots, q_{(B)}$. Second, find the rank $r_1, r_2, \cdots, r_B$ of the original bootstrap statistics T_{nj}^b from the smallest to the largest values. The quantile transformed null distribution thus is given by:

$$\tilde{T}_n^b(j) = q(r_b),\ \text{where } b = 1, 2, \cdots, B. \tag{5}$$

Simulation studies have shown that this mapping method is not only computationally fast but it also provides good finite sample error control.

3 Simulation Study

This section compares the practical performance of the three MTP methods described in the proceeding section under various data-generating distributions. Although in theory the bootstrap methods (the NCSB and QTBD methods) should perform superior asymptotically to the permutation method when the assumptions of the permutation method are violated, we want to examine whether the QTBD procedure can be more generally recommended in practice through simulation studies. Specifically, we performed a set of simulations where the numbers of true and false hypotheses are known, and we evaluate both the accuracy and power of each competing MTP method as the subset pivotality condition is met and is violated.

3.1 Synthetic Cases

Suppose gene expressions $X_{1i}(\cdot)$ $(i = 1, 2, \cdots, n_1)$ for Population 1 has the $p-$variate normal distribution with mean vector μ_1 and variance-covariance matrix Σ_1, and gene expressions $X_{2i}(\cdot)$ $(i = 1, 2, \cdots, n_2)$ for Population 2 has the $p-$variate normal distribution with mean vector μ_2 and variance-covariance matrix Σ_2. By varying vectors μ_1 and μ_2 and matrices Σ_1 and Σ_2, we form four different synthetic cases.

3.1.1 Case 1a: Balanced design with complete null hypotheses

Let $p = 100$, $n_1 = n_2 = 15$, $\mu_{1j} = \mu_{2j} = 0.0$, $\sigma^2_{1j} = \sigma^2_{2j} = 1.0$, where $j = 1, 2, \cdots, p$. The pairwise correlation between the expression of genes within Populations 1 and 2 in this simulation are zero, that is the gene expressions within an array are independent. In this case, all the null hypotheses are true and the joint distribution of the test statistics, $T_n(\cdot)$, satisfies the subset pivotality condition.

3.1.2 Case 2a: Unbalanced design with complete null hypotheses

Let $p = 100$, $n_1 = 5$, $n_2 = 25$, $\mu_{1j} = \mu_{2j} = 0.0$, $\sigma^2_{1j} = 0.1$, $\sigma^2_{2j} = 5.0$, where $j = 1, 2, \cdots, p$. We assigned a different correlation parameter (ρ) of gene expressions for genes within Populations 1 (ρ_1) and 2 (ρ_2). Specifically, $\rho_1 = 0.0$ and $\rho_2 = 0.6$ and the variance-covariance matrices are defined by the model $\Sigma_1 = \{a_{ij}\}$ and $\Sigma_2 = \{b_{ij}\}$, where $a_{ij} = \sigma^2_1 \rho_1^{|i-j|}$ and $b_{ij} = \sigma^2_2 \rho_2^{|i-j|}$. In this case, all the null hypotheses are still true but the data-generating distribution results in an extreme violation of the subset pivotality condition.

3.1.3 Case 1b: Balanced design with partial null hypotheses

To compare the power of the three competing MTPs, we modified Case 1a by having false null hypotheses for $j = 1, 2, \cdots, 10$ (i.e. $\mu_{1j} - \mu_{2j} = 2.0$), whereas the null hypotheses are true for $j = 11, 12, \cdots, 100$ (i.e. $\mu_{1j} - \mu_{2j} = 0.0$).

3.1.4 Case 2b: Unbalanced design with partial null hypotheses

Similar to Case 1b, we modified Case 2a to get Case 2b by simulating false null hypotheses for the first 10 rows (i.e. $\mu_{1j} - \mu_{2j} = 2.0$, for $j = 1, 2, \cdots, 10$), and true null hypotheses for all remaining rows.

We conduct 2,000 simulations for each of the above four cases by following the steps given below:

1. Generate synthetic data $X_{1i}(\cdot) \sim N(\mu_1 I_p, \Sigma_1)$ for $i = 1, 2, \cdots, n_1$ and $X_{2i}(\cdot) \sim N(\mu_2 I_p, \Sigma_2)$ for $i = 1, 2, \cdots, n_2$.

