BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 16 2006, pages 1979–1987
doi:10.1093/bioinformatics/btl328

Gene expression

Robust estimation of the false discovery rate

Stan Pounds* and Cheng Cheng

Department of Biostatistics, St. Jude Children's Research Hospital, 332 N. Lauderdale Street, Memphis, TN 38135 USA

Received on April 3, 2006; revised on May 17, 2006; accepted on June 9, 2006
Advance Access publication June 15, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Presently available methods that use *p*-values to estimate or control the false discovery rate (FDR) implicitly assume that *p*-values are continuously distributed and based on two-sided tests. Therefore, it is difficult to reliably estimate the FDR when *p*-values are discrete or based on one-sided tests.

Results: A simple and robust method to estimate the FDR is proposed. The proposed method does not rely on implicit assumptions that tests are two-sided or yield continuously distributed *p*-values. The proposed method is proven to be conservative and have desirable large-sample properties. In addition, the proposed method was among the best performers across a series of 'real data simulations' comparing the performance of five currently available methods.

Availability: Libraries of S-plus and R routines to implement the method are freely available from www.stjuderesearch.org/depts/biostats

Contact: stanley.pounds@stjude.org

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

1 INTRODUCTION

The false discovery rate (FDR; Benjamini and Hochberg 1995) is a widely used measure of statistical significance in the analysis of data collected in genomics experiments (Storey and Tibshirani 2003). In developing their original FDR-control procedure, Benjamini and Hochberg (1995) assumed that *p*-values testing true null hypotheses are independent observations from a continuous uniform (0,1) probability distribution. A family of related methods has been developed by proposing various modifications and extensions of Benjamini and Hochberg's (1995) procedure. Pounds (2006) provides an overview of this family of methods, which is now widely used in practice. Some questions have been raised about whether these methods would give trustworthy results when applied in the analysis of microarray gene expression data because the correlation between genes violated the assumption of independence. Fortunately, extensive theoretical work (Benjamini and Yekutieli 2001; Storey *et al*., 2004) has shown that these methods maintain their desirable statistical properties under specific forms of weak dependence. On the basis of mathematical and biological grounds, Storey and Tibshirani (2003) have argued that the correlation structure of most genome-wide microarray gene expression datasets satisfy this condition of weak dependence. However, no one has carefully considered how the performance of these methods are affected when the assumptions of continuity or uniformity are violated. Discrete *p*-values clearly violate the assumption of continuity. The *p*-values from one-sided tests (e.g. those with a one-sided alternative hypothesis) may be stochastically greater than uniform, thus violating the assumption of uniformity.

In practice, there are several applications that involve one-sided tests or discrete *p*-values. The present-absent *p*-values, commonly used to filter poorly hybridized Affymetrix probe sets, are based on one-sided Wilcoxon signed-rank tests comparing the expression of pefect match probes to that of mismatch probes (Pounds and Cheng 2005a). Gadbury *et al*. (2003) describe performing a randomization test when each of two groups is represented by only three replicates; in that setting, the possible *p*-values are 0.1, 0.2, . . . , 1.0. Apparently, experiments with such a small number of replicates are fairly common in practice (Lee *et al*., 2000). Also, the analysis of array CGH data often produces discrete copy number estimates (Lai *et al*., 2005) which may be used in later statistical analyses and thus lead to discrete *p*-values. Furthermore, discrete *p*-values will become more frequently encountered in practice as categorical genomic data, such as SNP chip genotype calls and copy number estimates (Lin *et al*., 2004), become more widely available. Therefore, it is critical to carefully determine whether currently available methods yield accurate results when applied to one-sided or discrete *p*-values. If these methods perform poorly, then development of alternative approaches is warranted.

Pounds (2006) has expressed concerns that many FDR methods will give inaccurate inferences when applied to discrete *p*-values. Methods that estimate the proportion of tests examining a true null hypothesis (i.e. the null proportion) rely heavily on the assumption of continuity through the use of beta-uniform mixture models (Allison *et al*., 2002; Pounds and Morris 2003), other types of models (Tsai *et al*., 2003; Liao *et al*., 2004), or smoothing (Storey 2002; Pounds and Cheng 2004; Cheng *et al*., 2004). Discrete *p*-values introduce additional problems that affect these and other methods, including the original procedure of Benjamini and Hochberg (1995). The FDR methods compute local FDR (*l*FDR; Benjamini and Hochberg 2000) estimates (or similar quantities) as an intermediate step to producing their final results. Discrete *p*-values introduces instability into the *l*FDR estimates that is propogated down to the reported *q*-values (Storey 2002), FDR estimates (Tsai *et al*., 2003) or FDR-adjusted *p*-values (Reiner *et al*., 2003).

Gilbert (2005) has developed an approach for controlling the FDR when *p*-values are discrete. Gilbert's method first identifies the subset of tests having a grid of possible *p*-values that satisfy a specific requirement described by Tarone (1990). Gilbert's

*To whom correspondence should be addressed.

approach applies Benjamini and Hochberg's (1995) method to the *p*-values from those tests that satisfy Tarone's requirement. Gilbert (2005) successfully applied his method in an application involving the comparison of two sets of aligned sequences. In that application, the relevant properties of the *p*-value grid varied substantially across the statistical tests, and therefore Tarone's requirement had a meaningful impact on how Gilbert's method operates. However, the *p*-value grid will vary little, if any, across the statistical tests performed in the analysis of microarray gene expression data. In microarray data analysis, discrete *p*-values arise from applying the same rank-based, permutation, or exact test to the expresion data for each gene. Gene expression data are typically continuous variables; therefore, tied data values are very rare. Hence, the vast majority (if not all) tests will have the same grid of possible *p*-values. Subsequently, Tarone's (1990) requirement will have little influence on how Gilbert's (2005) method operates. Thus, the results obtained from Gilbert's (2005) method will be very similar or identical to those of Benjamini and Hochberg's (1995) method. Therefore, Gilbert's (2005) method is also susceptible to the issues mentioned above.

The ramifications of one-sided testing have not yet been thoroughly explored. In the context of one-sided tests (i.e. those with a one-sided alternative), the *p*-values testing true null hypotheses may follow a distribution that is stochastically greater than uniform. Roughly speaking, the FDR is the ratio of the number of false positives to the number of significant results. Shifting the null distribution of *p*-values away from uniform will affect the numerator and denominator of this ratio in the same direction (increase or decrease); hence it is not obvious whether the FDR will increase or decrease. By analogous arguments, it is unclear how the *l*FDR estimates of currently available methods will behave when computed using one-sided *p*-values. Therefore, it is unclear whether any of the available methods will maintain their desirable statistical properties in the context of one-sided tests.

