BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 14 2014, pages 2018–2025
doi:10.1093/bioinformatics/btu124

Gene expression

Advance Access publication March 14, 2014

# A general method for accurate estimation of false discovery rates in identification of differentially expressed genes

Yuan-De Tan[1] and Hongyan Xu[2,*]

[1]College of Life Science, Hunan Normal University, Changsha, Hunan 410087, China and [2]Department of Biostatistics and Epidemiology, Georgia Regents University, Augusta, GA 30912-4900, USA

Associate Editor: Ziv Bar-Joseph

## ABSTRACT

**Summary:** The 'omic' data such as genomic data, transcriptomic data, proteomic data and single nucleotide polymorphism data have been rapidly growing. The omic data are large-scale and high-throughput data. Such data challenge traditional statistical methodologies and require multiple tests. Several multiple-testing procedures such as Bonferroni procedure, Benjamini–Hochberg (BH) procedure and Westfall–Young procedure have been developed, among which some control family-wise error rate and the others control false discovery rate (FDR). These procedures are valid in some cases and cannot be applied to all types of large-scale data. To address this statistically challenging problem in the analysis of the omic data, we propose a general method for generating a set of multiple-testing procedures. This method is based on the BH theorems. By choosing a C-value, one can realize a specific multiple-testing procedure. For example, by setting $C = 1.22$, our method produces the BH procedure. With $C < 1.22$, our method generates procedures of weakly controlling FDR, and with $C > 1.22$, the procedures strongly control FDR. Those with $C = G$ (number of genes or tests) and $C = 0$ are, respectively, the Bonferroni procedure and the single-testing procedure. These are the two extreme procedures in this family. To let one choose an appropriate multiple-testing procedure in practice, we develop an algorithm by which FDR can be correctly and reliably estimated. Simulated results show that our method works well for an accurate estimation of FDR in various scenarios, and we illustrate the applications of our method with three real datasets.

**Availability and implementation:** Our program is implemented in Matlab and is available upon request.

**Contact**: hxu@gru.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on September 12, 2013; revised on January 23, 2014; accepted on February 27, 2014

## 1 INTRODUCTION

The advance of 'omic' technologies has led to a great development of large-scale data such as microarray data, transcriptomic data, metabolomic data and proteomic data. These 'omic' data enable us to take a global insight into complex biological procedures, the interaction between genetic and environmental factors and pathological mechanisms of complex diseases, such as diabetes, heart disease, hypertension and various cancers, via identification of genes differentially expressed and cluster or classification of functional genes, pathway or network. Such large-scale data or high-throughput data challenge traditional statistical tests (Efron *et al.*, 2001; Newton *et al.*, 2001; Pan *et al.*, 2003; Tan *et al.*, 2006; Tusher *et al.*, 2001) because the traditional statistical tests are single tests. In a single hypothesis test, the threshold $\alpha$ is meaningful for determining whether a test is significant because $\alpha$ is a probabilistic threshold for the occurrence of an event in a population. However, threshold $\alpha = 0.05$ does not remain valid for multiple statistical analyses of omic data because, for example, in identification of 10 000 genes, at least 500 findings are expected by chance at the significance level of 0.05 (Nichols and Hayaska, 2003; Tusher *et al.*, 2001). Therefore, threshold $\alpha$ is required to be adjusted for multiple tests. Currently, several statistical procedures have been developed for adjusting threshold $\alpha$ or adjusting *P*-values in the multiple tests. Among them, the Bonferroni procedure, the Holm procedure (Holm, 1979), the Hochberg procedure (Hochberg, 1988), the Westfall and Young procedure (Westfal and Young, 1993), the Sidak procedure (Nichols and Hayasaka, 2003), the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995) and the Benjamini–Liu procedure (Benjamini and Liu, 1999) are typical multiple-testing approaches. Among these procedures, the BH procedure has been most broadly applied in practice. But we found that the true false discovery rate (FDR) does not always remain consistent with the estimated FDR of the BH procedure, and all existing multiple-testing procedures are actually special cases of the multiple-testing procedure family that ranges from the single-testing procedure to the Bonferroni procedure; in other words, these procedures give different rejection space (Fig. 1). In practice, our simulation also shows that true FDR is often significantly overestimated or underestimated by the BH procedure. To accurately estimate FDR in a large-scale statistical analysis, here we propose a method that is based on the BH theorems to generate a set of multiple-testing procedures in a range of the single-testing procedures to the Bonferroni procedure and a simulation algorithm to choose a desirable procedure.

## 2 METHODS

### 2.1 Rejection area for the hypothesis tests

Here we focus on discussing rejection areas of the single-testing procedure [A(s)], the Bonferroni [A(B)] and the BH multiple-testing procedures [A(BH)]. As seen in Figure 1, for *G* hypotheses to be tested, a single-

*To whom correspondence should be addressed.

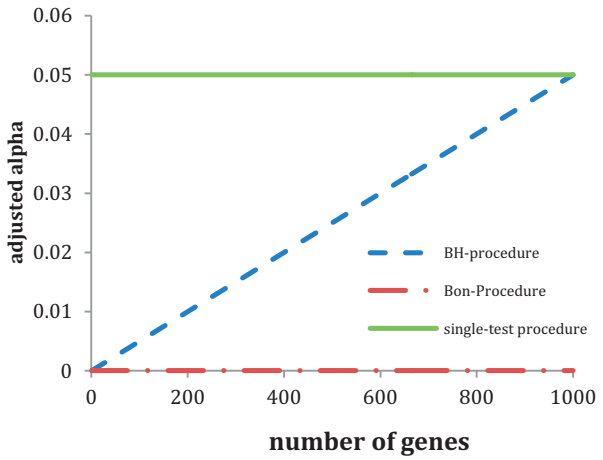


**Fig. 1.** Area of rejection. Single-testing procedure is a top (solid) line with $\varphi = \alpha$ for all genes to be tested. The Bonferroni (Bon-) procedureis bottom (dash and dot) line with $\varphi = \alpha/G$ for all genes and its area for reject is area under (dash and dot) line $(=\alpha(1/G) \times G = \alpha$ where $\varphi$ is an adjusted $\alpha$. The BH-procedure is a up-diagonal (dash) line with $\varphi_i = \alpha i/G$ for gene $i$, and its area of rejection is area under diagonal line $(= \alpha G/2)$ where $\varphi_i$ is an adjusted $\alpha$

