Estimating the Number of Genes That Are Differentially Expressed in Both of Two Independent Experiments

Megan ORR, Peng LIU, and Dan NETTLETON

A common procedure for estimating the number of genes that are differentially expressed (DE) in two experiments involves two steps. In the first step, data from the two experiments are separately analyzed to produce a list of genes declared to be DE in each experiment. Usually, each list is produced using a method that attempts to control the false discovery rate (FDR) in each experiment at some desired level α. In the second step, the number of genes common to both lists is used as an estimate of the number of genes DE in both experiments. A problem with this approach is that the resulting estimates can vary greatly with α, and the value of α that produces the best estimate for any given pair of experiments is difficult to predict. We propose a method that uses the p-values from both experiments simultaneously to produce one estimate—which does not depend on FDR level α—for the number of genes that are DE in both experiments. We use two simulation studies (one involving independent, normally distributed data and one involving microarray data) to compare the performances of our proposed method, the commonly used method, and another method proposed in literature to test for consistency of replicate experiments. The results of the simulation studies demonstrate the advantages of our approach. We conclude the article by estimating the number of genes that are DE in both of two experiments involving gene expressions in maize leaves.

Key Words: False discovery rate; λ-estimator; Microarray data analysis; Multiple testing.

1. INTRODUCTION

Comparing the results of two independent experiments is of common interest to many researchers. This becomes a difficult problem when large data sets are analyzed and hundreds or thousands of hypothesis tests are performed for each experiment. This problem

Megan Orr (✉) is recent Ph.D. graduate, Department of Statistics, Iowa State University, Ames, IA 50011, USA and Assistant Professor, Department of Statistics, North Dakota State University, Fargo, ND 58108, USA (E-mail: *megan.orr@ndsu.edu*). Peng Liu is Assistant Professor (E-mail: *pliu@iastate.edu*) and Dan Nettleton is Laurence H. Baker Endowed Chair and Professor (E-mail: *dnett@iastate.edu*), Department of Statistics, Iowa State University, Ames, IA 50011, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 17, Number 4, Pages 583–600
DOI: 10.1007/s13253-012-0108-8

is often encountered in the analysis of gene expression experiments where mRNA expression levels are compared between two or more treatment groups for each of thousands of genes. In many cases, one of the primary interests is to estimate how many genes exhibit differential expression (i.e., a difference in mean expression levels) in both experiments. For example, Covshoff et al. (2008) performed experiments for each of two cell types in maize leaves: bundle sheath (BS) and mesophyll (M). In each experiment, expression levels were measured in wild-type and mutant cells for a set of genes. The mutant cells lacked the PSII activity of the wild-type cells, and researchers were interested in observing the effects of this lack of activity on gene expressions. To understand whether PSII activity plays a similar role in both the BS and M cell types, the researchers were specifically interested in determining whether the impact of the mutation on gene expression was similar in both BS and M cell types. Thus, the researchers attempted to estimate the number of genes differentially expressed in both BS and M cell types, which is a numerical quantity essential for understanding the extent to which the genes differentially expressed in the BS cell type overlap with the genes differentially expressed in the M cell type.

Many methods have been proposed to estimate the number of equivalently expressed (EE) genes, and thus the number of differentially expressed (DE) genes, when performing a hypothesis test for each gene in one gene expression data set. The estimation problem becomes more complicated when there are two independent data sets to be compared and the number of DE genes in both experiments, a quantity we call m_{11}, is to be estimated.

In practice, m_{11} is typically estimated by creating a list of genes that are declared to be DE (i.e., the null hypotheses for these genes are rejected) separately for each experiment and then counting the number of genes that appear on both lists. We will call this method of m_{11} estimation the "intersection method". The intersection method is commonly used throughout the scientific literature to compare results from multiple gene expression experiments. We focus on the paper of Covshoff et al. (2008) as one representative example out of many similar examples. Other examples include Ianculescu et al. (2012), Voineagu et al. (2011), Buchanan-Wollaston et al. (2005), Wang et al. (2004), and Akopyants et al. (2004). Covshoff et al. (2008) used the intersection method to estimate m_{11} at two different levels of false discovery rate (FDR) control ($\alpha = 0.01$ and $\alpha = 0.05$). We performed the intersection method at an additional level of FDR control ($\alpha = 0.10$), as well as in conjunction with an α selection algorithm, and found that the four m_{11} estimates ranged from 168 to 2107.

As observed in Covshoff et al. (2008), an obvious flaw of the intersection method is that the estimate of m_{11} highly depends on what level α is chosen to control FDR in each experiment. The estimate of m_{11} using the intersection method is a non-decreasing function of α, and there is no way to know which value of α will produce the most accurate estimate. A low value for α can lead to underestimation of m_{11}, especially when the effect sizes are relatively small for many of the truly differentially expressed genes. This is due to the high number of Type II errors that can occur when controlling FDR. On the other hand, a large value for α may result in many Type I errors in each experiment, which can lead to overestimation of m_{11} by the intersection method.

The purpose of this paper is to introduce an improved method for estimating m_{11} which does not depend on FDR control. We first analyze the p-values from each data set sepa-

rately to estimate $m_0^{(1)}$ and $m_0^{(2)}$, the number of EE genes in the first experiment and second experiment, respectively, using the methods described in Liang and Nettleton (2012). Then we pair the p-values from both experiments by gene to estimate m_{00}, the number of genes that are EE in both experiments. We propose a bivariate extension of the λ-estimator (Storey 2002) in order to estimate m_{00}. Finally, from these three estimates we obtain an estimate for m_{11}.

The rest of the paper is organized as follows. In Section 2, we describe the proposed method for estimating m_{11}. In Section 3, two simulation studies are described and the results of these studies are presented. The results of the proposed method are compared to the intersection method when controlling FDR at 5 %, 10 %, and levels chosen based on the data, respectively. In addition, we also use a method proposed by Lai et al. (2007)—originally proposed with a different goal in mind—to estimate m_{11} and compare its results to those of the other methods. The results of the simulation studies show that the proposed method, when compared to both the intersection method and Lai's method, results in lower root mean squared error (RMSE) when estimating m_{11} for most simulation settings. In Section 4, the proposed method is used to analyze the data from the experiments described in Covshoff et al. (2008), and these results are compared to those of the intersection method and Lai's method. Finally, we provide some discussion in Section 5.

Please note that R code for the estimation of m_{11} is available on request.

2. METHODS

This section describes the proposed method for estimating m_{11}, the number of genes that are DE in both of two gene expression experiments. In Section 2.1, we illustrate how m_{11} can be estimated as a linear combination of m and the null counts $m_0^{(1)}$, $m_0^{(2)}$, and m_{00}. In Section 2.2, we review a method for estimating the number of EE genes in a single experiment, which we subsequently use to estimate $m_0^{(1)}$ and $m_0^{(2)}$. Section 2.3 describes our proposed method for estimating m_{00}. Finally, Section 2.4 discusses properties of $\hat{m}_{11}$, the estimator of m_{11}.

2.1. OVERVIEW OF m_{11} ESTIMATION

Consider the problem of testing m pairs of null hypotheses $(H_{11}, H_{21}), (H_{12}, H_{22}), \ldots, (H_{1m}, H_{2m})$, where H_{ij} is the null hypothesis for experiment i and gene j ($i = 1, 2$; $j = 1, \ldots, m$). Each hypothesis H_{ij} is either true, meaning that gene j in experiment i is EE, or false, meaning that this gene is DE. Table 1 is a contingency table which cross classifies the expression status (EE or DE) for each of the m genes by experiment and presents their frequencies.

