BIOINFORMATICS

Vol. 20 no. 11 2004, pages 1737–1745
doi:10.1093/bioinformatics/bth160

Improving false discovery rate estimation

Stan Pounds* and Cheng Cheng

Department of Biostatistics, MS 262 St Jude Children's Research Hospital, 332 N. Lauderdale Street, Memphis, TN 38105-2794, USA

Received on June 20, 2003; revised on January 27, 2004; accepted on January 28, 2004
Advance Access publication February 26, 2004

ABSTRACT

Motivation: Recent attempts to account for multiple testing in the analysis of microarray data have focused on controlling the false discovery rate (FDR). However, rigorous control of the FDR at a preselected level is often impractical. Consequently, it has been suggested to use the *q*-value as an estimate of the proportion of false discoveries among a set of significant findings. However, such an interpretation of the *q*-value may be unwarranted considering that the *q*-value is based on an unstable estimator of the positive FDR (pFDR). Another method proposes estimating the FDR by modeling *p*-values as arising from a beta-uniform mixture (BUM) distribution. Unfortunately, the BUM approach is reliable only in settings where the assumed model accurately represents the actual distribution of *p*-values.

Methods: A method called the spacings LOESS histogram (SPLOSH) is proposed for estimating the conditional FDR (cFDR), the expected proportion of false positives conditioned on having *k* 'significant' findings. SPLOSH is designed to be more stable than the *q*-value and applicable in a wider variety of settings than BUM.

Results: In a simulation study and data analysis example, SPLOSH exhibits the desired characteristics relative to the *q*-value and BUM.

Availability: The Web site www.stjuderesearch.org/statistics/splosh.html has links to freely available S-plus code to implement the proposed procedure.

Contact: stanley.pounds@stjude.org

INTRODUCTION

Microarray technology is a promising development in the life sciences that allows investigators to measure the expression of thousands of genes simultaneously. The statistical analysis of microarray data presents numerous challenges that must be addressed so that the potential of this technology can be fully realized (Tilstone, 2003). One of these challenges is assessing the significance of findings in a meaningful way. The large number of statistical hypothesis tests conducted in the analysis of microarray data can potentially lead to a large number of false discoveries, i.e. significant findings that arise solely by chance mechanisms. It is clear that meaningful interpretation of microarray data strongly depends on the ability to reliably estimate or control the occurrence of false discoveries.

An effective approach to this problem is the control of the false discovery rate (FDR) in multiple tests introduced by Benjamini and Hochberg (1995). They consider the random variables V that represent the number of false discoveries made, and the number of significant results R obtained in a multiple-testing scenario. Benjamini and Hochberg (1995) then define a random variable $Q \equiv V/R$ when $R > 0$ and $Q = 0$ when $R = 0$. They define the FDR as

$$\mathrm{FDR} = \mathrm{E}(Q). \tag{1}$$

Benjamini and Hochberg (1995) propose a procedure that operates on a set of p-values and proved that this procedure controls the FDR at a pre-specified level η when the true null hypotheses' p-values are independent uniform (0,1) random variables. In the proof, they show that the procedure actually controls the FDR at a level $\eta g_0/g$, where g_0 of g tested null hypotheses are true. Subsequently, Benjamini and Hochberg (2000) introduced an adaptive FDR-controlling procedure that is at least as powerful as the 1995 procedure by employing an estimate $\widehat{g}_0$ of g_0.

Storey (2002) noted that FDR control may be infeasible in some settings because it is difficult to pre-specify a meaningful FDR-control level. Storey (2002) also suggested that the positive FDR (pFDR)

$$\mathrm{pFDR} = E\left(\frac{V}{R} \,\middle|\, R > 0\right) \tag{2}$$

is a more appropriate error metric than the FDR. Storey (2002) also proposed

$$\widehat{\mathrm{FDR}}(\alpha) = \frac{\hat{\pi}\alpha}{\hat{F}(\alpha)} \tag{3}$$

and

$$\widehat{\mathrm{pFDR}}(\alpha) = \frac{\hat{\pi}\alpha}{\hat{F}(\alpha)[1-(1-\alpha)^g]} \tag{4}$$

as estimators of the FDR and pFDR, respectively, when one declares significance for all p-values less than a threshold α, where $\hat{\pi}$ estimates the proportion π of tested hypotheses that are true and $F(\alpha)$ represents the observed proportion of tests

*To whom correspondence should be addressed.

yielding a p-value less than α. Storey (2002) subsequently defined the q-value as a pFDR analog of the p-value and proposed

$$\hat{q}_{(i)} = \min_{j \geq i} \left[\widehat{\text{pFDR}}(p_{(j)}) \right] \tag{5}$$

as a q-value estimator for $i = 1, 2, \ldots, g$, where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$ are the ordered p-values computed in the g tests.

While the contributions discussed above represent important advances in statistical science, they do not directly address the question 'How many false positives can one expect among the k most significant findings?' Storey and Tibshirani (2003) suggest interpreting the q-value as the expected proportion of false positives among a set of findings. However, such an interpretation of the q-value may be unwarranted since 'the exact operating characteristics' of (5) have been 'left as an open research problem' (Storey, 2002). In particular, the cumulative minimization operation in (5) alters the ratio interpretation of (4).

Tsai *et al*. (2003) discuss the use of the conditional FDR (cFDR), defined as

$$\text{cFDR} = E\left(\frac{V}{R} \;\middle|\; R = r \right) = \frac{E(V|R=r)}{r}. \tag{6}$$

The cFDR is clearly a natural measure of the number of false positives among the r most significant findings. Under Storey's (2002) mixture model, Tsai *et al*. (2003) show that

$$\text{cFDR}(\alpha) = \text{pFDR}(\alpha) = \frac{\pi\alpha}{F(\alpha)}, \tag{7}$$

where $F(\alpha) = \Pr(p \leq \alpha)$. Therefore, under Storey's (2002) mixture model, it is natural to consider a ratio of the form

$$\hat{r}(\alpha) = \frac{\hat{\pi}\alpha}{\hat{F}(\alpha)} \tag{8}$$

as an estimator of the proportion of false positives among the set of findings with $p \leq \alpha$.

