Statistica Sinica **18**(2008), 881-904

SOME STEP-DOWN PROCEDURES CONTROLLING THE FALSE DISCOVERY RATE UNDER DEPENDENCE

Yongchao Ge[1], Stuart C. Sealfon[1] and Terence P. Speed[2,3]

[1] *Mount Sinai School of Medicine,* [2] *University of California at Berkeley*
[3] *The Walter and Eliza Hall Institute of Medical Research*

Abstract: Benjamini and Hochberg (1995) proposed the false discovery rate (FDR) as an alternative to the familywise error rate (FWER) in multiple testing problems. Since then, researchers have been increasingly interested in developing methodologies for controlling the FDR under different model assumptions. In a later paper, Benjamini and Yekutieli (2001) developed a conservative step-up procedure controlling the FDR without relying on the assumption that the test statistics are independent.

In this paper, we develop a new step-down procedure aiming to control the FDR. It incorporates dependence information as in the FWER controlling step-down procedure given by Westfall and Young (1993). This new procedure has three versions: lFDR, eFDR and hFDR. Using simulations of independent and dependent data, we observe that the lFDR is too optimistic for controlling the FDR; the hFDR is very conservative; and the eFDR a) seems to control the FDR for the hypotheses of interest, and b) suggests the number of false null hypotheses. The most conservative procedure, hFDR, is proved to control the FDR for finite samples under the subset pivotality condition and under the assumption that the joint distribution of statistics from true nulls is independent of the joint distribution of statistics from false nulls.

Key words and phrases: Adjusted p-value, false discovery rate, familywise error rate, microarray, multiple testing, resampling.

1. Introduction

The problems associated with multiple hypothesis testing have become greater with the advent of massively parallel experimental assays, especially microarrays. A naive approach of rejecting the hypotheses whose p-values are no greater than 0.05 will lead to 500 false positives on average for a microarray experiment that measures 10,000 genes. Therefore some form of error rate to control false positives is required. Traditionally, this error rate was represented by the familywise error rate (FWER), which is defined as the probability of erroneously rejecting at least one null hypothesis. Benjamini and Hochberg (1995) proposed the false discovery rate (FDR) as an alternative to the FWER. The FDR is defined to be the expected proportion of erroneously rejected null hypotheses among all rejected ones. The FDR can identify more putatively significant hypotheses than the FWER, and meaning of the FDR can be intuitively explained. Hence researchers are increasingly using the new error rate FDR, especially in exploratory analyses. If interest is in finding as many false null hypotheses as possible among thousands of tests, this new error measure seems more appropriate than the FWER. The FWER is probably more stringent than what most researchers want in the exploratory phase, as it permits no more than a single null hypothesis being erroneously rejected.

A particular problem in microarray data analysis is identifying differentially expressed genes among thousands of them. Dudoit, Yang, Callow and Speed (2002) and Westfall, Zaykin and Young (2001) used the maxT step-down approach to control the FWER. Ge, Dudoit and Speed (2003) introduced a fast algorithm implementing the minP step-down adjustment controlling the FWER. In another direction, Tusher, Tibshirani and Chu (2001), Efron et al. (2001), and Storey (2002) adopted the FDR approach in microarray experiments. This paper is mostly about a new algorithm to compute the FDR. Although we focus on microarray applications, as with the maxT and minP step-down adjustment in the previous paper Ge, Dudoit and Speed (2003) the new algorithm can be applied to other similar multiple testing situations as well.

Section 2 reviews the basic notions of multiple testing, and the concept of the FDR. Section 3 introduces some related work on procedures controlling the FDR. Section 4 presents three versions of our new step-down procedure aiming to control the FDR, gives some theoretical properties for the most conservative version (hFDR), and presents the resampling algorithm. Section 5 describes simulation results on the three versions of the new procedure introduced in Section 4. The microarray applications are in Section 6 and, finally Section 7 summarizes our findings and open questions.

2. Multiple Testing and False Discovery Rates

Assume that there are m pre-specified null hypotheses $\{H_1, \ldots, H_m\}$ to be tested. Given observed data $\boldsymbol{X} = \boldsymbol{x}$, a statistic t_i is used for testing hypothesis H_i, and p_i is the corresponding p-value (*a.k.a raw p-value, marginal p-value or unadjusted p-value*). In the microarray setup, the observed data $\boldsymbol{x}$ is an $m \times n$ matrix of gene expression levels for m genes and n RNA samples from treated and control groups, while t_i might be a two-sample Welch (1938) t-statistic computed from the expression levels of gene i. The null hypothesis H_i is that gene i is non-differentially expressed between the treated group and the control group. The biological question is how to find as many *differentially* expressed genes (*false*

Table 1. Summary table for multiple testing problems.

	Set	Number of decisions		
		reject	accept	total
true null	M_0	false positives: V	correct decisions: $m_0 - V$	m_0
false null	M_1	correct decisions: S	false negatives: $m_1 - S$	m_1
total	M	R	$m - R$	m

null hypotheses) as possible without having too many *false positives*. A false positive occurs when an unaffected gene is claimed to be differentially expressed (falsely rejecting a null hypothesis). In this paper, we adopt the notation that the observed values are denoted by lower case letters, t_i, p_i for example, and the corresponding random variables are denoted by upper case letters, T_i, P_i for example. For the sake of convenience, we always assume a two-sided test, i.e., $p_i = P(|T_i| \geq |t_i| \mid H_i)$. Let the set of true null hypotheses be M_0, the set of false null hypotheses be M_1, and the full set be $M = M_0 \cup M_1$. Let $m_0 = |M_0|$ and $m_1 = |M_1|$, where $| \cdot |$ denotes the cardinality of a set. Given a rejection procedure, let V denote the number of erroneously rejected null hypotheses and R the total number of rejected ones. Then $R - V$ is the number of correctly rejected hypotheses, denoted by S. The values of V, R and S are determined by the particular rejection procedure and the significance level α, say 0.05. Table 1 shows possible outcomes of a rejection procedure.

For any set $K = \{i_1, \ldots, i_k\} \subseteq M$, let H_K denote the partial joint null hypothesis associated with the set K. H_M is called the *complete null* as every null hypothesis is true. Traditionally, the *familywise error rate* (FWER) has been widely used. The FWER is defined as the probability of erroneously rejecting at least one null hypothesis, i.e., $P(V > 0)$. Since the above probability is computed under the restriction H_{M_0}, using the notations from Sarkar (2002), we have

$$\text{FWER} = P(V > 0 \mid H_{M_0}).$$

Westfall and Young (1993) gives a comprehensive introduction to this subject while focusing on a resampling-based adjusted p-values approach. Let Q be the false discovery proportion (Korn et al. (2004)): the proportion of erroneously rejected null hypotheses among all rejected ones,

$$Q = \frac{V}{R}, \tag{1}$$

where the ratio is defined to be zero when the denominator is zero. The *false discovery rate* (FDR) is defined to be the expectation of Q,

$$\text{FDR} = E(Q \mid H_{M_0}) = E\Big(\frac{V}{R} \Bigm| H_{M_0}\Big).$$

There are three kinds of FDR control.

- *Weak control*: control of the FDR under the complete null H_M, i.e., for any $\alpha \in (0,1)$,
$$E(Q \mid H_M) \leq \alpha.$$
Under H_M, we have $Q = I(V > 0)$, where $I(\cdot)$ is the indicator function, so weak control of the FDR is equivalent to weak control of the FWER.
- *Exact control*: control of the FDR for the true null set M_0, i.e., for any $\alpha \in (0,1)$,
$$E(Q \mid H_{M_0}) \leq \alpha.$$
The definition is applicable only when the true null set M_0 is known.
- *Strong control*: control of the FDR for all possible $M_0 \subseteq M$, i.e., for any $\alpha \in (0,1)$,
$$\max_{M_0 \subseteq M} E(Q \mid H_{M_0}) \leq \alpha.$$

In practice, we do not know the true null set M_0, and weak control is not satisfactory. It is important to have a strong control. In fact, the majority of existing FDR procedures offer strong control. The first known procedure with proof of strong control of the FDR was provided by Benjamini and Hochberg (1995).