The counts with two digits in the subscript represent interesting frequencies when looking at the experiments simultaneously. For example, m_{00} is the number of genes that are EE in both Experiment 1 and Experiment 2, while m_{01} is the number of genes that are EE in Experiment 1 but DE in Experiment 2. The marginal totals in Table 1 represent the number of EE genes, $m_0^{(i)}$, and the number of DE genes, $m_1^{(i)}$, for experiment i ($i = 1, 2$).

Table 1. Contingency table of frequencies based on cross classification of the expression status (EE or DE) for each of the m genes by experiment.

		Experiment 2		
		Gene EE	Gene DE	Total
Experiment 1	Gene EE	m_{00}	m_{01}	$m_0^{(1)}$
	Gene DE	m_{10}	m_{11}	$m_1^{(1)}$
	Total	$m_0^{(2)}$	$m_1^{(2)}$	m

We are ultimately interested in estimating m_{11}, the number of genes that are DE in both experiments. From Table 1, it is easy to see that

$$m_{11} = m - m_0^{(1)} - m_0^{(2)} + m_{00}. \tag{2.1}$$

The process of estimating m_{11} begins by estimating the marginal counts $m_0^{(i)}$ for each experiment. We will call these estimates $\hat{m}_0^{(i)}$ for $i = 1, 2$. Next, estimation of m_{00}, resulting in $\hat{m}_{00}$, is performed. Finally, we estimate the number of genes that are DE in both experiments by replacing the unknown counts in (2.1) by their estimates to get

$$\hat{m}_{11} = m - \hat{m}_0^{(1)} - \hat{m}_0^{(2)} + \hat{m}_{00}. \tag{2.2}$$

2.2. REVIEW OF THE λ-ESTIMATOR AND HISTOGRAM-BASED METHOD

Now consider the problem of simultaneously testing null hypotheses $H_{i1}, \ldots, H_{im}$ for experiment i based on corresponding p-values $p_{i1}, \ldots, p_{im}$. For $j = 1, \ldots, m$, we assume that $p_{ij} \sim \text{Uniform}(0, 1)$ when H_{ij} is true and that p_{ij} has a distribution stochastically smaller than uniform when H_{ij} is false. These are standard assumptions which imply that an unbiased size α test can be obtained for each j by rejecting H_{ij} if and only if $p_{ij} \le \alpha$.

For any fixed $\lambda_i \in [0, 1)$ in experiment i, Storey (2002) proposed

$$\hat{m}_0^{(i)}(\lambda_i) = \frac{\sum_{j=1}^{m} \mathbf{1}\{p_{ij} > \lambda_i\}}{1 - \lambda_i} \tag{2.3}$$

as an estimator of $m_0^{(i)}$, the number of true null hypotheses among $H_{i1}, \ldots, H_{im}$.

It follows from our uniformity assumption that

$$\begin{aligned} E\big(\hat{m}_0^{(i)}(\lambda_i)\big) &= \frac{1}{1-\lambda_i} \sum_{j=1}^{m} E\big(\mathbf{1}\{p_{ij} > \lambda_i\}\big) \\ &= \frac{1}{1-\lambda_i} \sum_{j=1}^{m} \Pr(p_{ij} > \lambda_i) \\ &= \frac{1}{1-\lambda_i} \left(\sum_{\{j: H_{ij} \text{ true}\}} \Pr(p_{ij} > \lambda_i) + \sum_{\{j: H_{ij} \text{ false}\}} \Pr(p_{ij} > \lambda_i) \right) \end{aligned}$$

$$= \frac{1}{1-\lambda_i}\left(\sum_{\{j:H_{ij}\text{ true}\}}(1-\lambda_i)\right) + \frac{1}{1-\lambda_i}\left(\sum_{\{j:H_{ij}\text{ false}\}}\Pr(p_{ij}>\lambda_i)\right)$$

$$= m_0^{(i)} + \frac{1}{1-\lambda_i}\sum_{\{j:H_{ij}\text{ false}\}}\Pr(p_{ij}>\lambda_i). \tag{2.4}$$

Clearly, $\hat{m}_0^{(i)}(\lambda_i)$ is a conservatively biased estimator of $m_0^{(i)}$ for all $\lambda_i \in [0, 1)$, where the bias is the second term in Equation (2.4). The degree of bias depends on the probabilities that p-values from tests with false null hypotheses are larger than λ_i. This directly relates to the power of the test for DE genes. The more powerful the tests for DE genes, the smaller the $\Pr(p_{ij} > \lambda_i)$ for DE genes and the smaller the bias.

The value of λ_i plays an important role in the estimation of $m_0^{(i)}$ as well as m_{00}, and ultimately m_{11}, which is described in Section 2.3. Storey (2002) investigated how the value of λ_i affects the bias and variance of $\hat{m}_0^{(i)}(\lambda_i)$. He concluded that as λ_i increases, the bias of $\hat{m}_0^{(i)}(\lambda_i)$ tends to decrease while the variance of $\hat{m}_0^{(i)}(\lambda_i)$ tends to increase. Thus, it is important to determine a λ_i with an appropriate trade-off between bias and variance.

There are many methods proposed for determining an appropriate value of λ_i (see Storey 2002; Storey and Tibshirani 2003; Mosig et al. 2001; Nettleton et al. 2006; or Liang and Nettleton 2012, for example), and it is important to note that the method we propose in Section 2.3 for estimating m_{11} can use any of these methods. However, to illustrate our method for m_{11} estimation and to evaluate its performance relative to the intersection method and Lai's method, we use the λ_i selection strategy recently proposed by Liang and Nettleton (2012). This is a "histogram-based" method that is closely related to a procedure originally proposed by Mosig et al. (2001) and studied in detail by Nettleton et al. (2006). Liang and Nettleton's (2012) version of this procedure performs well relative to competing approaches in simulation studies and has desirable theoretical properties, as demonstrated by Liang and Nettleton (2012).

The idea behind the histogram-based method is to select a value of λ_i from a set of candidates so that a histogram of p-values less than λ_i is approximately decreasing while a histogram of p-values greater than or equal to λ_i is approximately uniform. It makes sense to select such a λ_i because choosing a smaller value would lead to higher bias while choosing a larger value would lead to higher variance without an appreciable reduction in bias. A limitation of the histogram-based method is that it targets m times the height of the p-value density at 1 as its estimand. As discussed by Genovese and Wasserman (2004) and Langaas, Ferkingstad, and Lindqvist (2005), this estimand is an upper bound on $m_0^{(i)}$ that is chosen because $m_0^{(i)}$ is not identifiable without additional parametric assumptions. Because the histogram-based estimator targets an identifiable upper bound on $m_0^{(i)}$, it—like most other competing estimators—tends to be conservatively biased as an estimator of $m_0^{(i)}$, especially when the average power is low due to small sample sizes, large measurement error, and high variation in biological replicates, all of which are common in gene expression experiments.

The algorithm of Liang and Nettleton (2012) can be formally described as follows:

1. Partition the interval $[0, 1]$ into B bins of equal width. Let $c_b = (\frac{b-1}{B}, \frac{b}{B}]$ for $b = 1, 2, \ldots, B$.
2. Denote the number of p-values in the interval c_b as n_b for $b = 1, 2, \ldots, B$.
3. For each $b = 1, 2, \ldots, B$, calculate
$$\bar{n}_b = \frac{\sum_{k=b}^{B} n_b}{B - b + 1}. \tag{2.5}$$
4. Let $b^* = \min\{\min\{b : n_b \le \bar{n}_b\}, B - 1\}$.
5. Select $\lambda_i = \frac{b^*}{B}$.

Throughout this paper, we set $B = 20$ when using this algorithm, in accordance with the recommendations of Nettleton et al. (2006) and Liang and Nettleton (2012).

