Gene expression

How accurately can we control the FDR in analyzing microarray data?

Sin-Ho Jung1,* and Woncheol Jang2

1Department of Biostatistics and Bioinformatics, Duke University, NC 27710, USA and 2Institute of Statistics and Decision Sciences, Duke University, NC 27705, USA

Received on December 5, 2005; revised on April 11, 2006; accepted on April 23, 2006
Advance Access publication April 27, 2006
Associate Editor: David Rocke

ABSTRACT

Summary: We want to evaluate the performance of two FDR-based multiple testing procedures by Benjamini and Hochberg (1995, J. R. Stat. Soc. Ser. B, 57, 289–300) and Storey (2002, J. R. Stat. Soc. Ser. B, 64, 479–498) in analyzing real microarray data. These procedures commonly require independence or weak dependence of the test statistics. However, expression levels of different genes from each array are usually correlated due to coexpressing genes and various sources of errors from experiment-specific and subject-specific conditions that are not adjusted for in data analysis. Because of high dimensionality of microarray data, it is usually impossible to check whether the weak dependence condition is met for a given dataset or not. We propose to generate a large number of test statistics from a simulation model which has asymptotically (in terms of the number of arrays) the same correlation structure as the test statistics that will be calculated from the given data and to investigate how accurately the FDR-based testing procedures control the FDR on the simulated data. Our approach is to directly check the performance of these procedures for a given dataset, rather than to check the weak dependence requirement. We illustrate the proposed method with real microarray datasets, one where the clinical endpoint is disease group and another where it is survival.

Contact: jung0005@mc.duke.edu

1 INTRODUCTION

Microarrays are high throughput technology measuring the expression levels of a large number of genes simultaneously. Discovering the genes whose expression levels are associated with a clinical endpoint, such as disease type or survival, involves a serious multiple testing problem. Suppose that, for \( j = 1, \ldots, m \), we want to test the null hypothesis

\[ H_j : \text{gene } j \text{ has no association with the clinical outcome} \]

against the alternative hypothesis

\[ H_j : \text{gene } j \text{ has some association with the clinical outcome}. \]

Without an appropriate adjustment for the multiplicity, many discoveries will be false positives. Two multiple testing type I error rates have been used to tackle this issue: family-wise error rate (FWER) and false discovery rate (FDR). Usually, expression levels on different genes are correlated due to coexpressing genes and shared noises from experiment-specific and subject-specific conditions that are not adjusted for in analysis. The correlation, together with the high dimensionality, has a strong influence on the testing results controlling these type I errors. FWER is defined as the probability of at least one false positive. So, a multiple testing procedure controlling the FWER requires the null distribution of the test statistics while maintaining the correlation among test statistics. The permutation method proposed by Westfall and Young (1993) usually approximates the null distribution for FWER-control well. Refer to Huang et al. (2005) for the cases where the permutation method may not be appropriate.

Some investigators consider controlling the probability of one false rejection out of thousands of genes to be too strict and advocate an FDR-based multiple testing instead. Suppose that there are \( m \) genes among which the null hypotheses, \( H_j \), are true for \( m_0 \) genes, called non-prognostic genes, and the alternative hypotheses, \( H_j \), are true for \( m_1 = m - m_0 \) genes, called prognostic genes. Based on the calculated p-values, we may reject or accept the null hypotheses. Let \( R \) denote the number of genes for which the null hypotheses are rejected (discovery), and \( R_0 \) denote, among these \( R \) genes, the number of genes that the null hypotheses are true (false discovery). Then, the FDR is defined as the expected value of the proportion \( R_0/R \).

Benjamini and Hochberg (1995) propose a step-up procedure to control the FDR. They prove that this procedure conservatively controls the FDR if the test statistics for the \( m_0 \) non-prognostic genes are independent. The conservativeness becomes more serious as the proportion of prognostic genes \( m_1/m \) increases. Benjamini and Yekutieli (2001) loosen the independence assumption among \( m_0 \) non-prognostic genes to a weak dependence assumption called positive regression.

Pointing out the conservativeness of the Benjamini and Hochberg’s procedure, Storey (2002) proposes a procedure that is considered to be more accurately controlling the FDR when \( m \) is large as in most microarray data. Under the independence assumption on \( m \) genes, he shows that the FDR is approximated by

\[ \text{FDR} = \frac{m_0 \alpha}{R} \]

and replaces \( m_0 \) with an estimated value \( \hat{m}_0 \). He later loosens the independence assumption for a weak dependence assumption.
among whole \( m \) genes (Storey, 2003) or among \( m_0 \) non-prognostic genes (Storey and Tibshirani, 2001). Jung (2005) derives a sample size formula for the Storey’s procedure under the weak dependence assumption among \( m \) genes.