The BH Procedure: Let the indices of the ordered p-values be $d_1, \ldots, d_m$ such that $p_{d_1} \leq \cdots \leq p_{d_m}$. The $\{d_i\}$ are determined by the observed data $\boldsymbol{X} = \boldsymbol{x}$. Fix $\alpha \in (0,1)$. Define $i^* = \max\{i : p_{d_i} \leq c_i\}$, where $c_i = \alpha i/m$. Then reject $H_{d_1}, \ldots, H_{d_{i^*}}$ if i^* exists; otherwise reject no hypotheses.

The BH procedure is a step-up procedure because it begins with the largest p-value p_{d_m} to see if H_{d_m} can be accepted with the critical value c_m, and then $p_{d_{m-1}}$, until $p_{d_{i^*}}$ which can not be accepted any more as $p_{d_{i^*}} \leq c_{i^*}$. Benjamini and Hochberg (1995) proved that their procedure provides strong control of the FDR when the p-values from the true null hypotheses M_0 are independent. The ideas of the BH procedure appeared much earlier in seminal papers by Eklund (1961-1963) as mentioned in Seeger (1968), and later were rediscovered independently in Simes (1986). Eklund (1961-1963) even motivated the procedure by the definition of "proportion of false significances", which is equivalent to the Q of equation (1). Seeger (1968) and Simes (1986) also gave the proof that this procedure controls the FWER in the weak sense. Similar ideas for controlling the FDR also appeared in Sorić (1989). However, Benjamini and Hochberg's proof of strong control of the FDR has accelerated the usage of the BH procedure and the FDR concept.

3. Some Related Previous Works

Since the first groundbreaking paper on the FDR by Benjamini and Hochberg (1995), different procedures have been proposed to offer strong control of the

FDR under different conditions. A later work by Benjamini and Yekutieli (2001) relaxed the independence assumption to certain dependence structures, namely, when the underlying statistics are *positive regression dependent on a subset of the true null hypotheses* M_0 (PRDS). There are other works which modify the critical values c_i to produce a more powerful procedure, i.e., to reject more hypotheses while still offering strong control of the FDR at the same significance level α. For example, Kwong, Holland and Cheung (2002) used the same critical values $c_i = \alpha i/m$, for $i = 1, \ldots, m-1$, but with a different definition of c_m. Their c_m is always no less than the c_m defined in the BH procedure, and so their procedure is at least as powerful as the BH. Benjamini and Liu (1999) (hereafter called BL) derived a different set of critical values c_i for a step-down procedure when the underlying test statistics are independent.

The BL Procedure: Fix $\alpha \in (0,1)$. Let i^* be the largest i such that $p_{d_1} \leq c_1, \ldots, p_{d_i} \leq c_i$, where $c_i = 1 - [1 - \min\big(1, \alpha m/(m-i+1)\big)]^{1/(m-i+1)}$. Then reject $H_{d_1}, \ldots, H_{d_{i^*}}$ if i^* exists; otherwise reject no hypotheses.

Sarkar (2002) strengthened the Benjamini and Yekutieli (2001) results for the BH procedure in a much more general step-wise framework. He also relaxed the assumption of the BL procedure, from independence to a weak dependence condition: Sarkar assumed that the underlying test statistics exhibit the *multivariate total positivity of order* 2 (MTP_2) under any alternatives, and that the test statistics are exchangeable when the null hypotheses are true. There are other works which modify the BH procedure by multiplying the p-values with an estimate of $\pi_0 = m_0/m$ (Benjamini and Hochberg (2000), Storey (2002) and Storey and Tibshirani (2001)).

Most of the papers mentioned above deliver strong control of the FDR under the assumption that the test statistics are independent, or under the assumption that the expectation of some statistics of the dependent data can be bounded by that of independent data, as in the PRDS or in the MTP_2 cases. Hence, these works are able to generalize the results from independence to a weak dependence situation. However, there still seems to be a need to develop a procedure that can be applied to less independent data. For example, Troendle (2000) derived different step-up and step-down procedures for multivariate normal data. Yekutieli and Benjamini (1999) used a resampling procedure to compute FDR adjusted p-values under dependence. In their paper, they proposed FDR adjusted p-values, which are similar in concept to the FWER adjusted p-values in Westfall and Young (1993).

Given a particular rejection procedure, the FDR adjusted p-value for a hypothesis is the smallest level at which H_i is rejected while controlling the FDR.

i.e.,

$$\tilde{p}_i = \inf\{\alpha \mid H_i \text{ is rejected at FDR} = \alpha\}.$$

The FDR adjusted p-values are determined by a particular rejection procedure. For example, with the BH procedure, the adjusted p-values are

$$\tilde{p}_{d_i} = \min_{k=i,\ldots,m} \Big\{ \min\Big(p_{d_k} \cdot \frac{m}{k}, 1\Big)\Big\}.$$

On the other hand, if a procedure can assign adjusted p-values, then for any given $\alpha \in (0,1)$ we reject all hypotheses whose adjusted p-values are no greater than α. Yekutieli and Benjamini (1999) proposed a resampling-based FDR local estimator (see equations 9 and 10 of their paper):

$$\hat{Q}(p) = E\Big\{\frac{R^*(p)}{R^*(p) + \hat{s}(p)}\Big\}. \tag{2}$$

Here the expectation of equation (2) is computed by resampling under the complete null H_M, and $R^*(p)$ is a random variable defined as $\#\{i \in M : p_i \le p\}$, while $\hat{s}(p)$ is an estimate of the data dependent number $s(p) = \#\{i \in M_1 : p_i \le p\}$. The $s(p)$ is generally unobservable, as we do not have any information on M_1. Yekutieli and Benjamini (1999) used (2) to propose a rejection procedure aiming at strong control of the FDR.

Benjamini and Yekutieli (2001) (hereafter called BY) provided a conservative procedure by dividing each c_i of the BH procedure by a constant $\sum_{l=1}^{m} 1/l$.

The BY Procedure: Fix $\alpha \in (0,1)$. Define $i^* = \max\{i : p_{d_i} \le c_i\}$, where $c_i = \alpha i/(m\sum_{l=1}^{m} 1/l)$, then reject $H_{d_1}, \ldots, H_{d_{i^*}}$ if i^* exists; otherwise reject no hypotheses.

Benjamini and Yekutieli (2001) proved that this procedure controls the FDR in the strong sense without relying on the independence assumption. However, this procedure has limited use due to the extreme conservativeness by dividing by $\sum_{l=1}^{m} 1/l$, approximately $\ln(m)$.

Additional works studying the theoretical properties of the FDR include Finner and Roters (2001, 2002), Genovese and Wasserman (2002), and Sarkar (2002). Another important direction of the research on the FDR is to control the false discovery proportion (FDP), the random variable Q, rather than its expectation, the FDR (Genovese and Wasserman (2004), Korn et al. (2004), Meinshausen (2006), Romano and Shaikh (2006) and van der Laan, Dudoit and Pollard (2004)). Our work differs from this line of research in that we focus on the control of the expectation (the FDR). In the next section, we propose a new step-down procedure, which provides strong control of the FDR for dependent data.

4. A New Step-Down Procedure to Control the FDR

4.1. Motivation

In order to develop a step-down procedure providing control of the FDR for generally dependent data, we first review an elegant step-down procedure proposed by Westfall and Young (1993) based on the *sequential rejection principle* (pages 72−73 in their book).

For any $\alpha \in (0,1)$, consider the following.

1. If $P(\min(P_{d_1},\ldots,P_{d_m}) \le p_{d_1} \mid H_M) \le \alpha$, then reject H_{d_1} and continue; otherwise accept all hypotheses and stop.

$\vdots$

i. If $P(\min(P_{d_i},\ldots,P_{d_m}) \le p_{d_i} \mid H_M) \le \alpha$, then reject H_{d_i} and continue; otherwise accept $H_{d_i},\ldots,H_{d_m}$ and stop.

$\vdots$

m. If $P(P_{d_m} \le p_{d_m} \mid H_M) \le \alpha$, then reject H_{d_m}; otherwise accept H_{d_m}.