In our experience, currently available FDR methods have performed poorly in applications involving discrete or one-sided *p*-values. Therefore, we have developed a method that accurately computes FDR estimates for sets of *p*-values that are one-sided or discrete. Section 2 outlines the development of the method in general terms. In Section 3, the advantages of the proposed method are clearly shown in two specific applications (Gadbury *et al*., 2003; Pounds and Cheng, 2005a) and a series of simulation studies based on a study of gene expression in acute myeloid leukemia (AML; Ross *et al*., 2004). Finally, Section 4 gives some discussion and concluding remarks.

2 APPROACH

In this section, we develop the proposed method under the general framework of statistical testing. Here, we do not limit discussion to any specific statistical testing procedure. In addition, we assume that the tests may be one-sided or two-sided except when stated otherwise. Nevertheless, the ideas presented may seem less abstract if one keeps the applications of Section 3 in mind: a two-sided comparison of expression between two groups (Gadbury *et al*., 2003); filtering poorly hybridized Affymetrix probe sets with one-sided statistical tests (Pounds and Cheng, 2005a) and the simulation studies based on a variety of one- or two-sided comparisons of expression across two subtypes of AML (Ross *et al*., 2004).

2.1 The multiple-testing setting

Suppose that $i = 1,\ldots,m$ statistical tests are performed to examine whether a particular feature is differentially expressed. Each statistical test considers whether the observed data deviate significantly from what would be expected by chance under a stated null hypothesis. In two-sided tests, the null hypothesis typically asserts that no meaningful association exists; in one-sided tests, the null hypothesis states that no association of a given direction exists. For example, the null hypothesis may be two sided and state that two or more groups have equal mean or median expression. Also, the null hypothesis may be one-sided and state that the mean for one group or the difference of two groups' means is less than or equal to a specified value (often 0). Most statistical testing procedures report a *p*-value as a measure of how much the observed data deviates from what would be expected by chance if the equivalence statement of the null hypothesis is true. A smaller *p*-values indicate greater departure from what would be expected if the null hypothesis is true. For two-sided tests, the *p*-value is small for deviation in either direction; for one-sided tests the *p*-value is small only if there is significant deviation in the non-null direction. A threshold α must be specified to determine which results to report as significant. Each test giving a *p*-value less than or equal to the specified threshold α is declared statistically significant.

Each of the *m* hypothesis tests results in one of four outcomes: a false positive (erroneously declaring that a meaningful association exists or Type I error), a true positive (correctly declaring that a meaningful association exists), a false negative (not declaring that a meaningful association exists when one is truly present or Type II error) or a true negative (not declaring that a meaningful association exists when one truly does not exist). One challenge in this setting is to define appropriate statistical metrics of statistical significance and the occurrence of erroneous inferences.

2.2 The FDR methods

Benjamini and Hochberg (1995) introduced the FDR as a useful measure of statistical significance in the multiple-testing setting. Now, a family of methods that use *p*-values to estimate or control the FDR has been developed (Pounds, 2006). Given a set of *p*-values $\boldsymbol{p} = \{p_1, p_2, \ldots, p_m\}$ computed in *m* hypothesis tests, these methods first compute an estimate of the *l*FDR (Benjamini and Hochberg 2000). The *l*FDR estimates are expressed as ratios of the form

$$t_{(i)} = \frac{\hat{\boldsymbol{\upsilon}}(p_{(i)})}{\hat{\mathbf{F}}(p_{(i)})}, \tag{1}$$

where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ are the ordered *p*-values, $\hat{\boldsymbol{\upsilon}}(\alpha)$ estimates the expected proportion $\boldsymbol{\upsilon}(\alpha)$ of tests resulting in a false positive when α is used as the *p*-value threshold to determine significance, and $\hat{\mathbf{F}}(\alpha)$ estimates the expected proportion $\mathbf{F}(\alpha)$ of tests yielding a *p*-value less than or equal to α. Most methods express $\hat{\boldsymbol{\upsilon}}(\alpha)$ as

$$\hat{\boldsymbol{\upsilon}}(\alpha) = \hat{\boldsymbol{\pi}}\alpha, \tag{2}$$

where $\hat{\boldsymbol{\pi}}$ estimates the proportion π of tests with a true null hypothesis. Estimation methods simply report $t_{(i)}$ as an estimate of the proportion of tests with a *p*-value less than or equal to $p_{(i)}$ that have a true null hypothesis for $i = 1,\ldots,m$. Control methods perform

additional calculations. Control methods also find

$$q_{(i)} = \min_{j \geq i} t_{(i)} \tag{3}$$

for each i and reject the null hypothesis for each test with $q_{(i)}$ that is less than or equal to a prespecified level τ of FDR, positive FDR (pFDR; Storey 2002), or conditional FDR (cFDR; Tsai *et al*., 2003) control.

2.3 Assumptions of the FDR methods

Benjamini and Hochberg (1995) showed that their method controls the FDR at the desired level under the assumptions that p-values testing true null hypotheses were (1) independent, (2) uniformly distributed and (3) continuously distributed random variables. Later procedures are essentially modifications of that procedure that fit into the framework given in Equations (1)–(3). It is natural to question whether these methods will maintain their desirable control or estimation properties when applied to p-values that do not satisfy one of the three assumptions.

As mentioned in the introduction, early consideration was given to how violation of the independence assumption might affect the performance of the FDR methods. Benjamini and Yekutieli (2001) showed that the Benjamini and Hochberg (1995) procedure maintained its desirable control properties under certain forms of dependency. Also, Storey *et al*., (2004) showed that Storey's (2002) procedure maintained its control properties under these forms of dependency as well. In addition, Storey and Tibshirani (2003) argue for most applications that genome-wide microarray gene expression data naturally satisfies this form of dependency owing to the way that genes act in pathways. Finally, Pounds (2006) notes that most other FDR methods that fit the framework of (1), (2) and (3) should maintain desirable statistical properties (at least in an approximate sense) in the presence of these forms of dependency, although this remains to be rigorously shown for several of the methods.

In mathematical terms, the assumption of uniformity implies that

$$\Pr(p_i \leq \alpha \,|\, i\text{th null hypothesis is true}) \approx \alpha, \tag{4}$$

holds for all i. Specifically, (4) is used to derive (2). The proportion of tests yielding a false positive is the arithmetic average of the probability of a false positive across the tests. Under (4), π of the tests have a true null hypothesis and each of those has probability α of being declared significant. Additionally, methods that estimate π use (4) in computing $\hat{\boldsymbol{\pi}}$. Under (4), Pounds and Morris (2003) note that the minimum density of the set of p-values should be a conservative estimator of π (except when $\pi \approx 1$). Thus, for example, Storey (2002) assumes the minimum of the density occurs near 1 and smooths the empirical distribution function for large p to estimate π and Pounds and Morris (2003) use the minimum of the model fitted to the p-values to estimate π. Also, the distribution of p-values with a true null are a component of $\hat{\mathbf{F}}(\cdot)$, and therefore violation of (fassume2) can impact $\hat{\mathbf{F}}(\cdot)$ as well. Thus, violation of (4) impacts the components $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{v}}(\cdot)$, and $\hat{\mathbf{F}}(\cdot)$ of (1) and (2) and any biases are passed along to (3). Therefore, from a theoretical standpoint, the assumption of uniformity appears to be critical for FDR methods to maintain desirable control and estimation properties.