2. Obtain null distributions of test statistics using each of the three resampling based methods given in Section 2.

3. Obtain adjusted p-values using the single-step methods for Cases 1a and 2a and the step-down methods (see van der Laan et al. (2004) for details of the step-down procedures for deriving adjusted p-values from the null distribution matrix for any resampling based procedure) for Cases 1b and 2b.

4. Count the number of false rejections and the number of false acceptances for the cutoff values of $\alpha = 0.05$ and $\alpha = 0.10$.

5. Repeat Steps 1–4 2000 times.

3.2 Simulation Results

Table 1 summarizes the simulated results of Case 1a for controlling FWERs when the cutoff values are equal to 0.05 and 0.10, respectively. As shown in the table, permutation methods provide good control of Type I error rates (i.e. the simulated FWERs are close to their corresponding nominal values) because the subset pivotality condition is satisfied, and the simple Bonferroni method also does well for this data-generating distribution as the test statistics are in fact independent. The NCSB method provides relatively poor results for FWER control, even when the number of bootstrap samples increases to $B = 10,000$. However, for QTBD method, the simulation results are as good as the ones obtained from permutation methods. This is because the marginal distribution after quantile transformation of testing statistics is the t-distribution with the degree of freedom of $n - 2 = 28$, which in this case is the true distribution of test statistics.

Table 1: Comparison of FWER using different methods for Case 1a, in which $n_1 = 15$, $n_2 = 15$, $h_1 = 0$, $h_0 = 100$, $p = 100$, $\mu_1 = \mu_2 = 0.0$, $\sigma_1 = \sigma_2 = 1.0$, and $\rho_1 = \rho_2 = 0.0$.

Resampling Methods	B	$\alpha = 0.05$ (FWER)	$\alpha = 0.10$ (FWER)
Permutation			
	1000	0.0515	0.1050
	3000	0.0530	0.1035
	5000	0.0545	0.1085
	10000	0.0515	0.1100
NCSB			
	1000	0.0365	0.0845
	3000	0.0440	0.0835
	5000	0.0255	0.0775
	10000	0.0345	0.0895
QTBD			
	1000	0.0570	0.1125
	3000	0.0580	0.1190
	5000	0.0490	0.1060
	10000	0.0560	0.1160
Bonferroni			
		0.0515	0.0945

Table 2 shows the results of Cases 2a, where the subset pivotality condition is violated because $n_1 \neq n_2$ and $\sigma_1^2 \neq \sigma_2^2$, and the correlation structures within Populations 1 and 2 are very different, as discussed in Pollard and van der Laan (2003). Theoretically, the permutation methods should provide inaccurate control of FWERs. This is verified by the results that suggest overly conservative control of Type I error rates by an order of magnitude. Similar to Case 1a, the NCSB method again has poor performance, but better than the permutation method. The Bonferroni adjustment is overly conservative as well in this case because of the strong dependence among the test statistics. However, the QTBD method still provides accurate control for FWERs, taking advantage of the dependence in a protected way (protected as long as the marginal distribution is correctly specified).

Table 2: Comparison of FWER using different methods for Case 2a, where $n_1 = 5$, $n_2 = 25$, $h_1 = 0$, $h_0 = 100$, $\mu_1 = \mu_2 = 0.0$, $\sigma_1 = 0.1$, $\sigma_2 = 5.0$, $\rho_1 = 0.0$, and $\rho_2 = 0.6$.

Resampling Methods	B	$\alpha = 0.05$ (FWER)	$\alpha = 0.10$ (FWER)
Permutation			
	1000	0.0030	0.0100
	3000	0.0025	0.0105
	5000	0.0025	0.0120
	10000	0.0025	0.0110
NCSB			
	1000	0.0350	0.0880
	3000	0.0300	0.0780
	5000	0.0350	0.0835
	10000	0.0340	0.0875
QTBD			
	1000	0.0635	0.1285
	3000	0.0475	0.0960
	5000	0.0505	0.1095
	10000	0.0455	0.0995
Bonferroni			
		0.0130	0.0245

Table 3 shows the results of Case 1b, where there are false nulls (true positives), i.e., among the 100 genes, the first 10 null hypotheses are false with mean difference of $\Delta = 2.0$. As expected, the relative power of each multiple testing method for finding truly differentially expressed gene is very similar. However, as shown in Table 4, when the subset pivotality condition is violated and test statistics across genes present strong correlations, the average power for the permutation approach is very low (only 26%), compared to 58, 65 and 62% for the NCSB, QTBD and Bonferroni methods, respectively.

