BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 23 2005, pages 4280–4288
doi:10.1093/bioinformatics/bti685

Gene expression

A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data

Yang Xie[1,*], Wei Pan[1] and Arkady B. Khodursky[2]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA and
[2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, St Paul, MN 55108, USA

Received on June 30, 2005; revised on September 2, 2005; accepted on September 20, 2005
Advance Access publication September 27, 2005

ABSTRACT

Motivation: False discovery rate (FDR) is defined as the expected percentage of false positives among all the claimed positives. In practice, with the true FDR unknown, an estimated FDR can serve as a criterion to evaluate the performance of various statistical methods under the condition that the estimated FDR approximates the true FDR well, or at least, it does not improperly favor or disfavor any particular method. Permutation methods have become popular to estimate FDR in genomic studies. The purpose of this paper is 2-fold. First, we investigate theoretically and empirically whether the standard permutation-based FDR estimator is biased, and if so, whether the bias inappropriately favors or disfavors any method. Second, we propose a simple modification of the standard permutation to yield a better FDR estimator, which can in turn serve as a more fair criterion to evaluate various statistical methods.

Results: Both simulated and real data examples are used for illustration and comparison. Three commonly used test statistics, the sample mean, SAM statistic and Student's *t*-statistic, are considered. The results show that the standard permutation method overestimates FDR. The overestimation is the most severe for the sample mean statistic while the least for the *t*-statistic with the SAM-statistic lying between the two extremes, suggesting that one has to be cautious when using the standard permutation-based FDR estimates to evaluate various statistical methods. In addition, our proposed FDR estimation method is simple and outperforms the standard method.

Contact: yangxie@biostat.umn.ed

1 INTRODUCTION

DNA microarrays are biotechnologies that allow highly parallel and simultaneous monitoring of the whole genome (Brown and Botstein, 1999). Increasingly, they are used to detect genes expressed differentially under different conditions (Spellman *et al*., 1998). Typically, two steps are used to declare differentially expressed (DE) genes: first, one computes a summary or test statistic (e.g. the sample mean) for each gene and rank the genes in order of their test statistics; second, one chooses a threshold for the test statistics and call genes whose statistics are above the threshold 'significant' ones (Smyth *et al*., 2003). False discovery rate (FDR) introduced by Benjamini and Hochberg (1995) has become a popular way to formally assess the statistical significance level in microarray data analysis. FDR is defined as the expected percentage of false positives among the claimed positives. If we claim that r top ranked genes are significant DE genes, the expected percentage of equally expressed (EE) genes among these r genes is the FDR.

FDR can be used for several purposes in statistical analysis. First, FDR is related to the choice of cut-off for 'significance' to control the error rate in multiple tests. Benjamini and Hochberg (1995) introduced FDR as an error measure for multiple-hypothesis testing and proposed a sequential method based on *P*-values to control FDR. Storey (2002, 2003) proposed directly estimating FDR for a fixed rejection region, largely increasing the popularity of FDR in practice. Later, many authors (Tsai *et al*., 2003; Pounds and Cheng, 2004; Dalmasso *et al*., 2005) studied various issues related to FDR estimation, especially for microarray gene expression data. When FDR is used to provide an upper bound on the error one can tolerate, the conservativeness of FDR estimation is not an issue. Actually, Storey (2002, 2004) showed the conservative property of their FDR estimator. Second, some recent literature pointed out some connections between FDR and variable selection (Abramovich *et al*., 2000; Ghosh *et al*., 2004; Devlin *et al*., 2003; Bunea *et al*., 2003). Third, FDR can be used as a criterion to evaluate new statistical methods or compare different procedures: when claiming the same number of total positives, the method with the lowest FDR is regarded as the best. If the truths are known, such as in simulation studies or some calibration datasets derived from spike-in experiments, the use of FDR as a criterion to compare different methods is analogous to using sensitivity and specificity as criteria and is very straightforward. In typical biological experiments, the truth is unknown and an estimated FDR instead can be used. Tibshirani and Bair (2003) used both true and estimated FDR to evaluate the use of eigenarray in microarray data analysis (http://www-stat.stanford.edu/~tibs/research.html). Shedden *et al*. (2005) used estimated FDR to compare seven methods for producing expression summary statistics for Affymetrix arrays. Other authors (Broberg, 2003; Pan, 2003; Xie *et al*., 2004; Wu, 2005) also used estimated FDR to compare different methods in microarray data analysis. It is reasonable and fair only when the estimated FDR approximates the true FDR well, or at least, the estimated FDRs for various methods being compared reflect the same trend of the true FDRs; that is, even if an FDR estimator is biased, it should not improperly favor or disfavor any particular statistical method being compared. We emphasize that the 'fairness' of FDR estimation is a necessary property when it is used as a criterion; this paper will focus on this aspect of FDR estimation.

Knowing the distribution of a test statistic under the null hypothesis (called null distribution) is important for FDR estimation.

*To whom correspondence should be addressed.

Some regularized statistics, such as the SAM-statistic (Tusher *et al*., 2001; Efron *et al*., 2001; Pan *et al*., 2003), perform well for microarray data, but their null distributions are in general unknown; permutation methods have become popular to estimate null distributions owing to their flexibility and generality. However, there are some problems when using permutation to estimate null distributions for microarray data. Pollard and Van der Laan (2003, 2004) pointed out that when the number of replicates in two groups is different, the permutation test for two-sample comparison may not be valid. Other authors (Efron *et al*., 2001; Pan, 2003; Zhao and Pan, 2003) have noticed this problem and addressed it by modifying the test statistic so that the standard permutation can still estimate the null distribution well. Guo and Pan (2004) addressed the problem by using weighted permutation scores that down weight the influence of (predicted) DE genes on estimating the null distribution. Nevertheless, to our knowledge, there has been no consideration on whether the bias of the FDR estimator introduced by the standard permutation, if it exists, may depend on the test statistic being used; if true, it implies that the resulting FDR estimates cannot be used as a criterion to fairly compare various statistical methods. The purposes of this paper are (1) to investigate both theoretically and empirically whether the standard permutation-based FDR estimation method is biased, and if yes, whether this bias favors or disfavors any particular statistic; (2) to propose a new FDR estimator that can serve as a better criterion to evaluate various statistical methods.