Methods that estimate π use the assumption of continuity in computing $\hat{\boldsymbol{\pi}}$. The assumption of continuity is implied via the use of smoothing or modeling. For example, Storey (2002) smooths the slope of the empirical distribution function; Poundsand Cheng (2004) apply local regression to transformed p-value spacings; and Cheng *et al*., (2004) use spline-based estimates of the cumulative distribution function to compute $\hat{\boldsymbol{\pi}}$. Such uses of smoothing are implicitly based on the continuity assumption. Clearly, inaccurate values of $\hat{\boldsymbol{\pi}}$ will affect (??) and be carried forward into (1) and (3). Therefore, Pounds (2006) recommends that methods using $\hat{\boldsymbol{\pi}}$ estimators not be used in applications involving very discrete p-values.

2.4 Violation of the assumptions

Clearly, discrete p-values violate the continuity assumption. Violation of the continuity assumption can destabilize $\hat{\boldsymbol{\pi}}$ estimators. Therefore, as mentioned above, methods that estimate $\hat{\boldsymbol{\pi}}$ may not be conservative when applied to a set of highly discrete p-values. Of the methods mentioned above, only Benjamini and Hochberg (1995) and Gilbert (2005) do not use $\hat{\boldsymbol{\pi}}$ estimators. Each of these methods have been shown to have desirable control properties; however, these methods are not suitable for applications in which FDR estimation is preferred over FDR control. If interpreted as FDR estimates, the FDR-adjusted p-values (Reiner *et al*., 2003) computed by these methods will understate the actual prevalence of false positives (Pounds and Cheng, 2004; Pounds, 2006). Storey (2002) and Pounds (2006) argue that estimation is preferred to control in many microarray applications. Therefore, a method that can conservatively estimate the FDR for applications giving discrete p-values needs to be developed.

Also, applications involving a large number of one-sided tests can violate assumption (4). We note that (4) holds (in the sense of equality rather than approximation) if and only if the true sampling distribution of the test statistic is identical to the null distribution used to compute the p-value. Consider testing $H_0 : \theta = \theta_0$ against the alterative $H_A : \theta > \theta_0$. The null distribution is derived under the assumptions of the statistical test and the null hypothesis $H_0 : \theta = \theta_0$. Thus, the uniformity assumption holds only under those assumptions as well. If $H_A : \theta > \theta_0$ (e.g. the 'tested alternative'), then the true sampling distribution of the test statistic is shifted to the right of the null distribution. The p-value is computed by the area in the right tail of the null distribution. Subsequently, the p-value is stochastically less than uniform, a desirable property under the alternative hypothesis. However, if $\theta < \theta_0$ (for our purposes, this 'untested alternative' could be considered a true null), then the sampling distribution is shifted to the left of the null distribution and the p-value is stochastically greater than uniform, thus violating the assumption of uniformity. The violation occurs simply because the untested alternative is true. This type of violation of the uniformity assumption does not occur in the context of two-sided tests because both alternatives are tested.

The violation of uniformity owing to the 'untested alternative' in one-sided testing also occurs when non-parametric tests or permutation tests are used to compute the p-values. We illustrate by making the example above more specific. Suppose we wish to means of compare two populations μ_1 and μ_2. Let $\theta = \mu_2 - \mu_2$ and $\theta_0 = 0$. Thus, we are testing $\mathrm{H}_0 : \mu_2 - \mu_1 = 0$ against $H_A : \mu_2 - \mu_1 > 0$. Suppose we use the t-statistic as the test statistic, but evaluate the p-value via permutation. By symmetry, the distribution of test statistics derived by permutation center about 0. If $\mu_2 - \mu_1 < 0$, the true sampling distribution of the test statistic is centered around some negative value. The p-value is computed in the right side of the empirical null

distribution derived by permutation. Thus, the p-values center around some value >0.5 and are clearly non-uniform.

Now, consider how the distribution of a set of p-values computed from a large number of one-sided tests might appear. There are three cases: the null hypothesis is true, the tested alternative is true, and the untested alternative is true. Supposing an appropriate test was performed, these cases yield uniform, stochastically less than uniform, and stochastically greater than uniform distributions, respectively. Thus, the observed distribution is a mixture of these three components and has a U-shape. This is precisely what Pounds and Cheng (2005a) observed in their assessment and development of methods that use Affymetrix present–absent p-values. The present–absent p-values are based on a set of corresponding one-sided Wilcoxon signed-rank tests.

The U-shaped p-value distribution that occurs in the context of one-sided testing can impact FDR estimation or control in unpredictable ways. The increased density of p-values near $p = 1$ can lead to overly conservative estimates of $\hat{\boldsymbol{\pi}}$ that introduce conservative bias into $t_{(i)}$ via (2). However, the high density of p-values also increases $\hat{\mathbf{F}}(\alpha)$ for large α, which can actually cause $t_{(i)}$ to decrease for large $p_{(i)}$ via (1). The smaller $t_{(i)}$ for large $p_{(i)}$ could then introduce downward bias into $q_{(i)}$ via the minimization operation in (3). Thus, the conventional approaches defined by (1)–(3) may yield unreliable results (too liberal or too conservative) in the context of one-sided tests.

In Section 2.3, we noted that violation of the assumption (4) affects the $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\upsilon}}(\cdot)$, and $\hat{\mathbf{F}}(\cdot)$ components of FDR estimation or control. In this section, we have noted that one-sided testing can violate assumption (4) and lead to a p-value distribution that is U-shaped in practice. In addition, we have noted that discrete p-values can introduce instability into $\hat{\boldsymbol{\pi}}$ estimators. Therefore, development of $\hat{\boldsymbol{\pi}}$, $t\upsilon(\cdot)$ and $\hat{\mathbf{F}}(\cdot)$ estimators suitable for one-sided testing that may or may not yield discrete p-values is warranted. We now proceed by developing these estimators in Sections 2.5–2.7.

2.5 Estimator of the significant proportion

The method we propose uses the empirical distribution function as our estimate $\hat{\mathbf{F}}(\alpha)$ of the expected proportion $\mathbf{F}(\alpha)$ of tests that will yield a p-value less than or equal to α. Suppose that there are d distinct p-values among p. Let $\tilde{p}_1 < \tilde{p}_2 < \cdots < \tilde{p}_d$ represent those distinct p-values. For $j = 1, \ldots, d$, let m_j be the number of p-values among p that are equal to $\tilde{p}_j$. Thus,

$$\hat{\mathbf{F}}(\alpha) = \frac{1}{m}\sum_{j=1}^{d} \mathrm{I}(\tilde{p}_j \le \alpha) m_j, \tag{5}$$

where $\mathbf{I}(\cdot)$ is the indicator function that is defined as 1 if the enclosed statement is true and 0 otherwise. Note that (5) is well-defined for any set of observed p-values, continuous or discrete, one-sided or two-sided. Therefore, in the context of two-sided tests, if (5) is used in conjunction with a conservative estimator $\hat{\boldsymbol{\upsilon}}(\alpha)$ of $\upsilon(\alpha)$, then the resulting $t_{(i)}$ are conservative pFDR estimators when (2) is used for the numerator of (1), as shown by Storey (2002).