For an accurate control of the FDR, we need the joint distribution of \((R_0, R)\), which will be decided by the joint distribution of the test statistics, under the true effect size for each gene and the dependency among \( m \) test statistics. The aforementioned FDR-based procedures have been shown to reliably controlling the FDR through simulations based on the independence or weak dependence assumptions. However, because of high-dimensionality of microarray data and complexity of the real correlation structure, it is almost impossible to check whether the weak dependence condition holds for a given dataset or not. As an approach to tackling this difficulty, we propose to generate a large number of random vectors from the asymptotic distribution of the test statistics, to apply an FDR-based procedure and to see how accurately the procedure controls the FDR on the simulated data.

When a new dataset is given, we calculate the effect size (mean and standard deviation in case of a two-sample \( t \)-test) of each gene. For a chosen \( m_1 \) value within a range of interest, we identify the top \( m_1 \) genes in terms of effect size. The simulation is modeled as follows. We specify the effect sizes of these \( m_1 \) genes by the observed ones from the data. For the remaining \( m_0 = m - m_1 \) genes, the effect sizes are set at 0. Lin (2005) proposes a simulation algorithm to generate the null distribution of correlated test statistics conditioning on given data. By modifying his algorithm, we generate a large number of test statistics under the observed effect sizes and correlation structure. For a nominal FDR level to be chosen for analysis, we apply an FDR-based multiple testing procedure to each simulation sample and count the numbers of false and total discoveries. The true FDR is estimated by the empirical average of their proportions through a large number of simulations. By comparing the empirical FDR with the nominal one, we can decide how accurate the data analysis will be. Although the simulated test statistics have the same covariances as those observed from the data, the simulation scheme does not require the preliminary estimation of the covariance matrix. Our method is suggested when a sufficient number of arrays are available. The proposed method is applied to the real datasets by Golub et al. (1999) and Beer et al. (2002).

### 2 FDR-BASED MULTIPLE TESTING PROCEDURES

We assume that, for large number of arrays, the distribution of the test statistics is approximated by a multivariate normal distribution. Let \( p_1, \ldots, p_m \) denote the \( p \)-values for \( m \) genes that are calculated by a resampling method or the theoretical distribution of the test statistics. In this paper, we obtain them from the standard normal distribution based on large sample approximation. Let \( p_{(1)} \leq \cdots \leq p_{(m)} \) denote the ordered \( p \)-values for \( m \) genes, and \( H_{(j)} \) for \( j = 1, \ldots, m \) the corresponding null hypotheses. Benjamini and Hochberg (1995) propose to reject \( H_{(j)} \) for all \( j \leq J = \max\{j : p_{(j)} \leq \frac{q \alpha}{m}\} \). They prove that this procedure controls the FDR below \( q^* \) if the \( m_0 \) non-prognostic genes are independent.

Suppose that we reject \( H_j \) if \( p_j < \alpha \). Storey (2002), under independence or a weak dependence among \( m_0 \) null genes, claims that

\[
R_0 = \sum_{j=1}^{m} I(\text{H}_j \text{ true}, \text{H}_j \text{ rejected}) = \sum_{j=1}^{m} \Pr \left( \text{H}_j \text{ true} \right) \Pr \left( \text{H}_j \text{ rejected} \mid \text{H}_j \right) + o_p(m),
\]

which equals \( m_0 \alpha \) with the error term ignored. Here, \( m^{-1} o_p(m) \rightarrow 0 \) in probability as \( m \rightarrow \infty \). Hence, with \( R \) replaced by the observed value \( r \), we have

\[
\text{FDR}(\alpha) \approx \begin{cases} \alpha \frac{m_0}{r} & \text{if } r > 0 \\ 0 & \text{if } r = 0 \end{cases}
\]

(1)

Now, given \( \alpha \), estimation of \( \text{FDR}(\alpha) \) requires estimation of \( m_0 \) only. Storey (2002) proposes to estimate \( m_0 \) by

\[
m_0(\lambda) = \frac{\sum_{j=1}^{m} I(p_j > \lambda)}{(1 - \lambda)r} \text{ for a chosen constant } \lambda \text{ away from 0, such as 0.5. By combining this estimator with (1), we obtain}
\]

\[
\hat{\text{FDR}}(\alpha) = \begin{cases} \frac{\alpha + m_0}{r} & \text{if } r > 0 \\ 0 & \text{if } r = 0 \end{cases}
\]

For an observed \( p \)-value \( p_j \), Storey (2003) defines \( q \)-value, the minimum FDR level at which we reject \( H_j \), as \( q_j = \inf_{\alpha \geq p_j} \text{FDR}(\alpha) \). Jung (2005) shows that this formula is simplified to \( q_j = \text{FDR}(p_j) \) in a two-sample testing case. This procedure is implemented by a computer package called SAM (Significance Analysis of Microarrays) (Tusher et al., 2001).