The values of λ_i that result from applying this algorithm to the p-values from the experiments of Covshoff et al. (2008) are depicted in the marginal histograms of Figure 1. For the bundle sheath data, 890 p-values exceeded 0.70, the selected value of λ_1. Using (2.3), this yields $890/(1 - 0.70) = 2967$ as an estimate of $m_0^{(1)}$ for the bundle sheath data. Similarly, the estimate of $m_0^{(2)}$ for the mesophyll data is $1162/(1 - 0.55) = 2582$.

2.3. ESTIMATING THE NUMBER OF GENES THAT ARE EQUIVALENTLY EXPRESSED IN BOTH EXPERIMENTS

Let $(p_{11}, p_{21}), (p_{12}, p_{22}), \ldots, (p_{1m}, p_{2m})$ represent m pairs of p-values from testing the m pairs of null hypotheses mentioned in Section 2.1, $(H_{11}, H_{21}), (H_{12}, H_{22}), \ldots, (H_{1m}, H_{2m})$, where p_{ij} is the p-value for testing H_{ij}, the null hypothesis in experiment i for gene j ($i = 1, 2$; $j = 1, \ldots, m$).

We begin by estimating the number of true null hypotheses (EE genes) for each experiment using the histogram-based method as described in Section 2.2. Let λ_i denote the value selected by the algorithm in Section 2.2 for experiment i, and let $\hat{m}_0^{(i)} = \hat{m}_0^{(i)}(\lambda_i)$ denote the estimated number of true null hypotheses in experiment i ($i = 1, 2$).

Next, notice that if a pair of tests corresponding to gene j both have true null hypotheses, then the pair (p_{1j}, p_{2j}) is assumed to follow a product uniform distribution given by

$$\Pr\big((p_{1j}, p_{2j}) \in [a, b] \times [c, d]\big) = (b - a)(d - c) \tag{2.6}$$

for all $a < b, c < d$, and $a, b, c, d \in [0, 1]$. This follows from the assumption that p-values from tests with true null hypotheses are uniform and from the assumption that Experiments 1 and 2 are independent.

The next step is to estimate m_{00}, the number of genes with true null hypotheses in both experiments (i.e., the number of genes that are EE in both experiments). Define

$$n_{00} = \sum_{i=1}^{m} \mathbf{1}\big\{(p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1]\big\}. \tag{2.7}$$

From (2.6), we see that the probability that a p-value pair falls in $[\lambda_1, 1] \times [\lambda_2, 1]$ is $(1 - \lambda_1)(1 - \lambda_2)$ if its corresponding gene is EE in both experiments. Thus, a conservative

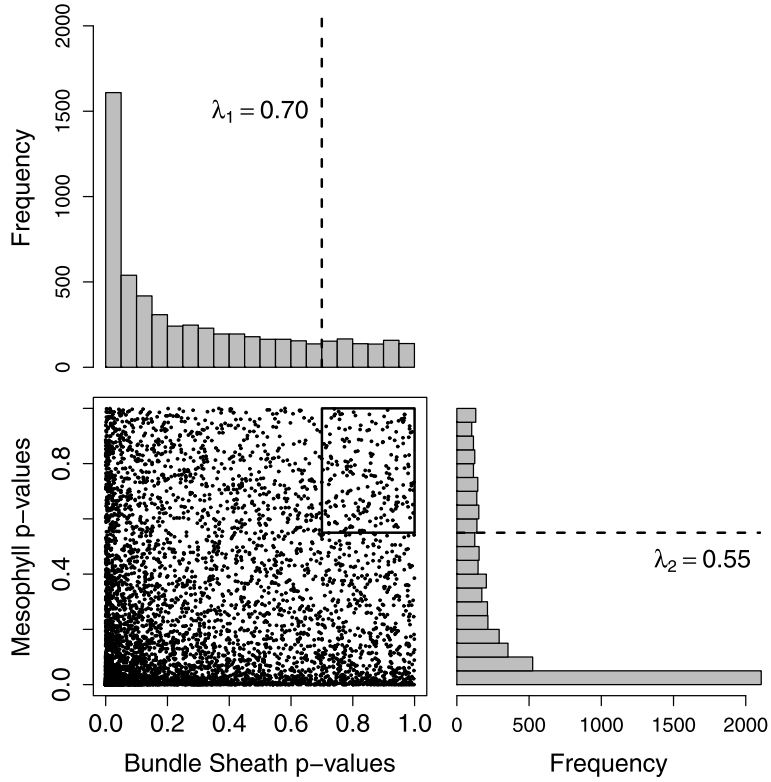

Figure 1. A scattterplot of p-values from the bundle sheath experiment versus p-values from the mesophyll experiment with histograms of marginal p-values. The values of λ_i selected via the algorithm in Section 2.2 are represented by dashed lines. The region $[\lambda_1, 1] \times [\lambda_2, 1]$ based on the values of λ_i selected for the individual sets of p-values is shown as a box in the upper right corner of the scatterplot.

estimate of m_{00} is

$$\hat{m}_{00} = \frac{n_{00}}{(1-\lambda_1)(1-\lambda_2)}. \tag{2.8}$$

This is a bivariate analog of Storey's λ-estimator.

2.4. PROPERTIES OF $\hat{m}_{11}$

Once the estimates $\hat{m}_0^{(1)}$, $\hat{m}_0^{(2)}$, and $\hat{m}_{00}$ have been obtained, we use (2.2) to estimate m_{11}.

Performing steps analogous to those used in (2.4), we derive the expected value of $\hat{m}_{11}$, for fixed $\lambda_1, \lambda_2 \in [0, 1)$, as

$$E(\hat{m}_{11}) = m_{11} - \frac{1}{(1-\lambda_1)(1-\lambda_2)} \sum_{\mathcal{F}} A_j, \tag{2.9}$$

where $\mathcal{F} = \{j : H_{1j}$ false and H_{2j} false$\}$ and

$$\begin{aligned} A_j = {} & (1-\lambda_1)\Pr(p_{2j} > \lambda_2) + (1-\lambda_2)\Pr(p_{1j} > \lambda_1) \\ & - \Pr(p_{1j} > \lambda_1)\Pr(p_{2j} > \lambda_2). \end{aligned} \tag{2.10}$$

Because the distribution of a p-value from a DE gene is stochastically smaller than uniform, $\Pr(p_{ij} > \lambda_i) < (1 - \lambda_i)$ for $i = 1, 2$ when gene j is DE in both experiments, making A_j positive for genes that are DE in both experiments. Furthermore, as the power of each test increases for all genes that are DE in both experiments, the values of A_j decrease, decreasing the bias of $\hat{m}_{11}$. As the power of each test decreases, both $\Pr(p_{1i} > \lambda_1)$ and $\Pr(p_{2i} > \lambda_2)$ increase toward $(1 - \lambda_1)$ and $(1 - \lambda_2)$, respectively, decreasing $E(\hat{m}_{11})$ toward zero. Thus, $\hat{m}_{11}$ is biased downward, and $E(\hat{m}_{11})$ is bounded between zero and m_{11}.

Ideally, we would also like to assess the uncertainty of the estimator $\hat{m}_{11}$ by providing a standard error or a confidence interval for m_{11}. Unfortunately, it is extremely difficult to evaluate the uncertainty through either analytical or numerical methods due to both the unknown complex correlation structure of the data as well as the potential for low average power of tests corresponding to DE genes. Our method for estimating m_{11} is a more complicated extension of the problem of estimating m_0. Many methods have been proposed for estimating m_0 including Storey (2002), Storey and Tibshirani (2003), Storey, Taylor, and Siegmund (2004), Langaas, Ferkingstad, and Lindqvist (2005), Nettleton et al. (2006), and Liang and Nettleton (2012). Langaas, Ferkingstad, and Lindqvist (2005) is the only one of these papers that considers the problem of estimating the variance of $\hat{m}_0$. Unfortunately, Langaas, Ferkingstad, and Lindqvist's (2005) expression for the variance of $\hat{m}_0$ applies only to the unrealistic case of independent p-values. To our knowledge, no one has developed a method for accurately evaluating the uncertainty associated with estimators of m_0 in realistic settings. Assessing the uncertainty associated with our estimator of m_{11} is an even more challenging open question.