2.6 Estimator of the null proportion

We propose an estimator $\hat{\boldsymbol{\pi}}$ of π for each of four cases: (a) p-values are two-sided and continuous, (b) p-values are two-sided and discrete, (c) p-values are one-sided and continuous and (d) p-values are one-sided and discrete. Briefly our estimator can be expressed as

$$\hat{\boldsymbol{\pi}} = \begin{cases} \min(1, 2\bar{p}) \text{ for cases (a) and (b)}, \\ \min(1, 2\bar{a}) \text{ for cases (c), and} \\ \min(1, 8\bar{a}) \text{ for case (d)}, \end{cases} \tag{6}$$

where

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m} p_i,$$

$$\bar{a} = \frac{1}{m}\sum_{i=1}^{m} a_i,$$

and

$$a_i = 2\ \min(p_i, 1 - p_i) \tag{7}$$

for each i. We note several practical and theoretical advantages of the estimators proposed in (6). First, the estimators are well defined and easily computed for any set of p-values and do not rely on an implicit assumption of continuity. Thus, the estimators should perform well even when p-values are discrete. Second, for π sufficiently < 1, the estimator has conservative bias, e.g. $\mathbf{E}(\hat{\boldsymbol{\pi}}) \ge \pi$, when π is sufficiently < 1. This later property is critical for the proposed method to maintain conservativeness in the ratios of (a).

The conservativeness of $\hat{\boldsymbol{\pi}}$ in cases (a) and (b) for sufficiently small π is obvious. Assuming that $E(p_i) \ge 0$ for all i and $E(p_i) \ge 1/2$ for each test i with a true null hypothesis, both of which hold by the definition of a two-sided p-value, it follows that

$$\mathrm{E}(\bar{p}) \ge \frac{\pi}{2},$$

which proves that $2\mathrm{E}(\bar{p}) \geqq \pi$. Therefore, if π is small enough so that $\Pr(2\bar{p} \le 1) \approx 1$, then $\mathrm{E}(\hat{\boldsymbol{\pi}}) \ge \pi$. In addition, if $E(p_i) \to 0$ as the statistical power of each test i with a false null increases, then $\mathrm{E}(2\bar{p}) \to \pi$ as well. If π is sufficiently small, then $\mathrm{E}(\hat{\boldsymbol{\pi}}) \to \pi$ as the statistical power approaches 1 for any fixed level. The conservativeness for cases (c) and (d) following by thinking of (9) as a p-value for a two-sided test. More detailed proofs are provided in the supplementary materials.

2.7 Estimator of the false discovery proportion

We propose an estimator $\hat{\boldsymbol{\upsilon}}(\alpha)$ of the false discovery proportion $\upsilon(\alpha)$ that is well-defined for one-sided or two-sided tests. In the two-sided case, we propose substituting $\hat{\boldsymbol{\pi}}$ from (6) into (2). As previously noted, $\hat{\boldsymbol{\pi}}$ from (6) is conservative for π sufficiently < 1, so that the estimate $\upsilon(\alpha)$ should be conservative as well.

For one-sided tests, we propose that

$$\hat{\boldsymbol{\upsilon}}(\alpha) = \begin{cases} \hat{\boldsymbol{\pi}}\alpha & \text{if } \alpha \le \frac{1}{2} \\ \frac{\hat{\boldsymbol{\pi}}}{2} + \hat{\mathbf{F}}(\alpha) - \hat{\mathbf{F}}\left(\frac{1}{2}\right) & \text{if } \alpha > \frac{1}{2} \end{cases} \tag{8}$$

be used to estimate $\upsilon(\alpha)$, with $\hat{\boldsymbol{\pi}}$ given by (6). The estimator is modified for $\alpha \ge 1/2$ to prevent the possibility that the U-shaped distribution lead to small values of (1) for large α (see Section 2.4). Essentially, the modification counts every p-value > 0.5 as though it is testing a true null hypothesis and counts it toward the numerator of the ratio in (1). Therefore, the estimator (8) is conservative for $\alpha \ge 1/2$ if it is conservative for $\alpha = 1/2$. In the Supplementary

Materials, we show that estimator (8) is conservative for $\alpha \leq 1/2$ when (6) is used to compute $\hat{\pi}$.

2.8 Smoothing local FDR estimates

Now, substituting (5) for $\hat{\mathbf{F}}(\alpha)$ and (2) or (8) for $\hat{\boldsymbol{\upsilon}}(\alpha)$ into (1) to yields rough (i.e. unsmoothed) *l*FDR estimates. Pounds and Cheng (2004) have noted in simulation studies that rough *l*FDR estimates can be unstable for small p and that smoothing can lead to more accurate and less variable cFDR estimates. We now propose a method to smooth the rough *l*FDR estimates that can be applied to continuous or discrete p-values. The method should be robust in the sense that its results are not heavily influenced by the discreteness of p-values or the instability of the rough *l*FDR estimates for small p.

Recall that $\tilde{p}_1 < \tilde{p}_2 < \cdots < \tilde{p}_d$ represent the d distinct p-values. Now, for $j = 1, \ldots, d$, define

$$\tilde{t}_j = \frac{\hat{\boldsymbol{\upsilon}}(\tilde{p}_j)}{\hat{\mathbf{F}}(\tilde{p}_j)}$$

as the rough *l*FDR estimate for the set of findings with a p-value less than or equal to $\tilde{p}_j$. For $j = 1, \ldots, d$, let $\tilde{t}_j^\star$ represent the rank of $\tilde{t}_j$ among $\tilde{t} = \{\tilde{t}_1, \ldots, \tilde{t}_d\}$ and $\tilde{p}_j^\star$ represent the rank of $\tilde{p}_j$ among $\tilde{p} = \{\tilde{p}_1, \ldots, \tilde{p}_d\}$. Given these definitions, we propose to compute smooth *l*FDR estimates as follows:

ALGORITHM 1. Smoothing Local FDR Estimates.

(1) Apply least trimmed-squares regression (Rousseeuw, 1984) with $\tilde{\mathbf{p}}^\star = \{\tilde{p}_1^\star, \ldots, \tilde{p}_d^\star\}$ as the x-variable observations and $\tilde{\mathbf{t}}^\star = \{\tilde{t}_1^\star, \ldots, \tilde{t}_d^\star\}$ as the corresponding y-variable observations.