This sequential rejection principle mimics researchers' verification procedures in practice. When people are faced with thousands of hypotheses, they will check the hypothesis with the smallest p-value to see if it is really a false null hypothesis. After a number of steps, say at step i, it might be reasonable to estimate the FWER using the null hypotheses $H_{d_i},\ldots,H_{d_m}$, since they have not been tested yet. The sequential rejection principle can also be written in the form of adjusted p-values. We define the *minP step-down adjustment*:

$$\check{p}_{d_i} = P(\min(P_{d_i},P_{d_{i+1}},\ldots,P_{d_m}) \le p_{d_i} \mid H_{d_i},\ldots,H_{d_m}). \tag{3}$$

and enforce the step-down monotonicity by assigning

$$\tilde{p}_{d_i} = \max_{k=1,\ldots,i} \check{p}_{d_k}.$$

This procedure is intuitively appealing. More importantly, Westfall and Young (1993) proved that the minP adjustments give strong control of the FWER under the subset pivotality property.

Subset pivotality: for all subsets $K \subseteq M$, the joint distributions of the subvector $(P_i, i \in K)$ are identical under the restrictions H_K and H_M, i.e.,

$$\mathcal{L}(P_i, i \in K) \mid H_K \overset{d}{=} \mathcal{L}(P_i, i \in K) \mid H_M.$$

Let

$$R_i = \#\{k \in \{d_i, d_{i+1}, \ldots, d_m\} : P_k \leq p_{d_i}\}.$$

The i-th step of the minP step-down procedure computes $P(R_i > 0 \mid H_{d_i}, \ldots, H_{d_m})$. This is essentially the FWER obtained by rejecting all untested hypotheses $(H_{d_i}, \ldots, H_{d_m})$ whose p-values are no greater than p_{d_i}. In other words, the minP step-down procedure has two stages. The first stage is to define a rejection rule at step i by using the same critical value p_{d_i} to reject hypotheses $H_{d_i}, \ldots, H_{d_m}$; the FWER adjusted p-value $\tilde{p}_{d_i}$ can then be computed by assuming that the hypotheses $H_{d_i}, \ldots, H_{d_m}$ are true. For any level α, the second stage is to reject all hypotheses whose adjusted p-values $\tilde{p}_{d_i}$ are no greater than α.

The sequential rejection principle can be applied to compute the FDR as with the minP step-down adjustment by a two-stage consideration. During the first stage, at step i, we define $R_i = \#\{k \in \{d_i, d_{i+1}, \ldots, d_m\} : P_k \leq p_{d_i}\}$. The critical issue is to compute the FDR related to this rejection procedure at step i. As with the minP step-down procedure at (3), at step i we can "naively" estimate the FDR under the assumption that the previous $i-1$ null hypotheses are false. We emphasize the "naively", as there might be a small proportion of the first i hypotheses that are true nulls.

According to the definition of FDR $= \mathrm{E}(V/R)$, where $0/0$ is defined to be zero, we can define the FDR adjusted p-values at step i to be

$$\check{p}_{d_i} = E\Big\{\frac{R_i}{(R_i + i - 1)} \;\Big|\; H_{d_i}, \ldots H_{d_m}\Big\}, \tag{4}$$

and enforce the monotonicity $\tilde{p}_{d_i} = \max_{k=1,\ldots,i} \check{p}_{d_k}$. In (4), $0/0$ is defined to be zero. This equation has some similarities to (2), which was used in the resampling procedure of Yekutieli and Benjamini (1999). Our procedure is different from theirs in two respects: (2) uses the same estimate $\hat{s}$ of s to compute every FDR adjusted p-value, while (4) has a different $\hat{s} = i - 1$ at each step i; (2) always computes the FDR under the complete null H_M, while we compute the FDR under different nulls H_K, where $K = \{d_i, \ldots, d_m\}$ for a particular step i. The aim of our new procedure is to provide strong control of the FDR under dependence just as the minP procedure provided strong control of the FWER (Westfall and Young (1993)).

It turns out that this procedure is too optimistic. In Section 5, our simulation results clearly show that it does not provide strong control of the FDR for the hypotheses with large FDR adjusted p-values. The reason is that, at step i of the first stage, the only rejected hypotheses are those whose p-values are no greater than p_{d_i}. However, at the second stage, we are very likely to reject a hypothesis whose p-value is much greater than p_{d_i} at a later step j $(> i)$. Therefore, the original definition is too optimistic. Denote the original definition of R_i by R_i^l,

$$R_i^l = \#\Big\{k \in \{d_i, \dots, d_m\} : P_k \leq p_{d_i}\Big\}. \tag{5}$$

We can define R_i more reasonably without using the *same* critical value p_{d_i} for all $P_{d_i}, \dots, P_{d_m}$. One way is to adapt the critical values according to the ranked p-values. For $k = 1, \dots, m-i+1$, let $P^i_{(k)}$ be the k-th smallest of the random variable p-values $P_{d_i}, \dots, P_{d_m}$, and let $p^i_{(k)} = p_{d_{i+k-1}}$ be the k-th smallest of the observed p-values $p_{d_i}, \dots, p_{d_m}$. The critical values $p^i_{(1)}, \dots, p^i_{(m-i+1)}$ can be used to compute the number of rejected hypotheses by a step-down procedure:

$$R_i^e = \max_{k=1,\dots,m-i+1}\Big\{k : P^i_{(1)} \leq p^i_{(1)}, \dots, P^i_{(k)} \leq p^i_{(k)}\Big\}. \tag{6}$$

The rejection procedure at step i to compute R_i^e is similar to the BH, BY, and BL procedures in Section 3. These procedures use constant critical values $c_1, \dots, c_m$, which depend solely on the significance level α. By contrast, our critical values $p^i_{(1)}, \dots, p^i_{(m-i+1)}$ depend on the data and the step index i as well: our critical values are more data-driven. In the simulation results of Section 5, the FDR seems to be controlled for the *hypotheses of interest*: the m_1 hypotheses with the smallest p-values, where m_1 can be suggested from the FDR curve.

We can use a more conservative strategy to compute R_i. At step i of the first stage, if we find $P^i_{(1)} \leq p^i_{(1)}$, then we stop. We naively think that we will reject hypothesis H_{d_i} and all of the later hypotheses $(H_{d_{i+1}}, \dots, H_{d_m})$, i.e., we put $R_i = m - i + 1$. In summary, a conservative definition of R_i can be given by

$$R_i^h = (m-i+1) \cdot I\Big(\min_{k=i,\dots,m} P_{d_k} \leq p_{d_i}\Big). \tag{7}$$

4.2. A step-down procedure

We now present a formal definition of the new step-down procedure in Section 4.1.

1. For each test statistic t_i, compute the p-value $p_i = P(|T_i| \geq |t_i| \mid H_i)$.
2. Order the p-values such that $p_{d_1} \leq \cdots \leq p_{d_m}$.
3. For $i = 1, \dots, m$, compute the FDR adjusted p-values as

$$\check{p}_{d_i} = E\Big\{\frac{R_i}{(R_i + i - 1)} \mid H_M\Big\}, \tag{8}$$

where R_i can be computed in any one of the three versions R_i^l, R_i^e, R_i^h from equations (5), (6) and (7), respectively.

4. Enforce monotonicity of the adjusted p-values, i.e., $\tilde{p}_{d_i} = \max_{k=1,\ldots,i} \check{p}_{d_k}$.

The step-down procedure for the three versions of R_i are called, respectively, lFDR (the lower adjusted), eFDR (the empirical adjusted), and hFDR (the higher adjusted).

Note that under the subset pivotality condition, (4) is equivalent to (8). This is very useful as we can compute all the FDR adjusted p-values under H_M instead of under different H_K, where $K = \{d_i, \ldots, d_m\}$ depends on step i. If we know the joint null distribution of $(P_1, \ldots, P_m)$ from model assumptions, then we can compute the expectation analytically. In the situations where we are not willing to make assumptions about the null joint distribution, the subset pivotality condition allows the expectation in (4) to be computed by simulating the complete null distribution under H_M by bootstrap or permutation resampling.

4.3. Finite sample results

Proposition 1. *For any of the three versions of R_i, the step-down procedure in Section 4.2 controls both the FWER and the FDR in the weak sense.*

By noting that the adjusted p-value $\tilde{p}_1$ in the first step is computed in the same way as that in equation (3) of the minP procedure, we have the proof.