2 METHODS

2.1 Test statistics

For the purpose of clarity, we only consider one-sample comparisons here, though extensions to two-sample comparisons and other more general settings are straightforward (Tusher *et al*., 2001; Broet *et al*., 2004). Suppose after preprocessing the data, we have observed gene expression levels (e.g. log ratios of the two channel intensities in cDNA arrays) $X_{i1},\ldots,X_{ik}$ for gene i, $i=1,\ldots,G$ from k arrays. The goal is to test H_i0: $E(X_{ij})=0$ for $i=1,\ldots,G$. We will consider three commonly used test statistics. The first one is the SAM-statistic (Tusher *et al*., 2001), shortened as S-statistic,

$$S_i=\frac{\bar{X}_i}{(V_i+V_0)/\sqrt{k}}, \tag{1}$$

where $\bar{X}_i=\sum_{j=1}^{k}X_{ij}/k$ and $V_i^2=\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2/(k-1)$ are the sample mean and sample variance of the expression levels for gene i, and V_0 is a constant used to stabilize the denominator of the test statistic. V_0 can be chosen in different ways; one is $V_0=\text{median}(V_1,\ldots,V_G)$.

The second is the mean statistic, $M_i=\bar{X}_i$, which corresponds to the early practice of simply using fold changes as a significance indicator (e.g. Broet *et al*., 2002). The third one is the Student's t-statistic, $t_i=\bar{X}_i/V_i$, which is a standardized mean statistic.

2.2 A standard method for FDR estimation

For a fixed cut-off value d for a test statistic Z_i, we can obtain the true or realized FDR and its estimate as (Storey and Tibshirani, 2003)

$$\text{FDR}(d)=\pi_0\text{FP}(d)/\widehat{\text{TP}}(d),\quad \widehat{\text{FDR}}(d)=\widehat{\pi_0}\widehat{\text{FP}}(d)/\widehat{\text{TP}}(d), \tag{2}$$

where π_0 is the proportion of EE genes among all genes, and $\widehat{\pi_0}$ is its estimator. FP is the number of true false positive genes, i.e., the number of genes which are EE genes but claimed as DE genes, $\widehat{\text{FP}}$ is the estimated number of false positive genes. $\widehat{\text{TP}}(d)$ is the total number of genes claimed as DE genes when the cut-off value is d.

2.2.1 Standard permutation method In order to obtain $\widehat{\text{FP}}$, we need to estimate the distribution of the test statistic Z_i under the null hypothesis H_{i0} (that gene i is an EE gene). Rather than assuming a parametric distribution for the null distribution of Z_i, a class of non-parametric methods have been proposed to estimate it empirically (Efron *et al*., 2001; Tusher *et al*., 2001; Xu *et al*., 2002; Pan *et al*., 2003). The idea is to permute the data and calculate the null statistic z_i, in the same way as calculating z_i, but based on the permuted data. Under the null hypothesis, the empirical distribution of the null statistics can be used to approximate the null distribution. In the current context of the one-sample test, under H_{i0}, we can permute the data by randomly keeping or flipping the sign of each of $X_{i1},\ldots,X_{ik}$. When k is small, we can consider all possible permutations; otherwise, a large number of random permutations, say B, can be used. Calculating the same test statistic from the b-th permuted data results in the null statistic $z_i^{(b)}$ for $b=1,\ldots,B$ and $i=1,\ldots,G$. For any given $d>0$, if we claim any gene i satisfying $|Z_i|>d$ to be significant, we estimate the true positive (TP) numbers and false positive (FP) numbers as

$$\widehat{\text{TP}}(d)=\#\{i:|Z_i|>d\},\quad \widehat{\text{FP}}(d)=\sum_{b=1}^{B}\#\{i:|z_i|^{(b)}>d\}/B. \tag{3}$$

We plug $\widehat{\text{TP}}(d)$ and $\widehat{\text{FP}}(d)$ into Equation (2) to calculate FDR(d) and $\widehat{\text{FDR}}(d)$. Other more sophisticated methods, such as SAM (Tusher *et al*., 2001) or mixture model (Pan *et al*., 2003; McLachlan and Peel, 2000) can be equally applied.

2.2.2 Proportion of EE genes Based on expression (2), we need to estimate π_0, the proportion of EE genes, to calculate FDR. Many authors have studied the issue based on the distribution of P-values (Storey, 2002; Allison *et al*., 2002; Pounds and Morris, 2003; Pounds and Cheng, 2004; Guan *et al*., 2004, http://www.biostat.umn.edu./rrs.php; Wu *et al*., 2004, http://www.biostat.umn.edu./rrs.php). However, owing to the difficulty of assigning P-values, the estimation of π_0 remains challenging. In fact, if the standard permutation method is used to estimate P-values, the same argument as to be discussed next implies that the P-values will be overestimated, leading to overestimation of π_0 (Guo and Pan, 2004). Other non-parametric approaches can only estimate an upper bound of π_0 (Dalmasso *et al*., 2005). Because estimation of π_0 is itself an unsettled research question, and more relevantly here, is not the focus of our current work, we bypass it in simulations: for simulated data, we use true π_0 in expression (2), which represents the ideal (but not practical) performance of the standard method. For real data, however, we use an estimated π_0.

2.2.3 Problem with the standard permutation: statistical theory The idea of using null statistics of all genes to construct the null distribution is based on the assumption that the null statistics of all genes are identically distributed. However, as shown next, the null statistic of a DE gene does not have the same distribution as that of EE genes. Hence, the empirical distribution of the null statistics of all genes may not approximate the true null distribution well.

Suppose for gene i, its observed gene expression level X_{ij} on array j has mean μi and variance σ_i^2; $\mu_i=0$ if it is an EE gene, and $\mu_i\neq 0$ otherwise. Define Bernoulli random variable Y_{ij} as: $Y_{ij}=1$ (corresponding to keeping the sign of X_{ij}) with probability $\pi=0.5$ and $Y_{ij}=-1$ (corresponding to flipping the sign of X_{ij}) with probability $1-\pi=0.5$, and assume that Y_{ij} and X_{ij} are independent. Then the random variable $W_{ij}=Y_{ij}X_{ij}$ represents the permuted gene expression level in the standard permutation method. It is simple to verify that $E(Y_{ij})=2\pi-1=0$, and we have

$$\begin{aligned}E(W_{ij})&=E(Y_{ij}E(X_{ij}))=(2\pi-1)\mu_i=0\\ \text{Var}(W_{ij})&=E(\text{Var}(Y_{ij}X_{ij}|Y_{ij}))+\text{Var}(E(Y_{ij}X_{ij}|Y_{ij}))\\ &=E(Y_{ij}^2\sigma_i^2)+\text{Var}(Y_{ij}\mu_i)\\ &=\sigma_i^2+\mu_i^2.\end{aligned}$$

If gene i is an EE gene, $\mu_i = 0$, and thus $\mathrm{Var}(W_{ij}) = \sigma_i^2 = \mathrm{Var}(X_{ij})$; otherwise, $\mu_i \neq 0$, and $\mathrm{Var}(W_{ij}) > \mathrm{Var}(X_{ij})$. The consequence is that permuted expression levels of DE genes inflate the variation of the distribution of all null statistics, as pointed out by previous authors (e.g. Pan, 2003). Although the heuristic argument is intuitively reasonable, it may not be equally transparent to everyone. Below we provide a more detailed, and hence more convincing discussion on this for each test statistic.