Storey (2002), Allison *et al*. (2002) and Pounds and Morris (2003) have proposed methods to compute (8). The methods utilize different techniques to obtain $\hat{F}(\alpha)$ and $\hat{\pi}$. Storey (2002) sets $\hat{F}(\alpha)$ to the observed proportion of p-values less than α and uses a smoothing spline technique to estimate $\hat{\pi}$. Pounds and Morris (2003) assume the p-values arise from a beta-uniform mixture (BUM) model and let $\hat{F}(\alpha)$ and $\hat{f}(\alpha)$, respectively, be the cumulative distribution function (CDF) and probability distribution function (PDF) corresponding to the maximum-likelihood estimates (Casella and Berger, 1990) of the model parameters and set $\hat{\pi} = \min_\alpha \hat{f}(\alpha) = \hat{f}(1)$. Allison *et al*. (2002) consider a more flexible version of the BUM model by allowing more than one beta component in the mixture. The method of Pounds and Morris (2003) is guaranteed to yield smooth and monotone $\hat{r}(\alpha)$ but the reliability of its estimates depends on the validity of the BUM model assumptions (i.e. the BUM model can accurately represent the actual distribution of p-values). The method proposed by Allison *et al*. (2002) is somewhat impractical due to the computational demands of using the bootstrap (Efron and Gong, 1983) and is not guaranteed to produce monotone estimates. Storey's (2002) estimator is not restricted by such model constraints but frequently produces a fluctuating $\widehat{\text{pFDR}}(\alpha)$ that is difficult to interpret because one would anticipate that the FDR and pFDR should be monotonically non-decreasing functions in α when p-values are computed in a reasonable fashion. The cumulative minimization operation in (5) ensures that the q-value is monotone. However, since the properties of (5) are not well established (Storey, 2002) it is quite possible that $E(\hat{q}(\alpha)) \leq \text{FDR}(\alpha) \leq \text{pFDR}(\alpha)$ in some scenarios. In such a scenario, the q-value would underestimate the proportion of false discoveries among the associated set of findings.

It may prove beneficial to obtain smooth estimates of the cFDR before enforcing monotonicity with a minimization operation. Smooth estimates of the cFDR will have less tendency to be biased downward by such an operation than rough estimates. A smooth estimate of the cFDR could be obtained by using a smooth $\hat{F}(\alpha)$ in (8). This could be accomplished by treating $\hat{F}(\alpha)$ as an estimate of the p-value CDF $F(\alpha)$. Conceptually, $F(\alpha)$ would represent the CDF of the variable obtained by randomly selecting a hypothesis from the considered set of hypotheses and generating a realization of the p-value for testing that hypothesis. A smooth estimate of $F(\alpha)$ could be obtained by integrating an estimate of its derivative, the p-value PDF $f(\alpha)$. The Methods section proposes the spacings LOESS histogram (SPLOSH) as a method that estimates the cFDR using this type of approach. The Results section uses an example and a simulation study to compare the proposed method with Storey's q-value and BUM in terms of the ability to accurately represent the expected proportion of false discoveries among the k most significant findings. The Discussion section elaborates on the important insights that SPLOSH brings to the problem of cFDR estimation.

METHODS

Suppose an analysis of the association of g genes with the characteristic of interest results in a set of g p-values $p_1, \ldots, p_g$. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$ represent the ordered p-values, and $a_{(i)} = (i - 1/2)/g$ their adjusted ranks. Suppose that among $p_1, \ldots, p_g$ there are $\tilde{g}$ unique p-values $\tilde{p}_{(1)} < \tilde{p}_{(2)} < \cdots < \tilde{p}_{(\tilde{g})}$. For $j = 1, \ldots, \tilde{g}$, let $\tilde{a}_j$ be the average of $a_{(i)}$ for all i such that $p_{(i)} = \tilde{p}_{(j)}$, which is the average of r_i for the p-values equal to $p_{(j)}$. Additionally, define $\tilde{p}_{(0)} = 0$ and $\tilde{a}_{(0)} = 0$ if $\tilde{p}_{(1)} > 0$, and define $\tilde{p}_{(\tilde{g}+1)} = 1$ and $\tilde{a}_{(\tilde{g}+1)} = 1$ if $\tilde{p}_{\tilde{g}} < 1$. Henceforth, i will be used to index the original p-values (or quantities computed from them), and j will be used to index the ordered unique p-values (or quantities computed from them, including $j = 0$ and $j = \tilde{g} + 1$

when defined). Let l and u, respectively, represent the lower and upper indices j of the set $\tilde{p}_{(j)}$:

$$l = \begin{cases} 0 & \text{if } p_{(1)} > 0, \\ 1 & \text{otherwise,} \end{cases} \tag{9}$$

and

$$u = \begin{cases} \tilde{g} & \text{if } p_{(\tilde{g})} < 1, \\ \tilde{g} + 1 & \text{otherwise.} \end{cases} \tag{10}$$

For $j = l, \ldots, u-1$, define

$$m_j = \frac{\tilde{p}_{(j+1)} + \tilde{p}_{(j)}}{2}, \tag{11}$$

$$\Delta_j = \tilde{p}_{(j+1)} - \tilde{p}_{(j)}, \tag{12}$$

$$\delta_j = \frac{\tilde{a}_{(j+1)} - \tilde{a}_{(j)}}{\tilde{p}_{(j+1)} - \tilde{p}_{(j)}}, \tag{13}$$

$$\tilde{x}_j = \arcsin\left[2 \times \left(m_j - \frac{1}{2}\right)\right], \tag{14}$$

and

$$\tilde{y}_j = \log(\delta_j), \tag{15}$$

and for $i = 1, \ldots, G$, define

$$x_i = \arcsin\left[2 \times \left(p_i - \frac{1}{2}\right)\right]. \tag{16}$$

Note that m_j is the midpoint of the interval $[\tilde{p}_{(j)}, \tilde{p}_{(j+1)}]$, and δ_j is the slope of the empirical distribution function (Mason *et al*., 1989) over the same interval. Hence, δ_j gives an estimate of the p-value PDF $f(\cdot)$ over the interval $[p_{(j)}, p_{(j+1)}]$ or, more specifically, at m_j. Therefore, applying a local regression technique such as LOESS (Cleveland, 1993; Cleveland and Devlin, 1988) to the pairs $(\tilde{x}_j, \tilde{y}_j)$ gives an estimate $\widehat{y}(\cdot)$ of the PDF in a transformed space. An estimate $\widehat{f}(\cdot)$ of the p-value PDF can then be obtained from $\widehat{y}(\cdot)$ by back-transformation. Therefore, one can estimate the cFDR by using the SPLOSH algorithm, which is described below:

(1) Compute the quantities defined in (11) and (16).