Theorem 2. *Consider the step-down procedure in Section 4.2 using the definition of R_i^h in (7). For any $\alpha \in (0,1)$, if we reject all hypotheses whose FDR adjusted p-values are no greater than α, and if we assume that subset pivotality holds and that the joint distribution of $P_{M_0} = \{P_i, i \in M_0\}$ is independent of the joint distribution of $P_{M_1} = \{P_i, i \in M_1\}$, then FDR $\leq \alpha$.*

Corollary 3. *Assume that P_{M_0} and P_{M_1} are independent, and that $P_1, \ldots, P_m$ satisfy the generalized Šidák inequality,*

$$P(P_1 \geq p_1, \ldots, P_m \geq p_m) \geq \prod_{i=1}^{m} P(P_i \geq p_i). \tag{9}$$

Then the BL procedure provides strong control of the FDR under the subset pivotality condition.

Lemma 4. *Let $X_1, \ldots, X_B$ be B samples of random variable X and let $\overline{X}$ be the sample average. If $P(0 \leq X \leq 1) = 1$, then $Var(\overline{X}) \leq 1/(4B)$.*

The proofs of Theorem 2, Corollary 3 and Lemma 4 are given in the Appendix.

Remarks:

1. Proposition 1 and Theorem 2 also hold when the FDR adjusted p-values are computed based on the test statistics rather than on the p-values. A resampling procedure based on the test statistics is described in Algorithm 1. The

Algorithm 1. Resampling algorithm for computing FDR adjusted p-values by using the test statistics only

```
Compute the test statistic $t_i, i = 1, \ldots, m$ on the observed data matrix $\boldsymbol{x}$ and,
without loss of generality, label them $|t_1| \geq \cdots \geq |t_m|$.
For the $b$-th step, $b = 1, \ldots, B$, proceed as follows.
1. Compute the resampled data matrix $\boldsymbol{x}^b$, for example by randomly permuting the columns of matrix $\boldsymbol{x}$.
2. Compute the test statistic $t_{i,b}$ for hypothesis $H_i$, $i = 1, \ldots, m$ on the data matrix $\boldsymbol{x}^b$.
3. For each $i = 1, \ldots, m$, mimic the step-down procedure:
    (a) for $j = 1, \ldots, m - i + 1$, let $t^{i,b}_{(j)}$ be the $j$-th largest member of $\{t_{i,b}, \ldots, t_{m,b}\}$ in absolute value;
    (b) define $R_{i,b}$ to be the unique integer $k$ such that
            $|t^{i,b}_{(1)}| \geq |t_i|, \ldots, |t^{i,b}_{(k)}| \geq |t_{i+k-1}|$ and $|t^{i,b}_{(k+1)}| < |t_{i+k}|$.
4. Compute $f_{1,b} = I(R_{1,b} > 0)$ and, for $i = 2, \ldots, m$, compute $f_{i,b} = R_{i,b}/(R_{i,b} + i - 1)$.
Steps 1-4 are carried out $B$ times and the adjusted p-values are estimated by
$\check{p}_i = \sum_{b=1}^{B} f_{i,b}/B$, with monotonicity enforced by setting
        $\tilde{p}_1 \leftarrow \check{p}_1, \qquad \tilde{p}_i \leftarrow \max(\tilde{p}_{i-1}, \check{p}_i) \qquad$ for $i = 2, \ldots, m$.
```

independence assumption between P_{M_0} and P_{M_1} is replaced by the independence between T_{M_0} and T_{M_1}. Note that this independence assumption is the same as in Yekutieli and Benjamini (1999), and does not make the further assumption that all null test statistics are independent.

2. For the problem of identifying differentially expressed genes considered in this paper, if the $m \times n$ matrix $\boldsymbol{x}$ is normally distributed, specifically, the first n_1 and remaining n_2 columns of $\boldsymbol{x}$ are independently distributed as $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ respectively, where μ_1 and μ_2 are vectors of size m, Σ_1 and Σ_2 are $m \times m$ matrices, then the subset pivotality property is satisfied.

Here is the proof. Let T_i be the statistic for gene i, e.g. the two-sample t-statistic. For any subset $K = \{i_1, \ldots, i_k\}$, let its complement set be $\{j_1, \ldots, j_{m-k}\}$. The joint distribution of $(T_{i_1}, \ldots, T_{i_k})$ only depends on $i_1, \ldots, i_k$ components of the vectors μ_1 and μ_2 and the corresponding submatrices of Σ_1 and Σ_2. This joint distribution does not depend on how $(H_{j_1}, H_{j_2}, \ldots, H_{j_{m-k}})$ is specified. This proves subset pivotality for the test statistics. The subset pivotality also holds for the p-values, which are constructed from these test statistics. However, the subset pivotality property fails if we are testing the correlation coefficient between two genes; interested readers are referred to Example 2.2 on Page 43 of Westfall and Young (1993).

3. Sarkar (2002) generalized the results of the BL procedure in another direction. He assumed that the underlying test statistics are MTP_2, and that the test statistics are exchangeable under the null hypotheses. The MTP_2 property is similar to the $PRDS$ condition or the generalized Šidák inequality, which implies that weakly dependent test statistics behave like independent ones. However, the *exchangeability* assumption of Sarkar (2002) is not required in our generalization, and our proof is much simpler than that of Sarkar (2002).

4. We assume that the expectation in (8) can be computed without error. When the expectation is computed by B resamplings, using the fact that the random variable $R_i/(R_i+i-1)$ falls on the interval [0, 1], Lemma 4 implies that the estimate for $\check{p}_{d_i}$ in (8) has a standard error no greater than $\sqrt{1/(4B)}$. This quantity $\sqrt{1/(4B)}$ is therefore the maximum standard error of the estimate for the FDR adjusted p-values $\tilde{p}_{d_i}$ in Step 4 of Section 4.2. Since B=10,000 in most computations of this paper, the standard error is at most 0.005.

4.4. Resampling algorithms

In this section, we use resamplings to compute (8) so that we can obtain the FDR adjusted p-values. In general, there are two strategies for resampling the observed data $\boldsymbol{x}$ (an $m \times n$ matrix) to get the resampled data matrix $\boldsymbol{x}^b$: permuting the columns of matrix $\boldsymbol{x}$, and bootstrapping the column vectors. The application of these resampling strategies is in Westfall and Young (1993); more bootstrap methods can be seen in Davison and Hinkley (1997) and Efron and Tibshirani (1993). In the simulation study and applications results of this paper, we focus on comparing two groups, and $\boldsymbol{x}^b$ is obtained by permuting the columns of the matrix $\boldsymbol{x}$ to assign the group labels.

The complete algorithm for the empirical procedure (R_i^e) is described in Algorithm S1 in the supporting material. For other versions of R_i, the algorithm is similar and so will be omitted. If the p-values have to be computed by further resampling, then we have the same problem as in the double permutation algorithm of the minP procedure of Ge et al. (2003). In this situation, we can also have an analogous algorithm to Box 4 of Ge et al. (2003). Here, however, we cannot use the strategy in that paper to reduce the space, i.e., we need to compute the whole matrices T and P described in that paper. The details of this algorithm are omitted here.

Another approach is to compute FDR adjusted p-values based on the test statistics only. As we saw with the maxT procedure for controlling the FWER in Ge et al. (2003), the advantages and disadvantages of the maxT procedure compared with the minP procedure are also relevant to FDR adjusted p-values.

The FDR adjusted p-values computed from the test statistics are described in Algorithm 1, and this algorithm will be used in the remaining of this paper.

5. Simulation Results

5.1. Data generation strategy

For all figures in this section, there are 1,000 genes with 8 controls and 8 treatments. We first simulate $1,000 \times 16$ errors $\epsilon_{i,j}$, $i = 1, \ldots, 1,000$, $j = 1, \ldots, 16$, where the $\epsilon_{i,j}$ are block independent; specifically in our simulations the $\epsilon_{i,j}$ are independently and identically distributed as N(0, 1), except that $\mathrm{cor}(\epsilon_{10i+k,j}, \epsilon_{10i+l,j}) = \rho$ for $i = 0, \ldots, 99$, $k \neq l \in \{1, \ldots, 10\}$, $j = 1, \ldots, 16$. Lastly, we add δ to the treated group, so

$$X_{i,j} = \epsilon_{i,j} + \delta \text{ if } i = 1, \ldots, m_1, \quad j = 9, \ldots, 16; \qquad X_{i,j} = \epsilon_{i,j} \text{ otherwise.}$$

Note that the multiple testing problem for this simulation is to find the differentially expressed genes based on one observed data matrix $\boldsymbol{X} = \boldsymbol{x}$. The data matrix can be parametrized by (m_1, δ, ρ). For each gene i, we compute a two-sample Welch t-statistic t_i. Algorithm 1 is applied with the resampled data $\boldsymbol{x}^b$ generated by randomly permuting the columns of matrix $\boldsymbol{x}$ ($B = 10{,}000$). When we apply Algorithm 1, we do not assume that the data have normal distributions, and we do not know anything about the values of (m_1, δ, ρ) in the process that generates $\boldsymbol{x}$.