3. SIMULATION STUDIES

In order to evaluate the performance of our proposed method for estimating m_{11}, simulation studies were performed. For each simulated data set, two null hypotheses

$$H_{1j} : \mu_{1j1} = \mu_{1j2} \quad \text{and} \quad H_{2j} : \mu_{2j1} = \mu_{2j2} \tag{3.1}$$

were tested (each against the two-sided alternative) for each gene j of $m = 10000$ genes, where μ_{ijk} represents the population mean expression for treatment k of gene j in experiment i $(i = 1, 2;\ j = 1, \ldots, m;\ k = 1, 2)$. These tests were performed by calculating a test statistic and its corresponding p-value, p_{ij}, for each gene j in experiment i using the moderated t-test approach proposed by Smyth (2004). This method of testing for DE genes was specifically designed for microarray experiments and involves borrowing information across all genes in order to better estimate the error variance of each gene and to obtain a t-distributed test statistic that performs better than the regular t-test with respect to ranking genes for differential expression. These p-values calculated using the moderated t-test were used to estimate m_{11} using Lai's method, the intersection method, and the proposed method.

The intersection method was performed by first using the histogram-based method to calculate $\hat{m}_0^{(i)}$ for each experiment i. Then $\hat{m}_0^{(i)}$ was used to convert the p-values to q-values (Storey 2002) separately for each experiment. A list of genes declared to be DE

at a specific level α of FDR control was created for each experiment. Finally, m_{11} was estimated as the number of genes common to both lists. In addition to the proposed and intersection methods, a method described in Lai et al. (2007) was used to estimate m_{11}. The original purpose of Lai's method was to determine if the results of two large replicate experiments are consistent enough for their data to be combined for a more powerful analysis. Although this purpose is not to estimate m_{11} directly, the most general mixture model proposed in Lai et al. (2007) can be used to estimate m_{11} by summing the estimated mixing proportions for components in the mixture model that correspond to genes that are DE in both experiments and multiplying the resulting proportion by m. Lai's method was performed using R code available at *http://home.gwu.edu/~ylai/research/Concordance*.

Our proposed method was implemented as described in Section 2. It should be noted that the estimators $\hat{m}_0^{(1)}$, $\hat{m}_0^{(2)}$, $\hat{m}_{00}$, and $\hat{m}_{11}$ are all random variables that can result in values outside the range of their estimands. We estimated m_{11} under many possible constraint combinations. For example, in one combination, we estimated m_{11} by constraining all estimates to values within the range of their estimands. Thus, any estimates of $m_0^{(1)}$, $m_0^{(2)}$, or m_{00} above m were replaced by m, and any negative estimates of m_{11} were replaced by 0. After considering many different constraint methods and examining the performance of the resulting m_{11} estimators, we concluded that constraining only the final estimate of m_{11} (and not constraining $\hat{m}_0^{(1)}$, $\hat{m}_0^{(2)}$, or $\hat{m}_{00}$ in intermediate calculations) produced slightly better results. Thus, this approach was used to estimate m_{11} in the following simulation studies.

Two simulation studies were performed. The first set of simulations used independent, normally distributed data. This allowed us to evaluate our method under ideal conditions that are consistent with the assumptions used to derive the moderated t-test (Smyth 2004). The second set of simulations used real microarray data in order to evaluate how our method performed when data have a distribution and correlation structure that we cannot model precisely but will encounter in practice. For each simulation setting, 100 data sets were randomly generated.

Aside from the distribution of the data, we also varied the sample size for each treatment (n), the magnitude of effect sizes for differentially expressed genes (controlled by a parameter μ_δ defined in Section 3.1), and the quantities defined in Table 1. Sample sizes of $n = 4$, 10, and 20 were chosen for the simulation studies because typical gene expression experiments have small sample sizes, usually due to the high cost of experimentation. For example, approximately 90 % of all data sets available on the Gene Expression Omnibus (Edgar, Domrachev, and Lash 2002) have a total sample size of $N = 40$ or less, with 40 % of available data sets having at most $N = 8$. These sample sizes correspond to $n = 20$ and $n = 4$, respectively, in our simulation studies. Also, due to the high variation in both measurement error and biological replicates in most gene expression experiments, the average power for detecting DE genes, and thus the average relative effect size, is generally small. Because of this, we chose small mean relative effect sizes of $\mu_\delta = 1$ and 2.

We compare the results of the proposed method to those of Lai's and the intersection methods when controlling FDR at various levels. Three different estimates are obtained using the intersection method. The first two estimates control FDR at an α-level determined *a priori* for each experiment. Motivated by a reviewer's comment, the final estimate for m_{11}

is obtained by choosing a separate α *a posteriori* for each experiment. For this intersection method estimator, α is chosen so that the number of genes declared to be DE is equal to the estimated number of DE genes for the given experiment.

All methods used to estimate m_{11} are evaluated both visually and using the root mean squared error (RMSE).

3.1. SIMULATIONS USING INDEPENDENT, NORMALLY DISTRIBUTED DATA

In the first simulation study, data consisting of $m = 10000$ genes were simulated from independent normal distributions with gene-specific variances. The gene-specific variances $\{\sigma_{ij}^2 : i = 1, 2;\ j = 1, \ldots, m\}$ were drawn independently from an inverse gamma distribution. The parameters of this distribution were estimated from a microarray data set using the methods of Smyth (2004). This data set consists of gene expressions from patients suffering from different types cardiomyopathy and can be obtained from the Gene Expression Omnibus (GEO, *http://www.ncbi.nlm.nih.gov/geo/*) under accession number GSE5406. For a description of the experiment performed to obtain these data, see Hannenhalli et al. (2006). Conditional on $\{\sigma_{ij}^2 : i = 1, 2;\ j = 1, \ldots, m\}$, treatment mean values $\{\mu_{ijk} : i = 1, 2;\ j = 1, \ldots, m;\ k = 1, 2\}$ were determined as follows. For an EE gene, $\mu_{ij1} = \mu_{ij2} = 0$. For a DE gene, μ_{ij1} was set to zero, and μ_{ij2} was drawn from a $\mathrm{N}(\mu_\delta \sigma_{ij}, \sigma_{ij}^2)$ distribution. Given the value for μ_{ijk} and σ_{ij}^2, n observation for experiment i, gene j, and treatment k were independently drawn from the $\mathrm{N}(\mu_{ijk}, \sigma_{ij}^2)$ distribution.

3.2. SIMULATIONS USING REAL MICROARRAY DATA

For the second simulation study, a microarray data set was used that consists of gene expressions from the bone marrow or peripheral blood of subjects with cytogenetically normal acute myeloid leukemia (CN-AML) and is described in Metzeler et al. (2008). This data set is also available on GEO under the accession number GSE12417. Only data from subjects in the training cohort were used. In this experiment, the tissues were prepared and then hybridized with individual Affymetrix HU133A arrays and the raw expression values were then transformed and normalized. There are 22284 genes on the Affymetrix HU133A array, but $m = 10000$ genes were randomly selected to be included in the simulations. The $N = 163$ total subjects were randomly split into two subsets representing two independent experiments with total sample sizes $N_1 = 82$ and $N_2 = 81$.