(2) Use the regression fit obtained in step 1 to compute the predicted values $\hat{\mathbf{t}}^\star = \{\hat{t}_1^\star, \ldots, \hat{t}_d^\star\}$ of $\tilde{\mathbf{t}}^\star$ at $\tilde{\mathbf{p}}^\star$.

(3) Use the linear interpolation scheme described by Iman and Conover (1979) to transform the estimates $\hat{\mathbf{t}}^\star$ of $\tilde{\mathbf{t}}^\star$ into smoothed *l*FDR estimates $\hat{\mathbf{t}}$ of $\tilde{\mathbf{t}}$. That is, for $j = 1, \ldots, d$, determine i_j such that $\hat{t}_j^\star$ falls in the interval $(\tilde{t}_{i_j}^\star, \tilde{t}_{i_j+1}^\star]$ and let

$$\hat{t}_j = \tilde{t}_{i_j} + \left(\frac{\tilde{t}_{i_j+1} - \tilde{t}_{i_j}}{\tilde{t}_{i_j+1}^\star - \tilde{t}_{i_j}^\star}\right) \times \left(\hat{t}_j^\star - \tilde{t}_{i_j}^\star\right).$$

Under the estimation paradigm, each entry $\hat{\mathbf{t}}$ is an estimate of the proportion of results with a p-value less than or equal to the corresponding entry of $\tilde{\boldsymbol{p}}$ that are false discoveries. Under the control paradigm, the values of $\hat{\boldsymbol{t}}$ can be used in place of $\boldsymbol{t}$ in (3), then the resulting values of $\boldsymbol{q} = \{q_{(1)}, \ldots, q_{(m)}\}$ can be compared to a preselected threshold τ.

The smoothed *l*FDR estimates computed by Algorithm 1 should have desirable statistical properties. First, the estimates should be conserative, i.e., they should tend to accurately state or overstate the occurrence of false discoveries, except when $\pi \approx 1$, as previously noted. The rough estimates use Storey's (2002) estimator for the denominator and a conservative estimate of $\upsilon(\alpha)$ for the numerator. Storey (2002) has shown his pFDR estimator, e.g. his method of computing $t_{(i)}$ in (1), to be conservative when the numerator is conservative; therefore, the proposed rough *l*FDR estimator should also be conservative. The regression procedures mentioned in Algorithm 1 operate directly on rough *l*FDR estimates that are conservative, and thus the predicted values produced by the regression procedures will tend to be conservative. When $\pi \approx 1$, the smoothed *l*FDR estimates should tend to be quite large, indicating that most significant results are false discoveries. In practice, such results would either not be reported or be reported with appropriate disclaimers. The regression procedures will effectively smooth the *l*FDR estimates and avoid the problems of unstable rough *l*FDR estimates observed by Pounds and Cheng (2004). The least trimmed-squares regression procedure is applied to the ranks of the distinct p-values and distinct rough *l*FDR estimates. If the rough *l*FDR estimates are already monotone in the p-value, then the fit of the ranks will be perfect, and the smooth *l*FDR estimates will equal the rough *l*FDR estimates. Otherwise, the least trimmed-squares regression will fit a curve to the ranks, which is then transformed into smooth local FDR estimates. Conover (1999) notes that such regression on ranks gives very good estimates for monotone functions. It is reasonable that the FDR should be a monotone increasing function of the p-value threshold α if all statistical tests are well-constructed and correctly performed (Liao *et al*., 2004). Rousseeuw (1984) has shown that the least trimmed-squares regression procedure is robust against outliers; thus, $\hat{\mathbf{t}}$ should not be heavily influenced by the potential instability of the rough *l*FDR estimates for very small p.

3 RESULTS

We now examine the performance of the proposed method via two case studies and a simulation study based on resampling from a collection of modified real datasets. We also compare the performance of the proposed method with those of five earlier methods. For our purposes, we note those methods as follows: the method of Benjamini and Hochberg (1995), BH95; Storey (2002), St02; Pounds and Morris (2003), PM03; Pounds and Cheng (2004), PC04 and Cheng *et al*., (2004), Ch04. Each method was implemented using our S-plus FDR routine library with default values for all optional tuning parameters. The library is freely available from www.stjuderesearch.org/depts/biostats.

3.1 Randomization tests with small sample size

Gadbury *et al*. (2003) describe an experiment comparing the expression of 12 625 features across two groups. Only three replicate arrays were produced for each experimental group. A randomization test was used to compute the p-value for each feature represented on the array. As a consequence, the observed p-values were very discrete (Table 1 and Figure S1, Supplementary Materials): the possible p-values were 0.1, 0.2, ..., 1.0. It is clear that the observed distribution is not a discrete uniform distribution, indicating that a high proportion of features are indeed differentially expressed across the two experimental groups. In fact, Gadbury *et al*. (2003) report $\chi^2 = 2\,795$ with nine degrees of freedom for testing whether all null hypotheses are true. Therefore, BH95 will be overly conservative in this setting, because it uses $\hat{\pi} = 1$ in its calculations. Likewise, St02 estimates $\hat{\pi} = 0.94$, which also appears conservative given the observed p-value distribution. The density estimates produced by PM03, PC04 and Ch04 poorly represent the observed p-value density (Figure S1, Supplementary Materials). The proposed method estimates that 38.4% of the results with $p = 0.1$ are false discoveries. This is the least conservative estimate among all methods examined. The proposed method should produce estimates that either accurately represent or overstate the actual

Table 1. Example from Gadbury *et al*. (2003)

P	*m*(*p*)	BH95	St02	PM03	PC04	Ch04	Prop.
0.1	2975	0.848	0.799	0.779	0.465	0.743	0.385
0.2	1457	0.682	0.642	0.821	0.502	0.743	0.517
0.3	1261	0.748	0.705	0.846	0.543	0.761	0.603
0.4	1137	0.806	0.760	0.864	0.586	0.805	0.671
0.5	1058	0.858	0.808	0.879	0.624	0.854	0.726
0.6	1002	0.903	0.851	0.891	0.658	0.898	0.773
0.7	934	0.944	0.890	0.901	0.688	0.931	0.816
0.8	931	0.982	0.925	0.910	0.715	0.952	0.852
0.9	957	1.000	0.953	0.918	0.738	0.961	0.880
1.0	913	1.000	0.978	0.925	0.758	0.964	0.907

The table above gives the *p*-value, the number *m*(*p*) of tests giving the *p*-value, and the FDR estimates obtained by various methods and the proposed method.

