Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 3 2012 *Article* 12

Borrowing Information Across Genes and Experiments for Improved Error Variance Estimation in Microarray Data Analysis

Tieming Ji, *Iowa State University*
Peng Liu, *Iowa State University*
Dan Nettleton, *Iowa State University*

Recommended Citation:
Ji, Tieming; Liu, Peng; and Nettleton, Dan (2012) "Borrowing Information Across Genes and Experiments for Improved Error Variance Estimation in Microarray Data Analysis," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 3, Article 12.
DOI: 10.1515/1544-6115.1806

Borrowing Information Across Genes and Experiments for Improved Error Variance Estimation in Microarray Data Analysis

Tieming Ji, Peng Liu, and Dan Nettleton

Abstract

Statistical inference for microarray experiments usually involves the estimation of error variance for each gene. Because the sample size available for each gene is often low, the usual unbiased estimator of the error variance can be unreliable. Shrinkage methods, including empirical Bayes approaches that borrow information across genes to produce more stable estimates, have been developed in recent years. Because the same microarray platform is often used for at least several experiments to study similar biological systems, there is an opportunity to improve variance estimation further by borrowing information not only across genes but also across experiments. We propose a lognormal model for error variances that involves random gene effects and random experiment effects. Based on the model, we develop an empirical Bayes estimator of the error variance for each combination of gene and experiment and call this estimator BAGE because information is Borrowed Across Genes and Experiments. A permutation strategy is used to make inference about the differential expression status of each gene. Simulation studies with data generated from different probability models and real microarray data show that our method outperforms existing approaches.

KEYWORDS: BAGE variance estimator, empirical Bayes, false discovery rate, permutation test, shrinkage estimator

Author Notes: This material is based upon work supported by the National Science Foundation under Grant Numbers 0714978, CCF-0811804 and 0701736. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. We would like to thank Takahashi. H., etc. at Laboratory of Plant Molecular Genetics at the University of Tokyo for providing the microarray data.

1 Introduction

Microarray technology is used to measure expression levels of thousands of genes simultaneously. Based on these expression levels, the ordinary t- or F-test can be applied to identify differentially expressed genes (genes whose expression distribution differs across treatments). However, the power of such tests is limited because there are usually only a few observations for each gene.

To improve the power of the ordinary t- or F-test, several groups have developed approaches for borrowing information across genes. Examples include Efron's t-test (Efron et al., 2001), the regularized t-test (Baldi and Long, 2001), the B-statistic (Lönnstedt and Speed, 2002), the tests of Wright and Simon (2003), the moderated t-test (Smyth, 2004), the F_S test (Cui et al., 2005), the tests of Tong and Wang (2007), Lo and Gottardo (2007), and Hwang and Liu (2010). These methods modify the t- or F-test by shrinking the REML (Restricted Maximum Likelihood) estimates of error variances using information from all genes in one experiment and show improved power over the ordinary t- or F-test.

Typically, many experiments are conducted using the same microarray platforms to study the same biological system. Thus, there is a potential to further borrow information across both genes and experiments. One example, out of many examples, is a study on aerenchyma formation in maize roots (Nakazono et al., 2009). This study contained four independent experiments conducted with the same microarray platform (GEO Platform GPL4521). Each experiment compared two different treatments using two-color microarrays, and treatments were different across experiments. Details about these experiments are presented in the Appendix A. Because there were only a few slides in each experiment, the REML estimator of error variance for each combination of gene and experiment is highly variable. However, because the four experiments used the same platform, the observations for the same set of genes were repeatedly collected in each experiment. Though these experiments did not compare the same set of treatments, the error variance for a given gene may be similar across treatments in multiple experiments. Thus, there is a potential benefit to utilize observations from all genes across all experiments to improve the estimation of the error variance for each combination of gene and experiment.

In order to borrow information across genes and across experiments, we model the log of each error variance as the sum of a random gene effect, a random experiment effect, and a random error. We use data from all genes and all experiments to estimate the distributions of the gene, experiment, and error effects. We use these estimates to obtain an improved variance estimator

that we refer to as the BAGE estimator because information is Borrowed Across Genes and Experiments. The amount and direction of information borrowing for a given gene depends on the relative values of the variances of gene effects, experiment effects, and error effects for the log variances in our model.

Replacing the REML estimates with the BAGE estimates in the ordinary F-test statistic results in a new statistic that we call F_{BAGE}. We develop a permutation test based on the F_{BAGE} statistic to detect differentially expressed genes. Simulations based on both hypothetical distributions and real data show that F_{BAGE} provided better gene rankings compared with the ordinary F-test and the moderated t-test (Smyth, 2004). When using the procedure for FDR control proposed by Storey (2002) in conjunction with the q-value computation method of Storey and Tibshirani (2003a), our method also identified more true positives while controlling the false discovery rate (FDR) at nominal levels.

The remainder of the paper is organized as follows. Section 2 introduces the data structure, the standard linear model with fixed error variances, the REML estimator of error variances, and the ordinary F-test for detecting differentially expressed genes. In the context of the framework established in Section 2, Section 3 introduces the proposed lognormal model for the variances as well as its resulting BAGE estimator. Section 4 introduces a permutation test for detecting differentially expressed genes based on BAGE estimates. Section 5 presents results comparing our BAGE estimator with the REML estimator and the estimator used in the popular LIMMA R package (Smyth, 2004) through simulations based on hypothetical data distributions. Section 6 presents results from simulations based on real experiments. Our proposed method is applied to the example experiments of Nakazono et al. (2009) in Section 7. Section 8 provides a summary and future work of our study. The R code used in the paper is available upon request.

2 Standard Linear Modeling of Expression Data and Tests of Interest

In this study, we consider I microarray experiments using the same microarray platform such that each experiment contains the same set of J genes. These experiments might have different designs with n_i observations for each gene in experiment i (i=1,2,...,I). We model $\mathbf{y}_{ij}$, the vector of normalized log signal intensities (or log ratio of signal intensities for the case of two-color

microarrays) with length n_i for gene j (j=1,...,J) in experiment i (i=1,...,I) as

$$\mathbf{y}_{ij} = \mathbf{X}_i\boldsymbol{\beta}_{ij} + \boldsymbol{\epsilon}_{ij}, \tag{1}$$

where $\mathbf{X}_i$ is the design matrix for the ith experiment, $\boldsymbol{\beta}_{ij}$ is a vector of fixed parameters, and $\boldsymbol{\epsilon}_{ij}$ is a vector of independent and identically distributed errors with mean 0 and variance σ^2_{ij}. We are interested in knowing if gene j is differentially expressed across two or more treatments, which can be equivalently represented as testing $H_{0,ij} : \mathbf{C}_i^T\boldsymbol{\beta}_{ij} = \mathbf{0}$ versus $H_{a,ij} : \mathbf{C}_i^T\boldsymbol{\beta}_{ij} \neq \mathbf{0}$ for an appropriately chosen matrix $\mathbf{C}_i$. The ordinary F-test statistic for testing $H_{0,ij}$ versus $H_{a,ij}$ is

$$F_{ij} = \frac{\hat{\boldsymbol{\beta}}_{ij}^T\mathbf{C}_i[\mathbf{C}_i^T(\mathbf{X}_i^T\mathbf{X}_i)^-\mathbf{C}_i]^{-1}\mathbf{C}_i^T\hat{\boldsymbol{\beta}}_{ij}/r_i}{S^2_{ij}}, \tag{2}$$

where S^2_{ij} is the REML estimator of σ^2_{ij}, $\mathbf{C}_i^T\hat{\boldsymbol{\beta}}_{ij}$ is the best linear unbiased estimator of $\mathbf{C}_i^T\boldsymbol{\beta}_{ij}$, and r_i is the rank of $\mathbf{C}_i$. With a normality assumption for $\boldsymbol{\epsilon}_{ij}$, the statistic F_{ij} can be compared to an F distribution to get a p-value. Alternatively, a permutation test can be used as suggested in Cui et al. (2005).

Obtaining a good estimate of σ^2_{ij} for use in the denominator of F_{ij} is crucial for effective statistical inference regarding $\mathbf{C}_i^T\boldsymbol{\beta}_{ij}$. In the following section, we will introduce a new strategy for estimating error variances by borrowing information both across genes and across experiments.

3 A Proposed Model for Error Variances and the Resulting Bayes Estimates

We model the error variance σ^2_{ij} of gene j in experiment i as

$$\log\sigma^2_{ij} = \mu + E_i + G_j + \varepsilon_{ij}, \tag{3}$$

where $\mu \in \Re$ is an unknown fixed parameter; $E_1, ..., E_I$ are random experiment effects distributed as $N(0, \sigma^2_E)$; $G_1, ..., G_J$ are random gene effects distributed as $N(0, \sigma^2_G)$; and $\varepsilon_{11}, ..., \varepsilon_{IJ}$ are random effects that allow non-additive effects of experiments and genes, and are distributed as $N(0, \sigma^2_\varepsilon)$. We assume that all random effects are mutually independent and that the parameters σ^2_E, σ^2_G, and σ^2_ε are unknown non-negative variance components. Model (3) is a natural extension of a single-experiment model that, as shown by Hwang and Liu

(2010), can be used to derive the shrinkage estimator of error variance used in the denominator of the F_S test proposed by Cui et al. (2005).