Although the simulated results of the above four cases seem not very sensitive to the number of either permutations or bootstrap sampling when the total number of genes are 100, we expect that as p, the number of tests gets larger, the number of bootstrap runs becomes more important because

Table 3: Comparison of FWER using different methods for Case 1b, in which $n_1 = 15$, $n_2 = 15$, $h_1 = 10$, $h_0 = 90$, $p = 100$, $\mu_1 = \mu_2 = 0.0$, $\sigma_1 = \sigma_2 = 1.0$, and $\rho_1 = \rho_2 = 0.0$.

Resampling Methods	B	$\alpha = 0.05$ (FWER)	$\alpha = 0.10$ (FWER)	Average Power
Permutation				
	1000	0.0545	0.1030	0.92
	3000	0.0535	0.1055	0.92
	5000	0.0535	0.1060	0.93
	10000	0.0535	0.1070	0.93
NCSB				
	1000	0.0350	0.0805	0.90
	3000	0.0430	0.0830	0.90
	5000	0.0270	0.0775	0.90
	10000	0.0355	0.0855	0.90
QTBD				
	1000	0.0510	0.1060	0.93
	3000	0.0450	0.1025	0.92
	5000	0.0530	0.1075	0.93
	10000	0.0530	0.1145	0.93
Bonferroni				
		0.0455	0.0850	0.92

more extreme tail probabilities must be estimated from the bootstrap distribution to control at a fixed FWER (say 0.05). Since the estimation of the marginal tail probabilities using the NCSB method is much more sensitive to the number of bootstrap samples than the QTBD method, it is a compelling alternative.

Table 4: Comparison of FWER using different methods for Case 2b, where $n_1 = 5$, $n_2 = 25$, $h_1 = 10$, $h_0 = 90$, $\mu_1 = \mu_2 = 0.0$, $\sigma_1 = 0.1$, $\sigma_2 = 5.0$, $\rho_1 = 0.0$, and $\rho_2 = 0.6$.

Resampling Methods	B	$\alpha = 0.05$ (FWER)	$\alpha = 0.10$ (FWER)	Average Power
Permutation				
	1000	0.0025	0.0080	0.26
	3000	0.0025	0.0090	0.26
	5000	0.0020	0.0090	0.26
	10000	0.0020	0.0090	0.26
NCSB				
	1000	0.0335	0.0875	0.59
	3000	0.0285	0.0725	0.59
	5000	0.0340	0.0810	0.57
	10000	0.0335	0.0865	0.58
QTBD				
	1000	0.0540	0.1110	0.65
	3000	0.0435	0.0980	0.64
	5000	0.0535	0.1180	0.65
	10000	0.0510	0.1090	0.65
Bonferroni				
		0.0125	0.0245	0.62

4 Real Case Study: Occupational exposure to Benzene

4.1 Data Description

We used the competing multiple testing procedures to analyze Affymetrix microarray data, which were collected (as part of a study of benzene exposure) from blood samples of a population of shoe-factory workers in China. Benzene exposure, an industrial chemical and component of gasoline, is an established potential factor in developing leukemia. However, the mechanisms of benzene-induced hematotoxicity and leukemogenesis remain unclear, as does the risk benzene poses at low levels of exposure. The purpose of the

Table 5: Important genes found by various multiple testing methods based on Benzene microarray data

Cutoff Values	Bonferroni	Permutation	NCSB	QTBD
$\alpha = 0.05$	3	4	0	4
$\alpha = 0.10$	4	7	0	13

benzene data analysis is to shed light on these mechanisms and thus better understand the risk that benzene poses, by examining the effects of benzene exposure on peripheral blood mononuclear cell (PBMC) gene expression. RNA was isolated from the PBMC of 8 high-exposed workers along with 8 unexposed controls (Forrest et al. (2005)). Given the data is balanced in this case, the subset pivotality assumption of the permutation method is satisfied. Because a two-sample t-test for each gene was performed, the dominating marginal null distribution used by the QTBD method, for all genes, is the t-distribution with 14 degrees of freedom.