Table 2. Example from Pounds and Cheng (2005a)

α	EDF(α)	BH95	St02	PM03	PC04	Ch04	Prop.
0.001	0.422	0.002	0.002	0.001	<.001	0.004	0.001
0.005	0.458	0.011	0.011	0.006	<.001	0.013	0.003
0.010	0.475	0.021	0.021	0.011	<.001	0.023	0.006
0.050	0.528	0.095	0.095	0.044	0.001	0.102	0.027
0.100	0.559	0.179	0.179	0.079	0.002	0.193	0.051
0.250	0.608	0.411	0.411	0.165	0.006	0.426	0.118
0.500	0.668	0.749	0.749	0.273	0.012	0.736	0.214
0.750	0.732	1.000	1.000	0.356	0.017	0.942	0.284
0.900	0.787	1.000	1.000	0.398	0.021	0.993	0.333
1.000	1.000	1.000	1.000	0.423	0.023	1.000	0.475

The table above shows α, the EDF at α, and the FDR estimates (or FDR-adjusted *p*-values) obtained by each method.

occurrence of false discoveries, as noted in Section 2.8. Therefore, in this example, it appears that the proposed method gives the most accurate estimate of the proportion of results with $p = 0.1$ that are false discoveries.

3.2 Filtering poorly hybridized Affymetrix probe sets

Pounds and Cheng (2005a) develop two filtering methods to remove poorly hybridized Affymetrix probe sets from subsequent analysis. They describe how to pool the present–absent *p*-values for a particular probe set across chips into a single summary *p*-value. They call this summary *p*-value the pooled *p*-value. There is a pooled *p*-value for each probe set. In this application, the pooled *p*-value is based on a one-sided test because we wish to include only those probe sets with high perfect-match signals relative to the mismatch signals. A smaller pooled *p*-value indicates statistical evidence that the perfect match signal is significantly greater than the mismatch signal and that the probe set should be called 'present' and included in later analysis. Pounds and Cheng (2005a) compute the pooled *p*-value for one of the experimental groups in the study described by Pounds and Morris (2003). The resulting distribution was U-shaped because the statistical tests were one-sided (null suggests that perfect match signal tends to be less than or equal to the mismatch signal). In addition, for $\sim 11\%$ of the probe sets, the pooled *p*-value was numerically equal to 0. Pounds and Cheng (2005a) note that neither PC04 or St02 in their published forms can adequately address the challenges posed by this unusual distribution. Pounds and Cheng (2005a) utilize components of each of these two procedures to implement their error-minimizing pooled *p*-value filter.

The proposed method and the other methods defined above were applied to this challenging set of *p*-values. The results for small *p*-values are shown in Table 2. All methods agree that very few false positives are included with a very small α; but they estimate very different rates for the accrual of false positives as α increases. Only PC04 obtains less conservative results than the proposed method. However, as observed by Pounds and Cheng (2005a), PC04 does not provide a good fit to the observed *p*-value distribution (Figure S2, Supplementary Materials). Therefore, since the proposed method should be conservative, it appears to obtain the most accurate results in this example as well.

3.3 Resampling-based simulation studies

Ross *et al*. (2004) used Affymetrix U133A chips to profile the gene expression in pediatric AML cells. They explored numerous objectives in that work, including comparisons of expression across various AML subtypes. In this work, we examined the performance of the proposed method in a (two-group) comparison of expression between the *t*(8;21) and MLL subtypes of the disease (the most well-represented subtypes in that study). Therefore, we extracted their data, normalized using the Affymetrix MAS 5.0 algorithm, for the patients diagnosed with the *t*(8;21) and MLL subtypes of that disease ($n = 21$ and 23, respectively). We then applied shift and scale transformations to the log-transformed data of each disease subtype, as described in greater detail below, to create modified datasets to perform resampling-based simulation studies to explore the performance of the proposed method and compare it with those of BH95, St02, PM03, PC04 and Ch04. We note that shift and scale transformations maintain the Pearson and Spearman measures of correlation between the expression levels of the probe sets.

We created five datasets. The first dataset was derived by *z*-transforming (subtracting the mean, then dividing by the standard deviation) the log-expressions of each probe set within each group (i.e. disease subtype). Thus, the first dataset represents an all null case: for each probe set, the mean log-expression is equal across the two groups. The second dataset was created by randomly selecting 10% of the probe sets, and then randomly choosing to add or subtract $\delta = 1/2$ to the log-expression of each subject in the *t*(8;21) group. Thus, in the second dataset, the two group means are equal for $\pi = 0.9$ of the probe sets and the mean log-expression differs by $\delta = 1/2$ of a standard deviation for the other probe sets. The third dataset was created by taking the first dataset and adding or subtracting $\delta = 1$ to the log-expression of each subject in the *t*(8;21) for 10% of the probe sets that were randomly chosen. The fourth dataset was created so that the mean log-expression differed by $\delta = 1/2$ across the two groups for 50% of the probe sets. In the fifth dataset, the mean log-expression differed by $\delta = 1$ in 50% of the probe sets.

A series of resampling-based simulations was performed using these datasets. In each simulation, a given number ($n = 3$, 5 or 10) of samples was randomly chosen without replacement from each disease subgroup to create a dataset. Next, the exact Wilcoxon

rank-sum procedure (one-sided or two-sided) was applied to the data of each feature to test for differential expression between the two groups. The FDR methods (BH95, St02, PM03, PC04, Ch04 and the proposed method) were applied to the resulting p-values. In addition, the number of false discoveries observed at each p-value threshold was noted. The process was repeated 100 times, and the results from each repetition were recorded. The Supplementary material shows the results of all simulations (Figures S32–S33).

The simulations of one-sided testing in the all-null setting support our theoretical concerns outlined in Section 2.4 (Supplementary Section S3 and Figure S3). In particular, the p-values from tests with a true null hypothesis roughly satisfy (4), p-values from tests with a true alternative hypothesis are stochastically less than uniform, and p-values from tests with a true 'untested alternative' are stochastically greater than uniform. We note that one-sided p-values with true 'untested alternatives' violate assumption (4).

Figure 1 shows the results of the simulation that resampled from the first dataset (no differential expression), chose $n = 10$ subjects from each subgroup, and applied an exact left-sided rank-sum test to the acquired datasets. In this setting, the p-values are fairly discrete; there are 101 possible p-values when the expression data have no ties. With over 100 possible p-values, one might be inclined to believe that the discreteness would not introduce major problems into the analysis. However, the discreteness causes St02, PC04 and Ch04 to produce very misleading results. The discreteness of p-values leads the π-estimator to become unstable, resulting to unrealistically optimistic estimates that half or more of the tested null hypotheses are false. In addition, St02 and BH95 apply the right-sided minimization in (3) prior to using an operation to obtain a smooth $\hat{\mathbf{F}}(\cdot)$. Pounds and Cheng (2004) have noted that applying the minimization operation prior to smoothing can introduce substantial liberal bias into the final results. The problems are most severe for Ch04 and PC04; each of these methods assumes continuity in computing both $\hat{\mathbf{F}}(\alpha)$ and $\hat{\boldsymbol{\pi}}$. St02 suffers less in this setting because it assumes continuity only in computing $\hat{\boldsymbol{\pi}}$. Still the instability of $\hat{\boldsymbol{\pi}}$ leads to a dramatic understating of the actual prevelance of false positives. PM03 performs well in this particular setting, because it fits a uniform model to the observed p-values in many of the replications. However, PM03 would likely not be routinely used in practice, because the discreteness of the p-values clearly violates the model assumptions (Pounds, 2006). The proposed method obtains more conservative estimates than even BH95 in this setting.