To facilitate discussion, we suppose that gene i is a DE gene throughout this section, and rewrite $X_{ij} = X_{ij}^* + \mu_i$; X_{ij}^* can be regarded as the expression level of gene i if gene i were equally expressed.

The mean statistic The null statistic for DE gene i is

$$m_i = \sum_{j=1}^{k} \frac{W_{ij}}{k} = \sum_{j=1}^{k} \frac{Y_{ij}X_{ij}^*}{k} + \mu_i \sum_{j=1}^{k} \frac{Y_{ij}}{k},$$

while if gene i were an EE gene, its null statistic would be

$$m_i^* = \sum_{j=1}^{k} \frac{Y_{ij}X_{ij}^*}{k}.$$

Because $\mu_i \neq 0$, it can be shown that $\mathrm{Var}(m_i) = \mathrm{Var}(m_i^*) + \mu_i^2/k$. Therefore, the distribution of the null statistic of a DE gene has heavier tails than that of an EE gene. In other words, because of the presence of both DE and EE genes, the distribution of the null statistics of all genes, as adopted in the standard permutation method, has heavier tails than that of only EE genes. Note that the difference between $\mathrm{Var}(m_i)$ and $\mathrm{Var}(m_i^*)$ depends on both μ_i and k, the difference will get smaller when k increases.

The t-statistic The null statistic for DE gene i is

$$t_i = \frac{\sum_{j=1}^{k} Y_{ij}X_{ij}^*/k + \mu_i \sum_{j=1}^{k} Y_{ij}/k}{V(Y_{ij}X_{ij}^* + \mu_i Y_{ij})/\sqrt{k}}.$$

In contrast, if gene i were an EE gene, its null statistic would be

$$t_i^* = \frac{\sum_{j=1}^{k} Y_{ij}X_{ij}^*/k}{V(Y_{ij}X_{ij}^*)/\sqrt{k}},$$

where $V(R_{ij})$ is the sample standard deviation of $\{R_{i1}, \ldots, R_{ik}\}$. Although, as shown earlier, the variance of the numerator of t_i is larger than that of t_i^*, the variance of the denominator of t_i may be also larger than that of t_i^*. Hence, we cannot simply conclude that $\mathrm{Var}(t_i) > \mathrm{Var}(t_i^*)$. Although it seems non-trivial to establish analytically, we use simulation to compare the variances of t_i, t_i^*, m_i and m_i^* under the assumption that Xij has a normal distribution.

We simulated X_{ij}^* from a standard normal distribution (i.e. with mean 0 and variance 1), and Y_{ij} from a Bernoulli distribution specified earlier, with $i = 1, \ldots, 100\,000$ and $j = 1, \ldots, k$. With $\mu_i = 2$ and $\mu_i = 0.5$, we calculated each m_i, m_i^*, t_i and t_i^*. Table 1 gives the sample variances of the four statistics with $k = 3, \ldots, 6$. It can be seen that m_i has a larger variance than m_i^*s; and the difference between the two is larger for a smaller k. In most cases, t_i has a larger variance than t_i^*s, but there is an exception when $\mu_i = 0.5$ and $k = 3$. So we cannot get a simple conclusion that variance of t_i is always bigger than variance of t_i^*, which is different from the situation of mean statistic. Of course, by the central limit theorem, the asymptotic distribution of z_i is the same as that of z_i^* as k tends to infinity for both the mean statistic and t-statistic. To compare the impact of DE genes on different test statistics, we calculated the relative difference for the mean statistic, $[\mathrm{var}(m_i) - \mathrm{var}(m_i^*)]/\mathrm{var}(m_i^*)$ and similarly that for the t-statistic. Table 1 shows that the relative difference of mean statistic is larger than that of t-statistic, so the discrepancy between the distribution of m_i and m_i^* is larger than that of t_i and t_i^*.

The SAM statistic As a modified t-statistic with a constant V_0 being added to the denominator, the behavior of the SAM statistic lies between the mean statistic and the t-statistic: if $V_0 = 0$, the SAM statistic is the same as the t-statistic; as V_0 tends to infinity, the SAM statistic reduces to the mean statistic (Efron *et al*., 2001). Therefore, we expect that the discrepancy between the distribution of the null statistic of EE genes and that of DE genes lies between that for the mean statistic and that for the t-statistic.

Table 1. Variances of the null mean statistics for a DE gene (m_i) and a corresponding EE gene (m_i^*), variances of the null t-statistics for a DE gene (t_i) and a corresponding EE gene (t_i^*) with various numbers of replicates k and true difference of the means between an EE gene and a DE gene (μ_i)

μ_i	k	3	4	5	6
2	$\mathrm{Var(m_i)}$	1.67	1.26	1.00	0.84
	$\mathrm{Var}(m_i^*)$	0.33	0.25	0.20	0.17
	Relative difference	4.00	4.02	3.98	4.07
	$\mathrm{Var(t_i)}$	32.97	7.19	3.66	2.49
	$\mathrm{Var}(t_i^*)$	12.4	2.93	1.98	1.64
	Relative difference	1.66	1.42	0.84	0.52
0.5	$\mathrm{Var(m_i)}$	0.42	0.31	0.25	0.21
	$\mathrm{Var}(m_i^*)$	0.33	0.25	0.20	0.17
	Relative difference	0.25	0.26	0.25	0.27
	$\mathrm{Var(t_i)}$	9.63	2.98	1.99	1.68
	$\mathrm{Var}(t_i^*)$	12.40	2.93	1.98	1.64
	Relative difference	–0.22	0.01	0.01	0.02

Relative difference for the mean statistic is defined as $[\mathrm{Var}(m_i) - \mathrm{Var}(m_i^*)]/\mathrm{Var}(m_i^*)$, and that for the t-statistic is $[\mathrm{Var}(t_i) - \mathrm{Var}(t_i^*)]/\mathrm{Var}(t_i^*)$.

