*Statistical Applications in Genetics and Molecular Biology*

*Volume* 7, *Issue* 1 2008 *Article* 13

# Adaptive Choice of the Number of Bootstrap Samples in Large Scale Multiple Testing

Wenge Guo* Shyamal Peddada†

*National Institute of Environmental Health Science, wenge.guo@gmail.com

†National Institute of Environmental Health Science, peddada@niehs.nih.gov

# Adaptive Choice of the Number of Bootstrap Samples in Large Scale Multiple Testing*

Wenge Guo and Shyamal Peddada

## Abstract

It is a common practice to use resampling methods such as the bootstrap for calculating the p-value for each test when performing large scale multiple testing. The precision of the bootstrap p-values and that of the false discovery rate (FDR) relies on the number of bootstraps used for testing each hypothesis. Clearly, the larger the number of bootstraps the better the precision. However, the required number of bootstraps can be computationally burdensome, and it multiplies the number of tests to be performed. Further adding to the computational challenge is that in some applications the calculation of the test statistic itself may require considerable computation time. As technology improves one can expect the dimension of the problem to increase as well. For instance, during the early days of microarray technology, the number of probes on a cDNA chip was less than 10,000. Now the Affymetrix chips come with over 50,000 probes per chip. Motivated by this important need, we developed a simple adaptive bootstrap methodology for large scale multiple testing, which reduces the total number of bootstrap calculations while ensuring the control of the FDR. The proposed algorithm results in a substantial reduction in the number of bootstrap samples. Based on a simulation study we found that, relative to the number of bootstraps required for the Benjamini-Hochberg (BH) procedure, the standard FDR methodology which was the proposed methodology achieved a very substantial reduction in the number of bootstraps. In some cases the new algorithm required as little as $1/6^{th}$ the number of bootstraps as the conventional BH procedure. Thus, if the conventional BH procedure used 1,000 bootstraps, then the proposed method required only 160 bootstraps. This methodology has been implemented for time-course/dose-response data in our software, ORIOGEN, which is available from the authors upon request.

**KEYWORDS:** BH procedure, bootstrap, confidence interval, false discovery rate, multiple testing, p-value

---

*This research is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The authors are grateful for several helpful and insightful comments from two referees that led to a much improved presentation of our manuscript. The authors also thank Doctors Pierre R. Bushel and David M. Umbach for carefully reading this manuscript and for their useful comments, and Shawn Harris for the implementation of the proposed algorithm in the software, ORIOGEN.

# 1 Introduction

Statistical analysis of high dimensional data often involves the problem of simultaneously testing a large number of null hypotheses. For instance, gene expression microarray data analysis involves comparisons of gene expressions of several thousand genes under two or more conditions (e.g. normal and tumor cells). A traditional approach to dealing with this problem is to control the familywise error rate (FWER), the probability of falsely rejecting at least one true null hypothesis (Hochberg and Tamhane, 1987). However, when the number of statistical hypotheses to be tested is large, control of FWER is so stringent that often only a few false null hypotheses are rejected. Consequently, less conservative alternative measures of error rates have been proposed in the literature (Benjamini and Hochberg, 1995; Efron et al, 2001; Hommel and Hoffmann, 1987; Korn et al, 2004; Lehmann and Romano, 2005; Sarkar, 2007; van der Laan, Dudoit and Pollard, 2004). One such measure is the false discovery rate (FDR), which is defined as the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses. This widely used measure was introduced by Benjamini and Hochberg (1995). In the seminal paper, the authors also introduced a simple FDR controlling procedure (popularly called Benjamini-Hochberg or BH procedure) and proved that it controls the FDR when the underlying test statistics are independently distributed. Later Benjamini and Yekutieli (2001) and Sarkar (2002) strengthened the result of Benjamini and Hochberg by showing that the BH procedure controls the FDR for positively dependent test statistics.

When some of the null hypotheses are not true, the BH procedure is conservative by a factor $\pi_0$, i.e., $FDR \leq \pi_0\alpha$, where $\pi_0$ is the proportion of true null hypotheses among all null hypotheses. Storey (2002) and Storey, Taylor and Siegmund (2004) introduced a more powerful adaptive FDR controlling procedure (Storey procedure) in which $\pi_0$ is estimated by using a simple estimate $\hat{\pi}_0$ (which will be introduced in Section 2.4) and then the BH procedure is applied at level $\alpha/\hat{\pi}_0$. For other adaptive FDR procedures, see Benjamini and Hochberg (2000) and Benjamini, Krieger and Yekutieli (2006). The BH and Storey procedures have been extensively used in many applications such as the analysis of data from microarray experiments, quantitative trait locus (QTL), functional magnetic resonance imaging (fMRI) and clinical trials (Reiner, Yekutieli and Benjamini, 2003; Storey and Tibshirani, 2003; Benjamini and Yekutieli, 2005; Genovese, Lazar and Nichols, 2002; Mehrotra and Heyse, 2004).

The computation of the BH and Storey procedures requires the exact $p$-value corresponding to each hypothesis. However, since the exact or even the

asymptotic distribution of the underlying test statistic is usually unknown or difficult to determine, the exact $p$-values are not always available. Quite often, the individual $p$-values are obtained by using the bootstrap methods. Since the total number of possible bootstrap samples is usually very large, it is practically impossible to derive $p$-values based on all possible bootstrap samples. Given the large number of null hypotheses to be tested, even a modest number of bootstrap samples to calculate individual $p$-values may result in a substantial increase in total computation time. For instance, if the microarray chip (e.g. Affymetrix chip) consists of 50,000 genes then the total number of desired bootstrap samples (at the rate of 10,000 per gene) can be as large as 500 million bootstrap calculations. Depending upon the underlying test statistic, these calculations could be very time consuming, despite the availability of super fast computing.

