BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 21 2014, pages 3044–3053
doi:10.1093/bioinformatics/btu432

Gene expression

Advance Access publication July 14, 2014

An improved method for computing q-values when the distribution of effect sizes is asymmetric

Megan Orr[1,*], Peng Liu[2] and Dan Nettleton[2]

[1]Department of Statistics, North Dakota State University, Fargo, ND 58102, USA and [2]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Asymmetry is frequently observed in the empirical distribution of test statistics that results from the analysis of gene expression experiments. This asymmetry indicates an asymmetry in the distribution of effect sizes. A common method for identifying differentially expressed (DE) genes in a gene expression experiment while controlling false discovery rate (FDR) is Storey's *q*-value method. This method ranks genes based solely on the *P*-values from each gene in the experiment.

Results: We propose a method that alters and improves upon the *q*-value method by taking the sign of the test statistics, in addition to the *P*-values, into account. Through two simulation studies (one involving independent normal data and one involving microarray data), we show that the proposed method, when compared with the traditional *q*-value method, generally provides a better ranking for genes as well as a higher number of truly DE genes declared to be DE, while still adequately controlling FDR. We illustrate the proposed method by analyzing two microarray datasets, one from an experiment of thale cress seedlings and the other from an experiment of maize leaves.

Availability and implementation: The R code and data files for the proposed method and examples are available at *Bioinformatics* online.

Contact: megan.orr@ndsu.edu

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

Received on October 18, 2013; revised on June 26, 2014; accepted on July 3, 2014

1 INTRODUCTION

Performing tens of thousands of hypothesis tests for one experiment has become a commonplace as technologies for producing high-dimensional data are becoming more prominent. This is especially the case in the field of statistical genomics, where technologies such as microarray and RNA-seq [see Brown and Botstein (1999) or Metzker (2010) for reviews of these technologies] measure the abundance of messenger RNA transcripts for thousands of genes in each subject of a gene expression study. Oftentimes, researchers are interested in comparing the gene expressions of subjects from two treatment groups. A major objective of such experiments is to identify genes that exhibit differential expression, i.e. a difference in the population treatment mean expression levels. Meaningful biological results depend on reliable detection of differentially expressed (DE) genes. Hence, the power of detecting differential expression for genes that are truly DE should be as high as possible, whereas genes that are equivalently expressed (EE)—i.e. genes that have no difference in the population treatment mean expression levels—should have a minimal chance of being declared DE. Thus, we must find an appropriate method for testing multiple hypotheses that provides good power while controlling some multiple testing errors.

When considering a traditional multiple testing problem, family-wise error rate (FWER) is often the preferred error rate used to control multiple testing error. The Bonferroni method and Holm's (1979) method are commonly used for this purpose. These methods, however, are not appropriate for high-dimensional gene expression data, as controlling FWER results in extremely low power for detection of differential expression. As a consequence, an alternative error rate known as false discovery rate (FDR; Benjamini and Hochberg, 1995) is usually used. FDR is simply the expected proportion of EE genes among all genes declared to be DE, where the proportion is defined to be zero if no genes are declared to be DE. Although FDR allows for more Type I errors than FWER when controlling these different error rates at the same numeric level α, the power for detecting differential expression is greatly increased, as discussed in Storey and Tibshirani (2003).

A common method for identifying DE genes while controlling FDR in a gene expression experiment is the *q*-value method, first proposed by Storey (2002). The *q*-value for a given gene represents the estimated FDR if the given gene and all genes with smaller *q*-values are declared to be DE. This method estimates FDR based on a set of *P*-values corresponding to m hypothesis tests. As previously mentioned, researchers are often interested in studying gene expression differences between groups of subjects. This is most often accomplished by performing a hypothesis test for a difference in treatment expression means for each gene, converting the *P*-values from the resulting tests to *q*-values, and then declaring genes with a *q*-value less than an FDR threshold to be DE. Potentially relevant information that this method does not take into account is the signs of the test statistics. Figure 1 shows a histogram of *t*-test statistics from a microarray experiment described in Jang *et al.* (2014). In this experiment, gene expressions from wild-type cells in thale cress seedlings were compared with those from mutant cells. Figure 1 shows the distribution of observed test statistics for 22 810 gene. Although it is not visually obvious, there is asymmetry in the distribution of test statistics, as more genes have negative test statistics than positive test statistics, which indicates asymmetry

*To whom correspondence should be addressed.

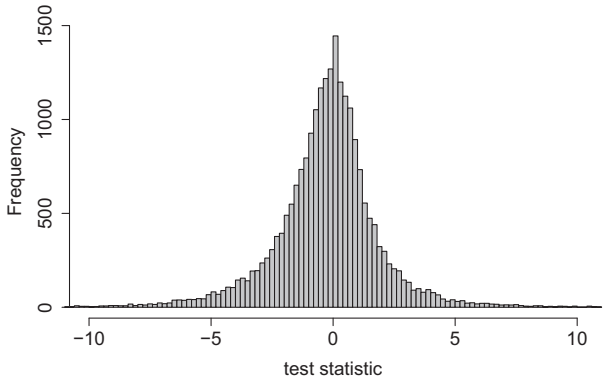

Fig. 1. A histogram of *t*-statistics from an experiment described in Jang *et al.* (2014) in which gene expressions from wild-type cells were compared with those in mutant cells in thale cress seedlings

in the distribution of effects sizes (i.e. the true differences in gene expression treatment means) as well.

We propose a new method for FDR estimation that alters the traditional *q*-value method by separating the two-sided *P*-values from an experiment into two subsets of *P*-values based on the sign of the test statistics, and then computing the *q*-values separately for each subset to create a better ranking of the genes with respect to differential expression. Through simulation studies using both independent normally distributed data and real gene expression data, we demonstrate how the proposed method can result in an improved ranking of genes with respect to differential expression over the traditional *q*-value method while still adequately controlling FDR.

The rest of the article is organized as follows. Section 2 reviews Storey's (2002) *q*-value method and introduces the proposed method for FDR estimation. Section 3 describes two simulation studies and uses the results of these studies to compare the performances of the proposed method and traditional *q*-value method by looking at the ranking of genes with respect to differential expression, the number of truly DE genes declared to be DE and also how well each method controls FDR. Section 4 presents analysis of two real microarray datasets, both from two-sample studies. Section 5 presents an analysis of a three-treatment gene expression experiment to illustrate how the proposed method can be generalized. Finally, Section 6 concludes the article with some discussion.

2 METHODS

This section describes the proposed method for estimating FDR when effect sizes in an experiment are asymmetric. Section 2.1 reviews Storey's (2002) *q*-value method, and Section 2.2 describes how the proposed method alters Storey's method to obtain a better FDR estimator when the distribution of treatment effects is asymmetric. Finally Section 2.3 discusses the advantages of the proposed method.

2.1 Review of Storey's *q*-value method

Suppose we wish to test m null hypotheses $H_1, \ldots, H_m$, where H_j is true if gene j is EE and false if gene j is DE. We will assume that p_j, the two-sided *P*-value that corresponds to H_j, follows a Uniform(0,1) distribution if gene j is EE and a distribution stochastically smaller than uniform if gene j is DE. These are standard assumptions made so that an unbiased size α test is obtained by rejecting H_j if and only if $p_j \leq \alpha$. In the case of a two-sample *t*-test, assuming a uniform null distribution for the *P*-values is equivalent to assuming a central *t*-distribution for the test statistic of any EE gene.

Benjamini and Hochberg (1995) defined FDR in a manner equivalent to $FDR = E(V/\max\{R, 1\})$, where V is the random variable representing the number of EE genes declared to be DE (or the number of Type I errors), and R is the random variable representing the total number of genes declared to be DE. Many methods have been proposed for estimating FDR, but the *q*-value method (Storey, 2002) is likely the most commonly used approach for gene expression experiments.