(2) Apply LOESS to $(\tilde{x}_j, \tilde{y}_j)$ for $j = l, \ldots, u-1$ to obtain an estimated curve $\widehat{y}(\cdot)$.

(3) For $j = l, \ldots, u$, let $\widehat{f}^*(\tilde{p}_{(j)}) = \exp[\widehat{y}(\tilde{x}_{(j)})]$ be an estimate of $f(\tilde{p}_{(j)})$ up to a unitizing constant c.

(4) Let $\widehat{f}(p_i) = 1/c\,\widehat{f}^*(p_i)$ estimate the PDF at p_i for $i = 1, \ldots, G$, where

$$c = \frac{1}{2}\sum_{j=l}^{u-1}\left[\widehat{f}^*(\tilde{p}_{(j)}) + \widehat{f}^*(\tilde{p}_{(j+1)})\right]\Delta_j \tag{17}$$

is determined by trapezoid rule integration (Anton, 1992).

(5) Let $\widehat{F}(\tilde{p}_{(l)}) = 0$, and for $k = l+1, \ldots, u$ let

$$\widehat{F}(\tilde{p}_{(k)}) = \frac{1}{2}\sum_{j=l}^{k-1}\left[\widehat{f}(\tilde{p}_{(j)}) + \widehat{f}(\tilde{p}_{(j+1)})\right]\Delta_j \tag{18}$$

be an estimate of $F(\tilde{p}_{(j)})$ obtained by trapezoid rule integration.

(6) Following Efron *et al*. (2001), let

$$\widehat{\pi} = \min_{1 \le i \le g} \widehat{f}(p_i). \tag{19}$$

(7) For $i = 1, \ldots, g$, obtain $r_{(i)} \equiv \hat{r}(p_{(i)})$ by substituting $p_{(i)}$, $\widehat{\pi}$ and $\widehat{F}(p_{(i)})$ in (8). Additionally, use L'Hospital's rule (Anton, 1992)

$$\lim_{\alpha \to 0} \hat{r}(\alpha) = \lim_{\alpha \to 0} \frac{\hat{\pi}\alpha}{\widehat{F}(\alpha)} = \lim_{\alpha \to 0} \frac{\hat{\pi}}{f(\alpha)} \tag{20}$$

to motivate $\widehat{\pi}/\widehat{f}(0)$ as an estimate of the cFDR for p-values that equal 0.

(8) Following Storey (2002), define

$$h_{(i)} = \min_{k \ge i}\left(r_{(k)}\right) \tag{21}$$

as a monotone quantity based on the cFDR estimates $r_{(i)}$ for $i = 1, \ldots, g$.

This algorithm is called SPLOSH because it applies LOESS to the p-value spacings (Pyke, 1965) to obtain a histogram (the PDF estimate). Note that when $r_{(i)}$ is a smooth, generally increasing sequence in i that $h_{(i)}$ will not differ substantially from $r_{(i)}$ for any i. Hence, little downward bias is introduced by the minimization operation in (21).

In practice, numerical difficulties can arise when consecutive unique ordered p-values are very close to one another. To avoid these difficulties, one may wish to round the p-values to a specified number of decimal places before determining the set of unique ordered p-values $\tilde{p}_{(1)}, \tilde{p}_{(2)}, \ldots, \tilde{p}_{(\tilde{g})}$.

The log and arc-sine transformations are important components of the SPLOSH algorithm. The log-transformation of $\delta_{(j)}$ ensures that the PDF estimate will be strictly positive after back-transformation (Kooperberg and Stone, 1991). Additionally, Pyke (1965) has shown that spacings are asymptotically independent exponential random variables whose mean is given by the PDF in the neighborhood of the observations. Therefore, the log-transformation of $\delta_{(j)}$ helps to make the distribution of $y_{(j)}$ more symmetric, which is crucial to the success of the LOESS algorithm in obtaining a reasonable $\widehat{y}(\cdot)$. The arc-sine transformation of $m_{(j)}$ prevents the LOESS algorithm from overborrowing information from the center of the p-value distribution in estimating $\widehat{y}(\cdot)$ near $p = 0$ and $p = 1$. Overborrowing from the center tends to bias estimates of $f(p)$ for small p-values downward and bias estimates of $f(p)$ for large p-values upward. This will decrease $\widehat{F}(p)$ for

small p-values and increase $\widehat{\pi}$ resulting in substantially overestimating $r(p)$ across the entire p-value distribution (i.e. the numerator is inflated and the denominator is deflated). Overestimating $r(p)$ will be detrimental to the statistical power of inferential decisions based on $\hat{r}$ or $\hat{h}$. The arc-sine transformation still allows adequate borrowing from the center of the p-value distribution to stabilize $\hat{r}$. Unlike the logit transformation, the arc-sine transformation has the advantage of being defined at $p = 0$ and $p = 1$. Moreover, the authors' experience suggests that the logit transformation overly restricts the LOESS algorithm in borrowing information from the center to yield stable $\hat{r}$ at the extremes of the p-value distribution. Certainly, more rigorous mathematical examination of these statements is warranted. Nevertheless, the SPLOSH algorithm has proven useful in practice and performs well in simulation studies as will be demonstrated later.

RESULTS

The performance of SPLOSH, BUM and the q-value are now examined in the context of an example analysis and a simulation study. Storey's q-value is computed as published, i.e. using (4) and (5). BUM is used as published (Pounds and Morris, 2003) to compute (6). SPLOSH is also used to compute (6). SPLOSH is applied using the S-plus defaults (Insightful, 2002) for the tuning parameters of the LOESS algorithm. The S-plus default for the bandwidth is 0.75. To avoid numerical difficulties in the SPLOSH algorithm, p-values were rounded to six decimal places before determining the set of unique ordered p-values.