Although the bootstrap methods have been extensively applied in calculating marginal $p$-values, determination of the required number of bootstrap samples has not been discussed in the context of multiple testing. In practice, this number is typically determined in somewhat ad hoc manner and it is often chosen to be the same in calculating each $p$-value. This may not be efficient because not all bootstrap $p$-values need to be estimated with the same amount of precision. The precision depends upon the bootstrap $p$-value. It seems that a better method is to adaptively choose different numbers of bootstrap samples based on different computed $p$-values. For example, if a $p$-value has been computed to lie in $[0.2, 0.3]$, then it may not be necessary to continue its computation while using the BH or Storey procedures, since we have had enough information to make our decision.

The above observation motivates us to propose in Section 2 a computationally simple algorithm that reduces the total number of bootstrap samples and ensures proper control of the FDR while applying the BH or Storey procedure. By exploiting the confidence intervals of ideal bootstrap $p$-values (with the number of bootstrap samples $B \to \infty$) estimated from a small number of bootstrap samples, decisions regarding some of the hypotheses could be made with fewer bootstrap samples.

For each hypothesis, the proposed sequential algorithm "calibrates" the number of bootstrap samples according to the estimated standard error of each bootstrap $p$-value. The number of bootstraps needed for making decisions using the proposed algorithm is substantially smaller than if one were to perform decisions using $p$-values based on the same pre-specified number of bootstrap samples for each hypothesis.

As seen from the simulation study described in Section 3, the reduction in the number of bootstrap calculations can be as much as a 6-fold reduction

without sacrificing the control of FDR. In Section 4 we illustrate the proposed algorithm by applying it to a gene expression microarray data obtained by Lobenhoffer et al (2001). Theoretical justification for the proposed algorithm is provided in the Appendix.

# 2 Methods

## 2.1 Notations

Suppose $H_1, H_2, \ldots, H_m$ are $m$ null hypotheses of interest to be tested of which $m_0$ are true null hypotheses and the remaining $m_1 = m - m_0$ are false. Let $M = \{H_i : 1 \leq i \leq m\}$ indicate the set of all $m$ null hypotheses, $R$ denote the total number of null hypotheses rejected by a multiple testing procedure of which $V$ denote the number of true null hypotheses rejected. The proportion of false discoveries is defined to be $Q = \frac{V}{R}$ (and equal to 0 if $R = 0$) and the false discovery rate (FDR) is defined to be the expectation of $Q$, i.e.,

$$FDR = E(Q) = E(\frac{V}{R}) \ . \tag{1}$$

Suppose $P_i$ is the $p$-value associated with the $i^{th}$ null hypothesis $H_i, 1 \leq i \leq m$ and $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ are the sorted $p$-values. Let $H_{(i)}$ denote the null hypothesis corresponding to $P_{(i)}, 1 \leq i \leq m$.

The well-known Benjamini-Hochberg (BH) procedure for controlling FDR is a linear step-up procedure with critical constants $\alpha_i = \frac{i}{m}\alpha, \ i = 1, \cdots, m$, where $\alpha$ is a pre-specified significance level. The BH procedure proceeds as follows. If $P_{(m)} \leq \alpha$, then reject all null hypotheses; otherwise, reject hypotheses $H_{(1)}, \ldots, H_{(r)}$ where $r$ is the smallest index satisfying $P_{(m)} > \alpha, \ldots, P_{(r+1)} > \frac{r+1}{m}\alpha$. If, for all $r$, $P_{(r)} > \frac{r}{m}\alpha$, then reject none of the hypotheses. More formally, if we define

$$i_0(P) = \max\{i : P_{(i)} \leq \frac{i}{m}\alpha, i = 1, \cdots, m\}, \tag{2}$$

where $P = (P_1, P_2, \cdots, P_m)'$ is the vector of $p$-values, then the BH procedure rejects $i_0$ null hypotheses $H_{(1)}, \cdots, H_{(i_0)}$ and not reject the remaining $m - i_0$ null hypotheses. We shall denote the set of rejected null hypotheses by $R(P) = \{H_{(i)} : 1 \leq i \leq i_0(P)\}$ and the set of accepted hypotheses by $A(P) = \{H_{(i)} : i_0(P) < i \leq m\}$.

The BH and other stepwise procedures, such as Benjamini-Liu and Benjamini-Yekutieli procedures (Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001),

assume that all $p$-values are exact and available to us when testing $H_1, H_2, \ldots, H_m$. However, this is not the case when the marginal $p$-values are estimated using bootstrap methodology.

Based on $B$ bootstrap samples, suppose $P_i^{(B)}$ denotes the estimated $p$-value for the $i^{th}$ null hypothesis $H_i, i = 1, 2, \cdots, m$. Further, suppose that $[c_i^{(B)}, d_i^{(B)}]$ is a "suitable" confidence interval for the ideal bootstrap $p$-value $P_i$, centered at $P_i^{(B)}$, which is used to describe accuracy of the practical bootstrap $p$-value $P_i^{(B)}$.

## 2.2 The algorithm

The algorithm, which is called Algorithm 1, consists of the following steps.

*Step 1:* Fix a non-decreasing sequence of numbers of bootstrap samples $B_0 \leq B_1 \leq \cdots \leq B_N$. Set the significance level as $\alpha$ and the confidence level of confidence intervals as $1 - \beta$. Initialize $i = 0, \widetilde{M} = M$, and $A^{(B)} = R^{(B)} = \emptyset$ (null set).

*Step 2:* Set $B = B_i$. Based on $B$ bootstrap samples, for each hypothesis $H_j \in \widetilde{M}$, obtain bootstrap $p$-value $P_j^{(B)}$ and the corresponding $(1 - \beta)$ level confidence interval $[c_j^{(B)}, d_j^{(B)}]$. For $H_j \notin \widetilde{M}$, that is hypotheses for which decisions have been made at an earlier step, set $P_j^{(B)} = P_j^{(B_{i-1})}$, $c_j^{(B)} = c_j^{(B_{i-1})}$ and $d_j^{(B)} = d_j^{(B_{i-1})}$.