Thus, in this simulation, the proposed method suggests more clearly than any other method the correct conclusion that there are no differentially expressed genes in this dataset. In the other simulations of no differential expression, the proposed method also clearly suggested this conclusion (Supplementary Figures S4–S9) but other methods did not always do so (Suplementary Figures S5 and S6).

The proposed method maintained its conservativeness (in the sense that the expected value of the estimator is greater than the simulation estimate of the actual FDR for all p-value cutoffs) in the simulations that resampled from the second dataset, in which the means of 10% of features differ by $\delta = 1/2$ (Supplementary Figures S10–S15). However, in some of these simulations, St02, PC04, and Ch04 did not maintain their conservativeness (Supplementary Figures S11 and S12). Additionally, in these simulations, the proposed method appears to be one of the most powerful among methods maintaining their conservativeness.

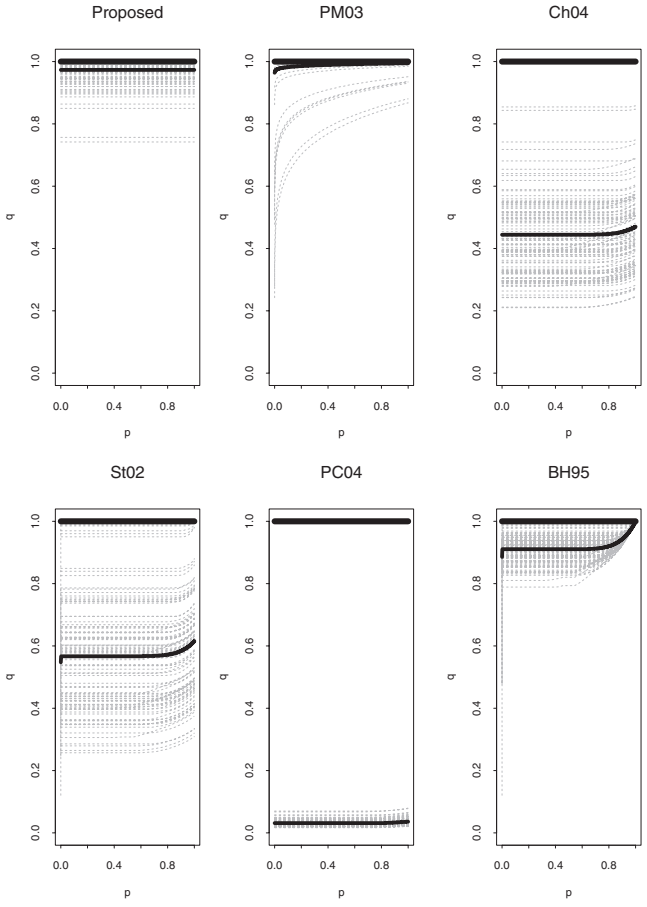

Fig. 1. Simulation results with one-sided Tests, $n = 10$, and $\pi = 1.0$. Each panel above depicts the simulation performance of one method with q from (3) on the y-axis and α on the x-axis. Each dashed line shows the results from one replication. The thin solid line shows the average across replications. The thick solid line indicates the simulation estimate of the actual false discovery ratio as a function of the p-value threshold α. An enlarged version is shown in Figure S6 (Supplementary Materials).

Patterns similar to those observed for the second dataset are also seen in the simulations from the third, fourth, and fifth datasets (Supplementary Figures S16–S21, S22–S27, and S28–S33, respectively). The proposed method maintains its conservativeness across all simulations. In some of these simulations, the other methods do not maintain their conservativeness. In each simulation, the power of the proposed method is among the best of methods that maintain their conservativeness.

In addition, the simulation results support our theoretical results regarding the relationship between the expected value of the proposed $\hat{\boldsymbol{\pi}}$ estimator and statistical power. For example, Supplementary Figures S31–S33 clearly show that the expected value of the proposed $\hat{\boldsymbol{\pi}}$ estimator (indicated by the height of the lines at $p = 1$) decreases as the sample size, and hence statistical power, increases.

4 DISCUSSION

We have proposed a simple, effective and conservative method to estimate the proportion of findings with a p-value less than a

threshold α that are false discoveries. The need for accurate estimation of FDR-type measures in the setting of one-sided or discrete p-values motivated us to develop this method. We have proven that the method has desirable theoretical properties, including conservative bias when π is sufficiently < 1. In addition, in simulation studies we have observed that the proposed FDR estimator has a high expected value when π is near 1. Through case studies and simulation, we have shown examples in which the proposed method gives more accurate results than do existing methods. Some methods have tuning parameters that may impact their performance; however, in our view, it is unlikely that tuning will resolve the gross deficiencies seen in some of the examples and simulations. In its current implementation, the proposed method does not use any tuning parameters. Furthermore, in simulation studies in which existing methods performed well, the proposed method yielded results comparable with those of the other methods. This shows that our method is robust against the the impact of one-sided or discrete p-values. We anticipate that the proposed method will yield conservative FDR estimates across the wide variety of p-value distributions that can arise in practice.

The proposed method may offer many advantages in applications with two-sided tests and continuously distributed p-values as well. Pounds and Cheng (2004) have noted that the rough *l*FDR estimates need some smoothing to yield stable *l*FDR estimates for small p. However, it is not obvious that the method they propose to obtain smooth *l*FDR estimates generally yields conservative estimates. The proposed method simply smooths a set of rough *l*FDR estimates that has already been shown to be conservative (Storey, 2002). Therefore, the smooth estimates should tend to be conservative; however, this has not been rigorously proven. Nevertheless, it is clear that the proposed method of smoothing should yield more conservative values for q_i than Storey's (2002) method, which does not employ smoothing prior to implementing (3). Unlike Allison *et al*. (2002) and Pounds and Morris (2003), the proposed method does not make model assumptions that are potentially problematic. The estimator $\hat{\pi}$ is clearly conservative so long as π is sufficiently <1. When $\pi \approx 1$ or $\pi = 1$, the smooth *l*FDR estimates produced by the proposed method are so large that they would not cause difficulties in practice. Also, by smoothing the rough *l*FDR estimates, the proposed method circumvents the problems encountered by Storey's (2002) method that occur when (3) is applied to unstable *l*FDR estimates (Pounds and Cheng 2004). Thus, the smooth *l*FDR estimates produced by the proposed method should be conservatively accurate in many settings, and the operation of (3) can safely be applied to the smooth *l*FDR estimates to perform a conservative pFDR control.