Analysis of the *Arf* targets

The SPLOSH algorithm and a comparison of its performance with that of the q-value estimator (Storey, 2002) and BUM (Pounds and Morris, 2003) are illustrated in the analysis of an experiment described by Kuo *et al*. (2003). The experiment was designed to screen for novel targets of the *Arf* gene on the Arf-Mdm2-p53 tumor suppressor pathway (Sherr 1998;Kuo *et al*. 2003). Both Affymetrix GeneChips and cDNA arrays were used. Here, we use the p-values from the cDNA array data for illustration.

The cDNA microarrays were printed from a murine clone library available at St Jude Children's Research Hospital. Arf and three known p53 targets on the pathway, $Mdm2$, $Cip1$ and $Btg2$ were spot-printed on the arrays. Samples from the reference and the *Arf*-induced cell lines were taken at 0, 2, 4 and 8 h. At each time point, three cDNA arrays were independently hybridized and scanned. There were 5776 probe spots on each array. Only the spots that passed a quality control of image analysis were used.

Expression levels were normalized and generated using ANOVA models (Wolfinger *et al*., 2001). Probe spots with less than two valid expression numbers at any time point were further filtered out, leaving 2936 spots for differential expression analysis. The ANOVA F-statistic was used to test for differential expression of each probe across the four time points. One concern was that the F-distribution may not approximate the null distribution accurately due to the small sample size at each time point; thus, a permutation test based on 1050 permutations (50 permutations in a 'test run' of the code + 1000 permutations in the 'analysis run' of the code) was applied to determine the p-values. The p-values were computed by counting the number of permutations yielding a larger ANOVA F-statistic than the observed F-statistic and dividing this number by 1050. Hence, p-values determined by this approach could equal $0 = 0/1050$, $1/1050$, $2/1050$, …,$1049/1050$ or $1050/1050 = 1$.

Only SPLOSH and BUM provide estimates of the p-value PDF. Figure 1 shows the BUM and SPLOSH estimates of the p-value PDF, $\widehat{f}(p)$, plotted against a histogram of the actual p-values. Clearly, SPLOSH more accurately represents the actual behavior of these p-values than does BUM. In particular, BUM grossly overestimates the null proportion $\widehat{\pi}$, which is roughly equivalent to the PDF at $p = 1$. The inadequate fit of the BUM model in this case causes one to question the quality of the corresponding estimates of error rates, such as the FDR. Figure 1 suggests that one can use SPLOSH in the context of the error-region approach proposed by Pounds and Morris (2003) to generate estimates of the error quantities they discuss. These error rate quantities include the FDR, empirical Bayes' posterior (Efron *et al*., 2001) and the number of each type of hypothesis testing outcome (false positives, true positives, false negatives, true negatives).

All three methods estimate the null proportion π. In this case, π is the proportion of the 2936 examined genes whose population mean expression is not altered by the experimental treatment. Table 1 gives the estimates of π obtained by the three methods. Figure 1 demonstrates that the null proportion should be close to 0.45 and hence Storey's method and SPLOSH provide the most reasonable estimates of π. As noted previously, BUM grossly overestimates π.

Figure 2 shows each method's estimates and monotone estimates as a function of α. BUM yields a monotone and smooth estimate, but as noted previously, the results are not reliable when such a great departure from model assumptions is observed. Storey's estimate $\widehat{\text{pFDR}}(\alpha)$ is not monotone in α over any interval of substantial length and exhibits considerable instability for $\alpha \leq 0.04$. Consequently, Storey's $\widehat{\text{pFDR}}(\alpha)$ differs substantially from his monotone $\hat{q}(\alpha)$ over most of the interval (0, 0.04). By contrast, the SPLOSH estimate $\hat{r}(\alpha)$ is monotone for all α and consequently $\hat{r}(\alpha) = \hat{h}(\alpha)$ for all α. The SPLOSH estimate $\hat{h}(\alpha)$ appears to be the most plausible of the three monotone estimates, because it is smooth and is equivalent to SPLOSH's $\hat{r}(\alpha)$, which is based on the excellent PDF estimate as seen in Figure 1.

The above example suggests that the SPLOSH method has many desirable properties compared with the other methods. However, one cannot truly appreciate the performance of the