*Step 3:* Apply the BH procedure to the vectors $c^{(B)} = (c_1^{(B)}, \cdots, c_m^{(B)})'$ and $d^{(B)} = (d_1^{(B)}, \cdots, d_m^{(B)})'$, respectively. Let the corresponding set of accepted hypotheses at level $\alpha$ be denoted by $A(c^{(B)})$ and the set of rejected hypotheses be denoted by $R(d^{(B)})$. For convenience, we rewrite $A(c^{(B)})$ as $A^{(B)}$ and $R(d^{(B)})$ as $R^{(B)}$.

*Step 4:* Set $i \leftarrow i + 1$ and $\widetilde{M} \leftarrow \widetilde{M} - \left\{A^{(B)} \bigcup R^{(B)}\right\}$. Repeat Steps 2 and 3 until $\widetilde{M} = \emptyset$ or $i = N$ (a pre-specified sequence length).

*Step 5:* If $\widetilde{M}$ is not empty, apply the BH procedure to the $p$-value vector $P^{(B_N)} = (P_1^{(B_N)}, \cdots, P_m^{(B_N)})$. The final sets of accepted and rejected hypotheses at level $\alpha$ are $A(P^{(B_N)})$ and $R(P^{(B_N)})$, respectively.

**Remark 1** Algorithm 1 may be viewed as a generalization of the pretest procedure of Davidson and McKinnon (2000, section 3). The basic difference between the two is that the Davidson and McKinnon procedure is designed for

standard single hypothesis testing problem, while the present methodology is derived for the multiple hypotheses testing problems.

**Remark 2** Algorithm 1 can be generalized to the situation where decisions on the acceptance and rejections of null hypotheses are made at different levels $\alpha_1$ and $\alpha_2$, respectively. In order to guarantee accuracy of Algorithm 1, we generally set $\alpha_1$ and $\alpha_2$ satisfying $\alpha_2 \leq \alpha \leq \alpha_1$.

**Remark 3** In implementation of Algorithm 1, choice of suitable confidence intervals for ideal bootstrap $p$-values is important. We generally choose the confidence intervals with high coverage probability at a pre-specified level $1-\beta$. For example, we might choose $\beta = 0.01$. In order to guarantee accuracy of Algorithm 1, we can also choose conservative simultaneous confidence intervals for all ideal bootstrap $p$-values at level $1 - \frac{\beta}{2m}$.

**Remark 4** In Algorithm 1, we apply the BH procedure to the less precise $p$-value intervals $[c_j^{(B)}, d_j^{(B)}]$, which construction will be discussed in details in next subsection. For those hypotheses in $A^{(B)} \bigcup R^{(B)}$, we stop computing $p$-values in all future calculations, since we are able to make decisions regarding these hypotheses with less precise $p$-values.

It is interesting to note that BH acceptance and rejection sets enjoy the following monotonicity properties.

**Proposition 1** *Using the notation in Algorithm 1, with* $c^{(B)} = (c_1^{(B)}, \cdots, c_m^{(B)})'$ *and* $d^{(B)} = (d_1^{(B)}, \cdots, d_m^{(B)})'$. *Assume that for each* $i = 1, \cdots, m$, $P_i^{(B)} \in [c_i^{(B)}, d_i^{(B)}]$, *then*

$$A(c^{(B)}) \subseteq A(P^{(B)}) \subseteq A(d^{(B)}) \tag{3}$$

$$R(d^{(B)}) \subseteq R(P^{(B)}) \subseteq R(c^{(B)}). \tag{4}$$

**Remark 5** Proposition 1 demonstrates, even though corresponding ideal bootstrap $p$-values of null hypotheses are unknown, by (3) and (4) we can still make correct decisions of rejection or acceptance on hypotheses in $R(d^{(B)})$ and $A(c^{(B)})$ with confidence on the basis of less precise $p$-value intervals $[c_i^{(B)}, d_i^{(B)}], i = 1, \cdots, m$. Algorithm 1 is based on this simple fact.

We now state the main theorem of this article, which justifies the use of Algorithm 1.

**Theorem 1** *Suppose Algorithm 1 is implemented to its completion then the following results hold:*

*(i) For all* $0 \leq i < N$, $R(d^{(B_i)}) \subseteq R(d^{(B_{i+1})})$ *and* $A(c^{(B_i)}) \subseteq A(c^{(B_{i+1})})$.

*(ii) Let, for each* $j = 1, \cdots, m$, $\widetilde{P}_j^{(B_N)}$ *be the bootstrap p-value for hypothesis* $H_j$ *on the basis of* $B_N$ *bootstrap samples. If* $\widetilde{P}_j^{(B_N)} \in [c_j^{(B_N)}, d_j^{(B_N)}]$, *then* $R(P^{(B_N)}) = R(\widetilde{P}^{(B_N)})$ *and* $A(P^{(B_N)}) = A(\widetilde{P}^{(B_N)})$, *where* $\widetilde{P}^{(B_N)} = (\widetilde{P}_1^{(B_N)}, \cdots, \widetilde{P}_m^{(B_N)})'$.

Some important implications of the above theorem are provided in the following remarks.

**Remark 6** From Theorem 1(i) we deduce that if a null hypothesis has been either accepted or rejected by the BH procedure at an earlier step, then that decision will not change in the subsequent steps of the algorithm.

**Remark 7** According to Theorem 1 (ii), in theory, Algorithm 1 will produce the same overall decisions of acceptance or rejection as the conventional BH procedure. Note that the BH procedure makes decisions based on $p$-values obtained from a pre-specified large number $B_N$ of bootstrap samples.