In light of our results, Pounds (2006) recommendations for choosing an appropriate FDR method for specific applications should be updated. When FDR estimation (i.e. reporting an estimate of the prevalence of false positives among results deemed significant) is preferred over FDR control (i.e. ensuring that the FDR is kept below a pre-selected threshold), we now recommend that the proposed method be used for applications that involve one-sided or discrete p-values. Storey (2002) argues that estimation is often preferred over control because it is difficult to pre-specify an appropriate control level. Therefore, Pounds (2006) recommends that the control paradigm only be applied when statistical power calculations (Pounds and Cheng, 2005b) are used in planning a study. The proposed method can be also considered a reasonable choice for applications giving two-sided, continuously distributed p-values as well. However, the methods of Yekutieli and Benjamini (1999) or Benjamini and Yekutieli (2001) should still be preferred for applications in which a large proportion of the genes have strongly correlated expressions; still, the interpretation should keep in mind that these methods are developed under the FDR control paradigm. Experiments using disease- or pathway-focused arrays may have the correlation structure mentioned above (Pounds, 2006); this issue is not likely to arise in genome-wide studies (Storey and Tibshirani, 2003).

With some minor modifications, the power of the proposed method can be improved when the expected value of each p-value under the null hypothesis is known. If this is the case, then one can replace $2\bar{p}$ in (6) with $\frac{1}{m}\sum_{i=1}^{m}\frac{p_i}{\mathrm{E}(p_i \mid i \in \mathcal{H}_0)}$. The revised estimator will still be conservative when π is sufficiently <1. The expected value for each term with $i \in \mathcal{H}_0$ is 1 and thus the expected value of the sum is greater than or equal to π. In the example of Section 3.1, the expected value of each p-value under the null hypothesis is 0.55; thus, the revised estimate would be $\hat{\pi} = 0.83$. Using the revised value of $\hat{\pi}$, our method estimates that 35% of the results with $p = 0.1$ in that example are false discoveries. In most applications, this simple revision should be possible. Computing a p-value requires that the null distribution of the test statistic is known or approximated. Therefore, the null distribution of the p-value can be determined or approximated as well. We chose to use (6) in describing the method and developing our software because it is computationally simple. Furthermore, we chose to use the estimatorin (6) because it may be very burdensome to determine the expected value of each p-value under the null hypothesis in some applications.

From a theoretical perspective, the p-values arising from a randomization test will become less discrete as the sample sizes increase. An interesting research topic is to determine a sample size at which the discreteness is negligible. In our simulation studies, the comparison of two groups with 10 replicates each still gave sufficiently discrete p-values to cause St02, PC04 and Ch04 to give very inaccurate and misleading results. Therefore, it appears that the discreteness is an issue even at this moderate sample size. A comparison of Figures S1–S3 (Supplementary Materials) reveals that the problems caused by discreteness actually worsened in the all-null setting as the per-group sample size increased from 3 to 5 to 10. Thus, the severity of the problems introduced by discreteness does not appear to be a monotonically related to sample size. As the sample size increases, the positions of possible p-values are shifted in a complex manner; thus, it is difficult to predict at what point the discreteness becomes negligible. At some point, the number of possible assignments is so large that permutation tests are used to approximate exact tests. In that setting, the possible p-values are equally spaced, and the discreteness did not appear to be an issue in the work of Pounds and Cheng (2004). Therefore, whenever exact testing is feasible, the proposed method is clearly preferred over the other methods examined in our simulation studies.

Some of our simulation results should remind investigators to interpret the results of a microarray experiment with caution even when the best available statistical tools are used in the analysis. Figure S2 (Supplementary Materials) shows one replication in which all five methods suggested that there may be some differentially expressed features in a dataset with no differential gene

expression. In addition, investigators should remember that the proposed method can perform well only when accurate *p*-values are computed, as Pounds (2006) has previously noted. In applications that involve large-scale multiple testing, such as the analysis of genomics data, it is imperative to obtain very accurate *p*-values. Many times, analytical approximations can be inaccurate for very small *p*-values. Therefore, exact approaches are often preferred approaches for computing *p*-values that will be used in subsequent FDR calculations. When exact approaches are infeasible, permutation tests can be applied with the number of permutations very large so that the *p*-values can be accurately estimated to many decimal places for performing multiple-testing adjustments.

For the case of discretely distributed *p*-values based on one-sided tests, the $\hat{\pi}$ estimator of (6) is very conservative. The Supplementary Materials describe a way to improve power in that case by replacing $8\bar{a}$ with $k\bar{a}$ where $k \leq 8$.

ACKNOWLEDGEMENTS

This research was supported in part by the NIH Cancer Center Support Grant (CC and SP; CA-21765), NIH/NIGMS Pharmacogenetics Research Network and Database (CC; U01 GM61393, U01 GM61374, http://pharmgkb.org/) and the American Lebanese Syrian Associated Charities. The authors thank Dr. Angela McArthur for editorial assistance.

Conflict of Interest: none declared.

REFERENCES

Allison,D. *et al*. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal*., **39**, 1–20.

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

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

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

Cheng,C. *et al*. (2004) Statistical significance threshold criteria for analysis of microarray gene expression data. *Stat. Appl. Genet. Mol. Biol*., **3**, 36.

Conover,W.J. (1999) *Practical Nonparametric Statitistics*. 3rd edn. John Wiley and Sons, New York, NY.

Iman,R.L and Conover,W.J. (1979) The use of the rank transform in regression. *Technometrics*, **21**, 499–509.

Gadbury,G.L. *et al*. (2003) Randomization tests for small samples: an application for genetic expression data. *Appl. Stat*., **52**, 365–376.

Gilbert,P.B. (2005) A modified false discovery rate multiplecomparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Appl. Stat*., **54**, 143–158.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lee,M-L T. *et al*. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Liao,J.G. *et al*. (2004) A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**, 2694–2701.

Lin,M. *et al*. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.

Pounds,S. (2006) Estimation and control of multiple testing error rates for microarray studies. *Brief.Bioinform.*, **12**, 25–36.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1745.

Pounds,S. and Cheng,C. (2005a) Statistical development and evaluation of gene expression data filters. *J. Comput. Biol*., **12**, 482–495.

Pounds,S. and Cheng,C. (2005b) Sample size determination for the false discovery rate. *Bioinformatics*, **21**, 4263–4271.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics*, **19**, 1236–1242.

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

Ross,M.E. *et al*. (2004) Gene Expression Profiling of Pediatric Acute Myelogenous Leukemia. *Blood*, **104**, 3679–3687.

Rousseeuw,P.J. (1984) Least median of squares regression. *J. Am. Stat. Assoc*., **79**, 871–881.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

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

Storey,J.D. *et al*. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Statist. Soc. B*, **66**, 187–205.

Tarone,R.E. (1990) A modified Bonferroni method for discrete data. *Biometrics*, **46**, 515–522.

Tsai,C-A. *et al*. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

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