In summary, by permuting expression levels of all the genes, both EE and DE genes, the standard permutation tends to overestimate the tails of the null distribution, leading to conservative inference, e.g. overestimating P-values, FP and FDR.

2.3 A new method for FDR estimation

We propose a new permutation based FDR estimation method. If we know which genes are EE genes, we only use these EE genes alone to construct the null distribution without using DE genes and thus avoid the trouble of the standard permutation method. In practice, we never know for sure which genes are EE genes, whose identification may be in fact the purpose of the whole analysis. However, we can use the predicted EE genes to do the permutation and construct the null distribution. First, we predict DE genes based on a summary statistic, then remove the predicted DE genes, and use the remaining genes in permutation. To do so, we have to address first which statistic to use to predict DE genes. A simple and natural way is to use the same statistic as the test statistic to predict DE genes. But if the performance of the test statistic itself is not good, this method may not work well. So an alternative way is to use a statistic that in general has a good performance; the SAM statistic seems to be a reasonable candidate (Tusher, 2001; Lonnstedt and Speed, 2002; Qin and Kerr, 2003; Xie *et al*., 2004). Based on our limited experience, we decided to use the S-statistic.

Another question is how many genes should be removed. Because predicting the number of DE genes is quite challenging, we propose removing the same number of genes as that of claimed significant DE genes. For example, if we identify top 50 genes as significant DE genes, we remove 50 most significant genes based on the S-statistic from the gene list, and then use the remaining genes to do the permutation, construct the null distribution and, therefore, estimate the FP and FDR. More specifically, the new FDR estimation procedure works as follows. Suppose z_i is our test statistic. For any given $d > 0$, we claim any gene i satisfying $|Z_i| > d$ to be significant, and we estimate TP as

$$\widehat{\mathbf{TP}}(d) = \#\{i : |Z_i| > d\}.$$

We define a set of non-significant genes $D(d)$ as $D(d) = \{i : |S_i| \leq d'\}$, where d' is chosen so that the number of genes not in set $D(d)$ is the same

as $\widehat{\mathbf{TP}}(d)$. As before, we permute observed expression levels B times; for each permuted dataset b, we calculate the null statistic $z_i^{(b)}$. Then, we use only the genes in $D(d)$ to estimate FP:

$$\widehat{\mathbf{FP}}(d) = \sum_{b=1}^{B} \#\{i \in D(d) : |z_i|^{(b)} > d\}/B.$$

Finally, FDR is estimated as $\widehat{\mathbf{FDR}}(d) = \widehat{\mathbf{FP}}(d)/\widehat{\mathbf{TP}}(d)$. Note that we do not use π_0 (or its estimate) in $\widehat{\mathbf{FDR}}(d)$ because we only use the genes in $D(d)$ to count false positives, which is equivalent to estimating π_0 as $1 - \widehat{\mathbf{TP}}(d)/G$.

3 RESULTS

3.1 Simulated data

To evaluate the performance of the standard FDR estimation method for different test statistics and whether our proposed FDR estimation method works, we used different simulation set-ups. For each simulation set-up, we simulated data 50 times, and used the mean FDR from these 50 replicates for comparisons. Because the variances of the results from these simulations were quite small, the Monte Carlo errors were negligible. In simulation set-up 1, a simulated dataset had $G = 4000$ genes, among which $G_1 = 400$ were DE genes and the other 3600 were EE genes on $k = 5$ arrays, so the proportion of EE genes was $\pi_0 = 0.9$. For EE gene i, its observed intensity log-ratios followed a normal distribution: $X_{ij} \sim N(0, 4)$ for $j = 1, \ldots, 5$; for DE gene i, $\mu_i \sim N(0, 16)$ and $X_{ij} \sim N(\mu_i, 4)$ for $j = 1, \ldots, 5$. Simulation set-up 2 was similar to set-up 1, but the standard deviation of gene i's expression level was not a constant; instead, it followed a continuous uniform distribution between 0 and 5. In simulation set-up 3, each simulated dataset was generated to mimic a real study (Tani *et al*., 2002), the purpose of which was to comprehensively define a family of genes whose transcription depends on the activity of leucine-responsive regulatory protein, or Lrp, in *Escherichia coli*. There were 4281 genes, 6 replicates and 800 DE genes randomly chosen (corresponding to $\pi_0 = 0.81$). For DE genes, the sample mean of each gene in real data was used as the true mean to generate simulated data; for EE genes, their means were all set at 0. For each gene, the sample variance was used as the true variance, and the expression level of each gene followed a normal distribution. Simulation set-up 4 was the same as set-up 3 except that the number of DE genes was increased to 2000 (leading to $\pi_0 = 0.53$); the purpose was to investigate how a small π_0 influences FDR estimation. Simulation set-up 5 was similar to set-up 1 but the number of DE genes was decreased to 200 ($\pi_0 = 0.95$). We applied the standard and new permutation methods to estimate FDRs using the mean (M), t and S test statistics. As mentioned earlier, when estimating FDR in the standard permutation method, we used the true π_0, an ideal but not practical case providing the best possible performance for the method; in contrast, we do not use the true π_0 for our new method.

Figure 1 compares the performance of the standard permutation and our new method when using the mean statistic as the test statistic under simulation set-ups 1–4. It shows that the standard permutation method largely over-estimates FDRs and the new method performs much better with its FDR estimates closer to the true ones. In simulation 4, after removing the DE genes predicted by the S-statistic, the FDR estimates based on our new method, though much better than that of the standard permutation, are still higher than the true ones. The reason is that there are a large number of DE genes (2000) in this set-up; because only relatively few DE genes are removed, the presence of many other remaining DE genes still affects the null distribution estimation.

Figures 2 and 3 present the results for the S-statistic and t-statistic, respectively. Again the standard permutation overestimates FDR. In general, the new method works better than the standard permutation, especially for the S-statistic. For the t-statistic, the new method gives larger biases than that of the standard method for simulation set-ups 3 and 4. The reason is that the standard method is implemented here using the true π_0 to estimate FDR, which is not possible in practice; in contrast, the new method always overestimates π_0 with $\hat{\pi}_0 = 1 - \widehat{\mathbf{TP}}/G$ when the number of removed genes ($\widehat{\mathbf{TP}}$) is fewer than the true number of DE genes, which is the case for the two plots in Figure 3. Nevertheless, the new method still works better than the standard permutation when a small number of genes are claimed to be significant, which often is of practical interest.