testing procedure has the largest rejection area: $A(s) = \alpha G$ where $\alpha$ is a probabilistic threshold, whereas the Bonferroni multiple-testing procedure has the smallest rejection area: $A(B) = (\alpha/G)G = \alpha$. These are two extreme rejection areas. In the single-testing procedure, $\alpha G$ hypotheses would be rejected by chance, whereas the Bonferroni multiple-testing procedure would have a chance of $\alpha$ to reject the $G$ hypotheses, which is called the family-wise error rate. The rejection area of the BH multiple-testing procedure is area under diagonal line, which is half of that of the single-testing procedure, $A(BH) = \int_0^G \alpha(i/G)di = aG/2$ where $i$ is the $i$-th hypothesis. But different from the single-testing procedure where $\alpha$ is set for all hypotheses and the Bonferroni multiple-testing procedure where $\alpha/G$ is given for all hypotheses, the BH multiple-testing procedure is a procedure where, from $G$ to 1, the adjusted $\alpha$ values are on a diagonal line, its rejection area is $\alpha G/2$. So the BH multiple-testing procedure falls in between the single-testing procedure and the Bonferroni multiple-testing procedure. Therefore, in theory, it is possible to give any multiple-testing procedure with a rejection area from $\alpha G$ to $\alpha$.

## 2.2 Benjamini-Hochberg theorems

For $m$ null hypotheses $H_1, H_2, \ldots, H_m$, testing them has corresponding $P$-values $p_1, p_2, \ldots, p_m$. Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the ordered $P$-values and their corresponding ordered null hypotheses be $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. Then, we have a BH procedure:

Let $k$ be the largest $i$ for which $p_{(i)} \le \frac{i}{m}\alpha$, then all $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ would be rejected.

THEOREM: for independent statistical tests and for any configuration of false null hypotheses, the above procedure controls the FDR at the significance level $\alpha$ (Benjamini and Hochberg, 1995).

This theorem is based on the following lemma:

LEMMA: define $Q = V/(V+S)$ as FDR where $V$ and $S$ are, respectively, numbers of findings in true and false null hypotheses, for any $m_0$ independent $P$-values corresponding to true null hypotheses where $0 \le m_0 \le m$ and for any $m_1$ $P$-values corresponding to false null hypotheses where $m_1 = m - m_0$, the BH procedure satisfies inequality

$$E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \le \frac{m_0}{m}\alpha \tag{1}$$

By integrating inequality Equation (1), we obtain

$$E(Q) \le \frac{m_0}{m}\alpha < \alpha \tag{2}$$

and the FDR is controlled (Benjamini and Hochberg, 1995). Because $0 \le m_0 \le m$, any procedure with function of $i$

$$p_{(i)} \le \frac{f(i)}{m}\alpha \tag{3}$$

also satisfies inequality Equation (2) and has control of FDR where $i$ is an ordered number from 1 to $m$, and $f(i)$ is any function of $i$ and $1 \le f(i) \le m$. From Equation (3), one can see that the aforementioned multiple-testing procedure becomes the single-testing procedure when $f(i) = m$, the Bonferroni procedure when $f(i) = 1$ or the BH procedure when $f(i) = i$.

## 2.3 Function $f(i)$

Suppose we have a set of observed $P$-values $p_1, \ldots, p_k, \ldots, p_G$ for G null hypotheses $H_1, H_2, \ldots, H_k, \ldots, H_G$. Let them be ranked as $p_{(1)} \le .... \le p_{(i)} \le ... \le p_{(G)}$ with corresponding ordered null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(i)}, \ldots, H_{(G)}$ where $i$ is the $i$-th position in the ranked space corresponding to $k$ in the unranked space. We declare all hypotheses $H_{j \le (i)}$ to be interesting when $p_{(i)} \le \varphi_i$ in a step-up fashion where $\varphi_i$ is an adjusted $\alpha$. To realize a general adjustment of $\alpha$ for any control of FDR across the whole observed $P$-values, we need a control sequence. To this end, let a set of $G$ q-values be a linear-rank sequence $q_1 < q_2 < \ldots < q_g < \ldots < q_G$ where $g = 1,.., G$ and $q_g = g/G$. We then define a ratio of difference between two adjacent q-values to the sum of them as

$$\frac{(q_g - q_{g-1})}{(q_g + q_{g-1})} \tag{4}$$

and take the sum of the ratios over subset $i$,

$$S_i = 1 + \sum_{g=2}^{i} \frac{(q_g - q_{g-1})}{(q_g + q_{g-1})} \tag{5}$$

where $S_1 = 1$, $i = 2, 3, \ldots, G$ and in the whole ranked set, we have

$$S = 1 + \sum_{g=2}^{G} \frac{(q_g - q_{g-1})}{(q_g + q_{g-1})}. \tag{6}$$

Thus, the control sequence consists of a set of ranked rates with a power C:

$$R_i = \left(\frac{S_i}{S}\right)^C \tag{7}$$

where C is referred to as control value. In the control sequence, we have $R_1 \approx 0$ for $S_1 = 1$ and $R_G = (S_G/S)^C = 1^C = 1$ for $S_G = S$. With the control sequence, function of $i$ is found to be