## 2.3 Implementation of algorithm 1

Implementation of the proposed algorithm requires the computation of confidence intervals $[c_i^{(B)}, d_i^{(B)}], i = 1, 2, \cdots, m$. For the $i^{th}$ hypothesis $H_i$, based on a random sample of size $n$, suppose $x_i$ denotes the observed value of the test statistic. Suppose $B$ bootstrap samples of size $n$ each are drawn for testing $H_i$ and suppose the bootstrap test statistics are $T_{i,1}^* \leq T_{i,2}^* \leq \cdots \leq T_{i,B}^*$. Then the estimated bootstrap $p$-value is given by

$$P_i^{(B)} = \frac{\#\{T_{i,j}^* \geq x_i, j = 1, \cdots, B\}}{B},$$

where the numerator is the number of bootstrap test statistics that are greater than or equal to the observed value $x_i$. The ideal bootstrap $p$-value $P_i$ is the conditional expectation of $P_i^{(B)}$ given $x_i$, i.e., $E(P_i^{(B)}|x_i) = P_i$. The random variable $\#\{T_{i,j}^* \geq x_i, j = 1, 2, \cdots, B\}$ is binomially distributed with $B$ trials and probability of "success" given by $P_i$. Several methods exist in the literature for estimating confidence interval of the form $[c_i^{(B)}, d_i^{(B)}]$ for the unknown ideal bootstrap $p$-value $P_i$. Some of the well-known methods include Wald asymptotic method without or with continuity correction (Vollset, 1993; Blyth and Still, 1983), Wilson score method without or with continuity correction (Wilson, 1927; Blyth and Still, 1983), Clopper-Pearson exact method

(Clopper and Pearson, 1934), and Likelihood-based method (Miettinen and Nurmine, 1985) etc.

Agresti and Coull (1998) and Newcombe (1998) have demonstrated that the Clopper-Pearson exact confidence interval tends to have a very high coverage probability, a desirable feature for our algorithm. Therefore, in implementation of Algorithm 1, we use the Clopper-Pearson exact method to derive the $(1-\beta)$ level confidence interval $[c_i^{(B)}, d_i^{(B)}]$, which is described as follows. Suppose that $X \sim$ Binomial $(B, p)$ and let $H(\cdot)$ be its cumulative distribution function, i.e.,

$$H(B, p, k) = Pr\{X \leq k\}, \quad \text{for} \quad 0 \leq k \leq B.$$

Given $B$ and $k$, an inverse function is defined as a map $H_{B,k}^{-1}(\cdot) : [0,1] \to [0,1]$ satisfying $H\left(B, H_{B,k}^{-1}(p), k\right) = p$ for any $p \in [0,1]$. Using the notation, the lower and upper limits $c_i^{(B)}$ and $d_i^{(B)}$ is expressed as

$$c_i^{(B)} = H_{B,r_i-1}^{-1}(1-\beta/2) \quad \text{and} \quad d_i^{(B)} = H_{B,r_i}^{-1}(\beta/2), \tag{5}$$

where $r_i = \#\{T_{i,j}^* \geq x_i, j = 1, 2, \cdots, B\} = BP_i^{(B)}$, and $c_i^{(B)}$ and $d_i^{(B)}$ can be numerically calculated using the bisection method at a pre-specified accuracy level $\gamma$. Typically, we might set $\gamma = 10^{-6}$.

In implementation of Algorithm 1, we generally choose a high confidence level for confidence intervals and use a simple rule to generate the non-decreasing sequence $B_0 \leq B_1 \leq \cdots \leq B_N$ such as setting $B_{i+1} = 2B_i, 0 \leq i < N$, where $B_N$ is the pre-specified number of bootstrap samples by standard bootstrap methodology (Andrews and Buchinsky, 2000; Davidson and MacKinnon, 2000).

## 2.4 Extension to adaptive procedure

In this subsection, we extend our proposed algorithm to the Storey adaptive BH procedure. Benjamini and Hochberg (1995) showed that for the BH procedure, $FDR \leq \pi_0\alpha$, where $\pi_0 = m_0/m$ is the ratio of the true null hypotheses among all hypotheses. By introducing a simple estimate $\hat{\pi}_0$ of $\pi_0$, Storey (2002) and Storey et al (2004) proposed a more powerful adaptive step-up procedure with critical constants $\alpha_i = \frac{i}{\hat{\pi}_0 m}\alpha, i = 1, \cdots, m$, and showed that the procedure controls the FDR at $\alpha$ under the assumption of independence of the underlying test statistics. Given $p$-value vector $P = (P_1, \cdots, P_m)$, the estimate $\hat{\pi}_0$ is defined as

$$\hat{\pi}_0(P) = \frac{m - Q(P) + 1}{m(1-\lambda)},$$

where $\lambda$ is a pre-specified positive constant, $Q(P) = \sum_{i=1}^{m} I(P_i \leq \lambda)$ is the number of hypotheses with $p$-values less than or equal to $\lambda$, and $I(\cdot)$ denotes the indicator function. Using the same notations as in Algorithm 1, let $P_i^{(B)}$ denote bootstrap $p$-values corresponding to hypotheses $H_i, i = 1, \cdots, m$ on the basis of $B$ bootstrap samples and $[c_i^{(B)}, d_i^{(B)}]$ denote associated confidence intervals. Suppose $A(P^{(B)})$ and $R(P^{(B)})$ respectively denote the sets of accepted and rejected null hypotheses while applying the Storey procedure to the $p$-value vector $P^{(B)} = (P_1^{(B)}, \cdots, P_m^{(B)})'$.

Based on vectors $c^{(B)} = (c_1^{(B)}, \cdots, c_m^{(B)})'$ and $d^{(B)} = (d_1^{(B)}, \cdots, d_m^{(B)})'$, two new estimates $\hat{\pi}_1^0 = \pi_0(c^{(B)})$ and $\hat{\pi}_2^0 = \pi_0(d^{(B)})$ of $\pi_0$ can be constructed. Based on $\hat{\pi}_1^0$ and $\hat{\pi}_2^0$, two different Storey procedures can be obtained. We apply the first Storey procedure to to $c^{(B)}$ and the second one to $d^{(B)}$. Suppose $A_1(c^{(B)})$, $R_1(c^{(B)})$ and $A_2(d^{(B)})$, $R_2(d^{(B)})$ are the corresponding sets of accepted and rejected null hypotheses, respectively. Similar to Proposition 1, we can obtain the following result.