More importantly, by comparing Figures 1, 2 and 3 we can see why estimated FDRs based on the standard permutation method cannot be used as a fair criterion to evaluate the performance of the test statistics. In simulation set-up 1, the mean statistic gives the lowest true FDR while the t-statistic gives the highest; we can draw the same conclusion when using the proposed new FDR estimates, however, we would incorrectly conclude that the mean statistic gives the highest FDR if the standard permutation method is used. In simulation set-up 2, the S-statistic and the t-statistic give lower true FDRs than the mean statistic; the standard FDR estimators give the same conclusion, but the degree of bias for the mean statistic is much higher than that for the other two statistics. Simulations 3 and 4 give the similar conclusion that the bias of the standard FDR estimator depends on the test statistic, favoring the t-statistic and the S-statistic. Our new proposed FDR estimator provide a more fair criterion to compare the various statistics.

The choices on which and how many genes should be removed in the new FDR estimation method will affect its performance. As shown for simulation set-up 4, removing far fewer genes than the true number of DE genes may still result in overestimating FDR, though often to a lesser degree than that of the standard permutation. As an extreme in the other direction, we consider simulation set-up 5 (with $\pi_0 = 0.95$). Here we consider removing top 50 genes, 100 genes, 200 genes and 400 genes, respectively. To facilitate comparisons, in addition, we include results based on permuting only true EE genes, which is ideal but not practical, providing the best scenario. As shown in Table 2 as expected, permuting only true EE genes leads to excellent estimates of FDR while permuting all genes overestimates FDR, especially for the mean statistic. The more genes we remove, the lower the FDR we estimate. If we remove 200 genes, the same number as the true number of DE genes, the FDR estimates are very close to the true FDRs. As expected, if we remove 400 genes, the FDRs are slightly underestimated. Ideally, if the estimated π_0 is close to the truth, we can remove the same number of estimated DE genes. But as discussed earlier, most current methods overestimate π_0. For example, we used Storey and Tibshirani's (2003) method to estimate π_0 in this simulation and obtained $\hat{\pi}_0 = 0.975$, which corresponds to about 100 DE genes; removing only 100 predicted DE genes still results in various biases of the FDR estimates in the standard permutation for the three statistics. On the other hand, we can also see from Table 2 that our proposed simple procedure (removing the same number of genes as TP genes) can work well; see the final section for a further discussion on this issue.

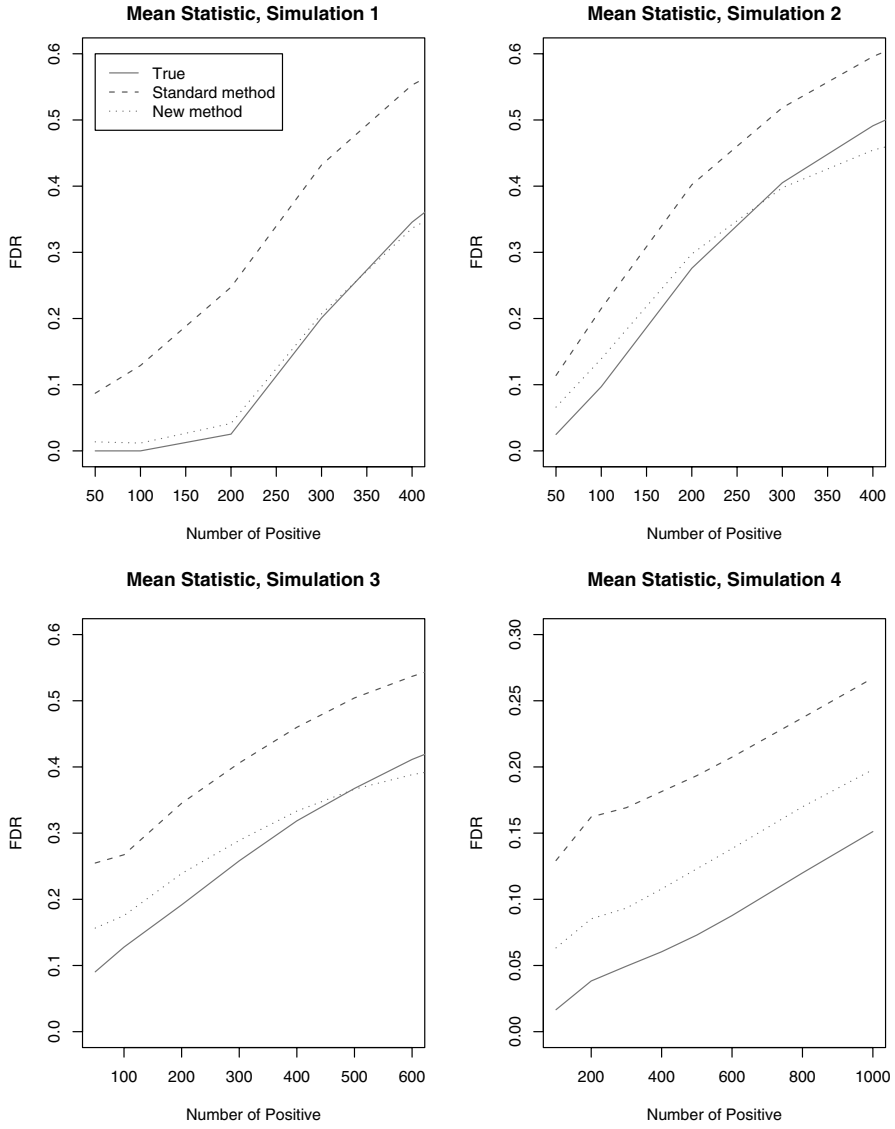

Fig. 1. FDR curves when using the sample mean as the test statistic under different simulation set-ups. Simulation 1, $X_{ij} \sim N(\mu_i, 4)$, the proportion of EE genes is $\pi_0 = 0.9$; Simulation 2, $X_{ij} \sim N(\mu_i, \sigma_i)$ and σ_i follows a uniform distribution, $\pi_0 = 0.9$; Simulation 3, mimicking the Lrp data, $\pi_0 = 0.81$; Simulation 4, mimicking the Lrp data, $\pi_0 = 0.53$.

As suggested by a referee, we also compared the performance of the method that downweights the influence of DE genes (Guo and Pan, 2004). From Table 2, we can see that the weighted method improves results over the standard permutation with less biased FDR estimates, especially for the *S*- and *t*-statistics, but may give a slightly larger bias of the FDR estimate for the mean statistic, thus slightly disfavoring the mean statistic. Larger studies are needed to draw a firm conclusion.