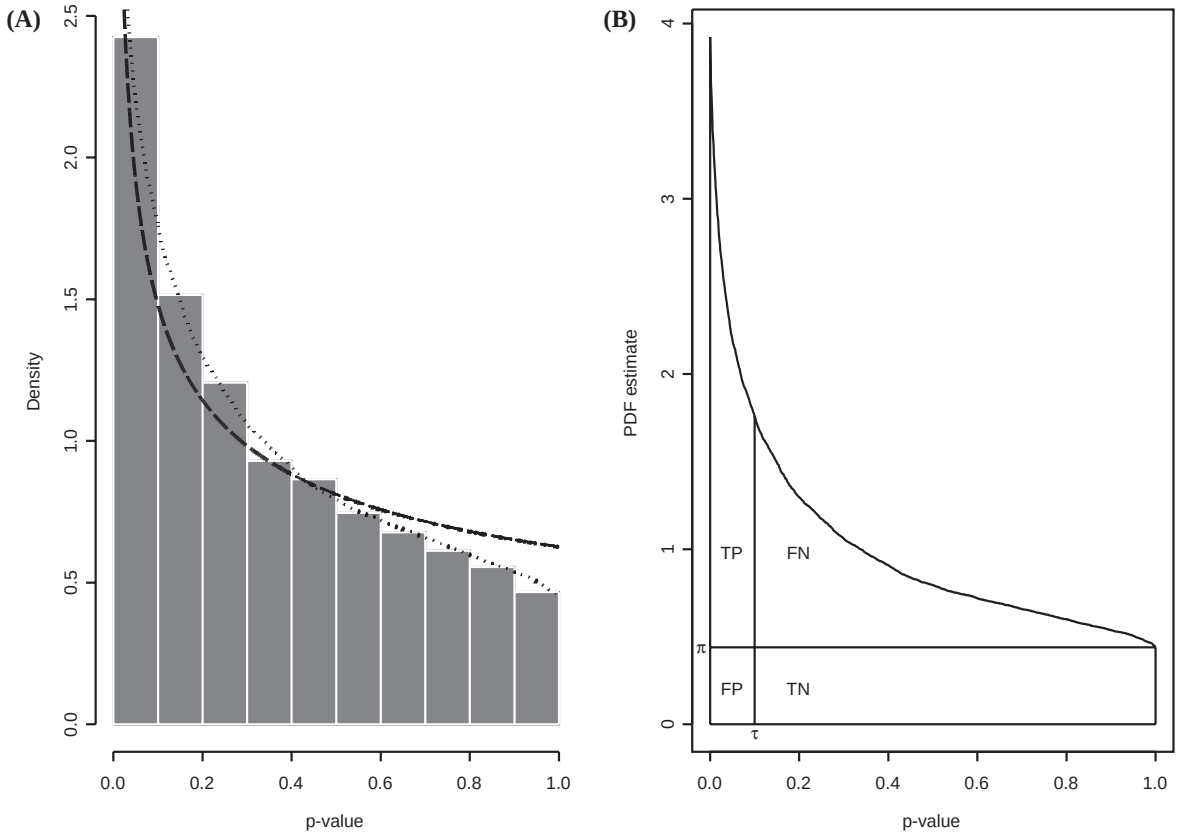

Fig. 1. Estimated densities and SPLOSH error region diagram from the *Arf* experiment. (**A**) BUM (dashed line) and SPLOSH (dotted line) estimates of the p-value PDF against a traditional histogram. (**B**) An error region diagram based on the SPLOSH estimate. The areas of the true positive (TP), false positive (FP), false negative (FN) and true negative (TN) regions give estimates of the proportion of tests resulting in their respective outcomes. The areas also provide geometric interpretations of error rate quantities. For example, the pFDR is the ratio of the area of the FP region to the sum of the areas of the TP and FP regions.

Table 1. Estimates of the null proportion π

Method	$\widehat{\pi}$
BUM	0.6271
SPLOSH	0.4399
Storey	0.4453

methods in this case-study analysis, because we do not know the true state of nature. In a simulation study, one can study the behavior of the estimates over a large number of repeated mimics of an experiment under an assumed state of nature. In this way, one can identify preferred methods when the true state of nature is identical or similar to the state of nature assumed in the simulation.

Simulation study

The simulation study considers 1000 independent realizations of a setting in which the expression of 3000 genes is studied in two groups. Each group is represented by three independent chips. A proportion, $\pi = 0.80$, of the 3000 genes (i.e. 2400 genes) has equal mean expression between the two groups. The mean expression levels of the other genes differ by 1.5 units. The expression levels of all genes are assumed to be normally distributed with variance 1. Moreover, all genes are assumed independent of one another in the simulation. One-way ANOVA (Casella and Berger, 1990) is used to assess the significance of each gene and compute a corresponding p-value. Storey's method (2002), BUM (Pounds and Morris, 2003) and SPLOSH are applied to the set of p-values obtained in each realization. In each realization, the simulation maintained a record of whether each p-value was testing a true or false null hypothesis. In this way, the simulation could estimate the actual cFDR as a function of the number of rejections by averaging the proportion of false discoveries within each rejection set across the replications. The expected value of the method's estimates of the cFDR (or pFDR) as a function of the number of rejections is estimated by averaging over the simulation replicates.