More generally, model (3) provides a natural framework for borrowing information across genes and experiments for improved error variance estimation. If the variance components σ_E^2, σ_G^2 and σ_ε^2 were all zero, then model (3) would imply a constant variance across all genes and all experiments. Although some of the earliest approaches to microarray data analysis assumed constant variance across genes, it is now well accepted that error variance differs from gene to gene. The random gene effects $(G_1, \ldots, G_J)$ allow for differences in error variance across genes.

If $\sigma_G^2 > 0$ but the variance components σ_E^2 and σ_ε^2 were both zero, then gene-specific error variances would be constant across experiments, and data from all experiments could be combined together and analyzed like a single microarray experiment with many treatment groups. However, differences from experiment to experiment in laboratory conditions, techniques, experimental materials, etc. may cause gene expressions to be more variable in some experiments than in others. The random experiment effects $(E_1, \ldots, E_I)$ allow for error variances to differ by experiment, and the random errors $(\varepsilon_{11}, \ldots, \varepsilon_{IJ})$ allow for non-additivity of gene and experiment effects. The assumed normal distributions for the gene, experiment, and error effects permit information sharing across both genes and experiments to improve individual estimates of error variance.

Our formulation of model (3) was motivated by an empirical investigation of error variance estimates computed separately for each combination of gene and experiment using data from multiple microarray experiments. While we do not expect model (3) to be precisely correct for all collections of microarray experiments, we expect it to be very useful for improving error variance estimates in most cases. As we shall demonstrate later in this section, our approach motivated by model (3) adapts to data by borrowing information from other genes and/or other experiments to a greater or lesser extent depending on the estimated values of σ_E^2, σ_G^2, and σ_ε^2.

If we let $\zeta_{ij} = \log \sigma_{ij}^2$ for all $i = 1, ..., I$ and $j = 1, ..., J$, then

$$\boldsymbol{\zeta} \equiv (\zeta_{11}, ..., \zeta_{1J}, ..., \zeta_{IJ})^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\mu} = \mathbf{1}\mu$ and

$$\boldsymbol{\Sigma} = \sigma_E^2 \mathbf{I}_{I\times I} \otimes \mathbf{J}_{J\times J} + \sigma_G^2 \mathbf{J}_{I\times I} \otimes \mathbf{I}_{J\times J} + \sigma_\varepsilon^2 \mathbf{I}_{IJ\times IJ}.$$

Here and throughout the remainder of the paper, we use $\mathbf{1}$ to denote a vector of ones, $\mathbf{I}_{m\times m}$ to denote an identity matrix with dimension m by m, and $\mathbf{J}_{m\times n}$ to

denote a matrix of ones with dimension m by n. If we assume $\boldsymbol{\epsilon}_{ij}$ in model (1) is multivariate normal, the conditional distribution of S_{ij}^2 given σ_{ij}^2 is $\sigma_{ij}^2 \frac{\chi^2_{n_i-d_i}}{n_i-d_i}$, where n_i is the number of slides in experiment i, d_i is the rank of the design matrix of experiment i, and $\chi^2_{n_i-d_i}$ denotes a χ^2 random variable with $n_i - d_i$ degrees of freedom that is independent of ζ_{ij}. It follows that

$$(\log S_{ij}^2|\zeta_{ij}) \stackrel{d}{=} (\zeta_{ij} + \log \frac{\chi^2_{n_i-d_i}}{n_i-d_i}|\zeta_{ij}). \tag{5}$$

As discussed in Hwang and Liu (2010), the distribution of $\log \frac{\chi^2_{n_i-d_i}}{n_i-d_i}$ can be approximated by a normal distribution with mean a_i and variance b_i, where a_i and b_i can be easily approximated to any desired degree of accuracy by simulation. Denoting $(\log S_{ij}^2 - a_i)$ by z_{ij}, if we assume approximate normality and conditional independence of the z_{ij}'s given $\boldsymbol{\zeta}$, it follows that

$$\mathbf{z}|\boldsymbol{\zeta} \quad \text{is approximately distributed as} \quad N(\boldsymbol{\zeta}, \mathbf{V}), \tag{6}$$

where

$$\mathbf{z} = (z_{11}, z_{12}, ..., z_{1J}, ..., z_{I1}, z_{I2}, ..., z_{IJ})^T,$$
$$\mathbf{V} = \text{diag}(\mathbf{b}), \text{ and } \mathbf{b} = (b_1, ..., b_I)^T \otimes \mathbf{1}_{J\times 1}.$$

Combining (4) and (6), it is straightforward to show that

$$E(\boldsymbol{\zeta}|\mathbf{z}) \approx (\boldsymbol{\Sigma}^{-1} + \mathbf{V}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{V}^{-1}\mathbf{z}) \tag{7}$$

We discuss how to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Appendix B. Plugging those estimates into (7) generates an empirical Bayes estimator of $\boldsymbol{\zeta}$, $\hat{\boldsymbol{\zeta}} \equiv (\hat{\zeta}_{11}, ..., \hat{\zeta}_{1J}, ..., \hat{\zeta}_{IJ})^T$. Our proposed estimation of σ_{ij}^2 is given by $\hat{\sigma}^2_{\text{BAGE},ij} \equiv \exp(\hat{\zeta}_{ij})$ for all i=1,...,I and all j=1,...,J.

Let $\hat{\boldsymbol{\sigma}}^2_{\text{BAGE}} \equiv (\hat{\sigma}^2_{\text{BAGE},11}, ..., \hat{\sigma}^2_{\text{BAGE},1J}, ..., \hat{\sigma}^2_{\text{BAGE},IJ})^T$. Computation of $\hat{\boldsymbol{\sigma}}^2_{\text{BAGE}}$ requires $(\boldsymbol{\Sigma}^{-1}+\mathbf{V}^{-1})^{-1}$, $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{V}^{-1}$ in (7). Because $\mathbf{V}$ is a diagonal matrix, finding its inverse is trivial. The inverse of $\boldsymbol{\Sigma}$ is described in Appendix C. The inverse of $(\mathbf{V}^{-1}+\boldsymbol{\Sigma}^{-1})$ is computed using a recursive algorithm illustrated in Appendix D.

It is straightforward to see that the posterior expectation $\text{E}(\boldsymbol{\zeta}|\mathbf{z})$ in (7) is a weighted average of the prior mean $\boldsymbol{\mu}$ and the data $\mathbf{z}$. Specifically, when the degrees of freedom are the same for all experiments, then $a_1 = \cdots = a_I = a$ and $b_1 = \cdots = b_I = b$. If the parameters are replaced with their estimates in (7), the log BAGE estimator of the error variance for gene j in experiment i is

$$\log \hat{\sigma}^2_{\text{BAGE},ij} = \widehat{E(\zeta_{ij}|\mathbf{z})} = wz_{ij} + (1-w)\hat{z}_{ij}, \tag{8}$$

where

$$\hat{z}_{ij} = \bar{z}_{..} + w_E(\bar{z}_{i.} - \bar{z}_{..}) + w_G(\bar{z}_{.j} - \bar{z}_{..}), \tag{9}$$

$$w = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + b},\ w_E = \frac{J\hat{\sigma}_E^2}{\hat{\sigma}_\varepsilon^2 + b + J\hat{\sigma}_E^2},\ \text{and}\ w_G = \frac{L\hat{\sigma}_G^2}{\hat{\sigma}_\varepsilon^2 + b + L\hat{\sigma}_G^2}. \tag{10}$$

Note that (8) involves a convex combination of z_{ij} and $\hat{z}_{ij}$. z_{ij} is an estimate of $\log \sigma_{ij}^2$ solely based on the REML estimate of σ_{ij}^2, while $\hat{z}_{ij}$ in (9) is an estimate of $\log \sigma_{ij}^2$ based on REML estimates of the error variances of all J genes in all I experiments. The weight coefficient w in (8) depends on the proportion of variation in the $\log \sigma_{ij}^2$ values that can be explained by the additive effects of experiments and genes in (3). If this proportion is small, the estimated variance σ_ε^2 for model (3) will be relatively large, and there is not much information to borrow either across genes or experiments. In this case, w in (10) is computed close to 1, and $\hat{\sigma}^2_{\text{BAGE},ij}$ is based mainly on the REML estimator of σ_{ij}^2. On the other hand, if $\hat{\sigma}_\varepsilon^2$ is relatively small, w is close to 0, and $\hat{\sigma}^2_{\text{BAGE},ij}$ is largely determined by $\hat{z}_{ij}$.