**Proposition 2** *Assume that for each* $i = 1, \cdots, m$, $P_i^{(B)} \in [c_i^{(B)}, d_i^{(B)}]$, *then*

$$A_1(c^{(B)}) \subseteq A(P^{(B)}) \subseteq A_2(d^{(B)}) \tag{6}$$

$$R_2(d^{(B)}) \subseteq R(P^{(B)}) \subseteq R_1(c^{(B)}). \tag{7}$$

Based on Proposition 2, if we use more powerful Storey procedure instead of BH procedure in our proposed algorithm, we only need to modify Step 3 of Algorithm 1 as follows.

*Step 3′.* Apply the first Storey procedure with estimate $\hat{\pi}_1^0$ to $c^{(B)} = (c_1^{(B)}, \cdots, c_m^{(B)})'$ and the second one with estimate $\hat{\pi}_2^0$ to $d^{(B)} = (d_1^{(B)}, \cdots, d_m^{(B)})'$, respectively. Let $A_1(c^{(B)})$ be the corresponding set of accepted hypotheses at level $\alpha$ for the first procedure and $R_2(d^{(B)})$ be the set of rejected hypotheses for the second one. For convenience, we rewrite $A_1(c^{(B)})$ as $A^{(B)}$ and $R_2(d^{(B)})$ as $R^{(B)}$, respectively.

# 3 Simulation study

We performed a simulation study to evaluate the reduction in the number of bootstrap replicates achieved by the proposed algorithm, relative to the standard BH procedure. To generate a gene expression data with a realistic correlation structure between gene expressions, we used the data of Lobenhofer et al (2002). The published data consisted of gene expression of 1,900 probes at 6 different time points with 8 cDNA microarray chips at each time point.

Thus, there were 48 chips with each chip containing 1,900 probes. We randomly selected 8 chips from 48 chips (with replacement) and assigned them to Group 1 and another random sample (with replacement) of 8 chips to Group 2. Thus, two new groups were created which preserved the underlying correlation structure among the 1,900 probes. Further, apart from any difference due to randomization, there was no difference between Group 1 and Group 2. To create a non-null data, we added a value of 1 to the last $\pi_1 = 5\%$ or $10\%$ probes in Group 2. Thus either 95% or 90% of the probes are not differentially expressed between the two groups.

In this simulation study we chose the FDR levels as $\alpha = 0.01$ or $0.05$ and performed two-sided tests with symmetric bootstrap $p$-values as

$$P_i^{(B)} = \#\left\{|T_{i,j}^*| \geq |x|, j = 1, \cdots, B\right\}/B, \text{ for } i = 1, \cdots, m.$$

We choose a confidence level $1 - \beta = 0.99$ for the Clopper-Peason exact confidence intervals of ideal bootstrap $p$-values, which is numerically solved by the bisection method with accuracy level, $\gamma = 10^{-6}$. We use a simple rule to generate the non-decreasing sequence $B_0 \leq B_1 \leq \cdots \leq B_N$. For $B_0$, we set $B_0 = 125, 250$ or 500; for $B_N$, we set $B_N = 2,000$; for other $B_i$'s, we set $B_{i+1} = 2B_i, 0 \leq i < N$. We repeated this simulation experiment 200 times for estimating various performance criteria described below.

In the simulation study, we compared the proposed Algorithm 1 with the conventional BH procedure based on $B_N = 2,000$ bootstrap samples. The performance criteria of our interest are:

- *Average number of bootstrap samples (aveB):* It is defined as $aveB = \frac{1}{m}\sum_{i=0}^{N} M_i B_i$, where $M_i$ is the number of tested hypotheses in the $i^{th}$ stage of Algorithm 1.

- *Fold reduction of number of bootstraps (Fold red.):* It is defined as $B_N/aveB$.

- *Degree of consistency with the standard procedure (Degree cons. ):* It is defined as $100 \cdot (1 - \frac{|FR_1 - FR_2|}{FR_2})$, where $FR_1$ and $FR_2$ are respectively the numbers of rejected hypotheses using Algorithm 1 and the conventional BH procedure. If $FR_1 \geq FR_2$, then $\frac{|FR_1 - FR_2|}{FR_2}$ is interpreted as the percent increase in false discoveries relative to the regular procedure; otherwise, it is interpreted as the ratio of missing discoveries relative to the regular procedure. Theoretically we expect that, whenever the number of rejections is same between the two procedures, the same sets of hypotheses are rejected by both procedures. This is because, the operation of multiple testing procedures involves first ranking the hypotheses

Table 1: Simulation results.

| | | $\pi_1 = 0.05$ | | | $\pi_1 = 0.1$ | | |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $B_0$ | $aveB$ | Fold red. | Degree cons. (%) | $aveB$ | Fold red. | Degree cons. (%) |
| 0.01 | 125 | 317 | 6.3 | 98 | 488 | 4.1 | 99 |
| | 250 | 431 | 4.6 | 100 | 589 | 3.4 | 99 |
| | 500 | 651 | 3.1 | 100 | 788 | 2.5 | 99 |
| 0.05 | 125 | 357 | 5.6 | 99 | 534 | 3.7 | 99 |
| | 250 | 478 | 4.2 | 95 | 640 | 3.1 | 100 |
| | 500 | 686 | 2.9 | 99 | 833 | 2.4 | 98 |

based on the individual $p$-values and then choosing cutoff based on the rankings.

The simulation results are reported in Table 1. The degree of consistency of our proposed algorithm with the BH procedure was very high. For almost all settings of parameters, it is more than 98%. Compared with the conventional BH procedure, the average number of the bootstrap samples was greatly reduced while applying the proposed algorithm. Although the gain in efficiency depends upon the initial number of bootstraps $B_0$, it was usually very high. According to our simulations, even in the worst case scenario it achieved almost a 2-fold reduction and in the best case scenario it was as much as 6-fold reduction. In addition, the degree of consistency of our proposed algorithm was not affected by the choice of $B_0$ (Table 1). Thus the proposed algorithm would require nearly one fifth to two fifths of the time the conventional BH procedure would take while guaranteing high degree of consistency of decisions with the BH procedure.