In order to simulate a data set for the first experiment using the microarray data, we started with the original data from the 10000 randomly selected genes and the $N_1 = 82$ randomly selected subjects. For the j^{th} gene, the standard deviation, s_{1j}, was calculated from the expression values of all $N_1 = 82$ subjects. Then, $2n$ subjects were randomly chosen as the subjects for the experiment. This group was further split, randomly, into two groups of size n, each representing a different treatment group. Differentially expressed genes were then created by adding a randomly generated treatment effect to the data from the second treatment group. Treatment effects, μ_{1j2}, were generated as in Section 3.1, except the calculated s_{1j} values were used in place of the simulated σ_{1j} values. Data were simulated for the second experiment in a similar manner.

3.3. RESULTS

Tables 2 and 3 give the mean estimates of m_{11} along with their RMSEs for each simulation setting and estimation method. Also, for settings with $m_{11} = 0$, the number of times

Table 2. Mean estimates of m_{11} and the RMSE (rounded to the nearest integer) for each simulation setting and estimation method when data are independent and normally distributed. Columns with "FDR" in their name refer to the intersection method with level of control given after the "$\leq$" sign. Here, $\hat{\alpha}_i$ indicates the α level was estimated separately for each experiment using the method described in Section 3. Within each row, the results for the method with the lowest RMSE are presented in bold font. For simulation settings with $m_{11} = 0$, the number of times m_{11} was correctly estimated as 0 is given in square brackets below the mean and RMSE, and the results with the highest number is given in bold font.

n	μ_δ	m_{00}	m_{11}	Proposed	Lai	FDR$\leq$ 0.05	FDR$\leq$ 0.10	FDR $\leq \hat{\alpha}_i$
4	1	9000	0	17 (29)	11 (12)	**0 (0)**	**0 (0)**	5 (6)
				[45/100]	[0/100]	**[100/100]**	**[100/100]**	[2/100]
		9000	500	**116 (387)**	50 (451)	0 (500)	0 (500)	37 (463)
		7000	1000	**291 (715)**	208 (793)	2 (998)	10 (990)	172 (828)
		5000	2000	**606 (1400)**	563 (1437)	17 (1983)	67 (1933)	469 (1532)
		3000	3000	943 (2065)	**1042 (1959)**	64 (2936)	201 (2799)	861 (2140)
	2	9000	0	16 (33)	11 (12)	**0 (0)**	0 (0)	9 (10)
				[48/100]	[0/100]	**[97/100]**	[70/100]	[0/100]
		9000	500	**303 (201)**	226 (274)	16 (484)	42 (458)	147 (354)
		7000	1000	**646 (362)**	548 (543)	121 (879)	231 (769)	464 (537)
		5000	2000	**1349 (660)**	1208 (793)	446 (1554)	747 (1253)	1122 (880)
		3000	3000	**2066 (944)**	1950 (1051)	981 (2019)	1538 (1463)	1918 (1085)
10	1	9000	0	10 (21)	7 (8)	**0 (1)**	1 (1)	6 (7)
				[60/100]	[0/100]	**[80/100]**	[26/100]	[0/100]
		9000	500	**217 (285)**	175 (326)	44 (456)	63 (437)	115 (385)
		7000	1000	**469 (535)**	433 (568)	146 (854)	208 (792)	343 (657)
		5000	2000	**1001 (1004)**	979 (1022)	397 (1603)	566 (1434)	825 (1176)
		3000	3000	1552 (1456)	**1636 (1366)**	755 (2245)	1081 (1919)	1420 (1581)
	2	9000	0	9 (17)	2 (2)	**1 (2)**	3 (3)	7 (8)
				[50/100]	[0/100]	[34/100]	[6/100]	[2/100]
		9000	500	**392 (112)**	365 (136)	230 (271)	267 (234)	286 (214)
		7000	1000	**792 (213)**	739 (262)	576 (424)	668 (332)	679 (322)
		5000	2000	**1638 (369)**	1519 (482)	1315 (685)	1533 (468)	1485 (517)
		3000	3000	2468 (537)	2318 (683)	2189 (812)	**2591 (411)**	2372 (630)
20	1	9000	0	9 (19)	5 (6)	**0 (1)**	2 (2)	6 (7)
				[55/100]	[0/100]	[49/100]	[15/100]	[1/100]
		9000	500	**290 (214)**	271 (229)	135 (365)	159 (341)	190 (311)
		7000	1000	601 (404)	**600 (401)**	344 (656)	413 (587)	481 (520)
		5000	2000	1223 (782)	**1278 (723)**	805 (1196)	983 (1017)	1075 (926)
		3000	3000	1864 (1142)	**2054 (947)**	1376 (1625)	1719 (1281)	1770 (1233)
	2	9000	0	10 (19)	**0 (1)**	2 (2)	4 (5)	4 (5)
				[55/100]	[0/100]	[14/100]	[1/100]	[7/100]
		9000	500	**430 (74)**	411 (90)	353 (148)	373 (128)	363 (137)
		7000	1000	**871 (136)**	818 (182)	776 (225)	839 (162)	792 (209)
		5000	2000	1755 (250)	1660 (341)	1654 (347)	**1815 (187)**	1655 (346)
		3000	3000	2645 (361)	2511 (489)	2635 (366)	**2961 (50)**	2580 (423)

Table 3. Mean estimates of m_{11} and the RMSE (rounded to the nearest integer) for each simulation setting and estimation method for simulations using microarray data. Columns with "FDR" in their name refer to the intersection method with level of control given after the "$\leq$" sign. Here, $\hat{\alpha}_i$ indicates the α level was estimated separately for each experiment using the method described in Section 3. Within each row, the results for the method with the lowest RMSE are presented in bold font. For simulation settings with $m_{11} = 0$, the number of times m_{11} was correctly estimated as 0 is given in square brackets below the mean and RMSE, and the results with the highest number is given in bold font.