3.2 Chromosomal evolution data

A cDNA microarray experiment with three replications was used to compare the standard and the new FDR estimation methods. The purpose of the experiment was to identify duplications and deletions in genomic DNA (gDNA) of *E.coli*; more details can be found in Zhong *et al*. (2004).

We used Storey and Tibshirani's (2003) method to estimate π_0 and obtained $\hat{\pi}_0 = 1.002$; hence, we decided to use $\widehat{\pi_0} = 1$ for the standard method. Table 3 shows that the *S*-statistic performs best compared to the mean and *t*-statistics in terms of giving the lowest false positive numbers based on both the standard and new methods; though the standard permutation method gives higher false positive numbers than that of the new method, and these differences are especially large for the mean statistic, these observations are in agreement with that of the simulations.

In this experiment, 63 genes have been confirmed to be duplications or deletion genes (i.e. true positives) by real-time PCR and Southern blots. Based on these 63 genes, we can calculate an upper bound for the true false positive number as the number of genes identified by the test statistic but not in the list of 63 true positive genes. Because the follow-up experiment mainly targeted the genes

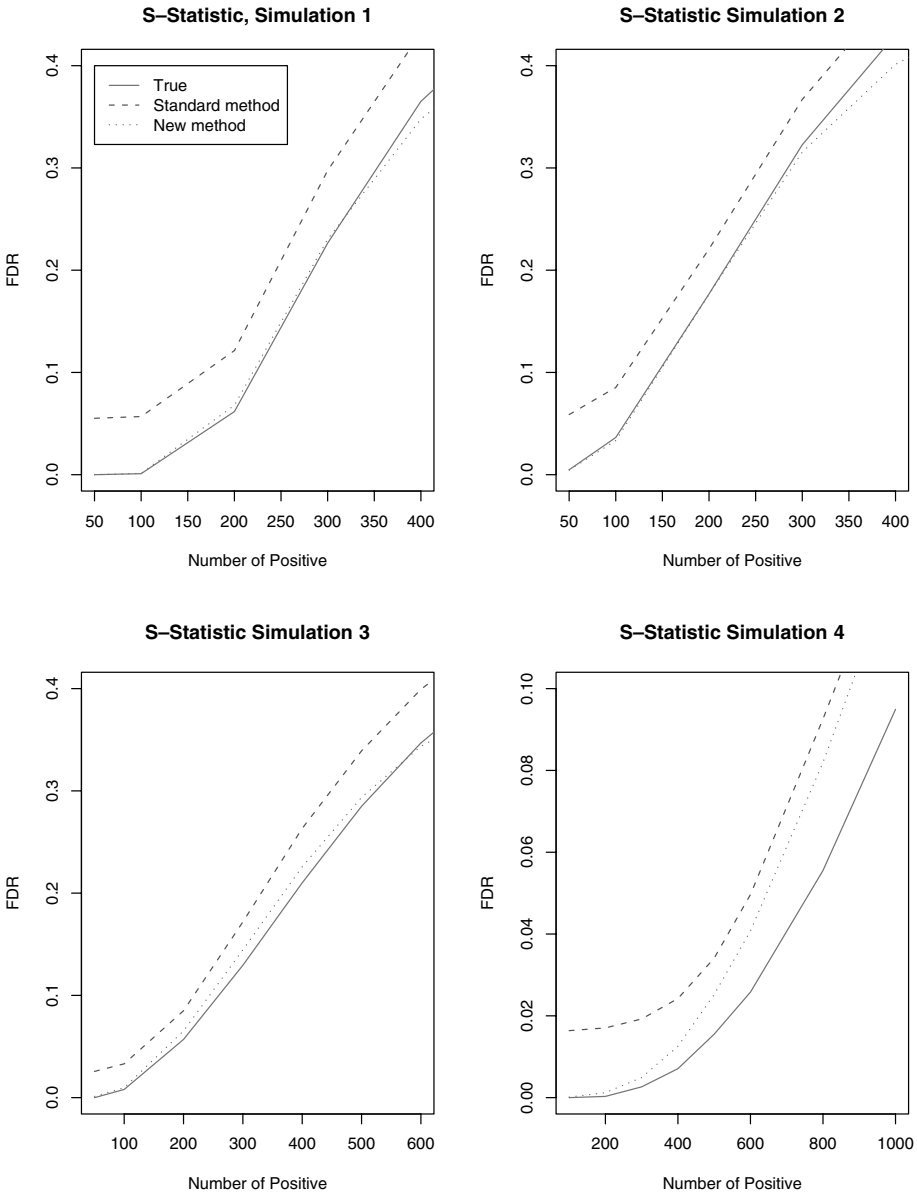

Fig. 2. FDR curves when using S as the test statistic under different simulation set-ups. Simulation 1, $X_{ij} \sim N(\mu_i, 4)$, $\pi_0 = 0.9$; Simulation 2, $X_{ij} \sim N(\mu_i, \sigma_i)$ and σ_i follows a uniform distribution, $\pi_0 = 0.9$; Simulation 3, mimicking the Lrp data, $\pi_0 = 0.81$; Simulation 4, mimicking the Lrp data, $\pi_0 = 0.53$.

with large absolute values of the mean statistics, the upper bound of the true false positive number should be most accurate for the mean statistic. Table 3 shows that if we use the mean statistic to identify 100 significant genes, there should be at most 39 false positive genes; the standard permutation estimates 84 genes as false positives out of 100 significant ones, while the new method gives 38. Hence, the standard permutation largely overestimates the FDR and the new method provides a better estimator. On the other hand, because many top genes ranked by the S-statistic or the t-statistic were not examined in follow-up, the upper bounds of the true false positive numbers for them are likely to be too loose, as evidenced by that the estimated FPs are all well under the bounds using either the standard or the new method.