# 4 Illustration

We illustrate the proposed methodology by applying it to the time-course data of Lobenhofer et al (2002), which was briefly described in the previous section. The goal is to select significant genes and cluster them by their expression patterns over the six time points, namely, 1, 4, 12, 24, 36 and 48 hours after the breast cancer cells were treated with estradiol. For each gene, we establish a null hypothesis versus an alternative as in Peddada et al (2003). We then applied the order-restricted inference methodology of Peddada et al (2003) and

Peddada et al (2005) to derive the $p$-value corresponding to each test needed for using the proposed methodology.

Briefly, the order-restricted inference methodology of Peddada et al (2003) and Peddada et al (2005) is useful for selecting significant genes and clustering them by their expression patterns. Unlike many of the methods available in the literature for analyzing time-course data (e.g. Storey et al, 2005; Liu et al, 2005), this methodology does not use a mathematical/statistical model to describe the relationship between gene expression and the explanatory variable such as time/dose values. Instead, it exploits the underlying mathematical inequalities between mean expression values. As a consequence, in addition to time-course/dose-response studies, the proposed methodology is also applicable to studies where the explanatory variable is an ordinal variable, such as, for clustering gene expressions by severity of a lesion (normal-hyperplasia-adenoma-carcinoma).

As in Peddada et al (2003) we considered 10 different patterns of mean expression associated with the Lobenhofer et al (2002) data: decreasing with time, umbrella pattern with peak at 4, 12, 24 and 36 hours, increasing pattern with time, inverted umbrella pattern with minimum at 4, 12, 24 and 36 hours. We set the FDR level to be 0.05, $B_0$ =1,250, and $B_N$ =10,000. As in Section 3 we use the same confidence level 99% and the same double rule to generate the sequence $B_i$'s.

The proposed methodology selected a total of 206 significant genes and the conventional BH procedure selected 201 significant genes with the number of bootstraps $B_N = 10,000$. Among these two lists, 199 genes are common and all of the top 50 genes selected by Peddada et al (2003) were also selected by the proposed algorithm. The average number of bootstraps required by the proposed algorithm was only 2,175 in comparison to $B_N$ =10,000. Thus the number of bootstraps required by the proposed adaptive method is about 22% of the "full" bootstrap of the conventional method, an almost 4.5 times reduction in the number of bootstraps, a very substantial reduction. On our computer, the conventional BH procedure took about 16 minutes. Whereas our procedure took 2 minutes.

# 5 Discussion

The FDR methodology has been extensively applied in large scale multiple testing problems where the BH and Storey procedures are two most commonly used methods for controlling the FDR. Similar to other stepwise procedures in multiple testing, usage of the BH and Storey procedures relies on the availabil-

ity of true $p$-values. In many instances, especially in the context of high dimensional data such as microarray data, the true $p$-values are usually unavailable and one has to contend with bootstrap $p$-values. Unfortunately, one cannot obtain the true bootstrap $p$-values (known as ideal bootstrap $p$-values) either, because that would entail drawing all possible bootstrap samples. Hence, typically one needs to contend with the practical bootstrap $p$-value, an estimate of ideal bootstrap $p$-value obtained from a finite number of bootstrap samples. The precision of the estimate depends upon the number of bootstraps drawn. Existing literature suggests that researchers have used a wide range of number of bootstrap samples to estimate the ideal bootstrap $p$-value. Depending upon the application and context some have used as few as 200 bootstraps and as many as 10,000 bootstraps (e.g. Suzuki and Shimodaira, 2006). Some even use as many as 100,000 bootstraps (Meuwissen and Goddard, 2004). Obviously, larger the number of bootstraps the more precise is the estimated bootstrap $p$-value.

Clearly, if the number of bootstraps is small then the list of null hypotheses rejected by a stepwise FDR controlling procedure may greatly change with repeated application of the procedure. More severely, it may lose the control of the FDR. On the other hand, increasing the number of bootstraps to a very large number will certainly come at the expense of computation time, especially in the face of high dimensional data. Therefore it is crucial to develop adaptive algorithms such as the one described here to reduce the total number of computations without sacrificing the control of the FDR. Computational issues raised in this article are relevant to any large scale multiple testing problem involving high-dimensional data, and is not limited to microarray data.

We provided a simple adaptive algorithm for determining the number of bootstrap samples in large scale multiple testing. The algorithm not only reduces the average number of bootstraps substantially but it also produces results consistent with the conventional BH procedure based on a large number of bootstraps. In addition, it also maintains the FDR to a desired level. As we extended the proposed idea to Storey's procedure, in a similar way the idea could also be applied to other multiple testing procedures to achieve significant reduction in the number of computations.

In passing, we like to comment about the difference between the nonparametric bootstrap and the permutation methodology for estimating the $p$-vlue of a test. While nonparametric bootstrap, considered in this paper, is a resampling procedure where the samples are drawn with replacement, the permutation methodology is a resampling procedure where samples are drawn without replacement. The permutation $p$-value based on B permutations is

given by $\hat{P}^{(B)} = \#\left\{1 + T_j^* \geq x, j = 1, \cdots, B\right\}/(B+1)$. Both, bootstrap as well as permutation procedures have some desirable features. For instance, the permutation procedure satisfies the required stochastic ordering under the null hypothesis, i.e. $Pr_{H_0}\left\{\hat{P}^{(B)} \leq x\right\} \leq x$, for all $0 \leq x \leq 1$. Although this is not necessarily true for bootstrap for small sample sizes, it is asymptotically satisfied. Furthermore, relative to bootstrap, the permutation tests are computationally less intensive. However, a distinct advantage of bootstrap is that it is more broadly applicable than the permutation tests. It is useful for testing the equality of various parameters of interest rather than the equality of distributions.