$$f(i) = G^{R_i} \tag{8}$$

$\alpha$ then is adjusted as

$$\varphi_i = \alpha \times \frac{f(i)}{G} = \alpha \frac{G^{R_i}}{G} \tag{9}$$

Equation (9) gives a family of multiple procedures from the single-testing procedure to the Bonferroni procedure. The C-value is a so-called control value because it has a control effect on $\varphi_i$; in other words, different C-values have different control levels of false-positive findings. For example, when $C = 0$, then $R_i = 1$ and $\varphi_i = \alpha G^1/G = \alpha$ for all $P$-values, which is the single-testing procedure; when $C \to \infty$, then $\lim_{C\to\infty} R_i = 0$ and $\lim_{C\to\infty} \varphi_i = \alpha/G$ or $C = G > 1000$, $R_i \approx 0$ and $\varphi_i \approx \alpha/G$ for all $P$-values, which is just the Bonferroni procedure; and if $C = 1.22$, we have $\varphi_i \approx \alpha\frac{i}{G}$, which is just the BH procedure (Fig. 2).

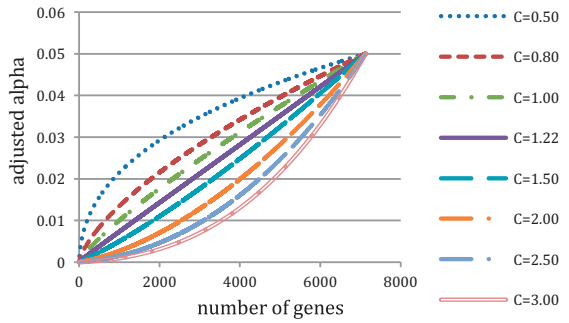


**Fig. 2.** Multiple procedures. These multiple-testing procedures for testing 7129 genes differentially expressed between two conditions are created by using C = 0.50, …1.22, …3.0 on Equation 9. Compared to the BH-procedure (dash line) in Figure 1, the procedure with C = 1.22 (deep purple and solid line) is the BH-procedure. Therefore, procedures with C > 1.22 have stronger control of FDR than the BH-procedure while those with C < 1.22 have weaker control of FDR than the BH-procedure

Thus, at $C > 1.22$, Equation (9) generates a set of multiple-testing procedures that have stronger FDR control than the BH procedure, whereas at $C < 1.22$, it generates another set of multiple-testing procedures that have weaker FDR control than the BH procedure (Fig. 2). Therefore, the method not only can realize the existing multiple-testing procedures but also can produce a new procedure by choosing an appropriate C-value (Fig. 2).

## 2.4 C-value choice for FDR control

To control FDR, we need to choose a C-value to generate a multiple-testing procedure. In theory, one can use a permutation approach (Reiner *et al.*, 2003) to determine a multiple-testing procedure. But as we will see in Section 3, the permutation is not a good way because it tends to significantly underestimate FDR. Therefore, we here propose an alternative approach to choose a desirable C-value for a real dataset. For the convenience of discussion, we focus our algorithm on traditional two-condition (or two-class) *t*-test methodology.

Our algorithm runs through the following nine steps, among which Step 2 is to calculate means, variances and distances of each gene, Steps 3 and 4 are to simulate data using parametric bootstrap and Steps 5 and 6 are to estimate FDRs using a set of multiple procedures:

Step 1: apply a statistical method and a set of multiple procedures to real data and calculate a real ratio set of findings by comparing the ordered *P*-values to *M* sets of adjusted $\alpha$ values: $\rho_m = N_m/G$ where $m = 1, \ldots, M$, and $N_m$ is number of findings by a statistical method in real data at the *m*-th C level.

Step 2: calculate mean $\bar{x}_{gk}$ and variance $\sigma^2_{gk}$ of expression values of gene $g$ ($g = 1, \ldots, G$) in condition $k$ ($k = 1, 2$) from a real dataset, distance $d_g$ between two means for each gene $g$, $d_g = \bar{x}_{g1} - \bar{x}_{g2}$, and sort d-values from the largest to the smallest.

Step 3: adjust distance value by

$$d^*_g = \begin{cases} d_g \times a, & \text{if } \omega_g < \theta \\ d_g, & \text{otherwise} \end{cases} \tag{10}$$

where $\theta$ is an input value in (0,1], $a$ is also an input parameter value and $\omega_g = \frac{|d_g|}{\max|d_g|}$. Here $\theta$ and $a$ are used to adjust the larger estimated ratios ($\hat{\rho}$s) of findings at lower C levels. Here $\hat{\rho}_m = \hat{N}_m/G$ where $\hat{N}_m$ is the number of findings by the same statistical method in a simulated dataset at the *m*-th C level.

Step 4: determine a distribution of the real dataset. For microarray data, normal distribution is good, and for transcriptomic data, negative binomial distribution or Poisson distribution may be considered. Here, as an example, we focus on microarray data and use a normal pseudorandom generator to generate expression noise from gene to gene in each condition:

$$w_{gkr} = N^{-1}(0, \sigma_{gk}, r) \tag{11}$$

where $r$ is the $r$-th replicate, $r = 1, \ldots, R_k$ where $R_k$ is replicate number in condition $k$, and $\sigma_{gk}$ is standard deviation for gene $g$ in condition $k$. We set $\sigma_{g1} = \sigma_{g2}$ to generate the same expression noise distribution in two conditions. Then the adjusted d-values are randomly assigned to $\rho_m G$ genes:

$$\begin{aligned} x_{g1r} &= w_{g1r} + bd^*_g \\ x_{g2r} &= w_{g2r} - bd^*_g, \end{aligned} \tag{12a}$$

if $U_g \le \rho_m$,

$$\begin{aligned} x_{g1r} &= w_{g1r} \\ x_{g2r} &= w_{g2r} \end{aligned} \tag{12b}$$

otherwise, where $U_g$ is the uniform random variable from gene to gene, and $b$ is a weight value. The b-value should be smaller if $\hat{\rho}_M$ (the smallest estimated ratio of findings) is larger than $\rho_M$ (the smallest real ratio of findings); the b-value should be larger, otherwise.