4 DISCUSSION

This paper investigates the performance of permutation based FDR estimators for the mean, S- and t-statistics. As predicted by our theoretical analysis, our simulation study has confirmed that the standard permutation method overestimates FDR, even when we assume that the proportion of true DE genes is known. The degree of overestimation is especially serious when using the sample mean as the test statistic, less so for the S-statistic, and the least for the t-statistic. Because the magnitude of the bias depends on the test statistic being used, we should be cautious when using estimated FDR as a criterion to evaluate the performance of various test statistics. Our proposed method can estimate the true FDR

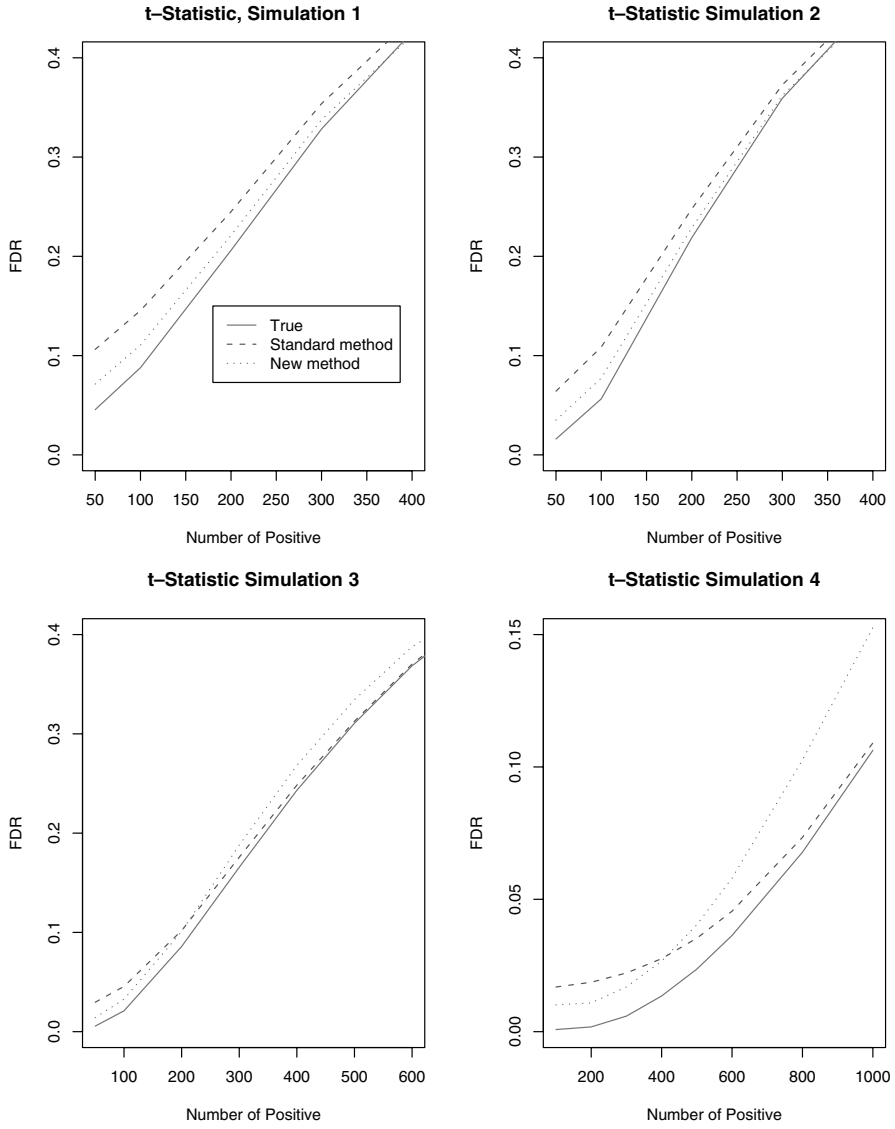

Fig. 3. FDR curves when using t as the test statistic under different simulation set-ups. Simulation 1, $X_{ij} \sim N(\mu_i, 4)$, $\pi_0 = 0.9$; Simulation 2, $X_{ij} \sim N(\mu_i, \sigma_i)$ and σ_i follows a uniform distribution, $\pi_0 = 0.9$; Simulation 3, mimicking the Lrp data, $\pi_0 = 0.81$; Simulation 4, mimicking the Lrp data, $\pi_0 = 0.53$.

better, hence providing a better means to evaluate various test statistics.

The basic idea underlying the new method is quite simple: because it is DE genes that cause the problem, removing the DE genes should improve the performance of the resulting FDR estimator. Our simulation and real data example show that the FDR estimation can be improved by permuting only predicted EE genes. We demonstrate that using the S-statistic to predict EE genes in the new method works well, though any other methods for detecting DE genes (Lonnstedt and Speed, 2002; Efron *et al*., 2001; Kendziorski *et al*., 2002; Newton and Kendziorski, 2003) that have proved useful can be also used.

An important parameter in our proposed method is the number of genes to be removed. In the current work, we have proposed removing the same number of genes as the number of identified significant DE genes. This method is simple and performs well in most cases. A justification is that FDR estimation depends more critically on the tails of the null distribution; Table 2 shows that removing a small number of the extreme genes effectively eliminates most of the bias. Nevertheless, if the number of DE genes is high, the current proposal may still overestimate FDR, although the degree of the bias is much less than that of the standard permutation method. On the other hand, when the true number of DE genes is smaller than that of claimed significant DE genes, the current proposal may underestimate FDR, which however is not really a serious issue. First, the biologists generally have a rough idea about the proportion of DE genes for the experiments. It is rare for one to try to identify more significant genes than the true ones because, with a smaller number of replicates and thus quite limited statistical power, the resulting FDR should be too high for the list of the identified genes to be

Table 2. True FDR (column T) and its estimates using the standard permutation (column S), the new method (column New) with various numbers of genes removed, permuting only true EE genes (column N) and weighted method (column W) for the mean-, *S*- and *t*-statistics for simulation set-up 5: there are $G = 4000$ genes, among which 200 are DE genes. The highlighted numbers are FDR estimates based on our proposed method: removing the same number of genes as the number of identified significant DE genes

						New (#genes removed)			
	$\widehat{\text{TP}}$	T	S	N	W	(50)	(100)	(200)	(400)
M	50	0.00	0.14	0.00	0.02	**0.01**	0.00	0.00	0.00
	100	0.05	0.28	0.05	0.11	0.12	**0.06**	0.05	0.04
	200	0.39	0.60	0.38	0.44	0.48	0.42	**0.37**	0.34
	300	0.56	0.74	0.56	0.60	0.65	0.60	0.55	0.49
	400	0.65	0.81	0.65	0.68	0.73	0.69	0.64	**0.58**
S	50	0.00	0.06	0.00	0.01	**0.00**	0.00	0.00	0.00
	100	0.10	0.17	0.11	0.12	0.14	**0.11**	0.10	0.10
	200	0.41	0.50	0.42	0.43	0.47	0.44	**0.40**	0.38
	300	0.57	0.66	0.58	0.57	0.62	0.60	0.56	0.52
	400	0.66	0.73	0.66	0.65	0.70	0.68	0.65	**0.59**
t	50	0.16	0.23	0.18	0.18	**0.18**	0.18	0.17	0.17
	100	0.30	0.37	0.32	0.32	0.33	**0.32**	0.31	0.30
	200	0.50	0.55	0.51	0.50	0.54	0.52	**0.50**	0.48
	300	0.62	0.66	0.62	0.60	0.65	0.64	0.61	0.58
	400	0.69	0.72	0.69	0.66	0.72	0.70	0.68	**0.64**