This procedure has been shown to reliably control the FDR under independence (Storey, 2002) or a weak dependence (Storey, 2003), such as block compound symmetry, by simulations. Given a real dataset, however, it is difficult to check whether the weak dependence requirement holds or not. In this paper, we propose a direct approach for evaluation of the FDR-based multiple testing methods in real data analysis. This approach uses a simulation algorithm by Lin (2005) modified for estimation of the true FDR.

### 3 A SIMULATION-BASED METHOD

In this section, we propose a simulation method to generate a large number of test statistics whose correlation structure, given the data, is asymptotically identical to that of the test statistics to be calculated from the given dataset. There are two versions of large sample approximations in this paper: one with respect to large \( m \) and the other with respect to large \( n \). All asymptotic results here and after are with respect to large \( n \).

#### 3.1 Two-sample \( t \)-test case: when the clinical outcome is dichotomous

Let \( x_{ij} \) denote the expression level of gene \( j (= 1, \ldots, m) \) from subject \( i (= 1, \ldots, n) \) in disease group \( k (= 1, 2) \). We assume that \( \{x_{i1}, \ldots, x_{in}\}, 1 \leq i \leq n_k \} \) are independent and identically distributed (i.i.d) random vectors from a multivariate distribution with means \( (\mu_{1j}, \ldots, \mu_{kj}) \), variances \( (\sigma_{1j}^2, \ldots, \sigma_{kj}^2) \) and correlation matrix \( \Gamma = (\rho_{ij})_{1 \leq i \neq j \leq m} \). Let \( (x_{11}, \ldots, x_{1n}) \) and \( (x_{21}^2, \ldots, x_{2n}^2) \) denote the sample means and pooled variances, respectively, estimated from the data. If the sample sizes are large, then the test statistics
We want to generate random vectors with asymptotically the same distribution as \((W_1,\ldots,W_m)\) using a simulation method. Let \((x_{ij}, 1 \leq i \leq n_k, k = 1,2)\) be IID \(N(0,1)\) random variables, which are independent of the dataset, and \(\{(x_{ij1},\ldots,x_{ijm}), 1 \leq i \leq n_k, k = 1,2\}\) with 
\[
\hat{x}_{ij} = \bar{x}_{ij} + (x_{ij} - \bar{x}_{ij}) \epsilon_{ij}
\]
denote a new dataset. Also, let 
\[
\hat{\bar{x}}_{ij} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}
\]
denote the sample means of the new dataset, and 
\[
\hat{W}_j = \frac{\hat{x}_{ij} - \hat{x}_{i2j}}{s_j \sqrt{n_1^{-1} + n_2^{-1}}}
\]
the resulting test statistics. Then, given the dataset \(\{(x_{ij1},\ldots,x_{ijm}), 1 \leq i \leq n_k, k = 1,2\}\), each \(\hat{W}_j\) is a weighted average of the independent normal random variables. It follows that, given the data, \((\hat{W}_1,\ldots,\hat{W}_m)\) is normal with means 
\[
\hat{\delta}_j = \frac{\bar{x}_{ij} - \bar{x}_{i2j}}{s_j \sqrt{n_1^{-1} + n_2^{-1}}}
\]
and covariance \(\hat{\Gamma}\). Hence, the conditional joint distribution of \((\hat{W}_1,\ldots,\hat{W}_m)\) given the data is asymptotically identical to the unconditional joint distribution of \((\hat{W}_1,\ldots,\hat{W}_m)\). See Appendix for a proof with general test statistics. Note that our simulation method requires calculation of sample means and variances from data, but not the correlation coefficients.