Step 5: apply the same statistical method to the simulated data and obtain a set of $G$ *P*-values from a given distribution. Apply Equation (7) and Equation (9) and a set of M C-values to create M multiple-testing procedures that have already been used in the real data and do multiple comparisons.

Step 6: because differentially expressed genes are given, count false findings and calculate FDR.

Step 7: $\theta$, $a$ and $b$ are determined by fitting the estimated ratio set of findings to the real ratio set of findings. We begin with $\theta = a = b = 1$ and roughly check whether the estimated ratio set of findings identified by M multiple-testing procedures is fitting to the real ratio set. If yes, go to Step 8; otherwise, reset $\theta$, $a$ and/or $b$ values (see Section 4) and again perform Steps 3–5 until the estimated ratio set is close to the real one.

Step 8: repeat Steps 4–6 about 15 times, and calculate means of estimated ratios of findings and estimate FDR values at each C level.

Step 9: smooth the estimated FDRs by using polynomial regression. The method is to plot M estimated FDRs versus M C-values (from the smallest to the largest) to yield an observed line, use the polynomial regression line to fit the observed line, then make solution for regression coefficients $\beta_1, \ldots, \beta_v$ and use the polynomial regression equation to calculate the FDR corresponding to each C-value. The order in polynomial regression depends on fitting. This is easily done in Excel.

Plot the real and estimated ratio sets of findings identified by M multiple-testing procedures, and fit a linear regression model with intersection $\beta_0$ and regression coefficient $\beta_1$. If $|\beta_0| \le 0.1$ and $0.9 \le \beta_1 \le 1.1$, then the estimated ratio set of findings is almost identical to the real ratio set across M multiple-testing procedures, and the true FDR set would also be given with the estimated FDR set; a C-value would be chosen by choosing a calibrated FDR. Otherwise, repeat Steps 4–7 by resetting parameters $a$, b and $\phi$ until $|\beta_0| \le 0.1$ and $0.9 \le \beta_1 \le 1.1$.

### 2.5 Performances of statistical methods

Empirical Bayesian analysis was performed using baySeq package; Fisher exact tests and a generalized linear model (GLM) were implemented in edgeR package. Packages baySeq and edgeR were downloaded from Bioconductor. These two packages use the BH procedure to adjust *P*-values (to estimate FDR). Beta *t*-tests were conducted in Matlab. We also used the BH procedure to adjust *P*-values. The q-value approach was performed using the qvalue package downloaded from Bioconductor. This package has GUI interface. We loaded *P*-values from two-sample *t*-tests into the qvalue.gui interface and executed q-value computation and obtained q-values. We chose DE genes at q-value $\leq 0.05$.

Our multiple-procedure approach and the algorithm for the choice of C-value were implemented in Matlab.

## 3 RESULTS

### 3.1 Simulation evaluation

We used Tusher *et al.*'s (2001) microarray data to illustrate this algorithm. We downloaded the data from the website http://www-stat.stanford.edu/~tibs/SAM/. The data consist of 7129 genes and eight human lymphoblastoid cell line samples, among which four were treated with ionizing radiation and the remainders treated with un-ionizing radiation were used as controls. From the real microarray data, means and variances of expression values of each gene in two conditions were obtained, and a normal pseudorandom generator, any one of two means, any one of two standard deviations for each gene and four replicates were used to create two noise expression datasets of 7129 genes with equal variances. Then, simulated data were generated by linearly and randomly assigning condition effects $\tau = 300U$ on differential expressions between the two conditions to 30% of genes in the two noise datasets where $U$ is a uniform variable in (0,1]. Note that our simulation procedure generates strong dependence among these 30% of genes in assigning their expression values. We applied a classical two-condition *t*-test method to this simulated dataset and set M = 140 C-values from 0.01 to 1.4 to create 140 procedures for multiple tests and obtained 140 ratios of findings. By applying the aforementioned algorithm and adjusting a, b and $\theta$ values, we found that when $a = 0.001$, $b = 5$ and $\theta = 0.054$, 140 estimated and real ratios of findings are close together. Thus, the results were averaged over 15 repeated simulations and displayed in Figure 3. Figure 3A shows that a plot line of estimated against real ratios of findings is almost a diagonal line with intersection $\beta_0 = -0.009 \approx 0$ and regression coefficient $\beta_1 = 1.0576 \approx 1$, so the estimated ratio set of findings is almost identical to the real one. From Figure 3B, one also can see that the plot dots of estimated FDRs versus true FDRs are almost on a diagonal line, suggesting that the estimated FDR is approximately identical to its true value at each FDR point.

Next, we simulated transcriptomic data with negative binomial distribution. Our simulations were conducted by setting 13 000 mRNA isoforms, two conditions (control and disease states), six replicated libraries in each condition, probability = $U$ where $U$ is random uniform variable, $U \in [0, 1]$ and size = $1000U$ for each isoform. The negative binomial pseudorandom generator and these setting parameters were used to create two noise transcriptomic datasets with equal variances, and then $\tau = 100u$ for conditional (or treatment) effect values on differential transcription of mRNA isoforms were randomly and linearly assigned to 10%