The term $\hat{z}_{ij}$ can be viewed as a prediction of $\log \sigma_{ij}^2$ as suggested by (9), where an estimate of the ith experiment effect $(\bar{z}_{i.} - \bar{z}_{..})$ and the jth gene effect $(\bar{z}_{.j} - \bar{z}_{..})$ are weighted according to estimates of experiment and gene variation respectively. For example, if $\hat{\sigma}_E^2$ and $\hat{\sigma}_G^2$ are both small, then w_E and w_G are both small, and the prediction of $\log \sigma_{ij}^2$ is obtained by heavily borrowing information across both genes and experiments to obtain $\hat{z}_{ij} \approx \bar{z}_{..}$. If $\hat{\sigma}_E^2$ is small and $\hat{\sigma}_G^2$ is large, then the prediction of $\log \sigma_{ij}^2$ based on model (3) is obtained by heavily using the jth gene effect to obtain $\hat{z}_{ij} \approx \bar{z}_{.j}$. Other scenarios can be interpreted similarly. The advantage of this approach is that the extent and direction of information borrowing (across genes or across experiments) is determined by the data.

4 Permutation Test

By replacing the REML estimator of error variance with the BAGE estimator in (2), we propose to use the test statistic

$$F_{\text{BAGE},ij} = \frac{\hat{\boldsymbol{\beta}}_{ij}^T \mathbf{C}_i [\mathbf{C}_i^T (\mathbf{X}_i^T \mathbf{X}_i)^- \mathbf{C}_i]^{-1} \mathbf{C}_i^T \hat{\boldsymbol{\beta}}_{ij} / r_i}{\hat{\sigma}^2_{\text{BAGE},ij}} \tag{11}$$

to test for differential expression.

Because the null distribution of $F_{\text{BAGE},ij}$ is unknown, we propose to approximate its null distribution through a permutation method within each experiment i. Since the number of observations in each treatment group is often small for microarray experiments, the total number of distinct permutations per gene is also small. This leads to highly discrete p-values. To overcome this problem, Storey and Tibshirani (2003b) proposed to pool the permutation-derived test statistics across all genes. However, as pointed out by Storey and Tibshirani (2003b), Xie, Pan, and Khodursky (2005), Fan et al. (2005), and Yang and Churchill (2007), permutation distributions for differentially expressed genes and non-differentially expressed genes may differ. Pooling permutation test statistics from all genes, including many differentially expressed ones, tends to increase the variation of the permutation distribution. Consequently, the approximated null distribution tends to have heavier tails than it should for some genes, and the p-values tend to be conservative. To alleviate this problem, Yang and Churchill (2007) proposed to only pool the permutation test statistics of a subset of genes that are most likely to be null. Through simulations, Yang and Churchill (2007) suggested using a cutoff of 0.1 for p-values obtained through ordinary F-tests when selecting a subset of null-like genes.

The idea of our permutation test is similar to Yang and Churchill (2007) except that, instead of throwing away the genes with p-values no larger than 0.1, we modify their observations (as described below) such that they become null-like genes, that is their p-values from the ordinary F-test become larger than 0.1. Then, we pool the permutation test statistics of all genes after this modification to approximate the null distribution. We illustrate the proposed modification and permutation method for an experiment comparing two treatments. Other cases can be handled similarly.

Suppose we test for differentially expressed genes between the two treatment groups of experiment i. First, we select the subset of null-like genes, G_{i0}, using the criterion that the p-value from the ordinary F-test is bigger than 0.1. Let G_{ia} denote the set of remaining genes. The next step is to modify observations for genes in G_{ia} so that these genes become null-like without affecting their REML estimates of error variances. Specifically, for each gene $j_a \in G_{ia}$, we randomly select a gene $j_0 \in G_{i0}$, and compute $\delta_{ij_0} = \frac{\triangle trt_{ij_0}}{\sqrt{S^2_{ij_0}}}$, where $\triangle trt_{ij_0}$ is the difference between two treatment means for gene j_0. Then we modify $\triangle trt_{ij_a}$ to be $\delta_{ij_0} \times \sqrt{S^2_{ij_a}}$ by adding a constant to all replicates of one treatment group of gene j_a. After this modification, all genes in experiment i are null-like. The following steps are similar to those suggested in Yang and Churchill (2007). We then permute the modified data in experiment i. The

$F_{\text{BAGE},ij}$ statistic is computed for each gene j and each permutation. The statistics pooled from all permutations and all genes in experiment i are used to approximate the null distribution of $F_{\text{BAGE},ij}$ for all j. P-values for gene j in experiment i are evaluated by comparing $F_{\text{BAGE},ij}$ computed by the original observations against the approximated null distribution.

Note that when approximating the null distribution of the test statistic $F_{\text{BAGE},ij}$, the REML estimates of error variances for the modified data remain the same as that of the original data for experiment i, and observations in other experiments are kept unchanged. Hence, the estimates of hyperparameters, σ_E^2, σ_G^2, σ_ε^2, μ, and $\mathbf{b}$ remain the same after modification, and the F_{BAGE} statistics for the null-like data can be computed without re-estimating $\boldsymbol{\Sigma}$, $\mathbf{V}$, and $\boldsymbol{\mu}$.

5 Simulation Studies Based on Probability Models

Through simulations based on different probability models, we compare the BAGE estimator of error variances with the REML estimator and the estimator of variance used in the popular LIMMA R package (Smyth, 2004). This latter estimator, developed by Smyth (2004) and referred to here as the LIMMA estimator, borrows information across all genes separately within each experiment. We also compare the F_{BAGE} test with the ordinary t-test based on REML estimates and the moderated t-test based on LIMMA estimates (Smyth, 2004).

For simulation studies in this section, we consider the case where each experiment compares two treatment groups using one microarray slide per experimental unit. We simulated 1,000 genes for each experiment where 500 genes were randomly chosen to be differentially expressed. In simulations 5.1 and 5.2, we generated data for 100 experiments with 6 slides in each experiment. In simulations 5.3 and 5.4, we simulated 100 sets of three experiments (one with 6 slides, one with 8 slides, and the other with 10 slides). In all simulations, slides in each experiment were evenly allocated to two treatment groups.

5.1 Simulated Cases with Equal Degrees of Freedom across Experiments

In simulation 5.1, we generated error variances under the lognormal model in (3). In order to set realistic hyperparameter values, we first analyzed data from the the maize root study (Nakazono et al., 2009) by the BAGE method. Based on the estimated hyperparameter values, we used σ_G^2=0.44, σ_E^2=0.20, σ_ϵ^2=0.05, and μ=-2.00 in simulation to generate true error variances, σ_{ij}^2's. Next, we simulated observations for gene j in experiment i independently from $\mathrm{N}(0, \sigma_{ij}^2)$. For each differentially expressed gene, we added a treatment effect to observations in one treatment group as described in Appendix E. We estimated error variances for simulated data using the REML, LIMMA and BAGE methods. For the BAGE method, we combined every 2, 5, and 10 experiments together. These variations of the BAGE method are subsequently denoted as BAGE(2), BAGE(5), and BAGE(10), respectively, with analogous notation for combining over other numbers of experiments.

Figure 1(a) shows the contour plots of REML, LIMMA, and BAGE estimates, where BAGE estimates were computed by combining 10 experiments together for analysis. The complete contour plots are in Figure 3(a) in Appendix H. These plots show that BAGE estimates are more concentrated around the diagonal line for all cases, which demonstrates that the BAGE estimates are closer to the true error variances.

Figure 1(c) shows box plots of differences between estimates and true error variances in log scale for all genes in all experiments. BAGE estimates have a smaller interquantile range than REML and LIMMA estimates. In addition, by combining more experiments, the box plots for BAGE estimates are more densely centered around the horizontal 0 line. Table 1 indicates that BAGE estimates have smaller biases and smaller mean square errors (MSEs) computed on the original scale as well averaged over all genes and all experiments.

In Figure 1(e), we plot the Receiving Operating Characteristic (ROC) curves from the three tests: the ordinary t-test, the moderated t-test, and the F_{BAGE} test. We only plot the region where the false positive rate (FPr) is below 0.05 because this is the most interesting region in practice. The ROC curves show that the BAGE method yields a higher true positive rate (TPr) than the other two methods at any given FPr. Furthermore, the more experiments combined in analysis by the BAGE method, the higher the ROC curve.