The above procedure is based on an equal variance-covariance assumption in two groups. However, some genes may possibly have unequal variances and covariances between the two groups even under \(H_j: \mu_{1j} = \mu_{2j}\). In this case, the null distribution of the \(t\)-test statistics based on equal variance-covariance assumption may not be approximated by the standard normal distribution. If the equal variance-covariance assumption is questionable, we can use the same simulation method with \(s_j \sqrt{n_1^{-1} + n_2^{-1}}\) replaced by 
\[
\sqrt{n_1^{-1} s_{1j}^2 + n_2^{-1} s_{2j}^2}
\]
in the denominators of \(W_j, \hat{W}_j\) and \(\hat{\delta}_j\), where \((s_{ijj}, j = 1,\ldots,m)\) are sample variances for group \(k(=1,2)\). In this case, the effect sizes will be expressed as 
\[
\hat{\delta}_j = \frac{\mu_{1j} - \mu_{2j}}{\sqrt{n_1^{-1} \sigma_{1j}^2 + n_2^{-1} \sigma_{2j}^2}},
\]
where \(\sigma_{kj}^2\) is the population variance for gene \(j(=1,\ldots,m)\) in group \(k(=1,2)\).

### 3.2 Cases for Cox regression and general test statistics

A large family of test statistics for \(H_j: \theta_j = \theta_j0\) versus \(H_j: \theta_j \neq \theta_j0\) can be written as \(U_j(\theta_j) = \sum_{i=1}^{n} U_{ij}(\theta_j)\), where 
\[
U_{ij}(\theta_j) = \int_{0}^{\infty} \left\{ z_{ij} - \sum_{l=1}^{n} Y_{lj}(z_l) \exp(z_l^\theta_{ij}) \right\} dN_j(t)
\]
and, for large \(n\), \(U_{ij} = U_{ij}(\theta_j0)\) can be expressed as a function of the data from subject \(i\) only, so that \(U_{1i},\ldots,U_{ni}\) are independent. Let \(\mu_j(\theta_j) = E(U_{ij})\) and \(\sigma_j(\theta_j) = \sum_{i=1}^{n} \mu_j(\theta_j)\). If \(E(U_{ij}(\theta_j))\) is a smooth function and \(E(U_{ij}(\theta_j)) = 0\) has a unique solution, then the solution \(\theta_j\) to \(U(\theta_j) = 0\) is consistent to \(\theta_j0\). Further, by the central limit theorem, \((U_{1i},\ldots,U_{ni})\) is approximately normal with means \(\mu_j(\theta_j)\) and covariances \(\sigma_j(\theta_j)\) that can be consistently estimated by 
\[
\hat{\mu}_j = \mu_j(\hat{\theta}_j)\]
and 
\[
\hat{\sigma}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (U_{ij} - \hat{\mu}_j),
\]
respectively, where \(\hat{\mu}_j = \mu_j(\hat{\theta}_j)\) and \(\hat{\sigma}_j = \sigma_j^2\).

For IID \(N(0,1)\) random variables \(e_1,\ldots,e_n\) that are independent of the data, let 
\[
\hat{U}_j = \hat{\mu}_j + \frac{1}{n_j} \sum_{i=1}^{n_j} e_i(\hat{\mu}_j - \hat{\mu}_j).
\]
Let \(W_j = \sigma_j^{-1} U_j\) and \(\hat{W}_j = \sigma_j^{-1} \hat{U}_j\). By Appendix, the conditional joint distribution of \((W_1,\ldots,W_m)\) given the data is asymptotically identical to the unconditional joint distribution of \((W_1,\ldots,W_m)\). Note that the standardized effect sizes of test statistics \(W_j\) are 
\[
\hat{\delta}_j = \sigma_j^{-1} \hat{\mu}_j.
\]

The two-sample \(t\)-test case discussed in Section 3.1 is a specific example of this formulation. As another example, we consider the cases to discover the genes associated with a survival endpoint. Let, for subject \(i(=1,\ldots,n)\), \(T_i\) denote the time to an event, such as tumor recurrence or death, and \((z_{i1},\ldots,z_{im})\) denote the gene expression data from \(m\) genes. Survival time may be censored due to loss to follow-up or study completion, so that we observe \(X_i = \min(T_i, C_i)\) together with a censoring indicator \(\Delta_i = I(T_i \leq C_i)\), where \(C_i\) is the censoring time that is assumed to be independent of \(T_i\) given the expression data \((z_{i1},\ldots,z_{im})\). A given dataset will be expressed as \(\{(X_i, \Delta_i, z_{i1},\ldots,z_{im}), 1 \leq i \leq n\}\). Let \(Y(t) = 1(X_i \geq t)\) and \(N_i(t) = \Delta_i I(X_i \leq t)\) be the at-risk and event processes for patient \(i\), respectively, and let \(Y(t) = \sum_{i=1}^{n} Y_i(t)\) and \(N(t) = \sum_{i=1}^{n} N_i(t)\).