Table 3. The confirmed upper bound of false positive number, estimates based on the standard permutation and new method for the chromosomal evolution data

Statistic	$\widehat{\text{TP}}$	Upper bound	Standard	New
Mean	20	1	5	2
	40	1	10	2
	60	1	22	7
	80	19	67	28
	100	39	84	38
S	20	1	5	0
	40	7	10	0
	60	19	18	3
	80	28	24	4
	100	44	37	13
t	20	9	6	3
	40	23	14	9
	60	39	26	17
	80	56	41	30
	100	72	59	46

useful. (Note that, as discussed earlier, if the number of replicates is high, the overestimation problem with the standard permutation method will largely diminish, and thus it is no longer compelling to correct the standard permutation.) Second, if the biologists have no idea about the number of DE genes, and to be conservative, we recommend the following procedure: first estimating π_0, and then only using the current proposal if the proportion of claimed significant genes is smaller than $1 - \hat{\pi}_0$, and using the standard permutation method otherwise. Because, using the same argument as before (and based on our experience with simulated data), the permutation method (with all genes) will tend to overestimate π_0 (see also Wu *et al*., 2004), this conservative approach is in general still no worse than the standard permutation method.

ACKNOWLEDGEMENTS

This work was supported by NIH grants HL65462 and GM066098 and a UM AHC Development grant.

Conflict of Interest: none declared.

REFERENCES

Abramovich,F., Benjamini,Y., Donoho,D. and Johnstone,I. (2000) Adapting to unknown sparsity by controlling the false discovery rate. *Technical Report*. Department of Statistics, Stanford University.

Albers,W. *et al*. (1976) Asymptotic expansions for the power of distribution free tests in the one-sample problem. *Annal. Stat*., **4**, 108–156.

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

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

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

Broet,P. *et al*. (2004) A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562–2571.

Broet,P. *et al*. (2002) Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comput. Biol*., **9**, 671–683.

Brown,P. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet*., **21**(suppl.), 33–37.

Broberg,P. (2003) Statistical methods for ranking differentially expressed genes. *Genome Biol*., **4**, R41.

Bunea,F. *et al*. (2003) The consistency of the FDR estimator. *Technical Report*. Department of Statistics, Florida State University, .

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dalmasso,C. *et al*. (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.

Devlin,B. *et al*. (2003) Analysis of multilocus models of association. *Gen. Epidemiol*, **25**, 36–47.

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

Iyer,V. *et al*. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Ghosh,D. *et al*. (2004) The false discovery rate: a variable selection perspective. *J. Stat. Plan. Infer*. (in press).

Guan,Z. *et al*. (2004) 'Model-Based Approach to FDR Estimation' *Research Report 2004-016*, Division of Biostatistics, University of Minnesota.

Guo,X. and Pan,W. (2004) Using weighted permutation scores to detect differential gene expression with microarray data. *J. Bioinformatics Comput. Biol*., **3**, 989–1006.

Hampel,F.R., Ronchetti,E.M., Rousseeuw,P.J. and Stahel,W.A. (1986) *Robust Statistics: The Approach Based on Influence Function*. John Wiley, NY.

Kendziorski,C. *et al*. (2002) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist. Med*., **22**, 3899–3914.

Khodursky,A.B. *et al*. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **97**, 12170–12175.

Khodursky,A.B. *et al*. (2003) *Escherichia coli* spotted double-strand DNA microarrays: RNA extraction, labeling, hybridization, quality control, and data management. *Methods Mol. Biol. USA*, **224**, 61–78.

Lambert,D. (1990) Robust two-sample permutation tests. *Ann. Stat*., **13**, 606–625.

Lehmann,E.L. and Stein,C. (1949) On the theory of some nonparametric hypotheses. *Ann. Math. Stat*., **20**, 28–45.

Lonnstedt,I. and Speed,T. (2002) Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models*. Wiley, NY.

Newton,M. and Kendziorski,C. (2003) Parametric empirical Bayes method for microarrays. *The Analysis of Gene Expression Data: Methods and Software,* Springer, New York.

Pan,W. (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, **19**, 1333–1340.

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

Pollard,K.S. and Van der laan,M.J. (2003) Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test. In *Proceedings of the 2003 International MultiConference in Computer Science and Engineering (METMBS'03)* Los Vegas, USA, pp. 3–9.

Pollard,K.S. and Van der Laan,M.J. (2004) Choice of null distribution in resampling based multiple testing. *J. Stat. Plan. Inf.*, **125**, 85–101.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1–9.

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.

Qin,L. and Kerr,K. (2004) Empirical evaluation of methodologies for microarray data analysis. *Nucleic Acids Res.*, **32**, 5471–5479.

Ren,B. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Shedden,K. *et al*. (2005) Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics*, **6**, 26.

Smyth,G.K. *et al*. (2003) Statistical issues in cDNA microarray data analysis. *Functional Genomics: Methods and protocols*, **224**, 111–136.

Spellman,P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the Yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Cell Biol.*, **9**, 3273–3297.

Spino,C. and Pagano,M. (1991) Efficient calculation of the permutation distribution of trimmed means. *J. Am. Stat. Assoc.*, **86**, 729–737.

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 genome-wide experiments. *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. Stat. Soc. B*, **66**, 187–205.

Tani,T. *et al*. (2002) Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis. *Proc. Natl Acad. Sci. USA*, **99**, 13471–13476.

Tibshirani,R. and Bair,E. (2003) Improved detection of differential gene expression through the singular value decomposition.

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

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

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.

Wu,B. *et al*. (2004) Parametric and nonparametric FDR estimation revisited *Research Report 2004-015*, Division of Biostatistics, University of Minnesota.

Wu,B. (2005) Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics*, **21**, 1565–1571.

Xie,Y. *et al*. (2004) A case study on choosing normalization methods and test statistics for two-channel microarray data. *Comp. Funct. Genom.*, **5**, 432–444.

Xu,X.L. *et al*. (2002) A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum. Mol. Genet.*, **11**, 1977–1985.

Zhao,Y. and Pan,W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **19**, 1046–1054.

Zhong,S. *et al*. (2004) Evolutionary genomics of ecological specialization. *Proc. Natl Acad. Sci. USA*, **101**, 11719–11724.