We applied the procedure for FDR control proposed by Storey (2002) in conjunction with the q-value computation method of Storey and Tibshirani (2003a) to control FDR level at 0.05. Table 1 lists the number of false

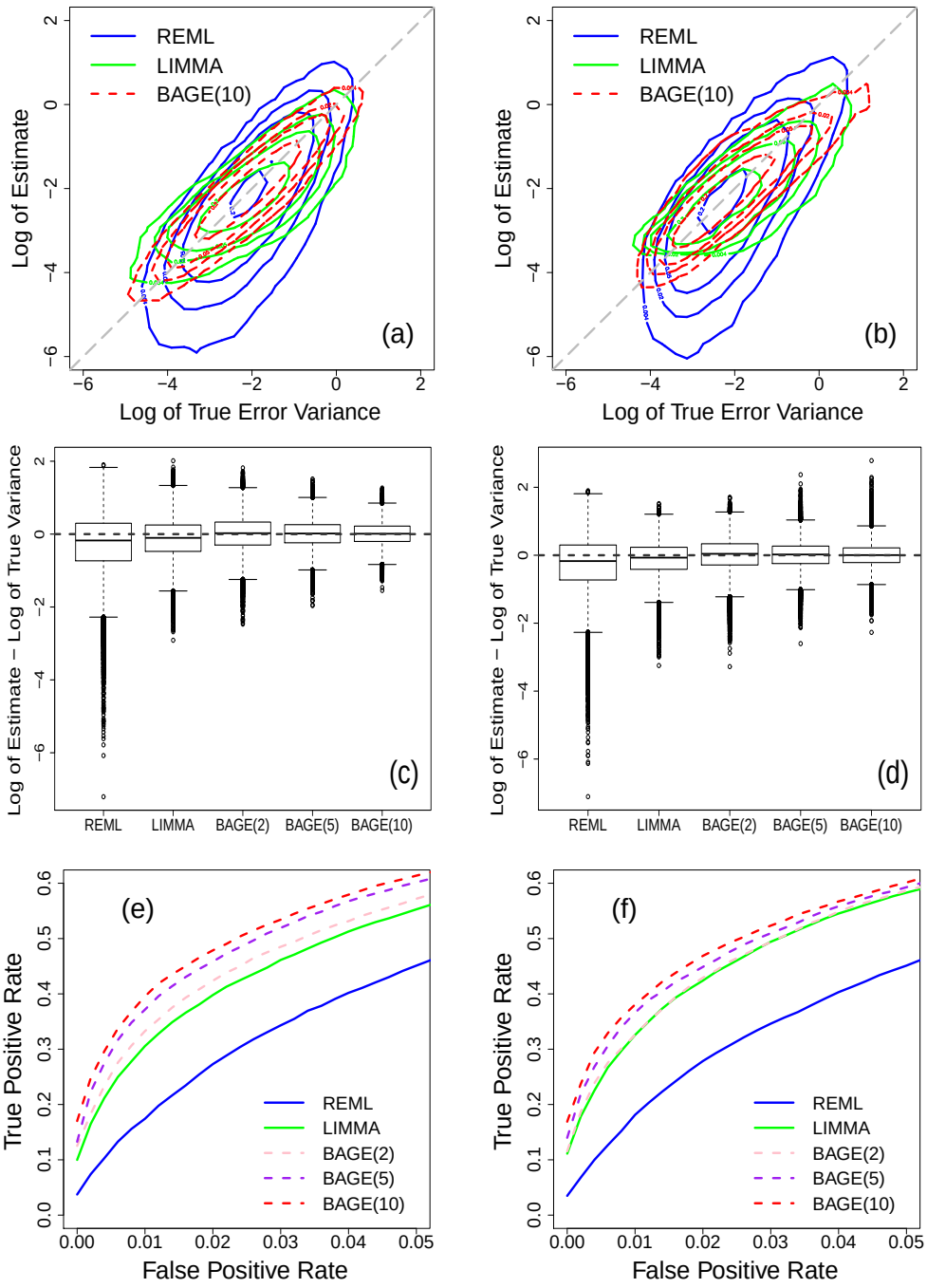

Figure 1: Comparison of REML, LIMMA, and BAGE methods. Left column: data are simulated under the lognormal model (simulation 5.1). Right column: data are simulated under the inverse gamma model (simulation 5.2). Top row: contour plots to compare estimates of error variances. Middle row: box plots of estimation errors in log scale. Bottom row: ROC curves.

Table 1: Comparison of bias, mean square error, area under ROC curves, false positive number (V), positive number (R), and V/R for simulations 5.1-5.4 and 6.1-6.2.

Simulations	Model	Degrees of freedom	Method	Bias	MSE	AUC	V	R	V/R
5.1	Log Normal	4	LIMMA	-.0387	.0173	.0203	9.58	196.22	.046
			BAGE(2)	-.0186	.0129	.0215	8.77	198.75	.044
			BAGE(5)	-.0112	.0086	.0232	8.11	215.08	.036
			BAGE(10)	-.0077	.0063	.0241	8.42	226.95	.036
5.2	Inverse Gamma	4	LIMMA	-.0396	.0282	.0216	8.79	200.94	.042
			BAGE(2)	-.0275	.0316	.0218	6.91	179.28	.037
			BAGE(5)	-.0215	.0250	.0228	7.47	202.49	.035
			BAGE(10)	-.0182	.0210	.0235	7.73	216.26	.035
5.3	Log Normal	Overall	LIMMA	-.0324	.0142	.0243	11.80	259.62	.045
			BAGE(3)	-.0126	.0086	.0257	10.97	268.35	.040
		8	LIMMA	-.0256	.0104	.0251	12.05	269.91	.044
			BAGE(3)	-.0114	.0071	.0260	11.84	277.28	.042
		6	LIMMA	-.0331	.0146	.0241	12.14	260.35	.046
			BAGE(3)	-.0131	.0092	.0256	11.66	271.41	.043
		4	LIMMA	-.0386	.0176	.0236	11.22	248.61	.045
			BAGE(3)	-.0132	.0095	.0255	9.41	256.37	.036
5.4	Inverse Gamma	Overall	LIMMA	-.0353	.0352	.0244	10.86	255.23	.042
			BAGE(3)	-.0241	.0690	.0254	10.13	260.19	.038
		8	LIMMA	-.0503	.0751	.0253	10.53	261.36	.040
			BAGE(3)	-.0464	.1151	.0257	10.57	266.50	.039
		6	LIMMA	-.0354	.0223	.0245	11.23	259.63	.043
			BAGE(3)	-.0167	.0223	.0255	10.80	266.19	.040
		4	LIMMA	-.0203	.0080	.0235	10.81	244.69	.043
			BAGE(3)	-.0093	.0073	.0250	8.67	247.87	.035
6.1	ALL	6	LIMMA	-	-	.0228	105.88	2534.81	.039
			BAGE(4)	-	-	.0238	116.49	2714.84	.040
6.2	Golub	6	LIMMA	-	-	.0218	45.12	1330.68	.033
			BAGE(3)	-	-	.0227	57.46	1528.09	.037

positives (V), number of positives (R), and V/R for the moderated t-test and the F_{BAGE} test averaged across 100 experiments. The FDR of these methods were all controlled below 0.05. The BAGE method reported more true positives than the LIMMA method regardless of the number of experiments that were combined (2, 5, or 10). In addition, when the number of experiments combined in the BAGE method increased, the F_{BAGE} test reported more positives and more true positives. By combining 10 experiments together for

analysis, the F_{BAGE} test identified nearly 15% more true positives than the moderated t-test. Table 1 also shows that, with respect to ROC curves, the area under the curve values (AUCs) of the F_{BAGE} test, averaged across 100 experiments, were larger than the moderated t-test for the region where FPr is no larger than 0.05. We also conducted a paired t-test to check whether the AUCs are significantly different between the BAGE methods and the LIMMA method. Each of the tests comparing BAGE(2), BAGE(5) or BAGE(10) with the LIMMA method yielded a p-value less than 0.001.

The LIMMA method proposed by Smyth (2004) models the error variances within a single experiment as draws from an inverse gamma distribution. In simulation 5.2, we simulated error variances of each experiment from an inverse gamma distribution with parameters $\alpha = d_0/2$ and $\beta = 1/(d_0 s_0^2)$, where d_0 and s_0^2 are defined in Smyth (2004). We estimated the pair of parameters $(d_0,\, s_0^2)$ for each of the four maize experiments (Nakazono et al., 2009) by the LIMMA method, and the estimates are (4.24, 0.16), (5.11, 0.12), (5.00, 0.06), and (4.59, 0.07). We generated data for 100 experiments with 25 experiments simulated using each pair of parameters. Appendix F provides further details about how error variances were simulated to be correlated across experiments.

Similar to simulation 5.1, we analyzed these 100 experiments by the REML, LIMMA and BAGE methods. Figure 1(b) compares the contour plots of REML, LIMMA, and BAGE(10) estimates. The complete contour plots are in Figure 3(b) in Appendix H. Similar to the results in simulation 5.1, the contour plots indicate that the BAGE estimates are more accurate and precise than estimates by the other two methods. Figure 1(d) shows box plots of estimating errors in log scale. The BAGE estimates have a smaller interquantile range than REML and LIMMA estimates. In addition, by combining more experiments, the interquantile range for the BAGE method becomes smaller and the median becomes closer to 0. Table 1 shows that the BAGE estimators exhibited smaller biases than the LIMMA estimator on average. Furthermore, the estimated MSEs for BAGE(5) and BAGE(10) were smaller than that of the LIMMA estimator on average.