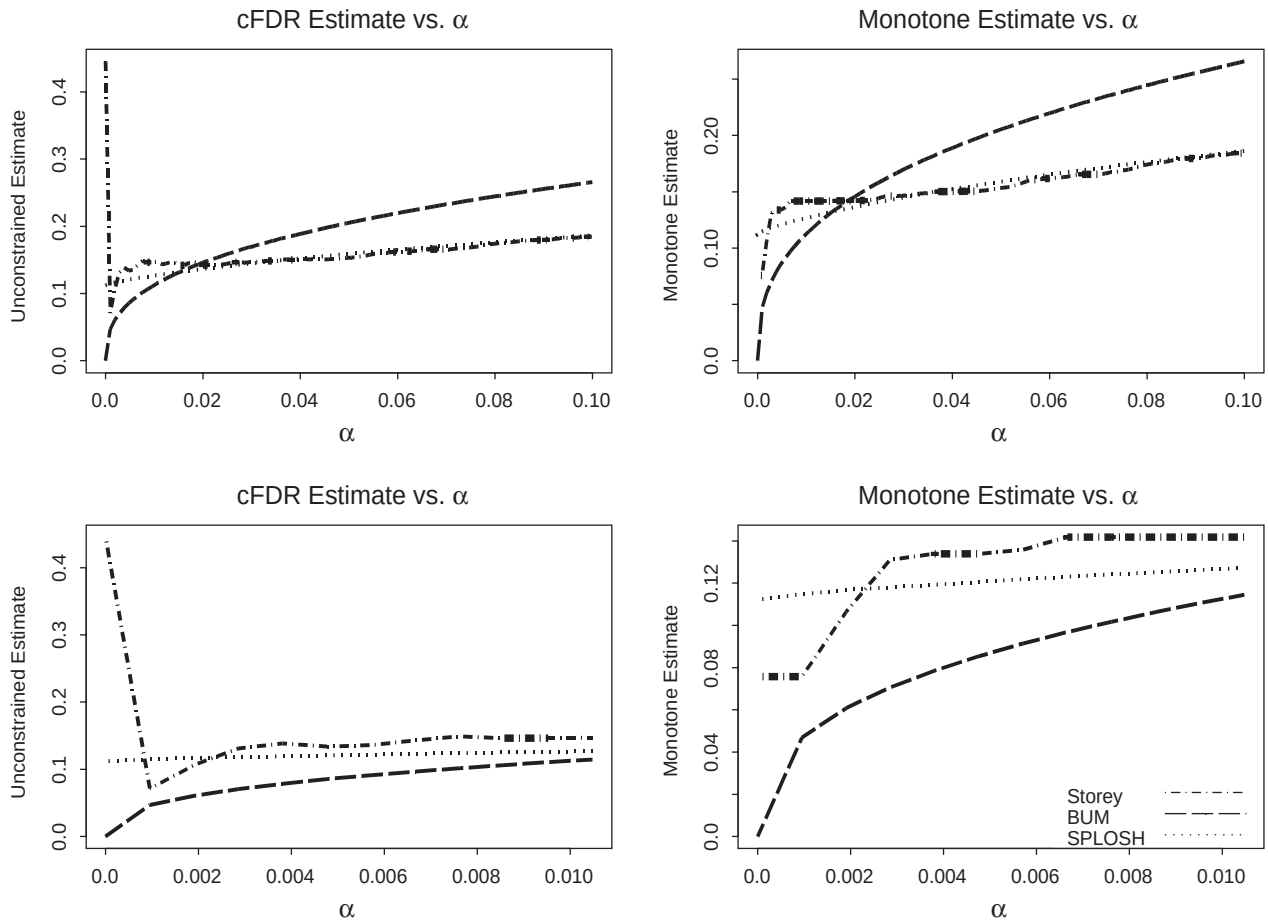

Fig. 2. Comparison of the *Arf* experiment results. The unconstrained and monotone estimates obtained using each method are displayed above. The figures on the left give the unconstrained estimates: Storey's $\widehat{\text{pFDR}}(\alpha)$, BUM's $\hat{r}(\alpha)$ and SPLOSH's $\hat{r}(\alpha)$. The figures on the right give the monotone estimates: Storey's $\hat{q}(\alpha)$, BUM's $\hat{r}(\alpha)$ and SPLOSH's $\hat{h}(\alpha)$. The bottom left panel gives a legend indicating the line patterns used to denote each method's results.

Table 2. Behavior of π-estimators in the simulation

Method	$\widehat{\text{E}}(\hat{\pi})$	$\widehat{\text{SE}}(\hat{\pi})$
BUM	0.790	0.014
SPLOSH	0.786	0.065
Storey	0.807	0.039

Table 2 illustrates the properties of the estimators of $\pi = 0.80$ observed in the simulation. BUM tends to underestimate π with low variance. Storey's estimator of π shows very good properties, and SPLOSH's estimator of π exhibits the greatest variation. However, the methods should be judged primarily on the quality of the cFDR estimates they produce.

Figures 3 and 4 summarize the simulation results regarding the properties of each method's estimates of the cFDR as a function of the number of rejections R. BUM systematically underestimates the cFDR for a small R due to model lack-of-fit. SPLOSH's $\hat{r}(\alpha)$ and Storey's $\widehat{\text{pFDR}}(\alpha)$ appear to be unbiased or to have little bias when used to estimate the cFDR as a function of k. However, Storey's estimator becomes unstable for small k. After using cumulative minimization to constrain the estimators to be monotone, Storey's $\hat{q}(p_{(i)})$ is considerably biased in the non-conservative direction, whereas the simulation estimate of the expected value of SPLOSH's $\hat{h}(\alpha)$ is only slightly below the simulation estimate of the actual FDR.

A major advantage of SPLOSH is that it is not greatly affected when cumulative minimization is used to enforce monotonicity. On average across the simulation, SPLOSH's $\hat{r}(\alpha)$ and $\hat{h}(\alpha)$ differed at only 2.96 of the 3000 p-values; in 877 of the 1000 replications the two quantities did not differ at any p-value. In contrast, Storey's $\widehat{\text{pFDR}}(\alpha)$ and $\hat{q}(\alpha)$ differed at an average of 788.5 of 3000 p-values, and the two estimators differed at some p-value in all 1000 replications.