In conclusion we believe that, with the influx of high-dimensional data from high-throughput technologies, there is a need for developing computationally efficient algorithms for multiple testing problems that control the FDR (e.g. Datta and Datta, 2005). Although a considerable literature exists on the multiple comparison procedures for high dimensional data, almost no work has been done on the computational challenges associated with this problem. One exception is Ge, Dudoit and Speed (2003), in which computational issues associated with implementation of Westfall and Young (1993)'s resampling-based minP method are discussed carefully and a fast algorithm for implementing the minP method are proposed. In addition, a few computational issues have also been discussed in the context of traditional small scale multiple testing problems, such as the closed testing method of Marcus, Peritz and Gabriel (1976). Although our solution seems to perform well for the problem considered in this article, more needs to be done along these lines.

# APPENDIX

## A Proofs

We now provide the proofs of the results stated in Section 2.

**Lemma 1** *For any $m$-dimensional vector $P$, $i_0(P)$ defined as (2) is decreasing in each component of $P$.*

*Proof of Lemma 1:* Consider two $m$-dimensional vectors $X = (X_1, \cdots, X_m)'$ and $Y = (Y_1, \cdots, Y_m)'$, we say $X \geq Y$ if and only if $X_i \geq Y_i$ for any $i = 1, \cdots, m$. Let the ordered forms of these two vectors be denoted by $\widetilde{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(m)})'$ and $\widetilde{Y} = (Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)})'$, where $X_{(1)} \leq X_{(2)} \leq$

$\cdots \leq X_{(m)}$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$. Note that, if $X \geq Y$, then $\widetilde{X} \geq \widetilde{Y}$. The result follows by recalling the definition of $i_0(P)$.

*Proof of Proposition 1:* Note that, for each $i = 1, 2, \cdots, m$, $c_i^{(B)} \leq P_i^{(B)} \leq d_i^{(B)}$ and hence $c^{(B)} \leq P^{(B)} \leq d^{(B)}$. Applying Lemma 1, we have

$$i_0(c^{(B)}) \geq i_0(P^{(B)}) \geq i_0(d^{(B)}). \tag{A.1}$$

For any $H_i \in A(c^{(B)})$, we note

$$c_i^{(B)} > \frac{i_0(c^{(B)})}{m}\alpha \geq \frac{i_0(P^{(B)})}{m}\alpha. \tag{A.2}$$

Thus $P_i^{(B)} > \frac{i_0(P^{(B)})}{m}\alpha$. Hence we deduce that $H_i \in A(P^{(B)})$, implying $A(c^{(B)}) \subseteq A(P^{(B)})$. Similarly, it can be demonstrated that $R(d^{(B)}) \subseteq R(P^{(B)})$.

Since $A(P^{(B)}) \bigcup R(P^{(B)}) = M$, therefore

$$\begin{aligned} R(d^{(B)}) &\subseteq R(P^{(B)}) = M - A(P^{(B)}) \\ &\subseteq M - A(c^{(B)}) = R(c^{(B)}). \end{aligned} \tag{A.3}$$

Similarly,

$$\begin{aligned} A(c^{(B)}) &\subseteq A(P^{(B)}) = M - R(P^{(B)}) \\ &\subseteq M - R(d^{(B)}) = A(d^{(B)}). \end{aligned} \tag{A.4}$$

*Proof of Theorem 1:* (i) For any $0 \leq i < N$, let the ordered values of $d_1^{(B_i)}, \cdots, d_m^{(B_i)}$ be denoted by $d_{(1)}^{(B_i)} \leq \cdots \leq d_{(m)}^{(B_i)}$. Then by the definition of $i_0(P)$ we have

$$d_{(k)}^{(B_i)} \leq \frac{i_0(d^{(B_i)})}{m}\alpha, \quad k = 1, \cdots, i_0(d^{(B_i)}).$$

From step 2 of Algorithm 1, we know that

$$d_{(k)}^{(B_{i+1})} = d_{(k)}^{(B_i)}, \quad k = 1, \cdots, i_0(d^{(B_i)}). \tag{A.5}$$

Therefore

$$d_{(k)}^{(B_{i+1})} \leq \frac{i_0(d^{(B_i)})}{m}\alpha, \quad k = 1, \cdots, i_0(d^{(B_i)}). \tag{A.6}$$

By the definition of $i_0(P)$ we have

$$i_0(d^{(B_i)}) \leq i_0(d^{(B_{i+1})}). \tag{A.7}$$

For any $H_j \in R(d^{(B_i)})$, we note $d_j^{(B_i)} \leq \frac{i_0(d^{(B_i)})}{m}\alpha$. By (A.6) and (A.7) we have

$$d_j^{(B_{i+1})} = d_j^{(B_i)} \leq \frac{i_0(d^{(B_i)})}{m}\alpha \leq \frac{i_0(d^{(B_{i+1})})}{m}\alpha. \tag{A.8}$$

Therefore, we deduce that $H_j \in R(d^{(B_{i+1})})$, implying $R(d^{(B_i)}) \subseteq R(d^{(B_{i+1})})$. Through a set of similar arguments, we conclude that $A(c^{(B_i)}) \subseteq A(c^{(B_{i+1})})$.

(ii) Note that for each $j = 1, 2, \cdots, m$, $\widetilde{P}_j^{(B_N)} \in [c_j^{(B_N)}, d_j^{(B_N)}]$. By Proposition 1 we have

$$R(d^{(B_N)}) \subseteq R(\widetilde{P}^{(B_N)}) \text{ and } A(c^{(B_N)}) \subseteq A(\widetilde{P}^{(B_N)}). \tag{A.9}$$

From Algorithm 1, we know that for any $H_j \notin R(d^{(B_N)}) \bigcup A(c^{(B_N)})$, $P_j^{(B_N)} = \widetilde{P}_j^{(B_N)}$. Based on the above facts, we show in the following that $i_0(P^{(B_N)}) = i_0(\widetilde{P}^{(B_N)})$.