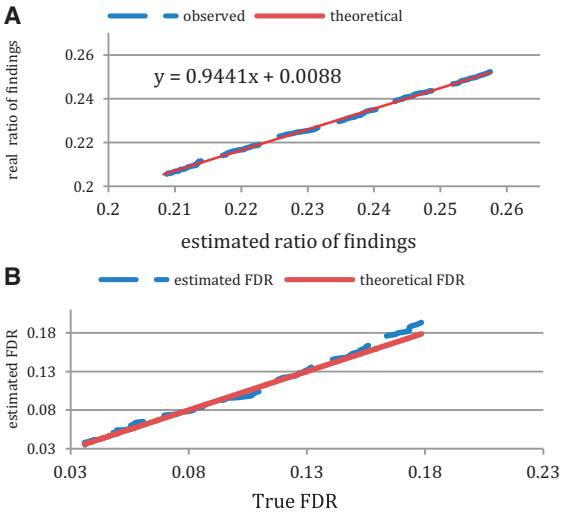


**Fig. 3.** A simulation example for FDR estimation. (**A**) Plot of estimated against real ratios of findings. The estimated and real ratios of findings were obtained by 140 testing procedures and our algorithm from a simulated dataset based Tusher *et al.* (2001) microarray data. Solid line is a regression line and dash line is observed value. (**B**) Plot of estimated against true FDRs obtained by 140 procedures from the pseudo real dataset. The solid line is a diagonal line on which estimated FDR is equal to its true value at each FDR point and the dash line is estimated DFR

of mRNA isoforms in two noise datasets where $u = (0,1]$ is the uniform variable. Thus, simulated data with 10% differentially transcribed isoforms, six replicates and two conditions were obtained. There are a batch of existing methods for finding mRNA isoforms of being differentially transcribed between two conditions: empirical Bayesian (Hardcastle and Kelly, 2010), Fisher exact test (Robinson and Smyth, 2008), GLM (McCarthy *et al.*, 2012) and Beta *t*-test (Baggerly *et al.*, 2003). These methods all use *the BH procedure* to estimate FDR. Similarly to Figure 3B, we here used the simulated transcriptomic datasets to evaluate the BH procedure based on the four statistical methods. We plotted true FDRs in findings of these methods against estimated FDRs using the BH procedure from cutoff = ~ 0 to ~ 0.21. Figure 4 shows that the BH-estimated FDR curves in the empirical Bayesian, GLM and Fisher exact test and especially, in the Beta *t*-test are much below their theoretical lines (true FDR versus true FDR), indicating that in these methods the BH procedure heavily underestimated FDRs.

To fully display statistical properties of our multiple-testing approach in various microarray experiments, we also generated simulation datasets in 27 scenarios. Our simulations were based on Tusher *et al.*'s (2001) microarray data. We set three levels of condition (or treatment) effect: A = 100, 200 and 300; three levels of proportions of differentially expressed genes: P = 10, 20 and 30% and three levels of sample sizes: n = 4, 6 and 18. Condition effect $\tau = Au$ was randomly assigned to P of genes. This assignment generated strong correlation expressions among genes. Our multiple-procedure approach and algorithm for FDR estimation were applied to these simulated datasets. Because of lots of the simulated datasets, we could not display the profiles of estimated and real ratios of findings and the estimated and true FDRs, as

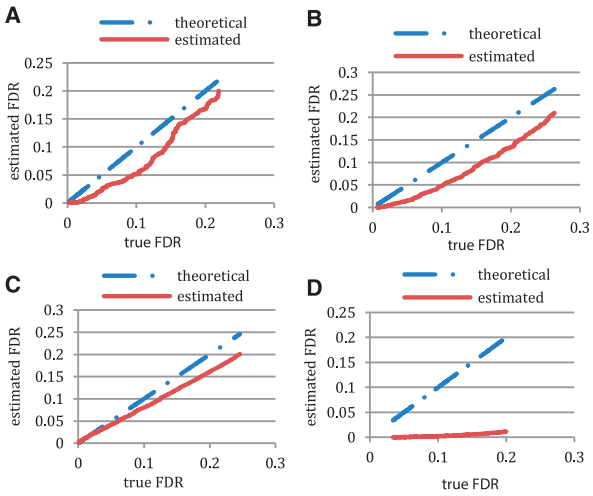


**Fig. 4.** Plots of true versus BH-estimated FDR in existing statistical methods. Estimated curve was made by plotting estimated FDR against true FDR along cutoff of ~0 to ~0.21 and theoretical line is a diagonal line made by plotting true FDR against true FDR along the same cutoff. The true FDR was calculated by counting false positives in findings of a statistical method at an FDR cutoff point in a simulated dataset and the estimated FDR was predicted by a statistical method. The true and estimated FDRs were averaged over three simulated datasets. (**A**) Exact tests, (**B**) Generalized linear Model, (**C**) Empirical Bayesian, (**D**) Beta t-tests

shown in Figure 3A and B, for each scenario; instead, we listed the numbers of findings obtained by traditional t-test method and the true and estimated FDRs obtained by a procedure chosen with its estimated FDR smaller than but closest to the cutoff of 0.05. To demonstrate our algorithm, we also applied permutation (Reiner *et al.*, 2003) and q-value (Storey and Tibshirani, 2003) approaches to these 27 stimulated datasets. The results are summarized in Supplementary Table S1.

It can be seen in Supplementary Table S1 that C-value chosen varies in the range of 0.9–1.32 including C = 1.22. As seen in Figures 1 and 2, C = 1.22 produces the BH procedure. Thus, the BH procedure is just one of our multiple procedures. The FDR estimated by our method at each C-value in each given scenario is close to its true value, except that in scenarios 3 and 17 where the true FDR values are little greater than their estimated values; all the other FDRs were slightly overestimated, meaning that our estimation of FDR is conservative. Supplementary Table S1 also gave the results obtained by performing the q-value approach (Storey and Tibshirani, 2003). Our method and the q-value approach have similar number of findings in most of simulated datasets. However, the q-value approach underestimated FDRs in nine scenarios (33.3%). Therefore, for conservativeness of FDR estimation, our method obviously outperforms q-value approach. By using the permutation algorithm the FDRs were severely underestimated in all 27 scenarios (Supplementary Table S1), indicating that the permutation is not a desirable approach for estimation of FDR.