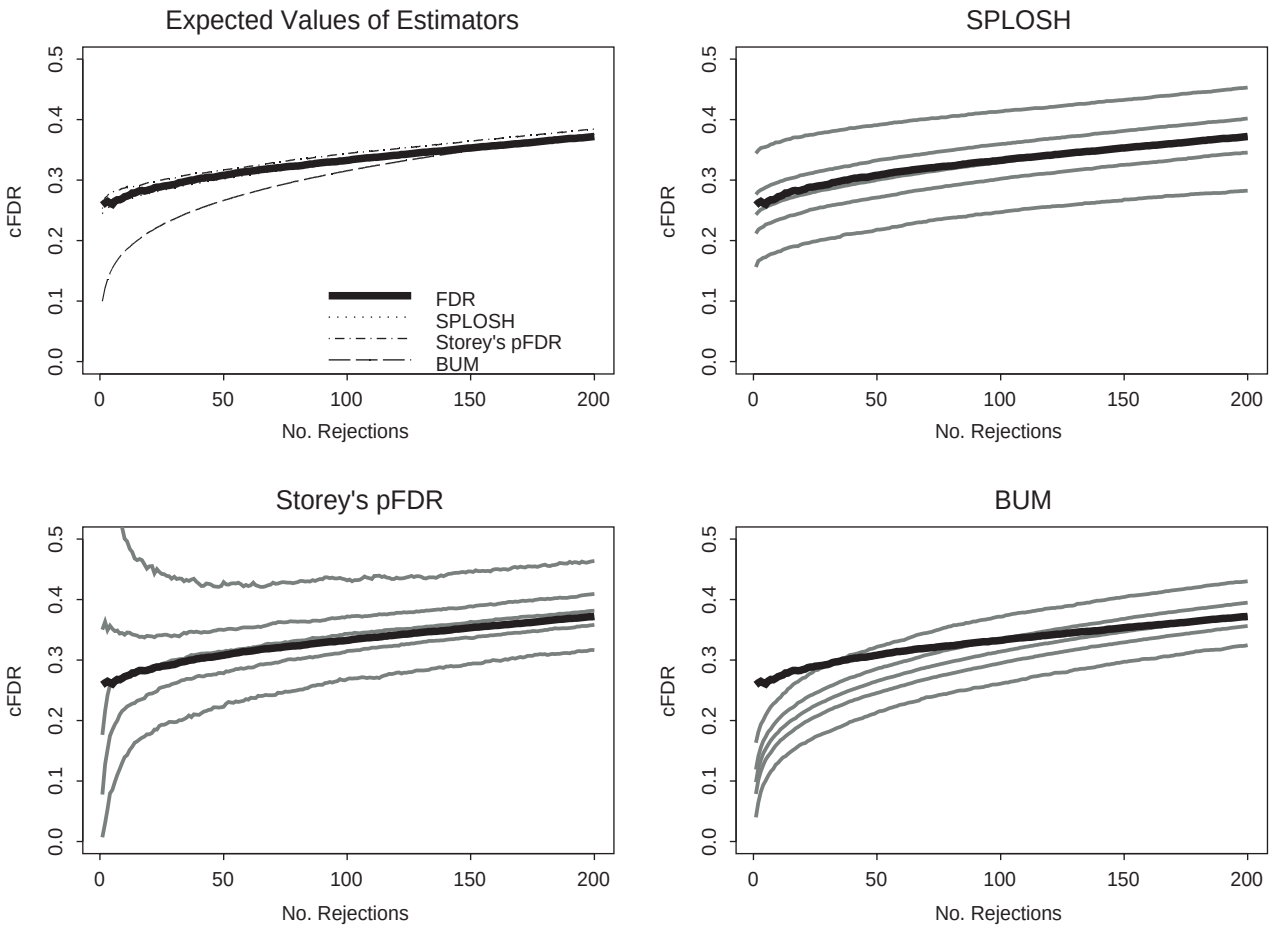

Fig. 3. Simulation results: unconstrained estimators. The top left panel shows the simulation estimate of the actual cFDR and the simulation estimates of the expected value of SPLOSH's $\hat{r}(p_{(K)})$, Storey's $\widehat{\text{pFDR}}(p_{(K)})$ and BUM's cFDR estimate. The top right panel shows the simulation estimate of the actual cFDR as a solid black line and uses light gray lines to represent the 2.5, 25, 50, 75 and 97.5 percentiles of SPLOSH's $\hat{r}$. The bottom left and bottom right panels, respectively, depict the corresponding quantities for Storey's $\widehat{\text{pFDR}}$ and BUM's cFDR estimate in a similar manner.

DISCUSSION

In the simulation study and example analysis, SPLOSH displayed desirable properties as a method used to estimate the proportion of false discoveries among the k-most significant findings. Unlike BUM, SPLOSH is not subject to the ability of an assumed model to accurately represent the observed p-values, hence SPLOSH is applicable in a wider variety of settings than BUM. In both the simulation and example, SPLOSH exhibited greater stability for small p-values than did Storey's method. In the cases examined here, SPLOSH was clearly the superior method of cFDR estimation.

It is reasonable to anticipate that SPLOSH will exhibit these desirable characteristics in other settings as well. In forming the estimate of $F(\alpha)$, Storey's method considers information only to the left of α whereas SPLOSH uses information on both sides of α. For small α, Storey's method is based on very limited information, hence it becomes unstable. However, SPLOSH maintains its level of stability, because it utilizes information to the right as well. SPLOSH's stability also reduces the tendency to bias estimates downward when cumulative minimization is used to enforce monotonicity. Storey's $\hat{q}(\alpha)$ is substantially biased downward, because $\widehat{\text{pFDR}}(\alpha)$ exhibits great fluctuation.

The properties of SPLOSH and other cFDR estimation techniques should be further examined in a mathematically rigorous manner. No mathematical proofs have established the behavior of SPLOSH or BUM in a general setting. Mathematical proofs have established Storey's $\widehat{\text{FDR}}$ and $\widehat{\text{pFDR}}$ as conservative estimators in some settings, but as already seen, these estimators can be too variable to be meaningfully interpreted in practice. In particular, none of these methods has yet been shown to estimate the cFDR reliably in a wide variety of settings. The primary insight SPLOSH provides for cFDR estimation is the importance of obtaining smooth

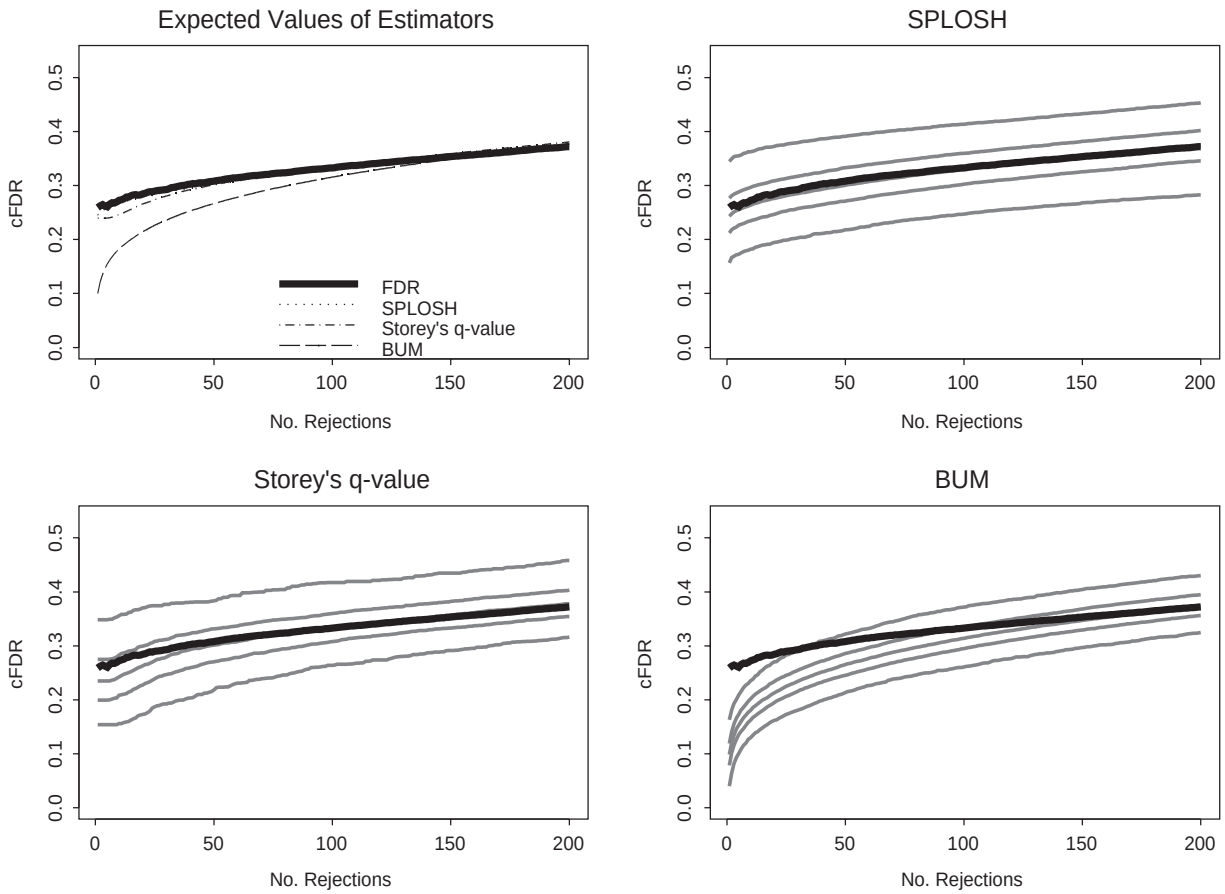

Fig. 4. Simulation results: monotone estimators. The top left panel shows the simulation estimate of the actual cFDR as a solid black line and the simulation estimates of the expected value of SPLOSH's $\hat{h}\left(p_{(K)}\right)$, Storey's $\hat{q}\left(p_{(K)}\right)$ and BUM's cFDR estimate. The top right panel shows the simulation estimate of the actual cFDR as a dark solid line and uses light gray lines to represent the 2.5, 25, 50, 75, and 97.5 percentiles of SPLOSH's $\hat{h}$. The bottom left and bottom right panels, respectively, depict the corresponding quantities for Storey's $\hat{q}$ and BUM's cFDR estimate in a similar fashion.

cFDR estimates before enforcing monotonicity by cumulative minimization.