n	μ_δ	m_{00}	m_{11}	Proposed	Lai	FDR ≤ 0.05	FDR ≤ 0.10	FDR $\leq \hat{\alpha}_i$
4	1	9000	0	68 (220)	1144 (2515)	**0 (0)**	0 (0)	30 (106)
				[39/100]	[0/100]	**[100/100]**	[98/100]	[83/100]
		9000	500	**181 (432)**	1381 (2816)	0 (500)	1 (493)	75 (462)
		7000	1000	**381 (726)**	957 (2066)	2 (998)	10 (990)	201 (840)
		5000	2000	**708 (1350)**	1173 (1727)	16 (1984)	63 (1938)	518 (1522)
		3000	3000	**1072 (1980)**	1762 (2120)	66(2935)	207(2797)	953 (2092)
	2	9000	0	155 (366)	657 (1561)	**0 (3)**	1 (7)	85 (229)
				[49/100]	[0/100]	**[94/100]**	[77/100]	[59/100]
		9000	500	**371 (343)**	670 (1490)	14 (486)	39 (462)	145 (425)
		7000	1000	**723 (501)**	785 (1015)	109 (892)	224 (781)	514 (589)
		5000	2000	**1506 (688)**	1500 (944)	412 (1595)	737 (1282)	1239 (900)
		3000	3000	**2258 (887)**	2153 (966)	962 (2049)	1560 (1470)	2069 (1026)
10	1	9000	0	50 (152)	325 (1404)	**0 (1)**	1 (4)	41 (144)
				[68/100]	[0/80]	**[87/100]**	[68/100]	[69/100]
		9000	500	**228 (365)**	412 (864)	43 (457)	63 (437)	112 (442)
		7000	1000	**510 (572)**	788 (774)	146 (854)	215 (787)	429 (641)
		5000	2000	**1097 (985)**	1362 (1053)	406 (1596)	593 (1415)	953 (1123)
		3000	3000	1628 (1431)	**2106 (1327)**	795 (2210)	1157 (1860)	1548 (1514)
	2	9000	0	116 (313)	320 (1440)	**2 (4)**	7 (19)	84 (246)
				[60/100]	[0/100]	[45/100]	[26/100]	[49/100]
		9000	500	415 (264)	559 (1170)	228 (273)	**265 (236)**	206 (381)
		7000	1000	**871 (319)**	884 (387)	576 (426)	675 (333)	728 (421)
		5000	2000	1739 (485)	1672 (549)	1325 (683)	**1561 (476)**	15865 (578)
		3000	3000	2681 (640)	2529 (665)	2247 (779)	**2710 (460)**	2575 (680)
20	1	9000	0	138 (429)	177 (551)	**1 (3)**	6 (19)	97 (308)
				[69/100]	[0/100]	[63/100]	[40/100]	[59/100]
		9000	500	277 (268)	**309 (257)**	134 (367)	158 (343)	110 (416)
		7000	1000	**632 (483)**	740 (545)	345 (656)	424 (581)	529 (570)
		5000	2000	1320 (773)	**1478 (696)**	817 (1185)	1010 (998)	1195 (879)
		3000	3000	1954 (1129)	**2241 (1018)**	1386 (1619)	1750 (1269)	1841 (1239)
	2	9000	0	40 (146)	35 (111)	**2 (4)**	5 (11)	34 (110)
				[70/100]	[0/100]	[41/100]	[16/100]	[59/100]
		9000	500	521 (302)	**465 (116)**	356 (145)	379 (123)	295 (346)
		7000	1000	990 (375)	940 (304)	785 (221)	**862 (179)**	845 (405)
		5000	2000	1857 (344)	1762 (381)	1666 (346)	**1846 (224)**	1740 (425)
		3000	3000	2817 (510)	2631 (558)	2680 (382)	**3058 (372)**	2761 (544)

m_{11} was correctly estimated as 0 is given in square brackets under the mean and RMSE. In each table for each setting, the estimate with the smallest RMSE is presented in bold font. Although the values of m_{01} and m_{10} as shown in Table 1 are not provided, $m_{01} = m_{10}$ for each simulation setting, and their common value can be obtained by taking half the value

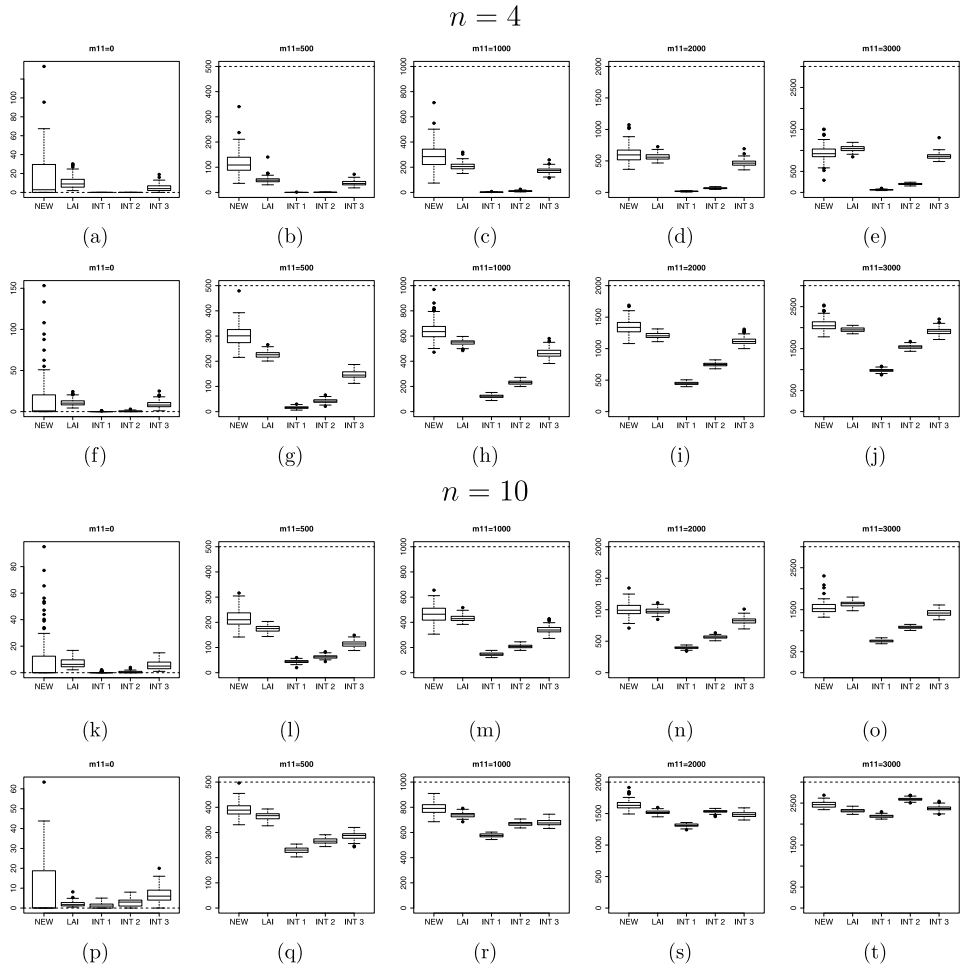

Figure 2. Boxplots of the m_{11} estimates for simulations involving independent, normally distributed data. The true value of m_{11} is given above each plot and is also represented by the horizontal dashed line in each plot. Plots (a)–(e) and (k)–(o) show the results for simulations with $\mu_\delta = 1$, and plots (f)–(j) and (p)–(t) show the results for simulations with $\mu_\delta = 2$. The first and second boxplots in each plot represent the estimates for the proposed and Lai's methods, respectively. The third and fourth boxplots represent the estimates for the intersection method when FDR was controlled at 0.05 and 0.10, respectively. The fifth boxplot represents the estimates for the intersection method when α was chosen *a posteriori*.

of $m - m_{00} - m_{11}$. Figures 2 and 3 illustrate the estimation results with boxplots for simulation settings with $n = 4$ and $n = 10$. Boxplots for settings with $n = 20$ look similar to those with $n = 10$ presented in Figures 2 and 3. For both sets of simulations, the proposed method outperforms the other methods for most simulation settings.

When $m_{11} = 0$, our proposed method either estimates m_{11} correctly with $\hat{m}_{11} = 0$, or overestimates it, due to the lower bound of zero placed on the estimate. As a consequence, the bounded estimator is positively biased. In this case, our method is outperformed in the simulations using both normal data and microarray data (see Tables 2 and 3) when we evaluate the estimation methods using RMSE. In general, the intersection method, when