## 3.2 Application

The first example for application of the multiple-procedure approach is Tusher *et al.*'s (2001) microarray data. As seen above, the data consist of 7129 genes. Tusher *et al.* (2001) tried to find genes differentially expressed between four human lymphoblastoid cell line samples treated by ionizing radiation and four samples without ionizing radiation. We set 140 C-values from 0.1 to 1.39 to create 140 multiple-testing procedures. Because C-value > 0.92 did not produce meaningful results (no findings), we obtained a set of 83 real ratios of findings from the real data using classical two-condition *t*-test method. After adjusting, we found $a = 0.02$, b = 2.9 and $\theta = 0.3$ and repeated our algorithm process for 15 times and generated a set of 83 estimated ratios of findings and a set of 83 estimated FDRs. We plotted real and estimated ratios of findings in Figure 5A1. The regression line is

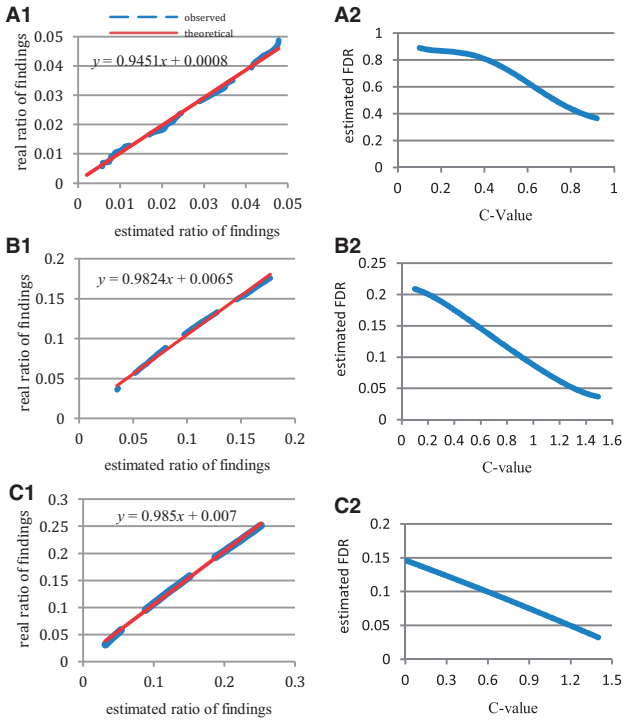


**Fig. 5.** Applications of our method to the real data. The real ratio set of findings was obtained by using a set of 140 procedures for multiple t-tests of gene differential expressions between two conditions from a real dataset. The estimated ratio set of findings was obtained by using our method from 15 simulated datasets each similar to real data. The estimated ratio of findings is compelled to be identical to its real value across all given procedures so that the true FDR generated by each of procedures would be reliably estimated. Column 1: plots of estimated versus real ratios of findings. Solid line is a regression line, dash line is observed line. Column 2: plots of estimated FDRs versus C-values (procedures). Row A: The ionizing irradiation microarray data provided by Tusher *et al.* (2001). 7129 genes of being expressed in 4 human lymphoblastoid cell line samples treated by ionizing radiation and in 4 samples without ionizing radiation were detected on human cDNA arrays. Row B: the data are acute leukemia microarray data that were downloaded from Broad Institute Cancer Program Data Set. This dataset consists of 12582 genes whose expression values in 24 acute lymphoblastic leukemia (ALL) cell lines and in 28 acute myelogenous leukemia (AML) cell lines were measured by oligonucleotide arrays. Row C: *Arabidopsis* microarray data (Lee *et al.* 2009). This dataset contains 22810 genes that were detected for differentially expression between the C58 infected and control stalks each with 3 replicates on plant *Arabidopsis* gene chip

almost a diagonal line with intersection $\beta_0 = 0.0008 \approx 0$ and regression coefficient $\beta_1 = 0.9451 \approx 1$. Hence, the estimated ratio of findings is proximately identical to its real ratio at each of C-values. Thus, the simulated FDR given in Figure 5A2 can be used to estimate the true FDR across the 83 C-values. Figure 5A2 shows that the smallest FDR is 0.36 at C-value = 0.92, which is much larger than cutoff of 0.05. Therefore, at FDR cutoff of 0.05, no gene in this dataset was found to be differentially expressed between human lymphoblastoid cell lines treated with ionizing radiation and control cell lines. This result is consistent with the report of Tusher *et al.* (2001) where the smallest FDR estimated is 0.12 in SAM.

The second example is leukemia microarray data downloaded from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. This dataset consists of 12 582 genes whose expression values in 24 acute lymphoblastic leukemia (ALL) cell lines, 20 myeloid lymphoid leukemia cell lines and 28 acute myelogenous leukemia (AML) cell lines were measured by oligonucleotide arrays. Here, we chose two leukemia cell types ALL and AML for differential expression tests. This is an example of two conditions with large and unequal sample sizes. For this dataset, we input 140 C-values from 0.1 to 1.39 and generated 140 real ratios from classical unpaired two-condition *t*-tests. After adjusting, we obtained $a = 0.5$, $b = 1.5$ and $\theta = 0.05$ and applied our algorithm to estimate a set of FDRs by the 140 procedures and repeated 15 times. The results are shown in Figure 5B. As seen in Figure 5B1, the plot dots of the estimated versus real ratios of findings almost fall on a diagonal line with intersection $\beta_0 = 0.0065 \approx 0$ and regression coefficient $\beta_1 = 0.9824 \approx 1$, suggesting that the estimated ratio of findings was constrained to be almost identical to the real one at each C-level. Thus, the true FDR set can be given by the estimated FDR set. Figure 5B2 shows that at C = 1.31, FDR = 0.049723 is close to 0.05 at which 582 genes were identified to be differentially expressed between ALL and AML cell lines.