SPLOSH demonstrates that smoothing substantially stabilizes cFDR estimation. The concept of smoothing may improve other methods used in the analysis of mircoarray data as well. Specifically, the significance analysis of microarrays (SAM) (Tusher *et al.*, 2001) approach may benefit from smoothing. SAM considers the set of ordered test statistics $T_1 \leq T_2 \leq \cdots \leq T_g$ and uses permutation techniques to generate estimates of the expected order statistics under the null hypothesis $T_1^* \leq T_2^* \leq \cdots \leq T_g^*$. SAM then determines the smallest i such that $T_i - T_i^*$ is greater than some threshold τ and declares T_j significant for all $j \geq i$. Smoothing $T_i - T_i^*$ before determining the threshold crossing index i may help to stabilize SAM by reducing the role of random fluctuations in determining the first threshold crossing point.

The use of L'Hospital's rule brings up an important theoretical question: 'What is the limiting behavior of the cFDR, pFDR and FDR as α approaches zero?' Benjamini and Hochberg (1995) naturally define the FDR associated with no rejections as zero. However, if one considers the procedure of rejecting all hypothesis with p-value less than α, it is unclear whether or not $\mathrm{FDR}(\alpha)$ is a smooth function at $\alpha = 0$. Further investigation into this topic is warranted. Additionally, it is important to examine mathematically the FDR estimation and control properties of the SPLOSH procedure for very small α.

All three methods considered in the example and simulation study assume that the p-values are reasonably modeled as independently and identically distributed random variables arising from a mixture model of the form proposed by Storey (2002). When this assumption is violated, these methods may yield unreliable results (Benjamini and Yekutieli, 2001). In such a scenario, one may consider using computationally intensive resampling procedures such as those described by Yekutieli and Benjamini (1999) or Ge *et al*. (2003) to account for dependencies between tests.

The SPLOSH algorithm could be modified or generalized by using smoothing techniques other than the LOESS algorithm. Fan and Gijbels (1996) and Hart (1997) describe alternative smoothing algorithms that could be used to obtain smooth PDF or CDF estimates. Although SPLOSH performs quite well, it may not be the optimal smoothing algorithm in the context of cFDR estimation based on *p*-value PDF or CDF estimates. An optimal smoothing algorithm for purposes of cFDR estimation and control is yet to be determined. Additional generalizations could be found by considering other transformations as well. Determining an optimal transformation is another open research topic.

The FDR, cFDR and pFDR are only a subset of metrics of the occurrence of errors in multiple testing settings. There are a variety of other techniques and measures of the occurrence of errors to address the multiplicity issue that arises in the analysis of microarray data. Dudoit *et al*. (2003) give an excellent overview of error occurrence metrics, types of control and algorithms used to control the occurrence of errors.

ACKNOWLEDGEMENTS

The authors thank Drs Mei-Ling Kuo and Martine Roussel of the Department of Genetics and Tumor Cell Biology at St Jude Children's Research Hospital for the use of their data in the example. The authors also thank the reviewers for their most helpful comments and feedback, which greatly improved the quality of this work. Additionally, Dr Angela McAuthor of Scientific Editing at St Jude provided useful suggestions to improve the manuscript. This research was supported in part by NIH Cancer Center Support Grant CA-21765 and the American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

Allison,D.B., Gadbury,G.L.,Heo,M., Fernández,J.R., Les,C-K., Prolla,J.A. and Weindruch,R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data. An.*, **39**, 1–20.

Anton,H. (1992) *Calculus*, 4th edn. John Wiley & Sons, New York.

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

Benjamini,Y. and Hochberg,Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Edu. 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.

Casella,G. and Berger,R. (1990) *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Cleveland,W.S. (1993) *Visualizing Data*. Hobart Press, Summit, NJ.

Cleveland,W.S. and Devlin,S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc*., **83**, 596–610.

Dudoit,S., Shaffer,J.P. and Boldrick,J.C. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Efron,B. and Gong,G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.*, **37**, 36–48.

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

Fan,J. and Gijbels,I. (1996) *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

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

Hart,J. (1997) *Nonparametric Smoothing and Lack of Fit Tests*. Springer-Verlag, New York.

Insightful (2002) *S-plus 6.1 Users Guide*, on-line.

Kooperberg,C. and Stone,C.J. (1991) A study of logspline density estimation. *Comput. Stat. Data. An.*, **12**, 327–347.

Kuo,M.-L., Duncavage,E.J., Mathew,R., den Besten,W., Pei,D., Naeve,D., Yamamoto,T., Cheng,C., Sherr,C.J. and Roussel,M.F. (2003) Arf induces p53-dependent and -independent anti-proliferative genes. *Cancer Res.*, **1**, 1046–1053.

Mason,R.L., Gunst,R.F. and Hess,J.L. (1989) *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*. John Wiley & Sons, New York.

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.

Pyke,R. (1965) Spacings. *J. R. Stat. Soc. B*, **27**, 395–449.

Sherr,C. (1998) Tumor surveillance via the ARF-p53 pathway. *Genes Dev.*, **12**, 2984–2991.

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.

Tilstone,C. (2003) Vital statistics. *Nature*, **424**, 610–612.

Tsai,C-A., Hsueh,H-M. and Chen,J.J. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

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

Wolfinger,R.D., Gibson,G., Wolfinger,E.D., Bennett,L., Hamadeh,H., Bushel,P., Afshari,C. and Paules,R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

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.