The formal definition of the *q*-value is given as

$$q_{(j)} = \min\left\{\frac{p_{(r)}\hat{m}_0}{r} : r = j, \ldots, m\right\}, \tag{1}$$

where $q_{(j)}$ is the *q*-value corresponding to $p_{(j)}$, the j^{th} smallest *P*-value, and $\hat{m}_0$ is the estimated number of EE genes among all m genes in the experiment. Specifically, q_j represents the estimated FDR if we declare gene j to be DE along with all other genes with *q*-values smaller than q_j.

Many approaches have been proposed for estimating m_0 by estimating the density of the *P*-values at $p_j = 1$ and multiplying this by m. Liang and Nettleton (2012); Nettleton *et al.* (2006); Storey (2002); Storey *et al.* (2004) have proposed approaches for doing this by developing methods for selecting a $\lambda \in (0, 1)$ and estimating m_0 as

$$\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} \mathbf{1}\{p_j > \lambda\}}{(1-\lambda)}. \tag{2}$$

Storey and Tibshirani (2003) developed an alternative method for estimating m_0 by first calculating $\hat{m}_0(\lambda)$ for a series of λ values between 0 and 1 using (2). Then the relationship between λ and $\hat{m}_0(\lambda)$ is estimated by fitting a natural cubic spline through the points $(\lambda, \hat{m}_0(\lambda))$. Finally, m_0 is estimated by evaluating this function at $\lambda = 1$. This approach will be used to estimate m_0 in the simulation studies and real data analyses in Sections 3, 4 and 5.

2.2 FDR estimation using two subsets of *P*-values

Consider the problem of identifying genes that are DE in an experiment. To do this, a test statistic t_j and corresponding two-sided *P*-value p_j is obtained for each gene $j = 1, \ldots, m$ by testing the null hypothesis $H_j : \mu_{j1} = \mu_{j2}$ against a two-sided alternative, where μ_{jt} is the population treatment mean expression for gene j under treatment t for $t = 1, 2$. We make the same assumptions about p_j as those described in Section 2.1.

Our proposed method begins by estimating m_0 for the entire set of m *P*-values. This can be done, for example, using any of the methods cited in Section 2.1. Then the *P*-values are divided into two subsets based on the sign of the corresponding test statistics. Let $\{p_k^{(1)} : k = 1, \ldots, m_1\}$ represent the subset of *P*-values corresponding to genes with negative test statistics, and let $\{p_k^{(2)} : k = 1, \ldots, m_2\}$ represent the remaining *P*-values, which correspond to genes with positive test statistics. Then the *q*-value method is applied separately to each subset of *P*-values. Therefore, for each gene k in each subset, the *q*-value is

$$q_{(k)}^{(i)} = \min\left\{\frac{p_{(r)}^{(i)}\hat{m}_0/2}{r} : r = k, \ldots, m_i\right\}, \tag{3}$$

where $p_{(r)}^{(i)}$ is the r^{th} smallest *P*-value in the i^{th} subset ($i = 1, 2$).

The estimators of FDR in (3) are based on the expectation that there are an equal number, $m_0/2$, of positive and negative test statistics corresponding to EE genes. This expectation follows from the assumption that EE genes have a Uniform(0,1) distribution, which also implies that the numerator in (3) is a natural estimate of the number of EE genes with

P-values less than or equal to $p^{(i)}_{(r)}$ among genes whose test statistics have sign $(-1)^i$. Thus, $q^{(i)}_{(k)}$ is a natural expression for the q-value associated with $p^{(i)}_{(k)}$.

2.3 Advantages of the proposed method

Sun and Cai (2007) showed that multiple testing methods that rank the significance of hypothesis tests solely on the resulting P-values, such as the traditional q-value method, are often inefficient in terms of minimizing the 'false non-discovery rate' (i.e. the expected proportion of DE genes declared to be EE). In many cases, additional information can be used to improve this ranking. When the distribution of effects sizes is asymmetric in an experiment, dividing the set of P-values from a gene expression experiment into two subsets based on the sign of the test statistics and calculating q-values separately for each subset, as described in Section 2.2, improves efficiency. Figure 2 helps illustrate this idea. The histogram on the left plots the two-sided P-values for genes that have negative test statistics from the microarray experiment in thale cress seedlings described in Jang *et al.* (2014). The histogram on the right plots the two-sided P-values for genes with positive test statistics from the same experiment. A horizontal dashed line is plotted at the estimated proportion of EE genes, $\hat{\pi}^{(i)}_0$, for each subset, and represents the estimated density for a P-value from an EE gene in subset i. This estimate is calculated as

$$\hat{\pi}^{(i)}_0 = \frac{\hat{m}_0/2}{m_i} \tag{4}$$

for $i = 1, 2$ and again is based on the expectation of an equal number of positive and negative test statistics from EE genes.

For each plot, the area of the left-most bar in each histogram represents the proportion of P-values that are < 0.05. The estimated proportion of EE genes among genes with test statistics of the same sign is represented by the proportion of area of the bar below the dashed line. Because the area of this bar is larger for the histogram that corresponds to negative test statistics than the one that corresponds to positive test statistics, and the area of the bar below the dashed line is relatively smaller, the estimated proportion of EE genes among genes with test statistics of the appropriate sign and P-values < 0.05 will be lower for the first histogram than the second. Thus, a gene with a P-value close to 0.05 will have a smaller q-value if this gene has a negative test statistic than if it has a positive test statistic. More generally, a gene with a small P-value will be more likely to be declared DE if it corresponds to a negative test statistic than if it corresponds to a positive test statistic, and it is possible for a gene with a higher P-value that corresponds to a negative test statistic to be ranked more significant (i.e. have a lower q-value) than a gene with a smaller P-value that corresponds to a positive test statistic. This reasoning agrees with (3), as the denominator in the formula corresponding to $i = 1$ will be larger than the denominator in the corresponding formula for $i = 2$ for the same two-sided P-value. Thus, two genes that have the same P-value but different signs of their corresponding test statistics will have different q-values, and the gene with the negative test statistic will have the lower q-value. We show via simulation in the next section that this strategy often results in a better significance ranking of genes when effect sizes are asymmetric.

3 SIMULATION STUDIES

To evaluate the performance of the proposed method and compare it with that of the traditional q-value method, we performed two sets of simulation studies. For each simulated dataset, each gene j of the $m = 10\,000$ total genes was tested for differential expression by testing $H_j : \mu_{j1} = \mu_{j2}$ against a two-sided alternative. For each test, a corresponding test statistic, t_j, and P-value, p_j, were computed using the moderated t-test proposed by Smyth

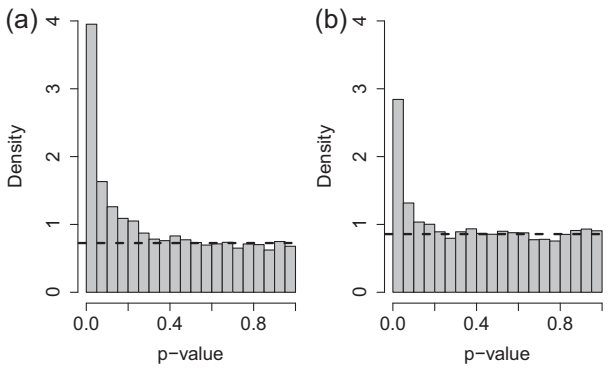

Fig. 2. A histogram of P-values for the thale cress data corresponding to (**a**) negative t-test statistics and (**b**) positive t-test statistics. The estimated proportion of EE genes, $\hat{\pi}^{(i)}_0$, $i = 1, 2$, is plotted as the dashed horizontal line in each plot

(2004). This method was developed specifically for analyzing data from microarray experiments and borrows information across all genes to more accurately estimate the error variances for each individual gene. The resulting P-values were then analyzed by the traditional q-value method and the proposed method described in Section 2.2. Note that both methods use the same estimate of π_0, obtained by the natural cubic spline approach of Storey and Tibshirani (2003), briefly described in Section 2.1.

Two sets of simulations studies were performed. The first simulation study involved generating datasets with independent normally distributed data to evaluate the methods with data that are consistent with assumptions used to derive the moderated t-test. The second set of simulation studies used real gene expression data to evaluate the methods under conditions that are not ideal, but are generally observed in gene expression data, namely, data with a complex correlation structure that cannot be perfectly modeled with common probability distributions.