5.2. Properties of different FDR procedures

In Figures 1 and 2, and Figures S1 and S2 in the supporting material, the BH procedure, the BY procedure and the FDR procedures with R_i^e, R_i^l, and R_i^h in Algorithm 1 (labelled as BH, BY, eFDR, lFDR and hFDR in these figures) have been applied to different simulated data. The raw p-values required for the BH and BY procedures are computed by $B = 10{,}000$ permutations. We also plotted p-val, Q and YB99 for comparisons, where p-val is the raw p-value, Q is the random variable for the false discovery proportion (V/R) and YB99 is the resampling-based FDR upper limit, at (10) of Yekutieli and Benjamini (1999). The y-axis plots the FDR adjusted p-values for each procedure when the top 1, 2, ... genes are rejected.

In these figures, for the lFDR procedure, the FDR adjusted p-values fall far below the false discovery proportion Q, so the lFDR procedure is too optimistic for FDR control. On the other hand, the BY and hFDR procedures are too conservative, they pay a huge price for allowing a dependence structure. The hFDR procedure has the advantage of giving smaller FDR adjusted p-values for the most extreme genes, while the BY procedure may reject more null hypotheses

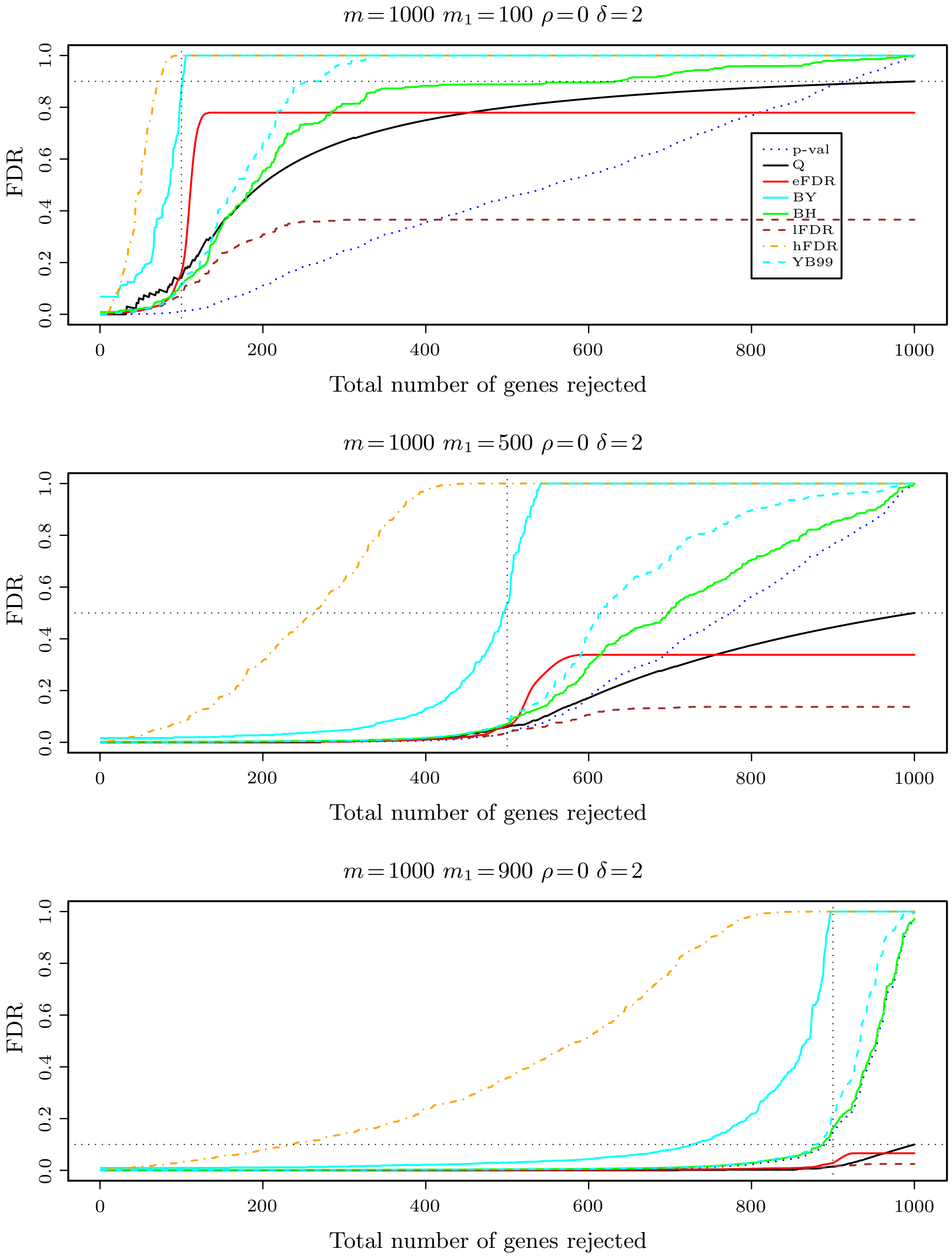


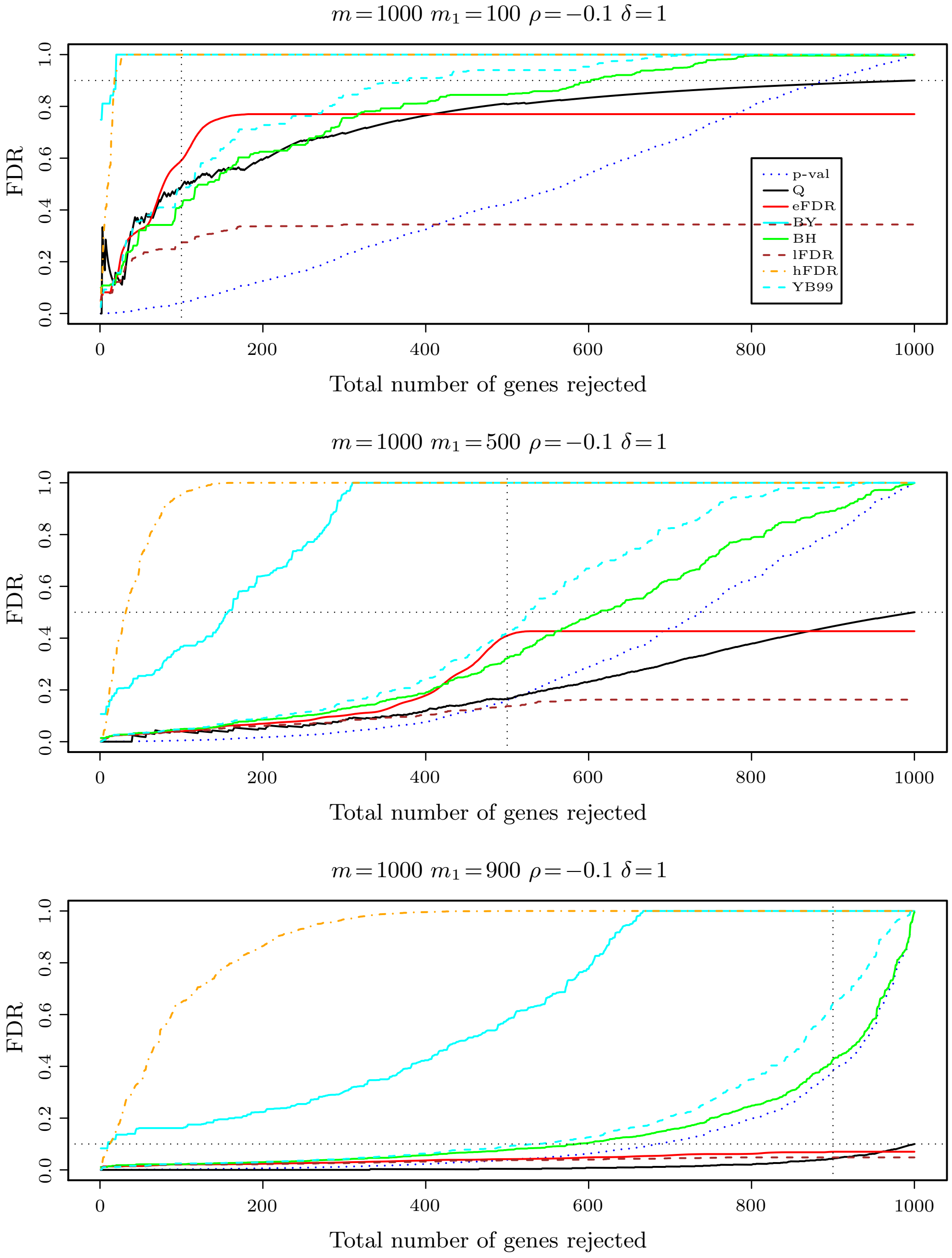
Figure 1. Different FDR procedures. The independent case: $\rho = 0$ and $\delta = 2$; the dotted vertical line is $x = m_1$; the dotted horizontal line is $y = m_0/m$, the overall proportion of false null hypotheses. Different panels are for different values of m_1 (100, 500, 900).