The ROC curves are plotted in Figure 1(f). The F_{BAGE} test yielded a higher TPr for any given FPr, and hence had larger AUCs than the moderated t-test as presented in Table 1. Similar to the study in simulation 5.1, paired t-tests were applied to test differences of AUCs for the BAGE methods and the LIMMA method, and they all yielded p-values less than 0.01. By combining more experiments in analysis, the BAGE method reported more positives and more true positives. By combining 10 experiments, the F_{BAGE} test reported nearly 10% more true positives than LIMMA, even though the data were simulated under the LIMMA model rather than the lognormal model used to derive

the BAGE estimator.

5.2 Simulated Cases with Different Degrees of Freedom across Experiments

The variance estimates obtained from an experiment with more degrees of freedom are more reliable than those from an experiment with fewer degrees of freedom. By appropriately combining experiments together for analysis, inferences of all experiments should improve, and the experiments with the fewest degrees of freedom are expected to benefit most.

In simulation 5.3, we generated microarray data with variances under the lognormal model using the same hyperparameter values as in simulation 5.1. We simulated 3 experiments with 6, 8, and 10 slides, respectively. BAGE estimates were computed by combining data from these 3 experiments. This 3-experiment setting was simulated 100 times.

Table 1 shows that the estimated biases and MSEs, averaged over all experiments and genes or averaged across genes over experiments with the same degrees of freedom, were smaller for the BAGE estimators than for the LIMMA estimators in all cases. The contour plots in Figure 3(c) of Appendix H also show that BAGE estimates are closer to the true error variances than REML and LIMMA estimates. In addition, Table 1 indicates that the F_{BAGE} tests yielded larger AUCs than the moderated t-tests on average. The improvement was more substantial for experiments with smaller degrees of freedom than experiments with larger degrees of freedom. A paired t-test comparing the BAGE(3) and LIMMA methods yielded a p-value less than 0.01 for all simulated experiments. Table 1 shows that by controlling FDR at the 0.05 level as discussed previously, BAGE(3) gave more true positives and less false positives on average.

In simulation 5.4, we generated microarray data similar to simulation 5.3 except that we used an inverse gamma distribution to simulate error variances. The simulation method and hyperparameter values are the same as in simulation 5.2.

Similar to the results in simulation 5.3, the contour plots in Figure 3(d) of Appendix H and the statistics in Table 1 show that the BAGE estimates and the F_{BAGE} test improved upon the LIMMA estimates and the moderated t-test, respectively. Paired t-tests indicated that the improvement of AUCs of the BAGE method over the LIMMA and REML methods is significant at level 0.01. Although the inverse gamma distribution differs from the lognormal

distribution under which BAGE was derived, BAGE still reported more true positives for all cases.

6 Real Data Simulation Examples

Instead of using hypothetical probability models for simulating error variances, in this section, we evaluate the performance of the BAGE method through simulations based on two real microarray data sets where the distribution of true error variances are not known.

Simulation 6.1 is based on the cancer data set ALL (Chiarentti et al., 2004). It contains 128 microarrays for 128 patients using the same Affymetrix microarray platform. Each microarray contains the same set of 12,625 genes. These 128 samples are of two major cell types (B and T). Each major cell type is further categorized into 5 sub cell types: B, B1, B2, B3, B4, T, T1, T2, T3, and T4. There are 19 samples of B1 type, 36 samples of B2 type, 15 samples of T2 type, and 10 samples of T3 type.

We simulated four experiments with 8 slides each, and evenly allocated slides in an experiment to two treatment groups. Specifically, we randomly selected 8 slides (samples) from each of B1, B2, T2, and T3 cell types for four experiments respectively. Within each experiment, we randomly picked 6,000 genes to be differentially expressed by adding treatment effects to observations of one treatment group.

We analyzed these 4 experiments individually by REML and LIMMA methods, and then we combined these 4 experiments together for analysis by the BAGE method. We simulated this 4-experiment setting 30 times.

The average ROC curves of 120 experiments in 30 simulations are in Figure 2 (a). These curves show that, for any fixed FPr under 0.05, the F_{BAGE} test detected more true positives than both the ordinary t-test and the moderated t-test. A paired t-test comparing AUCs of the BAGE(4) and LIMMA methods yielded a p-value less than 0.001. This indicates that the improvement in AUC for the BAGE(4) method over the LIMMA method is significant. On average, the F_{BAGE} test also reported more true positives when controlling FDR level at 0.05 as described previously (Table 1).

In simulation 6.2, we simulated data based on another cancer data set (Golub et al., 1999). This data set contains 38 microarray slides corresponding to 38 patients. Each slide follows the same Affymetrix microarray platform with 7,129 genes. Of the 38 samples, 11 arise from acute lymphoblastic leukemia (ALL), and 27 arise from acute myeloid leukemia (AML). In addition,

ALL has two different cell types (B and T).

We constructed 3 experiments with 8 microarray slides each, and evenly allocated slides to two treatments in each experiment. The 8 slides in three experiments were randomly selected from samples of ALL B type, ALL T type, and AML, respectively. In each experiment, we randomly selected 3,500 genes to be differentially expressed. BAGE estimates were computed by combining 3 experiments in analysis. We simulated this 3-experiment setting 30 times.

The ROC curves averaged over 90 experiments are plotted in Figure 2 (b), which show that the F_{BAGE} test found more true positives for any given FPr than both the ordinary t-test and the moderated t-test. A paired t-test showed that the improvement of AUCs of the BAGE(3) method over the LIMMA method is significant at the level 0.001. When using the procedure for FDR control proposed by Storey (2002) in conjunction with the q-value computation menthod of Storey and Tibshirani (2003a) to control FDR at the 0.05 level, Table 1 shows that the F_{BAGE} test reported more true positives than the moderated t-test while controlling FDR well under the 0.05 level.

Both of the two simulations based on real experiments suggest that the BAGE method works well and outperforms the LIMMA method for realistically generated data sets that do not follow our parametric assumptions.

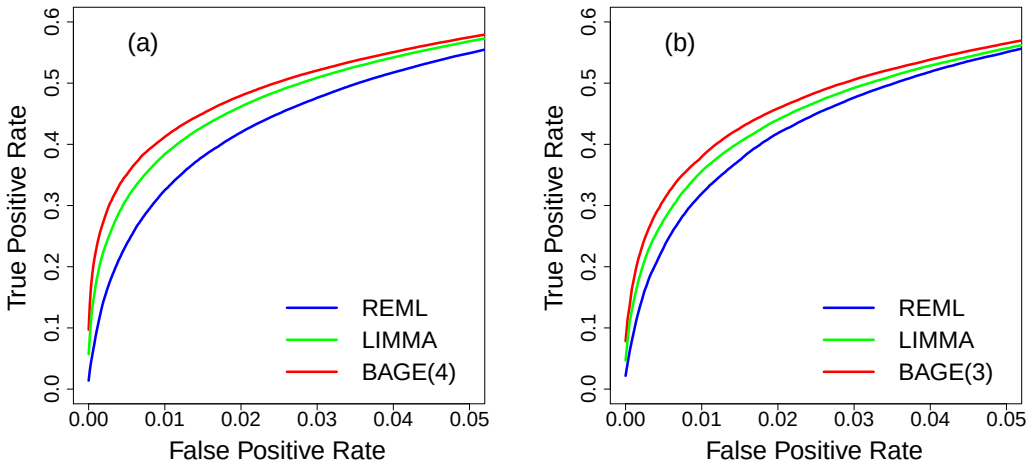

Figure 2: Comparison of the ordinary t-test based on REML estimates, the moderated t-test based on LIMMA estimates and the F_{BAGE} test based on BAGE estimates through simulations based on real data sets. (a) True error variances are simulated based on the ALL data set (simulation 6.1). (b) True error variances are simulated based on the Golub data set (simulation 6.2).

7 Analysis of the Study on Aerenchyma Formation in Maize Roots

The goal of the four maize experiments (Nakazono et al., 2009) is to identify genes involved in different stages of aerenchyma formation in maize roots when roots undergo a shortage of oxygen. All four experiments used 3-day-old seedlings of maize inbred line B73, and were conducted with the same two-color microarrays (GEO Platform GPL4521) of 14,118 total genes.

We selected 4 microarrays from each experiment for analysis as listed in Table 4 in Appendix A. We estimated error variances for each combination of gene and experiment by the REML, LIMMA and BAGE methods, and tested for differential expression by the ordinary t-test, the moderated t-test and the F_{BAGE} test, respectively. The number of reported differentially expressed genes while controlling FDR at the 0.05 level as previously described is listed in Table 2. The F_{BAGE} test reported more positives than the other two methods, which is not surprising given the outcomes of our simulation studies.

Table 2: Number of reported positives for each experiment in the four motivating experiments by three different methods.