For each simulation study, four variables were manipulated to evaluate the methods under different situations. Sample sizes of $n = 4$, $n = 10$ and $n = 20$ for each gene in each treatment were used. These sample sizes were chosen because gene expression experiments tend to have small sample sizes; ~90% of gene expression datasets on the Gene Expression Omnibus (Edgar *et al.*, 2002) have a total sample size (i.e. the number of subjects across all treatments for each gene) no >40. Thus, a maximum total sample size of $2n = 40$ was chosen for the simulations performed. The number of EE genes was also varied from $m_0 = 5000$ to $m_0 = 9500$ of $m = 10\,000$ total genes. Four levels of an asymmetry parameter (π_A) were considered to examine performance as the distribution of effect sizes ranged from symmetric about zero to highly asymmetric about zero. This asymmetry parameter is explained in more detail in the Section 3.1, but its impact on the effect size distribution can be seen in Figure 3, which plots the density for each value of π_A from which effect sizes, relative to the standard deviations of the genes, were randomly drawn for DE genes. Finally, two values of the mean of the effect sizes, relative to the standard deviation of the genes, were implemented in the simulations: $\mu_\delta = 1$ and $\mu_\delta = 2$. These

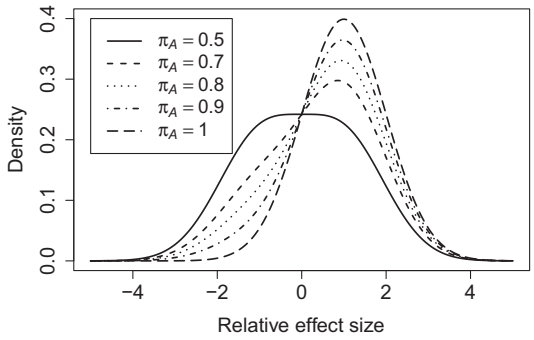

Fig. 3. The densities of effect sizes for DE genes for different values of π_A with $\mu_\delta = 1$

values were chosen so that the proportion of genes identified as significantly DE in our simulation study would range over values typically seen in practice.

For each simulation setting in each simulation study, 500 datasets were randomly generated and analyzed.

3.1 Simulations using independent normal data

In the first simulation study, data were randomly generated from independent normal distributions. For each dataset, data for each gene $j = 1, \ldots, 10\,000$ were generated as follows. The variance σ_j^2 was randomly selected from an inverse gamma distribution. The parameters of the inverse gamma distribution were calculated from the dataset of an experiment described in Hannenhalli *et al.* (2006) using methods proposed by Smyth (2004). If gene j was EE, then $\mu_{j1} = \mu_{j2} = 0$. If gene j was DE, then $\mu_{j1} = 0$ and the effect size $\mu_{j2} = \delta_j$ was randomly drawn from the mixture distribution

$$h_j(\delta) = \pi_A \phi(\delta;\, \mu_\delta \sigma_j,\, \sigma_j^2) + (1 - \pi_A)\phi(\delta;\, -\mu_\delta \sigma_j,\, \sigma_j^2), \qquad (5)$$

where $\phi(\delta;\, \gamma,\, \tau^2)$ is the normal density with mean γ and variance τ^2 evaluated at δ. Finally, n values were randomly drawn from a Normal(μ_{jt}, σ_j^2) distribution for each treatment $t = 1, 2$.

3.2 Simulations using gene expression data

The second simulation study used microarray data from an experiment in which gene expressions were measured from the heart tissue of $N = 108$ human subjects suffering from idiopathic dilated cardiomyopathy. This experiment is described in detail in Hannenhalli *et al.* (2006), and the data generated from this experiment are available from the Gene Expression Omnibus under accession number GSE5406. Although this dataset contains data from over 20 000 genes, $m = 10\,000$ genes were randomly selected for analysis in the simulation study. The expression values for each dataset were generated as follows. For each gene $j = 1, \ldots, m$, the variance s_j^2 was calculated from all N subjects. Then, data from two n subjects were randomly drawn, without replacement, from the microarray dataset. From this subset of data, the data from n subjects were randomly chosen to be in the first treatment, and the data from the remaining n subjects were assigned to the second treatment. Note that, at this point, because the data from both treatments were randomly drawn from the same population, the population treatment means were equal (i.e. $\mu_{j1} = \mu_{j2}$) for each gene. If gene j was EE, the data from this gene was not altered in any way. If gene j was DE, then the effect size, δ_j, was randomly chosen from mixture model (5), but by replacing σ_j with s_j. Then δ_j was added to the data from gene j in the second treatment. Note that this method of data generation did not change the correlation structure of the data in any way but only shifted the mean of the data between the two treatments for DE genes.

3.3 Results

Tables 1 and 2 as well as Supplementary Tables S1 and S2 (provided in the Supplementary Materials) present the results of the simulation studies. For each setting in each simulation study, the mean (based on the 500 simulated datasets) partial area under the receiver-operating characteristic (ROC) curve (pAUC) is given with its corresponding standard error in parentheses. These are partial areas because we considered only the most relevant region of the ROC curve where the false-positive rate was ≤ 0.10. The method that ranks the genes better with regard to differential expression will have a higher pAUC. For each simulation setting, a traditional paired t-test was performed to test for a difference in the mean pAUCs of the proposed method and the traditional q-value method. If this test was significant at 5%, then the higher mean pAUC is presented in bold font in Tables 1 and 2 and the supplementary tables. Additionally, if the higher mean pAUC is at least 5% greater than the lower mean pAUC, then the higher mean pAUC is also underlined, and this method was considered to 'outperform' the other method. The 5% improvement criterion was used to avoid situations where the mean pAUC was very similar between the proposed and traditional methods, in some cases equal when rounded to one decimal, but the t-test indicated a statistically significant (though not practically significant) difference. The combination of the two criteria allowed for the identification of simulation settings where both practical and statistical significance were present. In addition, mean S, the mean number of DE genes that were declared to be DE when FDR is nominally controlled at 5% is also given for each setting to observe the number of correctly identified DE genes, on average. Similar to the pAUCs, t-tests were used to determine whether one method produced a larger mean S and whether the higher mean S was at least 5% greater than the lower mean S. Also, to verify that each method adequately controls FDR, the empirical approximation of FDR was calculated for each method while nominally controlling FDR at 5%. We will call this quantity $\overline{V/R}$, which is the average of the dataset-specific fractions defined as the proportion of EE genes among all genes with q-values ≤ 0.05 or 0 if no genes have q-values ≤ 0.05.

Regardless of data type, both mean pAUC and mean S increase as the sample size increases as well as when the mean effect size increases for both the proposed and traditional q-value methods. This is unsurprising, as increasing the sample size or the mean effect size increases the power for detecting differential expression.

For the simulations involving normally distributed data with $\mu_\delta = 1$ (see Table 1), the proposed method performed better in 12 of 60 simulation settings with regard to mean pAUC, including