Suppose that, for subject \(i\), \(z_{ij}\) is related to the hazard rate by 
\[
\lambda_{ij}(t) = \lambda_{ij0}(t) \exp(\theta_{ij}),
\]
where \(\lambda_{ij0}(t)\) is the unknown baseline hazard specific to gene \(j\) (Cox, 1972). The hypotheses are expressed as \(H_j: \theta_j = 0\) versus \(H_j: \theta_j \neq 0\). The partial MLE, \(\hat{\theta}_j\), solves the partial score function 
\[
U_j(\theta_j) = \int_{0}^{\infty} \left\{ z_{ij} - \sum_{l=1}^{n} Y_{lj}(t) \exp(z_l^\theta_{ij}) \right\} dN_j(t).
\]
Let 
\[
\mu_{ij} = \int_{0}^{\infty} \left\{ z_{ij} - \sum_{l=1}^{n} Y_{lj}(t) \right\} Y(t) \exp(z_l^\theta_{ij}) dN_j(t)
\]
for the above simulation method.

1732
and
\[
\hat{\mu}_j = \int_0^\infty \left\{ z_j - \sum_{i=1}^n Y_i(t) \mu_{ij} \right\} Y_i(t) e^{\hat{\theta}_j t} \, dN(t),
\]
where \( dN(t) = \lambda(t) \, dt \) and
\[
d\Lambda(t) = \left\{ \sum_{i=1}^n Y_i(t) e^{\theta_j t} \right\}^{-1} \, dt.
\]
refer to Andersen and Gill (1982). Then the partial score test statistic \((U_1, \ldots, U_m) = (U_j(0), \ldots, U_m(0))\) is approximately normal with mean \( \mu_j = \sum_{i=1}^n \mu_{ij} \) and variance-covariances \( \sigma_{jj}', \) that can be consistently estimated by \( \hat{\mu}_j = \sum_{i=1}^n \hat{\mu}_{ij} \) and
\[
\hat{\sigma}_{jj}' = \sum_{i=1}^n (U_{ij} - \hat{\mu}_{ij})(U_{jj} - \hat{\mu}_{jj}),
\]
respectively. Let \( \sigma_{jj}' = \hat{\sigma}_{jj}' \). Note that \( \mu_j = 0 \) under \( H_j \).

The regression model (3) may not hold for some genes. By Lin and Wei (1989), whether model (3) is valid or not, the score statistic \( U_j \) is a meaningful measure of association between \( z_j \) and \( T_j \), and the test statistic \( W_j = \sigma_{jj}'^{-1} U_j \) is asymptotically \( \mathcal{N}(0,1) \) under \( H_j \). For robust testing against potential outliers in gene expression data, Jung et al. (2005) propose to use the rank of \( z_j \) among gene \( j \) observations, \( z_{1j}, \ldots, z_{nj} \), as the covariate of model (3), rather than the raw expression level.

4. NUMERICAL STUDIES

In this section, we take real microarray data and demonstrate our simulation-based method discussed in the previous section to check how well the FDR-control procedures will work under the dependency embedded in a given dataset.

An accurate estimation of the FDR, to be discussed below, requires identification of the prognostic genes and their effect sizes in addition to the correlation structure among the genes. While the second part in the right-hand side of (2), \( \sum_{i=1}^n \varepsilon_i(U_{ij} - \hat{\mu}_{ij}) \), is to approximate the correlation among \( m \) test statistics, the first part, \( \hat{\mu}_j \), is a function of their effect sizes. For an accurate calculation of the FDR, we need to know which genes are prognostic. Since none of the observed \( \hat{\mu}_j \) would be exactly zero, we first have to specify \( m_1 \) and identify \( m_1 \) genes with large effect sizes as follows. Given a dataset, we first calculate \( \hat{\mu}_j \) (= \( x_{ij} - x_{yj} \) in two-sample \( t \)-test case) and \( s_j \). For a chosen \( m_1 \), we identify the genes with top \( m_1 \) effect sizes in absolute value. The effect sizes for these \( m_1 \) genes will be set at the observed absolute values \( \delta_j = \sigma_{jj}'^{-1} | \hat{\mu}_j | \), but those for the remaining \( m_0 = m - m_1 \) genes will be set at \( \delta_j = 0 \). By the same arguments as in Appendix, the change in effect sizes does not change the correlation structure of the test statistics. In order to simplify the procedure, we take the absolute values of the effect sizes and use one-sided tests. We observe similar results by using two-sided tests and the raw effect sizes.