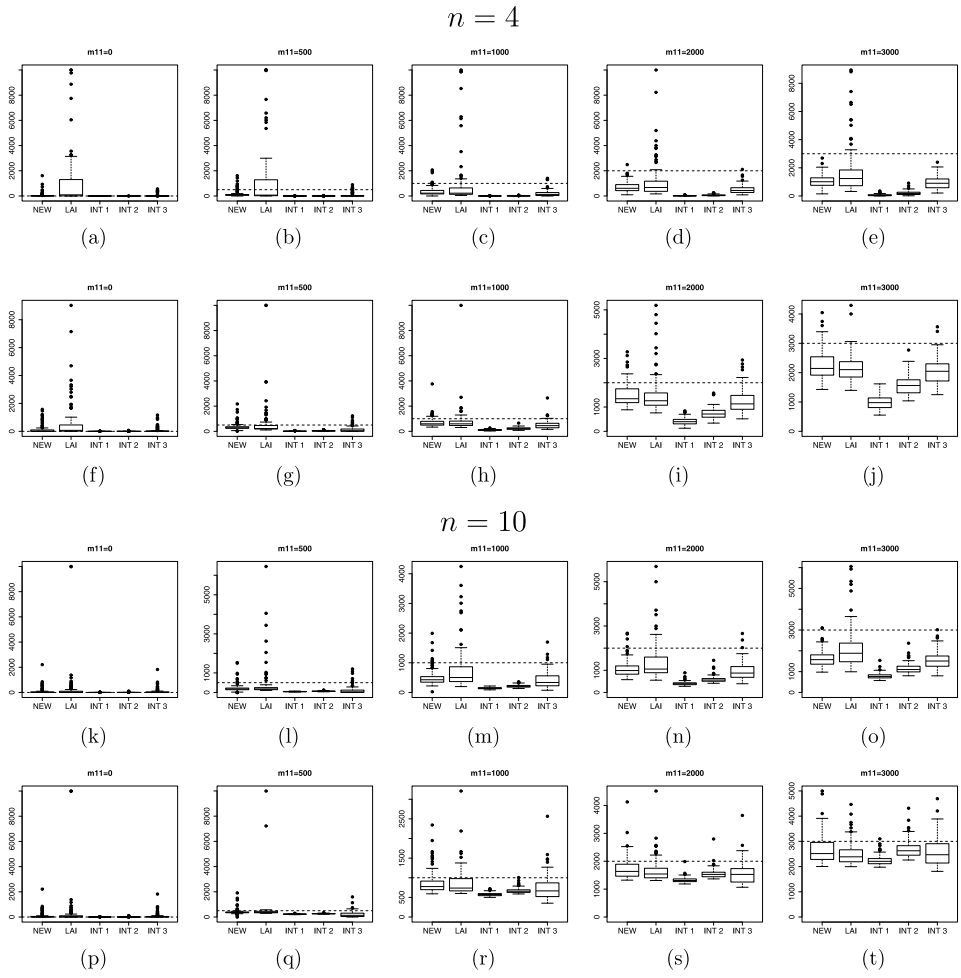

Figure 3. Boxplots of the m_{11} estimates for simulations involving microarray data. The true value of m_{11} is given above each plot and is also represented by the horizontal dashed line in each plot. Plots (a)–(e) and (k)–(o) show the results for simulations with $\mu_\delta = 1$, and plots (f)–(j) and (p)–(t) show the results for simulations with $\mu_\delta = 2$. The first and second boxplots in each plot represent the estimates for the proposed and Lai's methods, respectively. The third and fourth boxplots represent the estimates for the intersection method when FDR was controlled at 0.05 and 0.10, respectively. The fifth boxplot represents the estimates for the intersection method when α was chosen *a posteriori*.

FDR is controlled at 5 %, outperforms the other estimation methods for settings with $m_{11} = 0$. This is not surprising as this method uses the most stringent cutoff for declaring genes to be DE. For each setting with $m_{11} = 0$, we also evaluated the methods by counting the number of data sets (out of 100) whose analysis resulted in $\hat{m}_{11} = 0$. In half of these 12 simulation settings, the proposed method correctly estimates m_{11} for more data sets than any other method, including the intersection method with the lowest level of FDR control. This suggests that our method remains effective when $m_{11} = 0$.

In practice, researchers are more interested in estimating m_{11} when there are some genes DE in both experiments under study, i.e., when $m_{11} > 0$. For simulations using indepen-

dent, normally distributed data, in the 24 settings with $m_{11} > 0$ (see Table 2), the proposed method outperforms all other methods in 16 of the 24 simulation settings, including 13 of the 16 settings with $n = 4$ or $n = 10$. Lai's method estimates m_{11} best in five of the 24 settings, mostly for higher values of m_{11}. The intersection method (when FDR is controlled at either 5 % or 10 %) performs best in 3 of the 24 settings. See Figure 2 for a visual representation of the m_{11} estimation results.

For simulations using microarray data (see Table 3 and Figure 3), the proposed method outperforms the other methods in 13 of the 24 simulation settings with $m_{11} > 0$, and generally performs best when the average power of the test is low. When $n = 4$, the proposed method is best in all eight settings. Lai's method has the lowest RMSE in five of the 24 settings, four of which when $n = 20$. When FDR is controlled at 10 %, the intersection method performs best in the remaining six settings, all with $n = 10$ or $n = 20$. The intersection method never performs best when α is chosen *a posteriori* for each experiment, although it tends to outperform the intersection method when FDR is controlled at 5 % and 10 % for simulation settings with lower power. Also, in many settings, even though the intersection method outperforms the proposed method at one level of FDR control, it does not perform as well as the intersection method for the other levels of FDR control.

Unsurprisingly, m_{11} is underestimated when using our method in settings with lower power (except when $m_{11} = 0$ as explained earlier). This underestimation is unavoidable, as illustrated in Equations (2.9) and (2.10). However, the other estimation methods usually result in higher degrees of underestimation, except in the microarray simulations when Lai's method results in highly skewed estimates and high RMSEs, as illustrated in Figure 3.

4. REAL DATA ANALYSIS

In this section, we analyze the data described in Covshoff et al. (2008) using the proposed method and compare the results to those of Lai's method and the intersection method. The data come from two independent experiments in which the same two-color microarray platform was used to measure gene expressions in maize leaves. One experiment was performed on mesophyll (M) cells in the maize leaves and the other on bundle sheath (BS) cells. Each experiment had two treatments, wild type and mutant, with $n = 6$ two-color slides. Maize leaves in the mutant treatment had cells that lacked the PSII activity of the maize leaves in the wild-type treatment. As discussed in Section 1 researchers were specifically interested in estimating m_{11}, the number of genes that are DE expressed, due to lack of PSII activity, in both the BS and M cells.

Although the same platform was used for both experiments, only 7377 and 8463 genes were detected above background level for the M and BS experiments, respectively. For each experiment, the difference in mean expressions between the wild-type and mutant cells was tested against a null value of zero for each detected gene using the moderated t-test (Smyth 2004). The p-value from each test was converted to a q-value, and the intersection method was performed for each of three FDR levels. When controlling FDR at 1 %, 5 %, and 10 %, m_{11} was estimated to be 168, 573, and 1012, respectively. When the α selection algorithm described in Section 3 was performed, q-value cutoffs of 0.188 and 0.239 were used for

the M and BS experiments, respectively. This resulted in an m_{11} estimate of 2107 genes that are DE in both experiments. The most important observation from these results is that the estimates of m_{11} vary drastically for different levels of FDR control, and we have no way of knowing which estimate is the most appropriate.

One complication of this real data example is that different numbers of genes were analyzed in the BS and M experiments due to different number of genes detected above background in the two experiments. Because a gene that is not detected above background level in one experiment cannot be determined to be DE in this experiment and thus cannot be determined to be DE in both experiments, we only used the 5670 genes that were detected in both experiments for the purpose of estimating m_{11} using the proposed method.

Figure 1 shows histograms of the p-values for the BS and M experiments individually as well as the scatterplot of the p-values paired by gene. These p-values are available in Supplemental Tables S5 and S6 in Covshoff et al. (2008) in the columns labeled "pvalue_limma(FDR)". As discussed in Section 2.2, the dashed line in each histogram represents the value of λ_i $(i = 1, 2)$ selected using the algorithm in Section 2.2. For the BS experiment, $\lambda_1 = 0.70$, and there are 890 p-values larger than λ_1. For the M experiment, $\lambda_2 = 0.55$, and there are 1162 p-values larger than λ_2. Thus $\hat{m}_0^{(1)} = 2967$ and $\hat{m}_0^{(2)} = 2582$ as calculated in Section 2.2.