Table 1. The mean pAUC, mean S and $\overline{V/R}$ with corresponding standard errors in parentheses for the proposed and traditional q-value methods for each setting in the simulation study using independent normal data with $\mu_\delta = 1$

n	m_0	π_A	Mean pAUC (%)		Mean S		$\overline{V/R}$ (%)	
			Proposed	Traditional	Proposed	Traditional	Proposed	Traditional
4	9500	1	**39.4 (0.09)**	33.5 (0.09)	**7.7 (0.26)**	3.1 (0.14)	5.84 (0.49)	4.77 (0.59)
		0.9	**36.9 (0.09)**	33.4 (0.09)	**6.1 (0.21)**	2.7 (0.13)	5.80 (0.52)	3.91 (0.56)
		0.8	**35.3 (0.09)**	33.5 (0.08)	**5.5 (0.20)**	3.0 (0.14)	7.43 (0.62)	5.67 (0.60)
		0.7	**34.3 (0.09)**	33.4 (0.09)	**4.2 (0.16)**	2.8 (0.13)	7.00 (0.70)	5.27 (0.65)
		0.5	**33.5 (0.09)**	33.5 (0.09)	**3.9 (0.14)**	3.1 (0.14)	6.40 (0.67)	4.77 (0.63)
	9000	1	**39.4 (0.06)**	33.6 (0.06)	**38.7 (0.63)**	14.3 (0.42)	4.75 (0.16)	4.95 (0.16)
		0.9	**36.9 (0.07)**	33.6 (0.07)	**31.3 (0.53)**	14.2 (0.39)	4.73 (0.17)	4.68 (0.17)
		0.8	**35.2 (0.07)**	33.5 (0.06)	**25.1 (0.51)**	14.1 (0.40)	5.40 (0.22)	5.03 (0.29)
		0.7	**34.4 (0.07)**	33.7 (0.07)	**19.9 (0.45)**	14.3 (0.40)	5.25 (0.25)	4.50 (0.25)
		0.5	33.6 (0.06)	**33.7 (0.06)**	**15.9 (0.39)**	14.1 (0.40)	5.28 (0.29)	4.90 (0.29)
	7000	1	**16.5 (0.03)**	14.6 (0.03)	**38.9 (0.65)**	14.5 (0.41)	3.94 (0.14)	3.46 (0.14)
		0.9	**36.8 (0.04)**	33.5 (0.04)	**349.4 (1.54)**	227.9 (1.45)	4.25 (0.04)	4.26 (0.04)
		0.8	**35.3 (0.04)**	33.6 (0.04)	**298.6 (1.50)**	232.1 (1.42)	4.33 (0.05)	4.32 (0.06)
		0.7	**34.2 (0.04)**	33.5 (0.04)	**260.1 (1.44)**	229.5 (1.42)	4.33 (0.06)	4.32 (0.06)
		0.5	33.5 (0.04)	**33.5 (0.04)**	**227.7 (1.55)**	226.4 (1.54)	4.29 (0.06)	4.25 (0.06)
	5000	1	**11.9 (0.02)**	10.8 (0.02)	**39.6 (0.61)**	14.1 (0.40)	2.72 (0.12)	2.39 (0.12)
		0.9	**36.9 (0.04)**	33.6 (0.04)	**966.3 (2.20)**	720.2 (2.22)	3.54 (0.03)	3.49 (0.03)
		0.8	**35.2 (0.04)**	33.5 (0.04)	**846.1 (2.23)**	712.6 (2.19)	3.54 (0.03)	3.55 (0.03)
		0.7	**34.3 (0.04)**	33.5 (0.04)	**778.5 (2.30)**	719.0 (2.33)	3.58 (0.03)	3.60 (0.03)
		0.5	33.5 (0.04)	**33.5 (0.04)**	724.6 (2.32)	724.0 (2.33)	3.64 (0.03)	3.63 (0.03)
10	9500	1	**57.5 (0.09)**	54.9 (0.09)	**151.0 (0.56)**	131.8 (0.56)	5.00 (0.08)	4.97 (0.08)
		0.9	**56.4 (0.09)**	54.9 (0.09)	**143.4 (0.55)**	132.1 (0.54)	4.98 (0.08)	4.98 (0.08)
		0.8	**55.7 (0.10)**	54.9 (0.10)	**137.4 (0.58)**	131.5 (0.57)	4.99 (0.08)	5.08 (0.08)
		0.7	**55.1 (0.10)**	54.8 (0.09)	**133.9 (0.55)**	131.3 (0.55)	5.03 (0.08)	4.95 (0.09)
		0.5	**54.9 (0.09)**	54.9 (0.09)	**131.5 (0.55)**	131.3 (0.55)	5.11 (0.08)	5.08 (0.08)
	9000	1	**57.4 (0.07)**	54.9 (0.07)	**354.3 (0.78)**	315.5 (0.77)	4.92 (0.05)	4.85 (0.05)
		0.9	**56.4 (0.07)**	54.8 (0.07)	**337.7 (0.82)**	315.4 (0.83)	4.91 (0.06)	4.80 (0.06)
		0.8	**55.6 (0.07)**	54.8 (0.07)	**327.8 (0.84)**	315.9 (0.83)	4.89 (0.05)	4.88 (0.05)
		0.7	**55.2 (0.07)**	54.9 (0.06)	**321.4 (0.84)**	316.0 (0.84)	4.88 (0.05)	4.84 (0.05)
		0.5	54.9 (0.07)	**54.9 (0.07)**	315.9 (0.85)	316.0 (0.85)	4.89 (0.05)	4.88 (0.05)
	7000	1	**22.5 (0.03)**	21.6 (0.03)	**358.8 (0.81)**	319.1 (0.79)	3.83 (0.04)	3.78 (0.04)
		0.9	**56.3 (0.04)**	54.9 (0.04)	**1326.8 (1.52)**	1262.0 (1.51)	4.47 (0.03)	4.43 (0.03)
		0.8	**55.7 (0.04)**	54.9 (0.04)	**1300.1 (1.58)**	1264.7 (1.57)	4.55 (0.03)	4.52 (0.03)
		0.7	**55.2 (0.04)**	54.9 (0.04)	**1279.0 (1.47)**	1264.7 (1.49)	4.51 (0.03)	4.48 (0.03)
		0.5	54.9 (0.04)	**54.9 (0.04)**	1263.3 (1.55)	1263.3 (1.55)	4.49 (0.03)	4.49 (0.03)
	5000	1	**15.4 (0.02)**	14.9 (0.02)	**362.0 (0.89)**	322.2 (0.88)	2.73 (0.04)	2.71 (0.04)
		0.9	**56.4 (0.04)**	54.9 (0.04)	**2554.1 (2.05)**	2452.4 (2.08)	3.96 (0.02)	3.99 (0.02)
		0.8	**55.7 (0.04)**	54.9 (0.04)	**2505.8 (2.09)**	2450.8 (2.07)	3.97 (0.02)	3.96 (0.02)
		0.7	**55.2 (0.04)**	54.8 (0.04)	**2471.2 (2.01)**	2447.5 (2.02)	3.97 (0.02)	3.97 (0.02)
		0.5	55.0 (0.04)	**55.0 (0.04)**	2453.6 (2.11)	2453.6 (2.12)	3.96 (0.02)	3.96 (0.02)
20	9500	1	**68.8 (0.09)**	68.0 (0.09)	**256.0 (0.56)**	247.2 (0.58)	4.92 (0.06)	5.04 (0.06)
		0.9	**68.6 (0.08)**	68.0 (0.09)	**253.3 (0.54)**	248.0 (0.53)	4.96 (0.06)	4.98 (0.06)
		0.8	**68.2 (0.09)**	67.9 (0.09)	**249.9 (0.55)**	247.1 (0.55)	4.97 (0.06)	4.94 (0.06)
		0.7	**68.1 (0.09)**	68.0 (0.09)	**250.0 (0.59)**	248.4 (0.59)	5.16 (0.06)	5.13 (0.06)
		0.5	67.7 (0.09)	67.7 (0.09)	246.6 (0.56)	246.6 (0.57)	5.00 (0.06)	4.96 (0.06)
	9000	1	**68.8 (0.06)**	67.9 (0.06)	**549.6 (0.77)**	533.9 (0.73)	4.99 (0.04)	4.92 (0.04)
		0.9	**68.4 (0.06)**	67.9 (0.06)	**543.3 (0.78)**	533.3 (0.79)	4.98 (0.04)	4.94 (0.04)
		0.8	**68.2 (0.06)**	67.9 (0.06)	**539.5 (0.72)**	534.1 (0.72)	4.99 (0.04)	4.97 (0.04)
		0.7	**68.0 (0.06)**	67.9 (0.06)	**536.3 (0.81)**	533.6 (0.81)	4.90 (0.04)	4.92 (0.04)
		0.5	67.9 (0.06)	**68.0 (0.06)**	534.0 (0.77)	533.9 (0.77)	4.92 (0.04)	4.91 (0.04)
	7000	1	**26.3 (0.03)**	26.0 (0.03)	**556.0 (0.81)**	539.1 (0.82)	3.78 (0.03)	3.82 (0.03)
		0.9	**68.4 (0.04)**	67.9 (0.04)	**1855.8 (1.37)**	1829.8 (1.37)	4.66 (0.02)	4.66 (0.02)
		0.8	**68.2 (0.04)**	67.9 (0.04)	**1846.0 (1.36)**	1832.0 (1.36)	4.59 (0.02)	4.61 (0.02)
		0.7	**68.1 (0.04)**	68.0 (0.04)	**1837.3 (1.34)**	1831.2 (1.35)	4.61 (0.02)	4.60 (0.02)
		0.5	67.9 (0.04)	**67.9 (0.04)**	1830.7 (1.34)	1830.7 (1.34)	4.64 (0.02)	4.63 (0.02)