when target level α is much higher, say 0.5, or 1. Such levels are of limited use as researchers are more interested in smaller target levels, say 0.05 or 0.01. For example, in the middle panel of Figure 1 ($m_1 = 500, \delta = 2, \rho = 0$), at level 0.01, the hFDR procedure rejects 20 hypotheses, while the smallest FDR adjusted p-value for the BY procedure is 0.017. In contrast, at the 0.5 level, the hFDR procedure rejects only 262 hypotheses, whereas the BY procedure rejects 494 hypotheses.

The sample mean difference between the treatments and controls in Figure 1 ($\delta = 2$) is greater than that in Figure S1 in the supporting material ($\delta = 1$). In general, the larger the sample mean difference between the treatments and controls, the more powerful the procedure is to separate differentially expressed genes from non-differentially expressed ones. It is more interesting to look at our eFDR procedure, whose curve reaches its highest value around m_1 and then goes to a plateau. This displays a nice property of the eFDR procedure: it suggests an approximate value of m_1, the number of differentially expressed genes. If we reject fewer than m_1 genes, the adjusted p-value for the eFDR is higher than Q, i.e., the eFDR procedure provides strong control of the FDR. This feature is also displayed in the negatively dependent data in Figure 2 (and positively dependent data in Figure S2 in the supporting material).

The BH, BY and YB99 procedures may be too conservative for large m_1/m since they do not use any estimate of m_0 (see the middle and lower panels of Figures 1, 2, S1 and S2). This extreme conservativeness will not be a major concern in practice where m_1/m tends to be small.

Benjamini and Yekutieli (2001) proved that the BH procedure controls the FDR in the strong sense when the PRDS condition is satisfied. It might be interesting to construct negatively dependent data as a counterexample for the BH procedure, but we do not found have one so far. The BH procedure seems to work very well for the negatively dependent data of Figure 2. By noticing that the absolute values of the Student t-statistics are always positively dependent, we use one-sided tests to achieve more negative dependence, and the BH procedure still works well (data not shown). The BH procedure also works for other data generation strategies, such as using a finite mixture model for the errors $\epsilon_{i,j}$. The reason we could not find a counterexample for the BH is probably that the simulated data are still not strongly negatively dependent. For the data within each block of size 10, the statistics cannot be strongly negatively dependent, as the negative correlation coefficient cannot be less than $-1/9$, otherwise the constructed matrix violates the property that the correlation coefficient matrix must be non-negative definite. Therefore 1,000 genes are more or less independent. One may want to increase the number of blocks and decrease the size of each block to have a stronger negative correlation. However, with increasing


$m = 1000\ m_1 = 100\ \rho = -0.1\ \delta = 1$

$m = 1000\ m_1 = 500\ \rho = -0.1\ \delta = 1$

$m = 1000\ m_1 = 900\ \rho = -0.1\ \delta = 1$

Figure 2. The negatively dependent case $\rho = -0.1$ and $\delta = 1$: different FDR procedures and different values of m_1 (100, 500, 900).

Table 2. TThe FDR (sample average of Q) at the target level α when using 1000 samples of $\boldsymbol{X}$ with $m = 200$, $m_1 = 50$, $\delta = 1$, $\rho = 0$.

α	0.01	0.05	0.1	0.2	0.5
eFDR	0.0017	0.028	0.079	0.15	0.30
BH	0	0.0081	0.069	0.17	0.38

Table 3. The average number of genes rejected at the target level α when using 1000 samples of $\boldsymbol{X}$ with $m = 200$, $m_1 = 50$, $\delta = 1$, $\rho = 0$.

α	0.01	0.05	0.1	0.2	0.5
eFDR	0.29	1.8	5.2	14	42
BH	0	0.40	5.3	17	54

numbers of blocks, the overall dependence decreases. We have simulated the data with block size two and correlation coefficient $\rho = -0.7$ and the result is very similar to Figure 2.

Note that Figures 1, 2, S1 and S2 consider only one sample of the data $\boldsymbol{X} = \boldsymbol{x}$. The FDR can be estimated by computing the average of the Q for 1,000 samples of $\boldsymbol{X}$ when we reject the genes whose FDR adjusted p-values are no greater than α. Due to computational complexity, we only consider $m = 200$ and $m_1 = 50$ with block size two. As we consider a smaller block size, we can decrease the value of ρ from -0.1 to -0.7. The results are shown in Tables 2 and 3 for independent data and in Tables 4 and 5 for dependent data. Again the eFDR procedure performs better than the BH procedure for smaller values of α. The BH procedure, on the other hand, is better for large values of α.

6. Microarray Applications

Apo AI knock-out experiment: The Apo AI experiment (Callow et al. (2000)) was carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice. The apolipoprotein AI (Apo AI) is a gene known to play a pivotal role in HDL metabolism, and mice with the Apo AI gene knocked out have very low HDL cholesterol levels. The goal of the experiment was to identify genes with altered expression in the livers of these knock-out mice compared to inbred control mice. The treatment group consisted of eight mice with the Apo AI gene knocked out and the control group consisted of eight wild-type C57Bl/6 mice. For the 16 microarray slides, the target cDNA was from the liver

Table 4. The FDR (sample average of Q) at the target level α when using 1,000 samples of $\boldsymbol{X}$ with $m = 200$, $m_1 = 50$, $\delta = 1$, $\rho = -0.7$.

α	0.01	0.05	0.1	0.2	0.5
eFDR	0	0.035	0.066	0.13	0.32
BH	0	0.0068	0.060	0.14	0.38

Table 5. The average number of genes rejected at the target level α when using 1,000 samples of $\boldsymbol{X}$ with $m = 200$, $m_1 = 50$, $\delta = 1$, $\rho = -0.7$.

α	0.01	0.05	0.1	0.2	0.5
eFDR	0.30	2.0	5.1	15	49
BH	0	0.37	5.1	17	55

mRNA of the 16 mice. The reference cDNA came from the pooled control mice liver mRNA. Among the 6,356 cDNA probes, about 200 genes were related to lipid metabolism. In the end, we obtained a 6,356 × 16 matrix with 8 columns from the controls and 8 columns from the treatments. Differentially expressed genes between the treatments and controls are identified by two-sample Welch t-statistics.

Leukemia study: One goal of Golub et al. (1999) was to identify genes that are differentially expressed in patients with two types of leukemias: acute lymphoblastic leukemia (ALL, class 1) and acute myeloid leukemia (AML, class 2). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $m = 6{,}817$ human genes. The learning set comprises $n = 38$ samples, 27 ALL cases and 11 AML cases (data available at http://www.genome.wi.mit.edu/MPR). Following Golub et al. (personal communication, Pablo Tamayo), three preprocessing steps were applied to the normalized matrix of intensity values available on the website: (i) thresholding with a floor of 100 and a ceiling of 16,000; (ii) filtering with exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum intensities for a particular gene across mRNA samples; (iii) base 10 logarithmic transformation. Boxplots of the expression levels for each of the 38 samples revealed the need to standardize the expression levels within arrays before combining data across samples. The data were then summarized by a $3{,}051 \times 38$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level for gene i in mRNA sample j. Differentially expressed genes in ALL and AML patients were identified by computing two-sample Welch t-statistics.