The third example is *Arabidopsis* microarray data. To explore whether *Arabidopsis* genes respond to oncogenes encoded by the transfer-DNA (T-DNA) or to bacterial effector proteins codelivered by *Agrobacteria* into the plant cells, Lee *et al.* (2009) conducted microarray experiments at 3 h and 6 d after inoculating wounded young *Arabidopsis* plants with strains C58 and GV3101, a cognate of strain C58, which only lacks transfer-DNA, but possesses proteinaceous virulence (Vir) factors such as VirD2, VirE2, VirE3 and VirF (Vergunst *et al.*, 2000, 2003). Wounded, but uninfected, stalks were served as control. As an example, we just downloaded 6d postinoculation data from GEO (GEO accession: GSE14106) website (http://www.ncbi.nlm.nih.gov/geo/). The data consisting of 22 810 genes were obtained from the C58 infected and control stalks each with three replicates. We performed the classical two-condition *t*-test approach on this dataset and applied our multiple-procedure approach to do multiple tests. We set 140 C-values from 0.01 to 1.4 to create 140 procedures and obtained 140 real ratios of findings from two-condition *t*-tests. After adjusting, we found $a = 0.08$, $b = 15$ and $\theta = 0.035$. Using these parameters, we obtained 140 estimated ratios of findings and FDRs by 140 procedures for 15 repeats. Figure 5C1 shows that the observed plot line of estimated versus real ratios of findings completely overlapped with the predicted line. Because intersection $\beta_0 = 0.007 \approx 0$ and regression coefficient $\beta_1 = 0.985 \approx 1$, the observed line of estimated and real ratios of findings is a diagonal line, so that the estimated ratio of findings is one-to-one identical to the real one at each C-value. Thus, the true FDR set can be given by the estimated FDR set. Figure 5C2 shows that 1427 genes were identified to be differentially expressed between the young tumor and control stalks at C = 1.2 with FDR = 0.0497 (Table 1). However, using the fold change method, Lee *et al.* (2009) just identified 196 differentially expressed genes at $P < 0.01$. Both have a big difference. To verify our findings, we displayed heatmap of the raw data of the 1427 DE genes selected in Figure 6. Figure 6 shows that in the downregulation, all the 498 genes clearly had much lower expression values in young crown gall tumors than in

**Table 1.** The numbers of DE genes found by multiple-procedures from *Arabidopsis* microarray data (Lee *et al.*, 2009)

| C-value | Number of findings | Estimated FDR |
|---|---|---|
| 1.00 | 2121 | 0.0667 |
| 1.02 | 2071 | 0.0650162 |
| 1.17 | 1547 | 0.05227295 |
| 1.18 | 1499 | 0.0514162 |
| 1.19 | 1462 | 0.05055855 |
| 1.20 | 1427 | 0.0497 |
| 1.21 | 1389 | 0.04884055 |
| 1.22 | 1346 | 0.0479802 |
| 1.38 | 785 | 0.0340922 |
| 1.39 | 750 | 0.03321655 |
| 1.40 | 721 | 0.03234 |

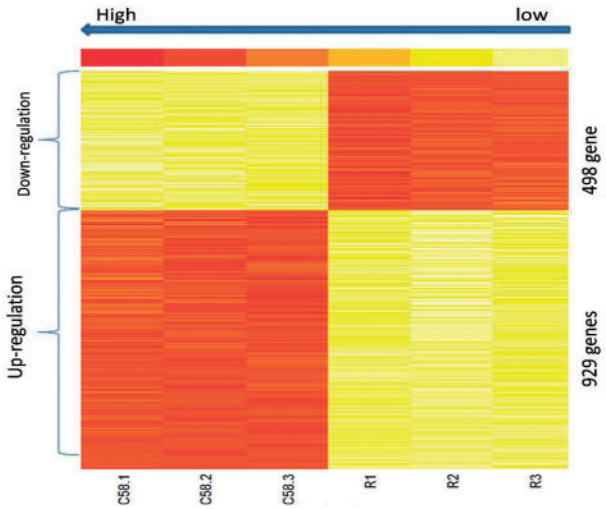


**Fig. 6.** Heatmap. Rows are 1427 genes found to be differentially expressed between young tumor (6d postinoculation), and control stalk and columns are three control (reference) stalk replicates (right side) and three young tumor replicates (left side), among which 498 genes were downregulated in C58 strain relative to control stalk, and 929 were upregulated in C58 strain relative to control stalks. Yellow is the lowest expression values, and deep red is the highest expression values

control (reference) stalks (as shown by deep red and yellow, respectively), whereas in the upregulation, all the 929 genes also had clearly much higher expression values in crown gall tumors than in control stalks. That is, in the 1427 genes, tumors and controls are explicitly separated by gene expressions. For the three replicates, the 1427 genes are differentially expressed.

## 4 DISCUSSION

The probabilistic threshold $\alpha$ is adjusted as $\varphi = \alpha\phi$ where $\phi$ is called rate for controlling false discoveries. The single-testing procedure and the Bonferroni procedure are two extreme procedures. As shown in Figure 1, the single-testing procedure is a top horizontal line with $\phi = 1$ for all G tests, whereas the Bonferroni procedure is a bottom horizontal line with $\phi = 1/\mathrm{G}$ for all G tests. The BH procedure has the rate of controlling false discoveries $\phi = i/\mathrm{G}$ where $i = 1 \ldots$, G, which is a diagonal line. They all are special cases of our method. Interestingly, if Equation (9) is changed as

$$\varphi_k = \alpha\phi_k = \alpha \frac{1}{G + 1 - G^{R_k}} \tag{13}$$

our multiple-procedures can then switch the step-up procedures given in Equation (9) to the step-down procedures given in Equation (13): for $R_1 = 0$, we have $\varphi_1 = \alpha/G$, whereas for $R_G = 1$, we have $\varphi_G = \alpha$. Like step-up procedures, by changing C-value, a set of $R_k$-values is also altered so that a set of new multiple-testing procedures is created. If $C = 0$, then $R_k = 1$ and $\varphi_k = \alpha$ for all hypotheses, which is the single-testing procedure. If $C > 1000$, then $R_k \approx 0$ and $\varphi_k \approx \alpha/\mathrm{G}$ for all hypotheses, which is just Bonferroni procedure. If $C = x$ so that $G^{R_i} \approx i$, $i = 1, \ldots, G$, then $\varphi_i = \alpha/(G - i + 1)$ is just the Holm procedure. If $C = y$ so that $\varphi_i = \alpha/(G - G^{R_i} + 1) \approx ai/G$, we get a proximate BH procedure. These indicate that in our multiple-procedure approach, the step-down procedures are equivalent to the step-up procedures.