(continued)

Table 1. Continued

n	m_0	π_A	Mean pAUC (%)		Mean S		$\overline{V/R}$ (%)	
			Proposed	Traditional	Proposed	Traditional	Proposed	Traditional
	5000	1	**17.8 (0.02)**	17.6 (0.02)	**562.9 (0.82)**	546.3 (0.82)	2.73 (0.03)	2.70 (0.03)
		0.9	**68.5 (0.03)**	68.0 (0.03)	**3338.1 (1.79)**	3303.4 (1.77)	4.20 (0.02)	4.21 (0.02)
		0.8	**68.3 (0.03)**	68.0 (0.03)	**3323.7 (1.79)**	3303.8 (1.80)	4.23 (0.02)	4.21 (0.02)
		0.7	**68.1 (0.03)**	67.9 (0.03)	**3310.6 (1.71)**	3301.4 (1.74)	4.21 (0.02)	4.22 (0.02)
		0.5	68.0 (0.03)	**68.0 (0.03)**	3304.2 (1.77)	3304.3 (1.76)	4.23 (0.02)	4.22 (0.02)

Note: For each setting, for both the mean pAUC and mean S, the highest values are given in bold font if a t-test has determined that there is a difference in the means of the proposed and traditional methods at 5% significance. If the t-test was not significant, bold font is not used. Statistically significant improvements that we deemed practically significant (at least 5% improvement) are indicated by underlining.

12 of 20 simulation settings with $n = 4$. In regard to mean S, the proposed method performed better in 25 of 60 simulation settings, including 18 of 20 settings with $n = 4$. The traditional method did not perform better than the proposed method in any simulation setting in terms of mean pAUC or mean S. When the mean effect size is increased to $\mu_\delta = 2$ (see Supplementary Table S1 in the Supplementary Materials), the proposed method outperformed the traditional method with regard to pAUC in 8 of 60 settings, all of which are in simulation settings with $n = 4$. For mean S, the proposed method outperformed the traditional method in 16 of 60 settings, including 14 of 20 settings with $n = 4$. As with $\mu_\delta = 1$, the traditional method did not outperform the proposed method in terms of mean pAUC or mean S in any simulation setting.

For simulations involving microarray data with $\mu_\delta = 1$, the proposed method performed better than the traditional methods in 17 of 60 (14 of 20 settings with $n = 4$) and 23 of 60 (17 of 20 settings with $n = 4$) simulation settings in regard to mean pAUC and mean S, respectively. The traditional method did not perform better than the proposed method in any simulation setting in terms of mean pAUC or mean S. When $\mu_\delta = 2$ (see Supplementary Table S2 in the Supplementary Materials), the proposed method outperformed the traditional method in mean pAUC in only 5 of 60 simulation settings, all when $n = 4$. In terms of mean S, the proposed method outperformed the traditional method in 15 of 60 settings, 13 of which when $n = 4$. The traditional method did not outperform the proposed method in any simulation settings in terms of mean pAUC or mean S.

When taking into account all simulation settings with normally distributed data, all of the 20 simulation settings in which the proposed method outperformed the traditional method in terms of mean pAUC were with $\pi_A \geq 0.8$. For mean S, 33 (80.1%) of the 41 simulation settings in which the proposed method performed better had $\pi_A \geq 0.8$. Among all simulation settings involving microarray data, 18 (81.8%) of the 22 simulation settings in which the proposed method outperformed the tradition method in terms of pAUC occurred when $\pi_A \geq 0.8$. For mean S, this occurred in 31 (81.6%) of 38 settings. Additionally, among the 42 simulation settings across all 240 simulations in which the proposed method outperformed the traditional method in terms of pAUC, 39 (92.9%) occurred in simulation settings with $n = 4$. Among the 79 simulation settings where the proposed method outperformed the traditional method in terms of mean S, 62 (78.5%) occurred in simulation settings with $n = 4$. The proposed method never outperformed the traditional method in simulation settings with $n = 20$ with regard to either mean pAUC or mean S. The traditional method did not outperform the proposed method in any simulation setting.

As shown in the last two columns of Tables 1 and 2 and Supplementary Tables S1 and S2, both the proposed and traditional methods adequately controlled FDR near the 5% level across all simulation settings and data types.

4 REAL DATA ANALYSIS

4.1 Thale cress seedlings

In this section, we analyze data from a study described in Jang *et al.* (2014) using both the proposed and traditional q-value methods. In this study, expressions from $m = 22\,810$ genes in thale cress seedlings were compared between two genotypes, wild-type and mutant, using six Affymetrix *Arabidopsis* Genome ATH1 Arrays, with $n = 3$ for each genotype. Mutant seedlings contained a transfer DNA insertion that the wild-type seedlings did not. Data from this experiment are available from the Gene Expression Omnibus under access number GSE48114.

The test statistics from this experiment are shown in the histogram in Figure 1. There is asymmetry in this histogram, as there are more negative test statistics than positive test statistics, suggesting asymmetry in the effect sizes. More specifically, there are $m_1 = 12\,355$ genes with negative test statistics and $m_2 = 10\,455$ genes corresponding to positive test statistics. The asymmetry becomes more evident in Figure 2, where P-values corresponding to negative test statistics follow a distribution that is stochastically smaller than the distribution of P-values corresponding to positive test statistics.

Using Storey and Tibshirani's (2003) natural cubic spline method, the estimated number of EE genes in this experiment is $\hat{m}_0 = 17\,933.75$. Because we expect there to be the same number of negative and positive test statistics among the EE genes, we estimate that there are $\hat{m}_0/2 = 8966.875$ EE genes with negative test statistics and $\hat{m}_0/2 = 8966.875$ EE genes with positive test statistics. From these estimates, we can also estimate

Table 2. The mean pAUC, mean S and $\overline{V/R}$ with corresponding standard errors in parentheses for the proposed and traditional q-value methods for each setting in the simulation study using microarray data with $\mu_\delta = 1$