Now, we generate a large number of sets (say, \( B = 4000 \)) of the test statistics \((\hat{W}_1, \ldots, \hat{W}_m)\), where
\[
\hat{W}_j = \delta_j + \sigma_{jj}'^{-1} \sum_{i=1}^n \varepsilon_i(U_{ij} - \hat{\mu}_{ij}).
\]
Let \((\bar{w}_b, \ldots, \bar{w}_{B})\) denote the set of \(m\) test statistics from the \(b\)-th simulation. Also, let
\[
r_{0b}(\alpha) = \sum_{j=1}^{m} I(\bar{w}_{bj} > z_{1-a}, \delta_j = 0),
\]
\[
r_{b}(\alpha) = \sum_{j=1}^{m} I(\bar{w}_{bj} > z_{1-a})
\]
denote the numbers of false rejections and total rejections, respectively, from the \(b\)-th simulation sample. Here, \(z_{1-a}\) is the \(100(1 - a)\)-th percentile of \(N(0, 1)\) distribution. For a large \(B\), the true FDR is approximated by
\[
\hat{q}(\alpha) = B^{-1} \sum_{b=1}^{B} r_{0b}(\alpha)/r_{b}(\alpha),
\]
If we want to control the FDR at \(q^*\) level, then we have to find the corresponding \(\alpha = \alpha^*\) by solving \(\hat{q}(\alpha) = q^*\) using the bisection method, and reject all genes with \(p\)-values smaller than \(\alpha^*\). We call this procedure ‘exact method’ since it exactly controls the empirical FDR from the simulations based on the true parameter values. We estimate the average true rejections and false rejections by
\[
\hat{r}_1 = B^{-1} \sum_{b=1}^{B} r_{1b}(\alpha^*),
\]
and
\[
\hat{r}_0 = B^{-1} \sum_{b=1}^{B} r_{0b}(\alpha^*),
\]
respectively, where \(r_{1b} = r_b - r_{0b}\). The calculated \(\hat{r}_0\) and \(\hat{r}_1\) will be compared with those by the Storey’s method, called SAM, and the method by Benjamini and Hochberg (1995), called BH, that are discussed in Section 2. The exact method uses a common critical value for \(m\) \(p\)-values like SAM, so that it can be used as a gold standard for SAM. Note that the exact method cannot be used in a real data analysis because we do not know which genes are prognostic. We can replace \(z_{1-a}\) with a resampling-based quantile. However, we use the theoretical standard normal quantile based on the asymptotic normality.

SAM and BH are applied to each simulated set of test statistics to control the FDR level at \(q^*\), calculate \(r_{0b}\) and \(r_{b}\), and estimate the FDR level, at which these procedures are really controlling, as the average of \(r_{0b}/r_{b}\).

\[
\hat{q} = B^{-1} \sum_{b=1}^{B} r_{0b}/r_{b},
\]
through the \(B\) simulations. If these procedures accurately control the FDR, then \(\hat{q}\) is expected to be close to \(q^*\).

### 4.1 An example for two-sample \(t\)-tests

An example dataset is taken from the golubTrain object in golubEsets package (version 1.0.1) in Bioconductor release 1.7. (Gentleman et al., 2004). Golub et al. (1999) explore \(m = 6810\) genes extracted from bone marrow in 38 patients, of which \(n_1 = 27\) with acute lymphoblastic leukemia (ALL) and \(n_2 = 11\) with acute myeloid leukemia (AML), in order to identify the genes with potential clinical heterogeneity being differentially expressed in the two subclasses of leukemia. Genes useful to distinguish ALL from AML may provide insight into cancer pathogenesis and patient treatment. We conduct simulation studies for \(m_1 = 20, 60, 100, 400, 1000\) and 2500, and \(q^* = 0.5, 0.4, 0.3, 0.2, 0.1, 0.05\) and 0.01. Using \(\lambda = 0.5\),

![Fig. 1. Distribution of \(R/R_0\) from 4000 simulations under \(q^* = 0.5\) and 0.01 and \(m_1 = 20\). (a) \(q^* = 0.5\) (SAM), (b) \(q^* = 0.5\) (BH), (c) \(q^* = 0.01\) (SAM), (d) \(q^* = 0.01\) (BH).](image-url)
we obtain \( \hat{m}_0 = 4278(\hat{m}_1 = m - \hat{m}_0 = 2532) \) from the original data.