Test \ Experiment	1	2	3	4
Ordinary t-test	255	0	0	0
Moderated t-test	6655	469	0	17
F_{BAGE} test	6973	521	2	20

The BAGE method extends the idea of sharing information across genes to across both genes and experiments. The amount of improvement by the BAGE method depends on the relative size of σ_G^2, σ_E^2, and σ_ε^2. The estimated hyperparameters for the maize root data are $\hat{\sigma}_G^2 \approx 0.48$, $\hat{\sigma}_E^2 \approx 0.18$, $\hat{\sigma}_\varepsilon^2 \approx 0.03$, and $\hat{\mu} \approx -2.32$. The small value of $\hat{\sigma}_\varepsilon^2$ relative to $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ for the maize microarray data indicates that the additive gene and experiment effects in model (3) account for much of the variability in log error variances across genes and experiments. Because $\hat{\sigma}_E^2$ is small relative to $\hat{\sigma}_G^2$, there are advantages to borrowing information across both genes and experiments rather than only across genes. In Appendix G, we simulated data under the lognormal model with 4 cases of different relative sizes of hyperparameters. Shown by the results summarized in Table 6 in Appendix G, when σ_ε^2 is smaller than σ_G^2 and σ_E^2,

the BAGE method improves substantially over the LIMMA method with respect to bias, MSE, AUC, and the number of reported true positives while controlling FDR well under a given level.

8 Conclusions and Discussion

Previous shrinkage methods only borrow information across genes within one experiment. Because multiple related experiments are often conducted with the same microarray platform, we can improve statistical inferences by incorporating information not only from genes within one experiment but also from other experiments. We have proposed a two-way hierarchical model for simultaneously estimating error variances of multiple microarray experiments when these microarray experiments use the same platform. Based on this model, we developed the BAGE estimator, an empirical Bayes estimator for estimating error variances of genes for each combination of gene and experiment. We also proposed the F_{BAGE} statistic and designed a permutation test procedure for finding differentially expressed genes with F_{BAGE} statistics.

Several simulations were done with different hypothetical probability models and real data sets. The BAGE error variance estimates were shown to be more accurate and precise on average compared to REML and LIMMA estimates. The F_{BAGE} statistic gave a better ranking for identifying differentially expressed genes than competing approaches. Furthermore, the F_{BAGE} test also yielded more true positives on average than the ordinary t-test and the moderated t-test while controlling FDR at the 0.05 level. The performance of the BAGE method improves when combining more experiments together for analysis. In addition, the improvement is more substantial for an experiment with a small number of replicates when it is analyzed by borrowing information from other experiments with more replicates.

Although we have focused on modeling gene expression data from microarray experiments, a similar idea could be adapted for analyzing data from next-generation sequencing (NGS) experiments. Like microarray experiments, NGS experiments typically involve many genes but only a few replicates for each gene within one experiment. Thus, there is clear potential for improving inference in NGS experiments by borrowing information across both genes and experiments.

Appendix A. Treatments in the Motivating Experiments

Aerenchymas are air channels formed in roots when plants undergo a shortage of oxygen. Aerenchyma formation enhances the oxygen exchange between roots and other parts of the plants. In this study, researchers are interested in discovering genes involved in different stages of aerenchyma formation in maize roots.

In order to approach this, four independent experiments were conducted. In all four experiments, 3-day-old maize seedlings of inbred line B73 were used, and tissues of interest were extracted. Each experiment compared two different tissues, conditions, or treatments. Amplified cDNAs were applied to the same two-color microarray platform (GEO Platform GPL4521) with a dye-swap design. The comparisons in the experiments are summarized in Table 3. Specifically, in experiment 1, seedlings were waterlogged, and basal and apical tissues were compared under the resulting hypoxic condition. Aerenchymas were formed at the basal region but not the apical region of roots. In experiment 2, seedling roots under hypoxic and aerated conditions were compared. For the basal region of maize roots, aerenchymas were formed under the hypoxic condition but not under the aerated condition. In experiment 3, two groups of seedlings were both grown under the aerated condition. Ethylene gas treatment was added to one group, where aerenchymas formed even under the aerated condition. Tissues from the basal regions with and without ethylene treatment were compared. In experiment 4, two groups of seedlings were both grown under the hypoxic condition. One group was treated with 1-methylcyclopropene (1-MCP), which is an ethylene perception inhibitor. No aerenchyma was formed with the 1-MCP treatment. Tissues from the basal regions with and without the 1-MCP treatment were compared.

The normalized data are downloadable from GEO (GSE26897). For the paper, we analyzed 4 slides from each experiment (see Table 4).

Table 3: Comparisons in the four motivating experiments.

Experiment	Treatment 1	Treatment 2	trt 1/trt 2[1]	Num[2]
1	Basal Tissue	Apical Tissue	Yes/No	6
2	Hypoxic condition	Aerated condition	Yes/No	5
3	No ethylene	Ethylene	No/Yes	4
4	No 1-MCP	1-MCP	Yes/No	4

[1] Aerenchyma formation under treatment 1 versus treatment 2.

[2] Number of microarray slides in each experiment.

Table 4: Selected slides for analysis in the paper.

Experiment	Slides (GSM number in GEO)
1	GSM662352; GSM662353; GSM662354; GSM662355
2	GSM662357; GSM662358; GSM662359; GSM662361
3	GSM662362; GSM662363; GSM662364; GSM662365
4	GSM662366; GSM662367; GSM662368; GSM662369

Appendix B. Parameter Estimation

The hyperparameters to be estimated in model (3) are μ, σ_E^2, σ_G^2 and σ_ε^2. This section describes how we estimate these parameters using a method-of-moments approach. First, note that based on (3) and (5), we have

$$z_{ij} = \zeta_{ij} + x_{ij} = \mu + E_i + G_j + \varepsilon_{ij} + x_{ij},$$

where $x_{ij} \stackrel{d}{=} \left(\log \frac{\chi^2_{n_i-d_i}}{n_i-d_i} - \mathrm{E}\left(\log \frac{\chi^2_{n_i-d_i}}{n_i-d_i}\right)\right)$, and all terms appearing in the sum are mutually independent. Our estimators are derived as follows.

B.1 Estimation of σ_ε^2

We have $z_{ij} - \bar{z}_{i\cdot} - \bar{z}_{\cdot j} + \bar{z}_{\cdot\cdot} = (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot j} + \bar{\varepsilon}_{\cdot\cdot}) + (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})$. Define

$$\mathrm{MS_{error}} = \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I}\sum_{j=1}^{J} (z_{ij} - \bar{z}_{i\cdot} - \bar{z}_{\cdot j} + \bar{z}_{\cdot\cdot})^2.$$

Then, $\mathrm{E}(\mathrm{MS_{error}}) = \sigma_\varepsilon^2 + \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I}\sum_{j=1}^{J} E(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2$. We further denote $\frac{1}{(I-1)(J-1)} \sum_{i=1}^{I}\sum_{j=1}^{J} \mathrm{E}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2$ by σ_x^2 and we estimate it by simulating x_{ij}'s. Thus, an unbiased estimator of σ_ε^2 is $\hat{\sigma}_\varepsilon^2 = \mathrm{MS_{error}} - \hat{\sigma}_x^2$.

B.2 Estimation of σ_E^2

We have $\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot} = (E_i - \bar{E}_\cdot) + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}) + (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})$. Define

$$\mathrm{MS_{exp}} = \frac{J}{I-1} \sum_{i=1}^{I} (\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2.$$

Then, $\mathrm{E}(\mathrm{MS}_{\mathrm{exp}}) = J\sigma_E^2 + \sigma_\varepsilon^2 + \frac{J}{I-1}\sum_{i=1}^{I}\mathrm{E}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$. We use $\sigma^2_{x,\mathrm{exp}}$ to denote $\frac{J}{I-1}\sum_{i=1}^{I}\mathrm{E}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$ and estimate it by simulating x_{ij}'s. Our moment estimator of σ_E^2 is $\hat{\sigma}_E^2 = \frac{1}{J}(\mathrm{MS}_{\mathrm{exp}} - \hat{\sigma}_\varepsilon^2 - \hat{\sigma}^2_{x,\mathrm{exp}})$.

B.3 Estimation of σ_G^2

We have $\bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot} = (G_j - \bar{G}_\cdot) + (\bar{\varepsilon}_{\cdot j} - \bar{\varepsilon}_{\cdot\cdot}) + (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})$. Define

$$\mathrm{MS}_{\mathrm{gene}} = \frac{I}{J-1}\sum_{j=1}^{J}(\bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot})^2.$$

Then, $\mathrm{E}(\mathrm{MS}_{\mathrm{gene}}) = I\sigma_G^2 + \sigma_\varepsilon^2 + \frac{I}{J-1}\sum_{j=1}^{J}E(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$. We denote $\frac{I}{J-1}\sum_{j=1}^{J}E(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$ by $\sigma^2_{x,\mathrm{gene}}$ and estimate it by simulation. Our moment estimator of σ_G^2 is $\hat{\sigma}_G^2 = \frac{1}{I}(\mathrm{MS}_{\mathrm{gene}} - \hat{\sigma}_\varepsilon^2 - \hat{\sigma}^2_{x,\mathrm{gene}})$.

B.4 Estimation of μ

We have $\mathrm{E}(\frac{1}{IJ}\sum_{i,j} z_{ij}) = \mu$. Thus, an unbiased estimator for μ is $\hat{\mu} = \frac{1}{IJ}\sum_{i,j} z_{ij}$.