n	m_0	π_A	Mean pAUC (%)		Mean S		$\overline{V/R}$ (%)	
			Proposed	Traditional	Proposed	Traditional	Proposed	Traditional
4	9500	1	**47.5 (0.62)**	37.5 (0.30)	**19.3 (0.95)**	10.9 (0.68)	4.36 (0.58)	4.26 (0.56)
		0.9	**44.7 (0.58)**	37.4 (0.29)	**13.7 (0.82)**	8.8 (0.69)	3.52 (0.48)	2.60 (0.45)
		0.8	**43.1 (0.55)**	37.6 (0.28)	**13.6 (0.72)**	10.0 (0.63)	4.02 (0.53)	3.70 (0.52)
		0.7	**42.6 (0.53)**	37.9 (0.29)	**12.3 (0.71)**	10.4 (0.70)	3.31 (0.52)	2.69 (0.44)
		0.5	**41.9 (0.52)**	37.9 (0.30)	**11.2 (0.65)**	10.5 (0.67)	4.22 (0.51)	3.68 (0.49)
	9000	1	**44.0 (0.43)**	37.5 (0.30)	**64.3 (2.52)**	36.5 (1.97)	3.79 (0.40)	3.46 (0.40)
		0.9	**41.9 (0.43)**	37.8 (0.29)	**53.4 (2.21)**	35.1 (1.89)	3.67 (0.39)	3.26 (0.39)
		0.8	**40.9 (0.44)**	37.6 (0.29)	**53.2 (2.37)**	42.0 (2.17)	4.23 (0.44)	3.84 (0.41)
		0.7	**40.0 (0.44)**	37.7 (0.29)	**48.7 (2.20)**	42.3 (2.12)	4.40 (0.43)	3.78 (0.43)
		0.5	**39.6 (0.46)**	37.2 (0.30)	**42.7 (2.08)**	41.6 (2.11)	5.10 (0.50)	4.47 (0.50)
	7000	1	**19.5 (0.34)**	15.9 (0.10)	**68.8 (2.58)**	38.7 (1.97)	2.71 (0.25)	2.58 (0.25)
		0.9	**39.6 (0.27)**	37.7 (0.29)	**439.5 (10.25)**	327.4 (9.71)	3.85 (0.25)	3.54 (0.25)
		0.8	**37.8 (0.27)**	37.2 (0.28)	**379.9 (9.92)**	316.5 (9.78)	3.68 (0.24)	3.36 (0.24)
		0.7	**37.5 (0.28)**	37.3 (0.29)	**345.2 (9.42)**	316.8 (9.44)	3.77 (0.26)	3.55 (0.26)
		0.5	37.6 (0.29)	37.7 (0.28)	**314.9 (10.26)**	313.8 (10.31)	3.66 (0.31)	3.55 (0.31)
	5000	1	**15.1 (0.36)**	11.6 (0.06)	**70.9 (3.26)**	40.6 (2.67)	2.03 (0.21)	1.99 (0.21)
		0.9	**39.7 (0.28)**	37.6 (0.31)	**1131.7 (19.19)**	906.3 (19.96)	3.60 (0.19)	3.42 (0.19)
		0.8	**38.6 (0.27)**	37.9 (0.28)	**1003.1 (19.68)**	880.3 (20.07)	3.35 (0.19)	3.19 (0.20)
		0.7	**38.0 (0.28)**	37.8 (0.29)	**979.0 (19.38)**	922.1 (19.67)	3.42 (0.18)	3.34 (0.18)
		0.5	38.1 (0.30)	**38.2 (0.30)**	899.4 (20.31)	898.8 (20.36)	3.25 (0.21)	3.21 (0.21)
10	9500	1	**60.5 (0.34)**	56.9 (0.25)	**155.1 (1.18)**	136.3 (1.23)	3.94 (0.33)	3.61 (0.33)
		0.9	**60.1 (0.37)**	56.9 (0.25)	**146.5 (1.20)**	135.7 (1.23)	4.24 (0.39)	3.82 (0.38)
		0.8	**59.6 (0.36)**	57.0 (0.24)	**142.3 (1.21)**	136.5 (1.23)	4.13 (0.37)	3.96 (0.38)
		0.7	**59.3 (0.38)**	56.8 (0.24)	**139.5 (1.22)**	136.8 (1.24)	4.30 (0.40)	4.10 (0.38)
		0.5	**59.9 (0.38)**	57.2 (0.24)	**137.4 (1.23)**	137.2 (1.24)	4.06 (0.38)	3.91 (0.36)
	9000	1	**58.9 (0.22)**	57.1 (0.24)	**363.0 (2.32)**	325.0 (2.50)	4.46 (0.32)	4.14 (0.32)
		0.9	**58.1 (0.24)**	57.0 (0.24)	**347.3 (2.44)**	324.7 (2.54)	4.54 (0.35)	4.31 (0.35)
		0.8	**57.8 (0.27)**	57.1 (0.24)	**337.3 (2.19)**	325.2 (2.27)	4.29 (0.32)	4.08 (0.31)
		0.7	**57.3 (0.27)**	56.8 (0.24)	**326.5 (2.26)**	321.3 (2.29)	4.10 (0.30)	3.99 (0.30)
		0.5	**58.0 (0.30)**	57.2 (0.24)	323.9 (2.34)	323.9 (2.35)	3.93 (0.34)	3.89 (0.34)
	7000	1	**23.4 (0.15)**	22.4 (0.08)	**371.7 (2.62)**	333.4 (2.83)	3.64 (0.26)	3.45 (0.26)
		0.9	**57.9 (0.22)**	57.0 (0.24)	**1346.0 (6.46)**	1282.4 (6.86)	4.42 (0.24)	4.31 (0.24)
		0.8	**57.6 (0.22)**	57.3 (0.24)	**1324.9 (6.45)**	1290.6 (6.66)	4.32 (0.24)	4.22 (0.24)
		0.7	**57.3 (0.23)**	57.1 (0.23)	**1305.4 (6.68)**	1290.3 (6.77)	4.32 (0.24)	4.27 (0.24)
		0.5	57.0 (0.24)	**57.0 (0.24)**	1276.7 (6.67)	1276.7 (6.67)	4.37 (0.25)	4.34 (0.25)
	5000	1	**16.5 (0.18)**	15.4 (0.05)	**379.9 (3.38)**	341.4 (3.59)	2.75 (0.20)	2.59 (0.20)
		0.9	**57.4 (0.21)**	56.5 (0.24)	**2571.0 (9.96)**	2468.4 (10.84)	4.23 (0.16)	4.09 (0.16)
		0.8	**57.4 (0.22)**	56.9 (0.23)	**2534.6 (10.27)**	2480.5 (10.70)	4.02 (0.17)	3.99 (0.18)
		0.7	**56.8 (0.23)**	56.7 (0.24)	**2514.1 (10.00)**	2492.2 (10.19)	4.29 (0.20)	4.25 (0.20)
		0.5	57.1 (0.24)	**57.1 (0.24)**	2480.8 (9.96)	2480.8 (9.97)	3.88 (0.18)	3.86 (0.18)
20	9500	1	**70.3 (0.25)**	69.0 (0.20)	**257.3 (0.82)**	248.6 (0.87)	5.03 (0.41)	4.95 (0.43)
		0.9	**70.8 (0.27)**	69.3 (0.19)	**254.7 (0.80)**	249.3 (0.81)	4.88 (0.41)	4.53 (0.39)
		0.8	**70.5 (0.25)**	69.1 (0.18)	**252.0 (0.80)**	249.3 (0.82)	4.30 (0.31)	4.21 (0.32)
		0.7	**70.4 (0.29)**	69.0 (0.20)	**250.6 (0.84)**	249.4 (0.84)	5.09 (0.41)	4.97 (0.41)
		0.5	**71.0 (0.28)**	69.4 (0.19)	248.2 (0.81)	248.1 (0.81)	4.17 (0.34)	4.08 (0.33)
	9000	1	**69.4 (0.16)**	69.2 (0.17)	**555.0 (1.39)**	538.2 (1.47)	4.65 (0.29)	4.49 (0.29)
		0.9	**69.3 (0.16)**	69.1 (0.17)	**548.0 (1.29)**	538.0 (1.32)	4.68 (0.29)	4.43 (0.29)
		0.8	**69.2 (0.20)**	68.9 (0.19)	**541.7 (1.41)**	536.4 (1.43)	5.18 (0.38)	5.14 (0.39)
		0.7	**68.9 (0.21)**	68.7 (0.20)	**540.2 (1.35)**	538.0 (1.37)	5.39 (0.39)	5.32 (0.39)
		0.5	**69.2 (0.22)**	69.0 (0.19)	536.5 (1.36)	536.6 (1.36)	5.14 (0.38)	5.07 (0.38)
	7000	1	**27.0 (0.13)**	26.5 (0.06)	**560.5 (1.82)**	543.3 (1.90)	3.63 (0.24)	3.45 (0.24)
		0.9	**69.0 (0.16)**	68.8 (0.18)	**1864.2 (3.39)**	1839.6 (3.57)	5.18 (0.25)	5.10 (0.25)
		0.8	**69.3 (0.17)**	69.2 (0.17)	**1856.4 (3.36)**	1843.4 (3.45)	4.80 (0.24)	4.71 (0.24)
		0.7	**69.0 (0.17)**	69.0 (0.18)	**1845.4 (3.60)**	1838.4 (3.61)	4.95 (0.26)	4.93 (0.26)
		0.5	68.9 (0.18)	**69.0 (0.18)**	1837.1 (3.31)	1836.9 (3.30)	4.85 (0.26)	4.82 (0.26)