Table 1 reports the estimated FDR, \( \hat{q} \), number of true and false rejections, \( \hat{r}_1 \) and \( \hat{r}_0 \), respectively, for SAM and BH methods. Only \( \hat{r}_1 \) and \( \hat{r}_0 \) are reported for the exact method since it controls the FDR accurately. BH is always more conservative than SAM, i.e. \( \hat{q}_{BH} < \hat{q}_{SAM} \). For example, when \( m_1 = 20 \) and the nominal FDR is set at \( q^* = 5\% \) level, SAM and BH control the FDR at \( q = 8.71 \) and 8.090%, respectively. The ratio \( \hat{q}_{SAM}/\hat{q}_{BH} \) increases in \( m_0 \) as in independent case (Storey, 2002). With \( m_1 \leq 400 \), SAM (BH) is conservative if \( q^* \geq 0.3 (q^* \geq 0.2) \) and anti-conservative otherwise.

SAM controls the FDR accurately when \( m_1 \geq 100 \) and \( q^* \leq 0.3 \). So does BH when \( m_1 \geq 400 \) and \( q^* \leq 0.2 \). However, the bias of SAM and BH in \( q^* \) is more serious with a smaller \( m_1 \) value.

There exists a similar trend in \( \hat{r}_1 \) to that in \( \hat{q} \). i.e. with \( m_1 \leq 400 \), both SAM and BH have smaller \( \hat{r}_1 \) than the exact method for \( q^* \geq 0.2 \) and larger \( \hat{r}_1 \) for \( q^* < 0.2 \). SAM and BH have almost the same \( \hat{r}_1 \) for \( m_1 \leq 400 \), but, with a larger \( m_1 \), the former tends to have a larger \( \hat{r}_1 \).

SAM always has higher false rejections, \( \hat{r}_0 \), than the other two methods. The discrepancy is more perceivable with smaller \( m_1 \). BH also has higher false rejections than the exact method when \( m_1 \leq 400 \). With \( m_1 \geq 1000 \), however, BH tends to have lower \( \hat{r}_0 \) than the exact method except for a small \( q^* \) such as \( 1 \) or \( 5\% \).

Figure 1 reports the distribution of \( R_0/R \) observed from the 4000 simulations for \( q^* = 0.5 \) or 0.01 with \( m_1 \) fixed at 20. When \( q^* = 0.5 \), \( R_0/R \) is highly distributed around 0 and 1. When \( q^* = 0.01 \), \( R_0/R \) has a large density around 0 and is widely distributed in the rest of the range. Note that the horizontal axes of Figure 1c and d are rescaled using a square root transformation to show the distribution of \( R_0/R \) around \( q^* = 0.01 \) better. At each \( q^* \) level, the distributions of \( R_0/R \) for SAM and BH look almost the same, so that it is difficult to observe any difference in the amount of conservativeness or anti-conservativeness between the two procedures. Furthermore, the distributions are widely spread over the range of \([0, 1]\), so that the figures do not clearly show the location shift of the distributions from the nominal \( q^* \).

### 4.2 An example with survival data

Beer et al. (2002) generated expression profiles of \( m = 4966 \) genes to discover the genes that can predict disease progression. The data include \( n = 86 \) stage I or III lung cancer patients, of whom 24 patients have disease progressions. By controlling the FWERM at 10% level, Jung et al. (2005) discovered two genes whose expression levels are significantly associated with the time to progression. In this section, we consider the same test statistics standardized by the standard errors as described in Section 3.2. Simulations are conducted at similar settings to those in the previous example.

The simulation results are reported in Table 2. We observe that both SAM and BH conservatively control the FDR in the simulation settings. BH is slightly more conservative than SAM, but they become similar as \( m_1 \) decreases. SAM controls the FDR very accurately for \( q^* \leq 0.1 \) which is the range of interest in usual data analyses. SAM becomes more accurate with a larger \( m_1 \). Using \( \lambda = 0.5 \), we obtain \( \hat{m}_0 = 4092(\hat{m}_1 = m - \hat{m}_0 = 874) \) from the original data.