In our simulations, $\hat{\sigma}_\varepsilon^2$ was rarely negative, and $\hat{\sigma}_E^2$ and $\hat{\sigma}_G^2$ were always positive. However, in case any estimate of them is negative, we use 0 as the estimate to replace the negative value.

Appendix C. Computation of $\mathbf{\Sigma}^{-1}$

Let $\mathbf{\Sigma} \equiv \sigma_E^2\mathbf{I}_{I\times I}\otimes\mathbf{J}_{J\times J} + \sigma_G^2\mathbf{J}_{I\times I}\otimes\mathbf{I}_{J\times J} + \sigma_\epsilon^2\mathbf{I}_{IJ\times IJ}$. If $\mathbf{\Sigma}$ is not singular,

$$\mathbf{\Sigma}^{-1} = a\mathbf{I}_{I\times I}\otimes\mathbf{J}_{J\times J} + b\mathbf{J}_{I\times I}\otimes\mathbf{I}_{J\times J} + c\mathbf{I}_{IJ\times IJ} + d\mathbf{J}_{IJ\times IJ},$$

where

$$a = -\frac{\sigma_E^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + J\sigma_E^2)},\ \ b = -\frac{\sigma_G^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + I\sigma_G^2)},\ \ c = \frac{1}{\sigma_\epsilon^2},\ \text{ and}$$

$$d = \frac{\sigma_E^2\sigma_G^2(2\sigma_\epsilon^2 + J\sigma_E^2 + I\sigma_G^2)}{\sigma_\epsilon^2(\sigma_\epsilon^2 + J\sigma_E^2)(\sigma_\epsilon^2 + I\sigma_G^2)(\sigma_\epsilon^2 + J\sigma_E^2 + I\sigma_G^2)}.$$

Appendix D. Computation of $(\mathbf{\Sigma}^{-1} + \mathbf{V}^{-1})^{-1}$

Corollary 1 of Theorem A.76 on page 377 of Rao and Toutenburg (1999) states that

$$(\mathbf{M} + \mathbf{c}\mathbf{c}^T)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{c}\mathbf{c}^T\mathbf{M}^{-1}}{1 + \mathbf{c}^T\mathbf{M}^{-1}\mathbf{c}}$$

for any symmetric and nonsingular $n \times n$ matrix $\mathbf{M}$ and any n-dimensional vector $\mathbf{c}$.

A recursive algorithm was adopted to compute $(\mathbf{\Sigma}^{-1} + \mathbf{V}^{-1})^{-1}$ based on Corollary 1. To illustrate this, let $b_{[k]}$ denote the kth element of $\mathbf{b}$. It is straightforward to show that $\mathbf{V}^{-1} = \sum_{k=1}^{I\times J} \mathbf{c}_k\mathbf{c}_k^T$, where $\mathbf{c}_k = (0, ..., \frac{1}{\sqrt{b_{[k]}}}, ..., 0)^T$ with only the kth element non-zero. Thus, we want to compute $(\mathbf{\Sigma}^{-1} + \mathbf{V}^{-1})^{-1} = (\mathbf{\Sigma}^{-1} + \sum_{k=1}^{I\times J} \mathbf{c}_k\mathbf{c}_k^T)^{-1}$. Furthermore, let $\mathbf{\Sigma}_{[k,k]}$, $\mathbf{\Sigma}_{[k,]}$, and $\mathbf{\Sigma}_{[,k]}$ denote the kth diagonal element, the kth row, and the kth column of matrix $\mathbf{\Sigma}$, with similar notations for other matrices. The following explains the recursive algorithm with the initial step and one-step forward recursion.

Initial step:

$$(\mathbf{\Sigma}^{-1} + \mathbf{c}_1\mathbf{c}_1^T)^{-1} = \mathbf{\Sigma} - \frac{\mathbf{\Sigma}_{[,1]}\mathbf{\Sigma}_{[1,]}}{b_{[1]} + \mathbf{\Sigma}_{[1,1]}}.$$

One-step forward recursion: let $\mathbf{A} = \mathbf{\Sigma}^{-1} + \sum_{k=1}^{K} \mathbf{c}_k\mathbf{c}_k^T$. Given $\mathbf{A}^{-1}$, we have

$$\left(\mathbf{\Sigma}^{-1} + \sum_{k=1}^{K+1} \mathbf{c}_k\mathbf{c}_k^T\right)^{-1} = \left(\mathbf{A} + \mathbf{c}_{K+1}\mathbf{c}_{K+1}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1})_{[,K+1]}(\mathbf{A}^{-1})_{[K+1,]}}{b_{[K+1]} + (\mathbf{A}^{-1})_{[K+1,K+1]}}.$$

Appendix E. Generating Microarray Expression Data with Treatment Effects

Given the error variances, we simulated observations of gene j in experiment i independently from $\mathrm{N}(0, \sigma_{ij}^2)$. For differentially expressed (DE) genes, we added some treatment effect to the expression values in one treatment group. The treatment effect for DE gene j in experiment i, γ_{ij}, was generated from the model

$$\gamma_{ij} = \delta \times \sigma_{ij} \times X_{ij}, \tag{12}$$

where δ is a tuning parameter and X_{ij} is a random variable following a Beta(θ_1, θ_2) distribution.

Typically in microarray experiments, a small number of genes have large treatment effects while a large number of genes have small treatment differences. Accordingly, we set up the parameter values of θ_1 and θ_2 such that the Beta distribution is right skewed. σ_{ij} is multiplied in formula (12) because we want to scale the treatment difference proportional to the standard deviation. The value of δ affects the magnitude of treatment difference and was set to produce a realistic number of DE genes.

This method was used to generate treatment effects for simulations 5.1 – 5.4 and 6.1 – 6.2 in the paper. The value of parameters δ, θ_1, and θ_2 in the above simulations are listed in Table 5.

Table 5: Parameter values for generating treatment effect for simulations 5.1 – 5.4 and 6.1 – 6.2.

Simulations	δ	θ_1	θ_2
5.1 and 5.2	5	9	10
5.3, 5.4, 6.1 and 6.2	8	2	4

Appendix F. Generating Error Variances Following Inverse Gamma Priors for Multiple Microarray Experiments

Our goal is to simulate error variances of gene expression data for multiple microarray experiments such that, within each experiment, the error variances are a random sample from an inverse gamma distribution, and the error variances of the same gene across experiments are correlated. This simulation method is used in simulations 5.2 and 5.4.

Let α_i and β_i be the two parameters for a gamma distribution with mean $\alpha_i\beta_i$ and variance $\alpha_i\beta_i^2$.

Let α, β be two constants such that $0 < \alpha < \min(\alpha_1, \alpha_2, ..., \alpha_I)$ and $\beta > 0$. We simulated Y_j independently from Gamma(α, β), and simulated Z_{ij} independently from Gamma$(\alpha_i - \alpha, \beta_i)$ for $j = 1, ..., J$ and $i = 1, ..., I$. It follows

that $\frac{\beta_i}{\beta}Y_j \sim \text{Gamma}(\alpha, \beta_i)$ and $\frac{1}{\sigma_{ij}^2} \equiv \frac{\beta_i}{\beta}Y_j + Z_{ij} \sim \text{Gamma}(\alpha_i, \beta_i)$. It can be easily verified that $\text{Cov}\left(\frac{1}{\sigma_{ij}^2}, \frac{1}{\sigma_{ik}^2}\right) = 0, \text{Cov}\left(\frac{1}{\sigma_{ij}^2}, \frac{1}{\sigma_{lj}^2}\right) = \alpha\beta_i\beta_l$, and $\text{Cov}\left(\frac{1}{\sigma_{ij}^2}, \frac{1}{\sigma_{lk}^2}\right) = 0$, where $j \neq k$, $i \neq l$, $j, k \in \{1, ..., J\}$, and $i, l \in \{1, ..., I\}$.

Given α_i's and β_i's, the tuning parameter α controls the correlation of the inverse of error variances of a gene across experiments. In simulations 5.2 and 5.4, we set $\alpha = 2$. This was set such that the simulated data in simulations 5.1 and 5.3 are comparable with that in simulations 5.2 and 5.4. That is, the BAGE method gives us about the same estimated values of σ_ε^2 for simulations 5.1–5.4. When α decreases, the estimated σ_ε^2 increases, and there is less information to borrow across genes or across experiments. In Appendix G, we discuss the cases where the BAGE method has more advantage or less advantage over the LIMMA method with different parameter values.

Appendix G. Results and Discussion When Changing Hyperparameter Values

In the lognormal model (3), σ_ε^2 quantifies the departure from additivity of gene and experiment effects. When σ_ε^2 is large, we are not able to borrow much information across genes or experiments. This can be understood more easily if we consider the situation where all experiments have the same degrees of freedom as stated in formulae (8)–(10). When σ_ε^2 is nearly as large as σ_E^2 and σ_G^2, the opportunity to borrow information across genes and experiments is limited. This is the case where the BAGE method has the least advantage over the LIMMA method. In the case where σ_ε^2 is small, regardless of the values of σ_G^2 and σ_E^2, according to formula (8) in the main paper, the BAGE method can borrow considerable information either across genes or across experiments or both depending on the values of σ_G^2 and σ_E^2. When σ_E^2 is small and σ_G^2 is large, BAGE estimates can be dramatically improved over LIMMA estimates by borrowing information provided in other experiments. When σ_E^2 is large and σ_G^2 is small, BAGE estimates still show substantial improvement over LIMMA though not as much as the previous case.