(continued)

Table 2. Continued

n	m_0	π_A	Mean pAUC (%)		Mean S		$\overline{V/R}$ (%)	
			Proposed	Traditional	Proposed	Traditional	Proposed	Traditional
	5000	1	**18.2 (0.11)**	17.8 (0.04)	**567.3 (2.15)**	550.2 (2.38)	2.68 (0.17)	2.52 (0.17)
		0.9	**69.3 (0.15)**	69.0 (0.16)	**3344.3 (5.29)**	3311.3 (5.59)	4.40 (0.16)	4.38 (0.16)
		0.8	**69.2 (0.18)**	69.1 (0.19)	**3335.1 (5.10)**	3315.1 (5.24)	4.49 (0.19)	4.46 (0.20)
		0.7	**68.9 (0.17)**	68.8 (0.18)	**3323.3 (5.23)**	3313.7 (5.29)	4.60 (0.18)	4.57 (0.18)
		0.5	68.8 (0.19)	**68.8 (0.18)**	3305.3 (5.18)	3305.1 (5.16)	4.52 (0.19)	4.51 (0.19)

Note: For each setting, for both the mean pAUC and mean S, the highest values are given in bold font if a t-test has determined that there is a difference in the means of the proposed and traditional methods at 5% significance. If the t-test was not significant, bold font is not used. Statistically significant improvements that we deemed practically significant (at least 5% improvement) are indicated by underlining.

that there are approximately $12\,355 - 8967 = 3388$ DE genes with negative effect sizes and $10455 - 8967 = 1488$ DE genes with positive effect sizes. This results in a final estimate of $3388/(3388 + 1488) \approx 69.4\%$ of DE genes with negative effect sizes. Therefore, although asymmetry in the test statistics is not visibly obvious, it is likely that there is a high degree of asymmetry in the distribution of effect sizes of DE genes. Based on the results of the simulation studies in Section 3.3, it is appropriate to use the proposed method for computing q-values to identify DE genes. This can be done by using (3) to estimate the FDR for each gene $k = 1, \ldots, 12\,355$ with a negative test statistic as

$$q^{(1)}_{(k)} = \min\left\{\frac{p^{(1)}_{(r)}(8\,966.875)}{r} : r = k, \ldots, 12\,355\right\}, \qquad (6)$$

and we can estimate FDR for each gene $k = 1, \ldots, 10\,455$ with a positive test statistic as

$$q^{(2)}_{(k)} = \min\left\{\frac{p^{(2)}_{(r)}(8\,966.875)}{r} : r = k, \ldots, 10\,455\right\}. \qquad (7)$$

Figure 4 plots the log ratio of the q-values versus the test statistics from the proposed and traditional method for the gene expression experiment in thale cress seedlings. Negative log ratios correspond to cases where the q-value from the proposed method was less than the q-value from the traditional method. Positive log ratios correspond to larger q-values for the proposed method. From this scatterplot, we can clearly see that the proposed method produces smaller q-values than the traditional method for genes with negative test statistics and larger q-values for genes with positive test statistics, with a clear separation when the test statistic is 0. This is not surprising based on Figure 2 and the discussion in Section 2.3.

Using the traditional q-value method, 490 genes were declared to be DE when controlling FDR at 5%. The proposed method declared 617 genes DE, a 25.9% increase over the traditional method. There were 417 genes that both methods declared to be DE. All 200 of the genes that were only declared DE by the proposed method had negative test statistics. Similarly, all of the 73 genes that were only declared DE by the traditional method had positive test statistics. This is what we would expect based on the higher number of negative test statistics than positive test statistics in this experiment.

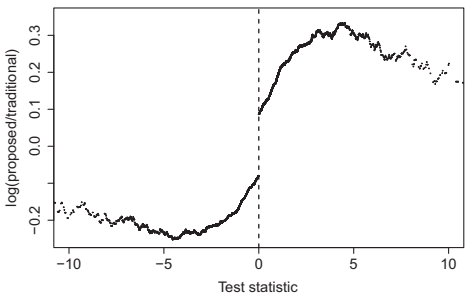

Fig. 4. Scatterplot of the log ratio of the q-values versus the test statistics for the experiment in thale cress seedlings. Negative log ratios correspond to cases where the q-value produced by the proposed method is less than the q-value from traditional method. Positive log ratios correspond to larger q-values for the proposed method

4.2 Maize leaves

The analysis of data from a microarray experiment described in Covshoff *et al.* (2008) is briefly summarized here. In this study, expressions from $m = 7377$ genes in the mesophyll cells of maize leaves were compared between two genotypes, wild-type and mutant, using $n = 6$ two-color microarray slides. Mutant plants lacked the PSII activity of wild-type plants, and researchers were interested in identifying genes that have different mean expressions due to this lack of activity. Data from this experiment are available at the Gene Expression Omnibus under accession number GSE9698.

The empirical distribution of test statistic values, shown in the histogram in Figure 5, is clearly asymmetric. There are $m_1 = 4141$ genes with negative test statistics and $m_2 = 3236$ genes corresponding to positive test statistics. Estimating the number of EE genes results in $\hat{m}_0 = 2907.11$. Using similar methods as in Section 4, we estimate that ~2687 genes are DE with negative effect sizes, and 1782 genes are DE with positive effect sizes, resulting in an estimated 60.1% of DE genes having negative effect sizes.

The proposed method declares fewer genes to be DE than the traditional method (2260 versus 2446). Similar to the analysis of

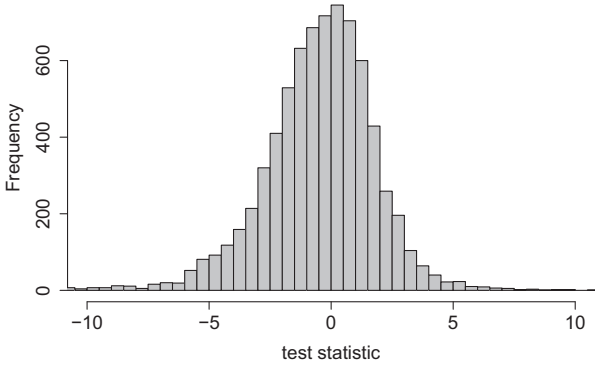

Fig. 5. A histogram of *t*-statistics from an experiment described in Covshoff *et al.* (2008) in which gene expressions from wild-type cells were compared with those in mutant cells in maize leaves

the experiment described in Jang *et al.* (2014), the 186 genes declared DE by the proposed method but not the traditional method all had negative test statistics. The 220 genes that were only declared DE by the traditional method all had positive test statistics.

5 EXTENSION OF THE TWO SAMPLE CASE

This section illustrates how the idea of partitioning *P*-values can be extended to gene expression experiments with more than two treatments. Consider the microarray experiment described in Lattanzi *et al.* (2007) in which researchers wished to identify genes associated with the presence of trimethyltin (TMT) in rats. Three rats were assigned to each of the three treatment groups: control, 1 μmol/L concentration of TMT and 5 μmol/L concentration of TMT. For each of the 15 866 genes, analysis of variance was used to determine whether significant differences among the treatment means existed. This dataset (available at the Gene Expression Omnibus under accession number GSE5073) will be re-analyzed in this section.

Using the moderated *F*-test (the generalization of Smyth's (2004) moderated *t*-test), a *P*-value for each gene was calculated to test the null hypothesis $H_j : \mu_{j1} = \mu_{j2} = \mu_{j3}$, where μ_{jt} is the population treatment mean expression for gene j in treatment t. For this experiment, $t = 1$ corresponds to the control treatment, $t = 2$ corresponds to the treatment of 1 μmol/L concentration of TMT and $t = 3$ corresponds to the treatment of 5 μmol/L concentration of TMT. Using these *P*-values, the methods described in Section 2 were used to estimate m_0 as $\hat{m}_0 = 10\,440.72$, corresponding to $\hat{\pi}_0 = 0.658$.