We now use the method described in Section 2.3 to estimate m_{00}. We have selected the value of λ_i for each experiment, so we can now count the p-value pairs in the region $[\lambda_1, 1] \times [\lambda_2, 1]$, as illustrated by the box in the upper right corner of the scatterplot in Figure 1. There are 216 p-values that fall within this region, so we estimate m_{00} as

$$\hat{m}_{00} = \frac{216}{(1 - 0.70)(1 - 0.55)} = 1600. \tag{4.1}$$

Using (2.2), we can now estimate m_{11}, or the number of genes that are differentially expressed between wild type and mutant cells in both the BS and M experiments, as

$$\begin{aligned} \hat{m}_{11} &= m - \hat{m}_0^{(1)} - \hat{m}_0^{(2)} + \hat{m}_{00} \\ &= 5670 - 2967 - 2582 + 1600 \\ &= 1721. \end{aligned} \tag{4.2}$$

This is approximately 30 % of the genes that were detected above background level in both experiments. Thus, lack of PSII activity is associated with changes in expression of many of the same genes in both bundle sheath and mesophyll cell types.

We also used Lai's method to obtain an estimate of m_{11} as 5223. This is an unrealistically high value relative to the $m = 5670$ genes that were analyzed. As discussed in Section 3.3 and observed in Figure 3, the use of Lai's method to estimate m_{11} can result in severe overestimation for microarray data, which may be the case for the data from the maize experiments.

5. DISCUSSION

As illustrated by the results of the simulation studies, the effectiveness of the intersection method can depend highly on the α-level chosen. Even when α is chosen *a posteriori*, it rarely outperforms the proposed method, although it tends to improve the performance of the intersection method when average power is low. The proposed method has a clear advantage over the intersection method because it does not depend on choosing a level of FDR control α, and we recommend using the proposed method over the intersection method for this reason. The proposed method also outperforms Lai's method in most simulation settings. Furthermore, we developed the proposed method for the purpose of analyzing gene expression experiments, the large majority of which have low sample size and low average power for detecting differential expression. Thus, we are mostly interested in how our method performs in such settings, and the simulation studies suggest that our method performs best in most of these settings. When sample sizes grow large, the differences among the methods diminish. However, we still recommend using the proposed method in experiments with larger n because it performs similarly when compared to the other methods and does not require choosing α as the intersection method does. Overall, our proposed method is effective, especially for experiments with low average power, but still performs adequately in experiments with larger sample sizes.

Additionally, the proposed method does not require that the two experiments use the same platform because it only requires two sets of p-values that correspond to a common group of genes. However, our method will likely produce the most biologically meaningful results if the same platform and experimental conditions are used for both of the two independent experiments being compared, and also if the same conditions or closely related conditions are compared. If experiments using different platforms to measure gene expression are compared, the platform could act as a confounding factor, and thus the results might be more of a comparison of platforms and less of biological effects. Therefore, when different experimental platforms are used, researchers should exercise caution when interpreting the results. This applies not only to our proposed method but also to all other methods used to estimate m_{11}.

ACKNOWLEDGEMENTS

This work was partially supported by the USDA NRICGP Microbial Functional Genomics program (grant no. 2008-3560418805).

[Received December 2011. Accepted July 2012. Published Online August 2012.]

REFERENCES

Akopyants, N. S., Matlib, R. S., Bukanova, E. N., Smeds, M. R., Brownstein, B. H., Stormo, G. D., and Beverley, S. M. (2004), "Expression Profiling Using Random Genomic DNA Microarrays Identifies Differentially Expressed Genes Associated With Three Major Developmental Stages of the Protozoan Parasite Leishmania Major," *Molecular and Biochemical Parasitology*, 136, 71–86.

Buchanan-Wollaston, V., Page, T., Harrison, E., Breeze, E., Ok Lim, P., Gil Nam, H., Lin, J.-F., Wu, S.-H., Swidzinski, J., Ishizzaki, K., and Leaver, C. J. (2005), "Comparative Transcriptome Analysis Reveals Significant Differences in Gene Expression and Signalling Pathways Between Developmental and Dark/Starvation-Induced Senescence in Arabidopsis," *The Plant Journal*, 42, 567–585.

Covshoff, S., Majeran, W., Liu, P., Kolkman, J. M., van Wilk, K. J., and Brutnell, T. P. (2008), "Deregulation of Maize C4 Photosynthetic Development in a Mesophyll Cell-Defective Mutant," *Plant Physiology*, 146, 1469–1481.

Edgar, R., Domrachev, M., and Lash, A. E. (2002), "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Repository," *Nucleic Acids Research*, 30, 207–210.

Genovese, C. R., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *Annals of Statistics*, 32, 1035–1061.

Hannenhalli, S., Putt, M. E., Gilmore, J. M., Wang, J., Parmacek, M. S., Epstein, J. A., Morrisey, E. E., Marguilies, K. B., and Cappola, T. P. (2006), "Transcriptional Genomics Associates FOX Transcription Factors With Human Heart Failure," *Circulation*, 114, 1269–1276.

Ianculescu, I., Wu, D.-Y., Siegmund, K. D., and Stallcup, M. R. (2012), "Selective Roles for cAMP Response Element-Binding Protein Binding Protein and p300 Protein as Coregulators for Androgen-Regulated Gene Expression in Advanced Prostate Cancer Cells," *The Journal of Biological Chemistry*, 287, 4000–4013.

Lai, Y., Adam, B., Podolsky, R., and She, J.-X. (2007), "A Mixture Model Approach to the Tests of Concordance and Discordance Between Two Large-Scale Experiments With Two-Sample Groups," *Bioinformatics*, 23, 1243–1250.

Langaas, M., Ferkingstad, E., and Lindqvist, B. H. (2005), "Estimating the Proportion of True Null Hypotheses, With Application to DNA Microarray Data," *Journal of the Royal Statistical Society, Series B*, 67, 555–572.

Liang, K., and Nettleton, D. (2012), "Adaptive and Dynamic Adaptive Procedures for False Discovery Rate Control and Estimation," *Journal of the Royal Statistical Society, Series B*, 74, 163–182.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., Sauerland, M.-C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S. P., Maharry, K., Paxchka, P., Larson, R. A., Berdel, W. E., Buchner, T., Wormann, B., Mansmann, U., Hiddemann, W., Bohlander, S. K., and Buske, C. (2008), "An 86-Probe-Set Gene-Expression Signature Predicts Survival in Cytogenetically Normal Acute Myeloid Leukemia," *Blood*, 112, 4193–4201.

Mosig, M., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann, A. (2001), "A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Percentage in Israeli-Holstein Cattle, by Means of Selective Milk DNA Pooling in a Daughter Design, Using an Adjusted False Discovery Rate Criterion," *Genetics*, 157, 1683–1698.

Nettleton, D., Hwang, J., Caldo, R., and Wise, R. (2006), "Estimating the Number of True Null Hypotheses From a Histogram of p Values," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 337–356.

Smyth, G. K. (2004), "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*, 3, 3.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

Storey, J. D., Taylor, J., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society, Series B*, 66, 187–205.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 9440–9445.

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., Mill, J., Cantor, R. M., Blencowe, B. J., and Geshwind, D. H. (2011), "Transcriptomic Analysis of Autistic Brain Reveals Convergent Molecular Pathology," *Nature*, 474, 380–386.

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., and Abruzzo, L. V. (2004), "Differences in Gene Expression Between B-Cell Chronic Lymphocytic Leukemia and Normal B Cells: A Meta-Analysis of Three Microarray Studies," *Bioinformatics*, 20, 3166–3178.