Although in applying our method to the aerenchyma study in maize roots, σ_ε^2 was estimated much smaller than σ_E^2 and σ_G^2, and σ_E^2 was estimated to be smaller than σ_G^2, we want to discuss more possible cases and outcomes. We simulated four cases as below (μ was set as 0), where

case 1 $\sigma_G^2 \approx \sigma_E^2 \approx \sigma_\varepsilon^2$;
case 2 $\sigma_G^2 > \sigma_E^2 > \sigma_\varepsilon^2$;
case 3 $\sigma_E^2 > \sigma_G^2 > \sigma_\varepsilon^2$;
case 4 $\sigma_E^2 \approx \sigma_G^2 > \sigma_\varepsilon^2$.

The parameter values for the four simulations are listed in Table 6. For each case, we simulated data for 40 experiments using the lognormal model for error variances. Each experiment contained 6 slides evenly allocated to two treatment groups. In each experiment, we simulated observations for 1,000 genes, of which 500 genes were randomly selected to be differentially expressed. We analyzed these data by the LIMMA method, and then by the BAGE method using four separate groups of 10 experiments each. The average results of 40 experiments for each method are listed in Table 6. AUCs are calculated for the region where FPr is no larger than 0.05. V, R, and V/R are calculated by using Storey and Tibshirani (2003a) procedure to control FDR at 0.05 level. In case 1, the BAGE method does not have much advantage over the LIMMA method with some minor improvement for estimation and AUC. However, in cases 2 and 4, BAGE has substantial improvement over LIMMA in both estimation and testing. In case 3, the BAGE method shows improvement over LIMMA in estimation and testing, but not as much as in cases 2 and 4. Based on the above discussion, we suggest to use the BAGE method to analyze experiments when σ_ε^2 is estimated to be smaller than σ_E^2 and σ_G^2.

Table 6: Parameter values and simulation results for four different cases.

Case	σ_G^2	σ_E^2	σ_ε^2	Method	Bias	MSE	AUC	V	R	V/R
1	.20	.20	.20	LIMMA	-.239	0.676	.021	11.18	216.38	.051
				BAGE(10)	-.120	0.526	.022	8.05	202.73	.039
2	.50	.20	.05	LIMMA	-.293	1.012	.019	10.83	200.43	.053
				BAGE(10)	-.076	0.384	.023	8.58	226.78	.037
3	.20	.50	.05	LIMMA	-.216	0.728	.022	11.43	227.78	.049
				BAGE(10)	-.063	0.360	.024	9.05	227.98	.039
4	.50	.50	.05	LIMMA	-.368	1.993	.019	10.15	188.38	.053
				BAGE(10)	-.080	0.718	.023	9.13	225.53	.039

Appendix H. Complementary Contour Plots

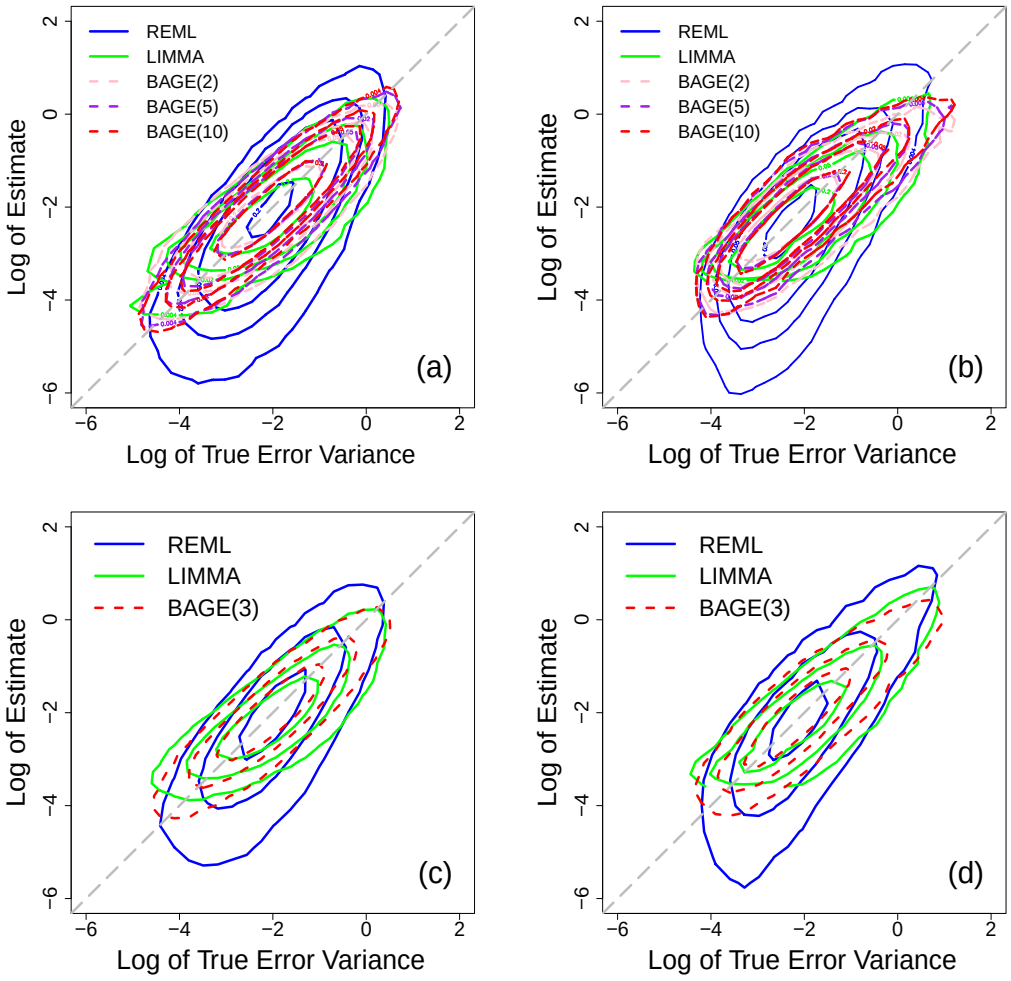

Figure 3: Contour plots of simulations 5.1 – 5.4. (a) simulation 5.1. Equal degrees of freedom across experiments with true error variances sampled from the lognormal model. (b) simulation 5.2. Equal degrees of freedom across experiments with true error variances sampled from the inverse gamma model. (c) simulation 5.3. Different degrees of freedom across experiments with true error variances sampled from the lognormal model. (d) Different degrees of freedom across experiments with true error variances sampled from the inverse gamma model.

References

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

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004), "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, 103, 2771–2778.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005), "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, 6, 59–75.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Fan, J., Chen Y., Chan H. M., Tam, P. K. H., and Ren, Y. (2005), "Removing intensity effects and identifying significant genes for Affymetrix arrays in macrophage migration inhibitory factor-suppressed neuroblastoma cells," *Proceedings of the National Academy of Sciences*, 102, 17751–17756.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, 531–537.

Hwang, J. T. G. and Liu, P. (2010), "Optimal tests shrinking both means and variances applicable to microarray data analysis," *Statistical Applications in Genetics and Molecular Biology*, 9, Article 36.

Lo, K. and Gottardo, R. (2007), "Flexible empirical Bayes models for differential gene expression," *Bioinformatics*, 23, 328–335.

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

Nakazono, M., Rajhi, I., Takahashi, H., Shiono, K., Ohtsu, K., Tsutsumi, N., Ji, T., Nettleton, D., Schnable, P., and Nishizawa, N. K. (2009), "Identification of genes expressed during aerenchyma formation in maize roots using laser microdissection and a microarray," *International Symposium of Root Research and Applications*, 5, 36–40.

Rao, C. R. and Toutenburg, H. (1999), *Linear Models: Least Squares and Alternatives*, 2nd edition. New York: Springer.

Smyth, G. K. (2004), "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 1, Article 3.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of Royal Statistical Society: Series B*, 64, 479–498.

Storey, J. D. and Tibshirani, R. (2003a), "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Storey, J. D. and Tibshirani, R. (2003b), *SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays*, New York: Springer.

Tong, T. and Wang, Y. (2007), "Optimal shrinkage estimation of variances with applications to microarray data analysis," *Journal of American Statistical Association*, 102, 113–122.

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

Xie, Y., Pan, W., and Khodursky, A. B. (2005), "A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data," *Bioinformatics*, 21, 4280–4288.

Yang, H. and Churchill, G. (2007), "Estimating p-values in small microarray experiments," *Bioinformatics*, 23, 38–43.