Because an *F*-test was performed on the data for each gene, all of the test statistics in the analysis of the TMT experiment are positive. Thus, the sign of the test statistic can not be used to partition the *P*-values, and a different partitioning rule needs to be implemented. The treatments in this experiment can easily be ranked based on the concentration of TMT. We might also expect that many genes that are DE will exhibit sample mean expressions that are monotone in the ranked treatments. In other words, we might expect to observe more DE genes with one of the following relationships:

(1) $\bar{y}_{j1} < \bar{y}_{j2} < \bar{y}_{j3}$ (monotone increasing) or

(2) $\bar{y}_{j1} > \bar{y}_{j2} > \bar{y}_{j3}$ (monotone decreasing),

where $\bar{y}_{jt}$ is the sample treatment mean expression for gene j in treatment t. Based on this reasoning, we will consider partitioning the *P*-values by two different rules.

The first partitioning rule will include two subsets of *P*-values: (i) $\{p_k^{(1)} : k = 1, \ldots, m_1\}$, the set of *P*-values corresponding to genes which have sample treatment means that are not monotone in the ranked treatments and (ii) $\{p_k^{(2)} : k = 1, \ldots, m_2\}$, the set of *P*-values corresponding to genes that do have sample treatment means that are monotone (increasing or decreasing) in the ranked treatments. In the TMT experiment, $m_1 = 9603$ and $m_2 = 6263$ for the first partitioning rule.

The second partitioning rule will include three subsets of *P*-values: (i) $\{p_k^{(1)} : k = 1, \ldots, m_1\}$, the set of *P*-values corresponding to genes which have sample treatment means that are not monotone in the ranked treatments, (ii) $\{p_k^{(2)} : k = 1, \ldots, m_2\}$, the set of *P*-values corresponding to genes that have sample treatment means that are monotone increasing in the ranked treatments and (iii) $\{p_k^{(3)} : k = 1, \ldots, m_3\}$, the set of *P*-values corresponding to genes that have sample treatment means that are monotone decreasing in the ranked treatments. In the TMT experiment, $m_1 = 9603$, $m_2 = 2383$ and $m_3 = 3880$.

If a gene is EE, the ranking of sample means is random. For a three sample experiment, there are 3! = 6 possible rankings of the sample means. Thus, an EE gene will have sample means that are monotone increasing or monotone decreasing in the ranked treatments with probability $2/6 \approx 0.333$, and this gene will have sample means that are not monotone in the ranked treatments with probability $4/6 \approx 0.667$. Therefore, for the first partitioning rule, $\hat{m}_0^{(1)} = 10\,440.72(4/6) = 6960.48$ and $\hat{m}_0^{(2)} = 10\,710(2/6) = 3480.24$, calculated using formula (3). Using formula (3), but replacing $\hat{m}_0/2$ with $\hat{m}_0^{(i)}$ (i = 1,2), *q*-values for each gene were calculated.

For the second partitioning rule, $\hat{m}_0^{(1)}$ remains the same, but $\hat{m}_0^{(2)} = \hat{m}_0^{(3)} = 10440.72(1/6) = 1740.12$. Again using altered formula (3) and an analogous formula for $i = 1, 2$ and 3, *q*-values for each gene were calculated.

The histograms of *P*-values corresponding to partitioning rule two are provided in Figure 6. The empirical distributions of *P*-values corresponding to genes with sample means that are monotone in the ranked treatments are noticeably stochastically smaller than the empirical distributions of *P*-values corresponding to genes where monotonicity is not observed. This indicates a higher proportion of genes are DE among the genes that exhibit monotonicity in the ranked treatment means than the genes that do not exhibit monotonicity.

The number of genes declared to be DE using the proposed method for each partitioning rule as well as the traditional *q*-value method is summarized in Figure 7. Although there is a large overlap in the genes declared to be DE among all three methods, partitioning the *P*-values results in a higher number of genes declared to be DE than if no partitioning is performed. Partitioning rule two resulted in a slightly higher number of significant genes than partitioning rule one (442 versus 434). Additionally, more genes are declared DE when partitioning is applied. Having a larger set of significant genes could aid researchers in more easily identifying the

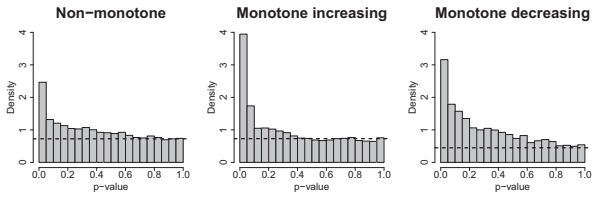

Fig. 6. A histogram of *P*-values for the TNT rat data corresponding to *P*-values partitioned using rule two. The estimated proportion of EE genes for each subset is plotted as the dashed horizontal line in each plot

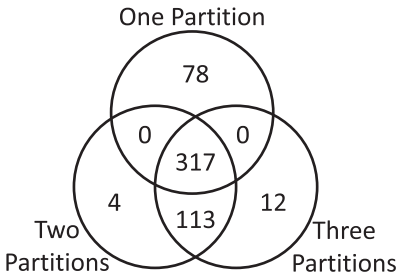

Fig. 7. Venn diagram of genes declared to be DE for the different partitioning rules

underlying biological processes impacted by the presence of TMT in rats.

6 DISCUSSION

The proposed method for estimating FDR by first estimating π_0 using all *P*-values and then analyzing two subsets of *P*-values separately based on the sign of the test statistics has advantages over the traditional *q*-value method when effect sizes are asymmetric. The proposed method was never outperformed by the traditional method in terms of mean pAUC or mean *S* in the two simulation studies, and ranked genes better with respect to differential expression in 17.5% of the simulation settings while adequately controlling FDR. The proposed method performed especially well for simulation settings with the smallest sample size ($n = 4$). Compared with the traditional method, the proposed method also declared more truly DE genes DE, on average, in 37.1% of simulation settings, and was never outperformed by the traditional method, even in simulation settings with symmetric effect sizes. Similar to the results for pAUC, the proposed method performed exceptionally well in simulation settings with $n = 4$. Therefore, we suggest that the proposed method be used to analyze datasets with small sample sizes ($n \leq 10$) when the distribution of test statistics is asymmetric.

Another advantage to the proposed method is that it can be generalized to analyze many gene expression experiments with more than two treatments, especially if the treatments in the experiment can be ordered by factors such as time or dose. Although the criteria used to partition the *P*-values differs depending to the treatment structure, the analysis is very similar once the partitioning has been performed.

Funding: This research was partially supported by the USDA NRICGP Microbial Functional Genomics program (grant no. 2008-3560418805) and the National Science Foundation (NSF) Plant Genome Research Program (grant no. IOS-1127017).

Conflict of interest: none declared.

REFERENCES

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.

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

Covshoff,S. *et al.* (2008) Deregulation of maize C4 photosynthetic development in a mesophyll cell-defective mutant. *Plant. Physiol.*, **146**, 1469–1481.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array repository. *Nucleic Acids Res.*, **30**, 207–210.

Hannenhalli,S. *et al.* (2006) Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation*, **114**, 1269–1276.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Jang,Y.H. *et al.* (2014) A homolog of splicing factor SF1 is essential for development and is involved in alternative splicing of pre-mRNA in *Arabidopsis thaliana*. *Plant J.*, **78**, 591–603.

Lattanzi,W. *et al.* (2007) Hypoxia-like transcriptional activation in TMT-induced degeneration: microarray analysis on PC12 cells. *J. Neurochem.*, **100**, 1688–1702.

Liang,K. and Nettleton,D. (2012) Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. B*, **74**, 163–182.

Metzker,M.L. (2010) Sequencing technologies–the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Nettleton,D. *et al.* (2006) Estimating the number of true null hypotheses from a histogram of *P* values. *J. Agr. Biol. Envir. St.*, **11**, 337–356.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol.*, **3**. Article 3.

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. *P. 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.

Sun,W. and Cai,T.T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.*, **102**, 901–912.