For the two datasets, we applied the BH, BY, eFDR, hFDR and YB99 procedures described in this paper to control the FDR. All of these resamplings were done by permuting the columns of the data matrix. Figure 3 gives the results for the Apo AI knock-out data. The x-axis is always the rank of p-values and y-axis is the FDR adjusted p-values. Note that the rank of different adjusted p-values is always the same as the rank of the raw p-values apart from the t-based procedures (eFDR, hFDR). In that case, the adjusted p-values have the same ranks as the two-sample Welch t-statistics. Similarly, Figure 4 gives the results of applying these procedures to the Leukemia dataset.

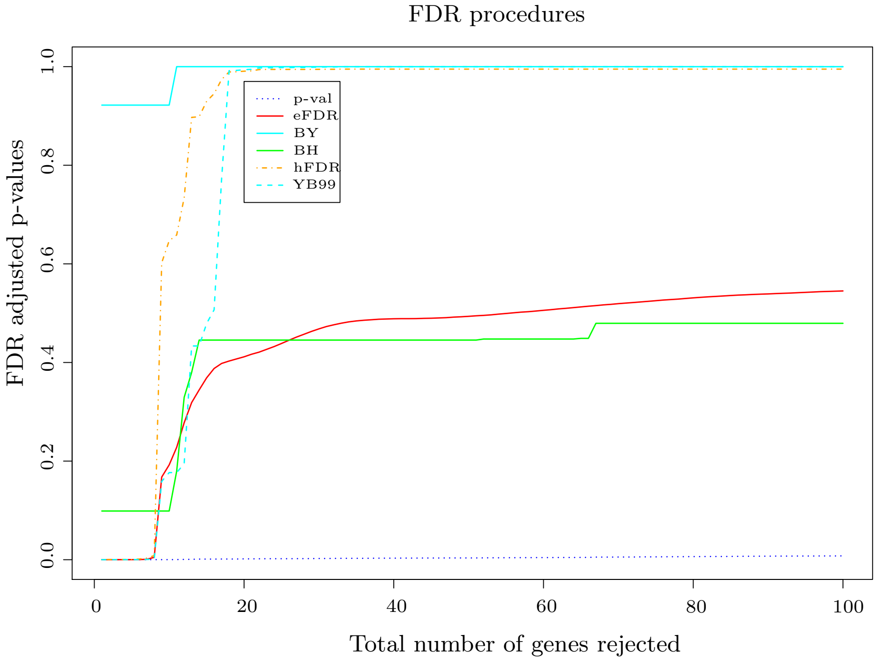


Figure 3. *Apo AI:* Plot of FDR adjusted p-values when the top 1, 2, ... genes are rejected. We only plot the total number of genes rejected up to 100 among 6,356 genes. The adjusted p-values were estimated using all $B = 16!/(8! \times 8!) = 12{,}870$ permutations.

7. Discussion

In this paper, we introduce a new step-down procedure aiming to control the FDR. This procedure uses the sequential rejection principle of the Westfall and Young minP step-down procedure to compute the FDR adjusted p-values. It automatically incorporates dependence information into the computation. We have essentially introduced three FDR procedures. The first, lFDR, is too optimistic for controlling the FDR from the simulated data. The second, hFDR, is shown to control the false discovery rate under the subset pivotality condition and under the assumption that joint distribution of statistics from true nulls is independent of the joint distribution of statistics from false nulls (see Theorem 2). From Remark 2 of Section 4.3, under some parametric formulation of the data, if each test statistic is generated within one gene, the subset pivotality property can be satisfied. As with the Westfall and Young minP step-down procedure, this procedure fails if the subset pivotality condition is not satisfied, for example testing the correlation coefficient between two genes. The hFDR procedure also extends the BL procedure from an independence condition to a generalized Šidák inequality condition, see (9). The third and most useful procedure, eFDR,

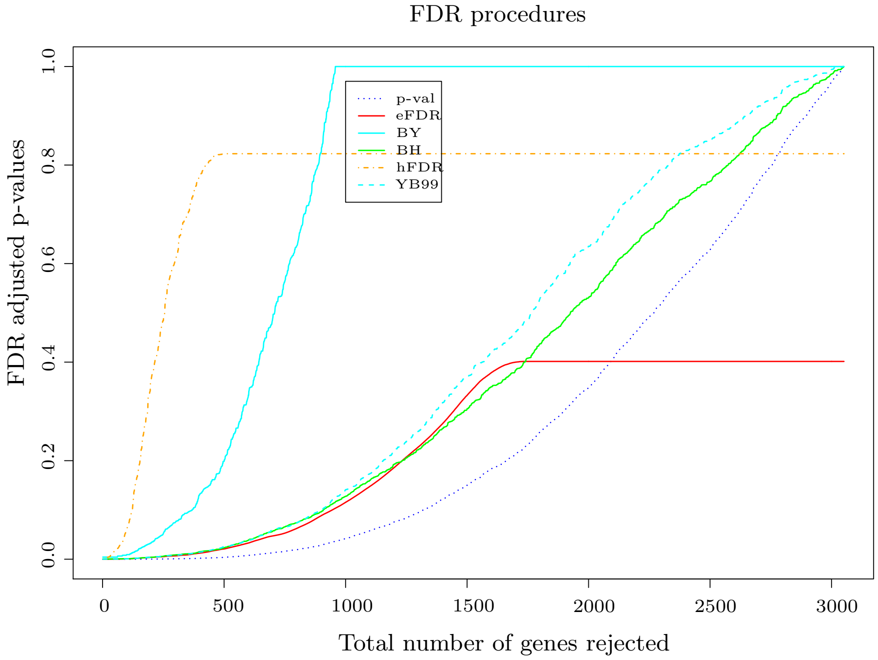


Figure 4. *Leukemia:* Plot of FDR adjusted p-values when the top 1, 2, . . . , genes among 3,051 ones are rejected. The adjusted p-values were estimated using $B = 10{,}000$ random permutations.

is recommended in practice. The theoretical properties of the eFDR, whether finite sample or asymptotic, are left to future research. The validity of the eFDR procedure is currently suggested by our simulation results and can be extended to a large number of hypotheses (Figure S3 in the supporting material shows the simulation for $m = 10,000$). One nice feature of the eFDR procedure is that it suggests the number of the false null hypotheses and the FDR adjusted p-value simultaneously. The FDR plot, see Figure 3 or 4, is also useful for diagnostic purposes.

8. Supporting Material

Figures S1, S2 and S3, and Algorithm S1.

Acknowledgements

We thank Juliet Shaffer for her helpful discussions on the subject matter of this paper. A number of people have read this paper in draft form and made valuable comments, and we are grateful for their assistance; they include Carol Bodian, Richard Bourgon, Gang Li and Xiaochun Li. We thank the referees for helpful comments that have led to an improved paper. This work was supported by NIH grants U19 AI 62623, RO1 DK46943 and contract HHSN 26600500021C.