In all the simulation settings, the exact method has the largest \( \hat{r}_1 \) and BH has the smallest \( \hat{r}_1 \), although the difference is small. When controlling the FDR at \( q^* = 10\% \) level, the exact method has only about \( \hat{r}_1/m_1 = 40\% \) of true rejections for \( m_1 \geq 60 \). The other two methods have lower true rejection rates. When \( m_1 \leq 200 \), the exact method has the smallest \( \hat{r}_0 \) and SAM has the largest \( \hat{r}_0 \) among the three methods. The discrepancy in \( \hat{r}_0 \) among the three methods becomes smaller as \( m_1 \) increases and \( q^* \) decreases.

### 5 DISCUSSION

Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) prove that their approach conservatively controls the FDR under independence and weak dependence of \( m_0 \) test statistics for which null hypotheses are true. However, from the first example of Section 4, we found that the BH method can be anti-conservative in a general correlation combined with a small number of differentially expressing genes, \( m_1 \), and a small nominal FDR level, \( q^* \). Storey et al. (2004) also claim that SAM procedure conservatively controls the FDR under weak dependence. From the first example, it is
shown that SAM can be anti-conservative too (with small \( m_1 \) and \( q^* \)).

If an FDR-based procedure does not control the FDR accurately in a real data analysis, there may be two potential reasons: (1) the test statistics are heavily correlated or (2) the null distribution of each test statistic cannot be approximated by the standard normal distribution because the sample size is not large enough for the normal approximation of the test statistics. In order to avoid issue (2) and focus our discussion on (1), we generated \( s \) from \( N(0, 1) \) distribution so that the simulated test statistics have exactly normal distributions. As mentioned in Appendix, however, if \( n \) is large enough, \( s \) can be generated from any distribution with mean 0 and variance 1, such as \( U(-\sqrt{3}, \sqrt{3}) \).

From simulations not reported in this paper, we observed that SAM closely estimates \( E(R_g(\alpha)|E(R^{-1}(\alpha)) = \alpha m_0 E(R^{-1}(\alpha)) \) rather than the FDR, \( E(R(\alpha)/R(\alpha)) \).

As mentioned previously, our procedure is to check the performance of the existing multiple testing methods for a given dataset, so that it is different from the simulation-based multiple testing procedures used for data analysis. However, a similar simulation method can be used to derive a testing procedure too, see Lin (2005). Given \( \gamma \in (0, 1) \), van der Laan et al. (2004) and Lehmann and Romano (2005) propose a conservative procedure to control the false discovery proportion, defined as \( P(R_g/R > \gamma) \), at a certain level. Our simulation method can be easily modified to evaluate the conservativeness of their procedure for a given dataset.

We have discussed our procedure in terms of microarray data, but it can be used for any high dimensional data involving multiple testing using dependent test statistics, e.g. proteomic data. The simulation programs are written in Fortran 77. Uniform random numbers were generated using RAN2 subroutine of Press numerical recipes. The Art of Scientific Computing. Cambridge University Press, NY.

APPENDIX

It suffices to show that the conditional joint distribution of \( (\bar{U}_1, \ldots, \bar{U}_m) \) given the data, \( D \), is asymptotically identical to the unconditional joint distribution of \( (\bar{U}_1, \ldots, \bar{U}_m) \). As discussed in Section 3.2, \( (\bar{U}_1, \ldots, \bar{U}_m) \) is asymptotically normal with means and covariances that can be consistently estimated by \( \mu_j \) and \( \sigma_{ij} \) for \( 1 \leq j, f \leq m \), respectively.

Given the data, \( \mu_j \) and \( U_{ij} \) are constants, so that \( \bar{U}_j = \mu_j + \sum_{i=1}^{n} (U_{ij} - \mu_j) \) for \( j = 1, \ldots, m \) are weighted sums of iid \( N(0, 1) \) random variables, \( \epsilon_1, \ldots, \epsilon_n \). Hence, \( (\bar{U}_1, \ldots, \bar{U}_m) \) is normal with means

\[
E(\bar{U}_j | D) = \mu_j + \sum_{i=1}^{n} (U_{ij} - \mu_j)E(\epsilon_i) = \mu_j
\]

and covariances

\[
\text{cov}(\bar{U}_j, \bar{U}_j | D) = \sum_{i=1}^{n} ((U_{ij} - \mu_j)(U_{ij} - \mu_j)) \text{var}(\epsilon_i) = \sigma_{jj}
\]

which concludes the proof.

In fact, \( \epsilon_i \)’s can be generated from any distribution with mean 0 and variance 1, such as \( U(-\sqrt{3}, \sqrt{3}) \). In this case, the conditional joint distribution of \( (\bar{U}_1, \ldots, \bar{U}_m) \) given \( D \) is approximated by the same distribution by the central limit theorem.