Assume that $i_0(P^{(B_N)}) \neq i_0(\widetilde{P}^{(B_N)})$. Without loss of generality, we suppose $i_0(P^{(B_N)}) < i_0(\widetilde{P}^{(B_N)})$. For simplicity's sake, let $i_0 = i_0(P^{(B_N)})$ and $i_0' = i_0(\widetilde{P}^{(B_N)})$. By the definition of $i_0(P)$, we have

$$P_{(i_0')}^{(B_N)} > \frac{i_0'}{m}\alpha \geq \widetilde{P}_{(i_0')}^{(B_N)}.$$

Thus,

$$H_{(i_0')} \in A(P^{(B_N)}) \bigcap \left( R(d^{(B_N)}) \bigcup A(c^{(B_N)}) \right). \tag{A.10}$$

Note that in Algorithm 1, $P^{(B_N)} \leq d^{(B_N)}$. By Proposition 1 and (A.10), we have $H_{(i_0')} \in A(c^{(B_N)})$. Similarly, by Proposition 1 and the definition of $i_0(\widetilde{P}^{(B_N)})$, we have

$$H_{(i_0')} \in R(\widetilde{P}^{(B_N)}) \subseteq R(c^{(B_N)}),$$

which leads logically to a contradiction. Therefore, $i_0(P^{(B_N)}) = i_0(\widetilde{P}^{(B_N)})$, and then the result follows.

*Proof of Proposition 2:* Note that $P_i^{(B)} \in [c_i^{(B)}, d_i^{(B)}]$ for all $i = 1, \cdots, m$, therefore $c^{(B)} \leq P^{(B)} \leq d^{(B)}$. Then

$$Q(d^{(B)}) \leq Q(P^{(B)}) \leq Q(c^{(B)}) \text{ for any given } \lambda.$$

It follows that $\hat{\pi}_1^0 \leq \hat{\pi}_0 \leq \pi_2^0$. By the definition of $i_0(P)$, we have

$$i_0(c^{(B)}) = \max \left\{ i : c_{(i)}^{(B)} \leq \frac{i}{\hat{\pi}_1^0 m}\alpha \right\}$$

$$i_0(P^{(B)}) = \max\left\{ i : P^{(B)}_{(i)} \leq \frac{i}{\hat{\pi}_0 m}\alpha \right\},$$

where $c^{(B)}_{(1)} \leq \cdots \leq c^{(B)}_{(m)}$ is the ordered values of $c^{(B)}$. Then we have $i_0(c^{(B)}) \geq i_0(P^{(B)})$. For any $H_i \in A_1(c^{(B)})$, we have

$$P^{(B)}_i \geq c^{(B)}_i > \frac{i_0(c^{(B)})}{\hat{\pi}^0_1 m}\alpha \geq \frac{i_0(P^{(B)})}{\hat{\pi}_0 m}\alpha. \tag{A.11}$$

Hence $H_i \in A(P^{(B)})$, implying $A_1(c^{(B)}) \subseteq A(P^{(B)})$. Similarly, we can show $R_2(d^{(B)}) \subseteq R(P^{(B)})$. The result follows by using $A(P) \bigcup R(P) = M$ for any vector $P$.

# References


[1] Agresti, A. and Coull, B. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52** 119-126.

[2] Andrews, D. and Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitons. *Econometrica* **68** 23-51.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57** 289-300.

[4] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25** 60-83.

[5] Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika* **93** 491-507.

[6] Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* **82** 163–170.

[7] Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29** 1165–1188.

[8] Benjamini, Y., Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics* **171** 783-790.

[9] Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association* **78** 108-116.

[10] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26** 404-413.

[11] Datta S and Datta S. (2005). Empirical Bayes screening of many $p$-values with applications to microarray studies. *Bioinformatics* **21** 1987-1994.

[12] Davidson, R. and MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews* **19** 55-68.

[13] Efron, B., Tibshirani, T., Storey, J. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96** 1151-1160.

[14] Ge, Y., Dudoit, S. and Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12** 1-77.

[15] Genovese, Lazar, N. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870-878.

[16] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

[17] Hommel, G. and Hoffmann, T. (1987). Controlled uncertainty. In *Multiple Hypothesis Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.) 154-161. Springer, Heidelberg.

[18] Korn, E., Troendle, J., McShane, L., and Simon, R. (2004). Controlling the number of false discovery: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* **124** 379–398.

[19] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics* **33** 1138–1154.

[20] Liu, H., Tarima, S., Borders, A., Getchell, T., Getchell, M and Stromberg, A. (2005). Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray. *BMC Bioinformatics* **6** (106).

[21] Lobenhofer, E., Bennett, L., Cable, P., Li, L., Bushel, P. and Afshari, C. (2002). Regulation of DNA replication fork genes by 17 beta-estradiol. *Molecular Endorcrinology* **16** 1215–1229.

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

[23] Mehrotra, D. V. and Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research* **13** 227–238.

[24] Meuwissen, T. and Goddard, M. 2004. Bootstrapping of gene-expression data improves and controls the false discovery rate of differentially expressed genes. *Genetics Selection Evolution* **36** 191–205.

[25] Miettinen, O. S. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4** 213-226.

[26] Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17** 857-872.

[27] Peddada, S., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C. and Umbach, D. (2003). Gene selection and clustering for time-course and dose response microarray experiments using order-restricted inference. *Bioinformatics* **19** 834-841.

[28] Peddada, S., Harris, S., Zajd, J. and Harvey, E., (2005). ORIGEN: order restricted inference for ordered gene expression data. *Bioinformatics* **21** 3933-3934.

[29] Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19** 368-375.

[30] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics* **30** 239–257.

[31] Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *The Annals of Statistics*, In press.

[32] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 479-498.

[33] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultanaeous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 187-205.

[34] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100** 9440-9445.

[35] Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22** 1540-1542.

[36] van der Laan, M., Dudoit, S., and Pollard, K. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3**(1) Article 15.

[37] Vollset, S. E. (1993). Confiidence intervals for a binomial proportion. *Statistics in Medicine* **12** 809-824.

[38] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons.

[39] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22** 1457-1483.