As seen in Table 1, our method is somewhat similar to SAM (Tusher *et al*., 2001) and RAM (Tan *et al*., 2006) because a set of C-values (procedures) is equivalent to a set of thresholds ($\Delta s$) given for declaring significant tests in comparison between the observed ranked statistics and the estimated ranked null statistics in SAM and RAM. However, as seen in Figure 3 and in Supplementary Table S1, our method ensures that estimated FDR is close to its true value by constraining that the estimated ratio of findings is identical or approximate to its real ratio at each of the given C-values, whereas the SAM and RAM do not have any constraint to ensure that estimation of FDR is reliable. 27 simulated datasets show that our method works well for accurate estimation of FDR in different sample sizes, different conditional effects and different ratios of genes differentially expressed in different biological systems of study. Three real datasets also demonstrated that our method performs well even in the cases of large and unequal sample sizes or very small sample sizes but large number of genes detected on arrays.

In microarray and transcriptomic experiments, the extremely large number of genes (the dimensionality of the feature space reaches tens or even hundreds of thousands) and extremely small number of biological conditions lead genes to be highly correlated in expression. This means that such genome-wide data are highly dependent data. Our simulated and real results show that the dependence among expressions of genes did not impact estimation of FDR even though the multiple procedures are based on independence of hypotheses (the BH theorems). This is because the estimation of ratio of findings is also based on highly correlative expressions of genes.

In theory, not only our step-up or step-down multiple-procedure can be applicable for any statistical methods but also our algorithm for choice of C-values (procedures) can be generalized to any other statistical tests or is suitable to any data by changing distributions. For example, if a real dataset is count data such as transcriptomic data or serial analysis of gene expression (SAGE) data, then the simulation in our algorithm can be conducted by choosing appropriate parameters on the Poisson, gamma, binomial or negative binomial distribution. But in practice, our method would not be available for cases in which distribution of statistics is unclear.

In addition, compared with existing large-scale statistical methods such as SAM (Tusher *et al*., 2001), RAM (Tan *et al*., 2006), ODP (Storey *et al*., 2007), Cyber T (Baldi and Long, 2001), Limma (Smyth, 2004; Smyth *et al*., 2005) and q-value (Storey and Tibshirani, 2003), our method needs more time to determine parameter values so that the estimated ratio of findings is identical or very close to its real ratio across a set of chosen C-values chosen, or say, $\beta_0$ is proximate to 0 and $\beta_1$ to 1. In effect, $\theta$, $a$ and $b$ values can easily be determined by noting that, as seen in Equation 10, a larger $\theta$-value would allow much larger d-values to be adjusted and a smaller $a$-value would make these d-values become smaller, then more $\hat{\rho}$-values near the M-end would become smaller; according to Equation 12, a larger b-value would allow differences in all differential expressions between two conditions to be larger, and as a result, more $\hat{\rho}$-values near the 1-end would be enlarged. Our experience is that first we glance at the difference between the $\rho$-value set and the $\hat{\rho}$-value set with $\theta = a = b = 1$ and determine b-value so that at the 1-end the $\hat{\rho}$-values are close to the $\rho$-values. Then we set a small $a$-value and adjust $\theta$ so that differences between the $\rho$-values and $\hat{\rho}$-values at the M-ends are smaller than a tolerant value, e.g. 0.01. Also, one can first fix a smaller $\theta$-value and then adjust $a$-value. Nevertheless, the constraint condition is that $\beta_0$ is close to 0 and $\beta_1$ to 1.

## ACKNOWLEDGEMENT

The authors would like to thank two anonymous reviewers for providing many constructive suggestions.

*Funding*: H.X. is supported by an intramural grant from the Georgia Regents University.

*Conflict of Interest*: none declared.

## REFERENCES

Baggerly,K.A. *et al.* (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**, 1477–1483.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

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

Benjamini,Y. and Liu,W. (1999) A distribution-free multiple test procedure that controls the false discovery rate. *J. Stat. Plan. Inference*, **82**, 163–170.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Hochberg,Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

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

Lee,C.W. *et al.* (2009) Agrobacterium tumefaciens promotes tumor induction by modulating pathogen defense in *Arabidopsis* thaliana. *Plant Cell*, **21**, 2948–2962.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

Newton,M.A. *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Nichols,T. and Hayaska,S. (2003) Controlling the family wise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.*, **12**, 419–446.

Pan,W. *et al.* (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integr. Genomics*, **3**, 117–124.

Reiner,A. *et al.* (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing dierential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3,** Article 3.

Smyth,G.K. *et al.* (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

Storey,J.D. *et al.* (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8**, 414–432.

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

Tan,Y.D. *et al.* (2006) Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes. *Genomics*, **88**, 846–854.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Vergunst,A.C. *et al.* (2000) VirB/D4-dependent protein translocation from *Agrobacterium* into plant cells. *Science*, **290**, 979–982.

Vergunst,A.C. *et al.* (2003) Recognition of the Agrobacterium tumefaciens VirE2 translocation signal by the VirB/D4 transport system does not require VirE1. *Plant Physiol.*, **133**, 978–988.

Westfal,P. and Young,L.S. (1993) *Resampling-Based Multiple Testing*. Wiley, New York, NY.