Appendix. The proof of Theorem 2 and Corollary 3 and Lemma 4

Proof of Theorem 2: Denote the α level critical value of $\min_{i\in K} P_i \mid H_M$ by $c_{\alpha,K}$, i.e.,

$$P\Big\{\min_{i\in K} P_i \le c_{\alpha,K} \mid H_M\Big\} = \alpha. \tag{10}$$

For given α at step i, when using the higher bound R_i^h to compute the FDR adjusted p-values, the hypothesis H_{d_i} is rejected if and only if $P_{d_i} \le c_{\alpha_i, M\backslash\{d_1,\ldots,d_{i-1}\}}$. Here $\alpha_i = \alpha m/(m-i+1)$. Let $\Lambda = (\lambda_1,\ldots,\lambda_{m_1})$ be a permutation of M_1. For $i = 0,\ldots,m_1$, let

$$B_{i,\Lambda} = \{P_{\lambda_1} \le c_{\alpha_1,M},\ldots,P_{\lambda_i} \le c_{\alpha_i,M\backslash\{\lambda_1,\ldots,\lambda_{i-1}\}}, P_{\lambda_{i+1}} > c_{\alpha_{i+1},M\backslash\{\lambda_1,\ldots,\lambda_i\}}$$
$$\text{and } P_{\lambda_1} \le \cdots \le P_{\lambda_{m_1}}\}.$$

The events $\{B_{i,\Lambda} : i, \Lambda\}$ are a mutually exclusive decomposition of the whole sample space, and so $\sum_{i,\Lambda} P(B_{i,\Lambda}) = 1$. In set $B_{i,\Lambda}$, the random p-values $P_{\lambda_1},\ldots,P_{\lambda_i}$ are not necessarily the i smallest ones among the random p-values from the set M. However, for given α and Λ, the critical value $c_{\alpha_k,M\backslash\{\lambda_1,\ldots,\lambda_{k-1}\}}$ is an increasing function of k, which implies that we can reject at least the i false null hypotheses $H_{\lambda_1},\ldots,H_{\lambda_i}$, so

$$\frac{V}{R} \le \frac{V}{(V+i)} \le \frac{(m-i)}{m}. \tag{11}$$

When $\boldsymbol{x} \in B_{i,\Lambda}$, if we have also erroneously rejected at least one null hypothesis by using the higher bound R_i^h for computing the FDR adjusted p-values, then the fact that $c_{\alpha_k,M\backslash\{\lambda_1,\ldots,\lambda_{k-1}\}}$ is an increasing function of k and that $P_{\lambda_{i+1}} > c_{\alpha_{i+1},M\backslash\{\lambda_1,\ldots,\lambda_i\}}$ implies

$$P_{(1),M_0} \le c_{\alpha_{i+1},M\backslash\{\lambda_1,\ldots,\lambda_i\}}. \tag{12}$$

Here $P_{(j),K}$ denotes the j-th smallest member of $\{P_i, i \in K\}$. As $\{\lambda_1,\ldots,\lambda_i\}$ is contained by M_1, the set $M\backslash\{\lambda_1,\ldots,\lambda_i\}$ contains M_0. Using the definition of $c_{\alpha,K}$ in (10), we have $c_{\alpha_{i+1},M\backslash\{\lambda_1,\ldots,\lambda_i\}} \le c_{\alpha_{i+1},M_0}$. Combining this with (12) gives

$$P_{(1),M_0} \le c_{\alpha_{i+1},M_0}. \tag{13}$$

Therefore

$$\text{FDR} = \sum_{i,\Lambda} E\Big\{\frac{V}{R}\cdot I(B_{i,\Lambda})\Big\}$$

$$
\begin{aligned}
&= \sum_{i,\Lambda} E\Big\{\frac{V}{R} \cdot I(B_{i,\Lambda}) \cdot I(P_{(1),M_0} \text{ is rejected})\Big\} \\
&\le \sum_{i,\Lambda} \frac{(m-i)}{m} \cdot P(B_{i,\Lambda} \cap \{P_{(1),M_0} \text{ is rejected}\}) \quad \text{(Apply (11))} \\
&\le \sum_{i,\Lambda} \frac{(m-i)}{m} \cdot P(B_{i,\Lambda} \cap \{P_{(1),M_0} \le c_{\alpha_{i+1},M_0}\}) \quad \text{(Apply (13))} \\
&= \sum_{i,\Lambda} \frac{(m-i)}{m} \cdot P(B_{i,\Lambda}) \cdot P(P_{(1),M_0} \le c_{\alpha_{i+1},M_0}) \qquad (14) \\
&= \sum_{i,\Lambda} \frac{(m-i)}{m} \cdot P(B_{i,\Lambda}) \cdot \frac{\alpha m}{(m-i)} \qquad (15) \\
&= \sum_{i,\Lambda} P(B_{i,\Lambda}) \cdot \alpha \\
&= \alpha.
\end{aligned}
$$

At (14), we use the assumption that P_{M_0} and P_{M_1} are independent. At (15), by subset pivotality,

$$
\begin{aligned}
P(P_{(1),M_0} \le c_{\alpha_{i+1},M_0}) &= P(P_{(1),M_0} \le c_{\alpha_{i+1},M_0} \mid H_{M_0}) \\
&= P(P_{(1),M_0} \le c_{\alpha_{i+1},M_0} \mid H_M) \\
&= \alpha_{i+1} = \frac{\alpha m}{(m-i)}.
\end{aligned}
$$

Proof of Corollary 3: Combining (7) and (4), we have

$$
\begin{aligned}
\check{p}_{d_i} = E\Big\{\frac{R_i^h}{(R_i^h + i - 1)}\Big\} &= P\Big(\min_{k=i,\ldots,m} P_{d_k} \le p_{d_i}\Big) \cdot \frac{(m-i+1)}{m} \\
&\le [1-(1-p_{d_i})^{m-i+1}] \cdot \frac{(m-i+1)}{m} \quad \text{(Apply (9))}.
\end{aligned}
$$

Therefore, if $p_{d_i} \le 1 - [1-\min(1, \alpha m/(m-i+1))]^{1/(m-i+1)}$, then $[1-(1-p_{d_i})^{m-i+1}] \cdot (m-i+1)/m \le \alpha$, and so $\check{p}_{d_i} \le \alpha$. Thus the BL procedure is more conservative than the step-down procedure in Section 4.2, which was shown to give strong control according to Theorem 2. Therefore the BL procedure also controls the FDR in the strong sense.

Proof of Lemma 4: Let $\mu = E(X)$. Construct a random variable $Y = I(X \ge \mu)$ and write $p = E(Y)$. We have $Var(X) + (\mu - p)^2 = E(X-p)^2 \le E(Y-p)^2 + (\mu-p)^2$, hence $Var(X) \le Var(Y) = p(1-p) \le 1/4$. Therefore $Var(\overline{X}) = Var(X)/B \le 1/(4B)$.

References

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

Benjamini, Y. and Hochberg, Y. (2000). The adaptive control of the false discovery rate in multiple hypotheses testing with independent statistics. *J. Behav. Educ. Statis.* **25**, 60-83.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82**, 163-170.

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

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* **10**, 2022-2029.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, Cambridge.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12**, 111-139.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, Boca Raton.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Eklund, G. (1961-1963). Massignifikansproblemet. Unpublished seminar papers, Uppsala University Institute of Statistics.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. *Biometrical J.* **8**, 985-1005.

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30**, 220-238.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12**, 1-44, with discussion 44-77.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499-517.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035-1061.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: Application to high dimensional genomic data. *J. Statist. Plann. Inference* **124**, 379-398.

Kwong, K. S., Holland, B. and Cheung, S. H. (2002). A modified Benjamini-Hochberg multiple comparisons procedure for controlling the false discovery rate. *J. Statist. Plann. Inference* **104**, 351-362.

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.* **33**, 227-237.

Romano, J. P. and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34**, 1850-1873.

Sarkar, S. K. (2002). Some results of false discovery rate in stepwise multiple testing procedure. *Ann. Statist.* **30**, 239-257.

Seeger, P. (1968). A note on a method for the analysis of significance en masse. *Technometrics* **10**, 586-593.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754.

Sorić, B. (1989). Statistical "discoveries" and effect-size estimation. *J. Amer. Statist. Assoc.* **84**, 608-610.

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

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

Troendle, J. F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rates. *J. Statist. Plann. Inference* **84**, 139-158.

Tusher, V., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116-5121.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statist. Appl. Genet. Mol. Biol.* **3**, Article 15.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment.* Wiley, New York.

Westfall, P. H., Zaykin, D. V. and Young, S. S. (2001). Multiple tests for genetic effects in association studies. *Methods in Molecular Biology* **184**, Biostatistical Methods (Edited by S. Looney), 143-168. Humana Press, Toloway, NJ.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82**, 171-196.

Department of Neurology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1137, New York, NY, 10029, U.S.A.

E-mail: yongchao.ge@mssm.edu

Department of Neurology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1137, New York, NY, 10029, U.S.A.

E-mail: stuart.sealfon@mssm.edu

Department of Statistics, University of California, 367 Evans Hall # 3860, Berkeley, CA, 94720-3860, U.S.A.

E-mail: terry@stat.berkeley.edu

(Received July 2007; accepted December 2007)