4.2 Results of multiple testing

Table 5 summarizes the numbers of important genes found using four different multiple testing methods according to the Affymetrix data set. We report the number of genes that are "significant" after control FWER using both the nominal values of $\alpha = 0.05$ and 0.10. Several observations can be made from Table 5:

1. Bonferroni methods provide small numbers of important genes.
2. The NCSB method finds no significantly differentially expressed genes.
3. Relative to the permutation method, QTBD method finds similar numbers of genes, although it appears to be less conservative.

Figure 1 shows the adjusted p-values versus the ordered genes for the benzene data. The adjusted Bonferroni p-values reach 1.0 quickly down the list. The adjusted p-values of other three resampling based methods more gradually reach the maximum value of 1.0. In this case, the additional potential power apparent from the QTBD method at an FWER=0.10 could

DOI: 10.2202/1544-6115.1310

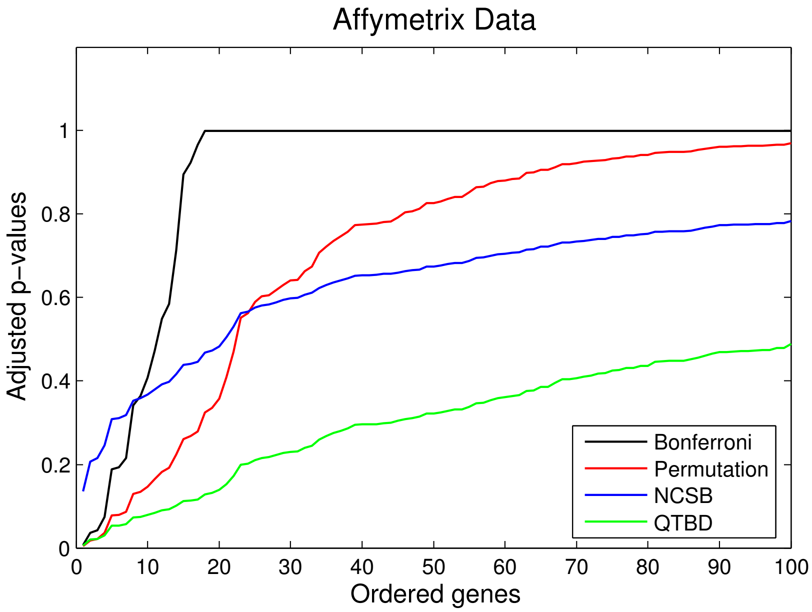

Figure 1: Adjusted p-values of various multiple testing methods, where the black, red, blue, and green lines show the p-values obtained using the Bonferroni, permutation, NCSB and QTBD testing methods, respectively.

be artificial, as it is based on the strong assumption that the marginal null distribution of the test statistics for all genes is t-distributed with 14 degrees of freedom, which, given the relatively small sample size, relies on a normality assumption on the original expression data.

5 Discussion

We have compared the practical performance of permutation methods, null-centered and scaled bootstrap (NCSB) method and the quantile-transformed-bootstrap-distribution (QTBD) method for controlling FWER in multiple testing procedures based on both synthetic and real microarray data sets, and compared the results with those obtained from traditional marginal p-value methods, specifically the Bonferroni's method. In short, the simulation results suggest that the QTBD method does not do worse than permutation methods even when the subset pivotality condition is met and do better (sometimes substantially better) when it is not met. From synthetic data analyses, we found the QTBD method provides accurate and relatively powerful control of FWER, by taking advantage of both the dependence structure and knowledge about the marginal null distribution. Given the easy implementation of the QTBD method, we expect that it should have wide applicability in large scale genomic studies and in other experiments that involve a large number of tested hypotheses.

References

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

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

M.S. Forrest, Q. Lan, A.E. Hubbard, L. Zhang, V. Vermeulen, X. Zhao, G-L. Li, Y-Y. Wu, M. Shen, S. Yin, S.J. Chanock, N. Rothman, and M.T. Smith. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. *Environmental Health Perspectives*, 113:801–7, 2005.

Katherine S. Pollard and Mark J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

Mark J. van der Laan and Alan E. Hubbard. Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006. URL `http://www.bepress.com/sagmb/vol5/iss1/art14`. Article 14.

DOI: 10.2202/1544-6115.1310

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Multiple Testing. Part II. Step-Down Procedures for Control of the Family-Wise Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art14`. Article 14.

P. H. Westfall and S. S. Young. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons, 1993.