Biometrics 62, 10–18
March 2006

DOI: 10.1111/j.1541-0420.2005.00397.x

Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples

Raphael Gottardo,[1,*] Adrian E. Raftery,[1] Ka Yee Yeung,[2] and Roger E. Bumgarner[2]

[1]Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.
[2]Department of Microbiology, University of Washington, Box 358070, Seattle, Washington 98195, U.S.A.
[*]*email:* raph@stat.washington.edu

Summary. We consider the problem of identifying differentially expressed genes under different conditions using gene expression microarrays. Because of the many steps involved in the experimental process, from hybridization to image analysis, cDNA microarray data often contain outliers. For example, an outlying data value could occur because of scratches or dust on the surface, imperfections in the glass, or imperfections in the array production. We develop a robust Bayesian hierarchical model for testing for differential expression. Errors are modeled explicitly using a t-distribution, which accounts for outliers. The model includes an exchangeable prior for the variances, which allows different variances for the genes but still shrinks extreme empirical variances. Our model can be used for testing for differentially expressed genes among multiple samples, and it can distinguish between the different possible patterns of differential expression when there are three or more samples. Parameter estimation is carried out using a novel version of Markov chain Monte Carlo that is appropriate when the model puts mass on subspaces of the full parameter space. The method is illustrated using two publicly available gene expression data sets. We compare our method to six other baseline and commonly used techniques, namely the t-test, the Bonferroni-adjusted t-test, significance analysis of microarrays (SAM), Efron's empirical Bayes, and EBarrays in both its lognormal–normal and gamma–gamma forms. In an experiment with HIV data, our method performed better than these alternatives, on the basis of between-replicate agreement and disagreement.

Key words: Affymetrix; Bayesian hierarchical model; Bonferroni adjustment; cDNA microarrays; Empirical Bayes; Heteroscedasticity; Markov chain Monte Carlo; Mixture distribution; Outlier; Singular distribution; t-distribution.

1. Introduction

cDNA microarrays (Schena et al., 1995) consist of thousands of individual DNA sequences printed on a high-density array on a glass microscope slide using a robotic arrayer. A microarray works by exploiting the ability of a given labeled cDNA molecule to bind specifically to, or hybridize to, a complementary sequence on the array. By using an array containing many DNA samples, scientists can measure—in a single experiment—the expression levels of hundreds or thousands of genes within a cell by measuring the amount of labeled cDNA bound to each site on the array. In a typical two-color microarray experiment, two messenger RNA (mRNA) samples, from control and treatment situations, are compared for gene expression. Both mRNA samples are reverse-transcribed into cDNA, labeled using different fluorescent dyes (red and green dyes), and then mixed and hybridized with the arrayed DNA sequences. The hybridized arrays are then imaged to measure the red and green intensities for each spot on the glass slide. The estimates of the red and green intensities are the starting point of any statistical analysis. Similarly, microarrays can be used to compare the mRNA levels of thousands of genes under several experimental or biological conditions by using several samples. One of the main research areas is to detect genes that are differentially expressed across the different conditions.

In recent years, there has been a considerable amount of work on the detection of differentially expressed genes. An early statistical treatment can be found in Chen, Dougherty, and Bittner (1997). A common approach is to test a hypothesis for each gene and then try to correct for multiple testing. Most of the statistics used are variants of t- or F-statistics. This was done by Dudoit et al. (2002) using Welsh's t-statistic with p-values estimated by permutations. Tusher, Tibshirani, and Chu (2001) and Baldi and Long (2001) used a modification of the t-statistic where the denominator was modified by adding a constant to improve the estimate of the standard deviation.

In each of these situations, two types of error can occur: a false positive (type I error) or a false negative (type II error). When many hypotheses are tested at the same time, the chance of making a type I error increases. One approach to

overcoming this problem is to try to control the total number of type I errors or false positives. This can be done using multiple testing procedures to control some measure of the overall type I error. The two most common measures in the area of microarrays are the familywise error rate (FWER), which is the probability of making at least one type I error, and the false discovery rate (FDR), which is the proportion of false positives among the total number of discoveries reported. One of the most used FWER-based adjustments is that of Bonferroni. Tusher et al. (2001) used a permutation technique to estimate and control the FDR.

Newton et al. (2001) and Kendziorski et al. (2003) introduced an empirical Bayes approach to detect changes in gene expression on cDNA slides. Efron (2004), extending the work of Efron et al. (2001), used an empirical Bayes approach to detect differentially expressed genes with a two-component mixture model. Ibrahim, Chen, and Gray (2002) and Tadesse, Ibrahim, and Mutter (2003) have introduced more fully Bayesian approaches where "exact" estimation is carried out by Markov chain Monte Carlo (MCMC).

In this article, we introduce a Bayesian hierarchical model to test for differentially expressed genes in a robust way based on the previous work of Gottardo et al. (2003b) to estimate the mean intensities from replicates. Robustness is achieved by using a hierarchical t-formulation (Besag and Higdon, 1999), which is more robust to outliers than the usual Gaussian model. The model includes an exchangeable prior for the variances, allowing each gene to have a different variance while still achieving some shrinkage. We elaborate this model by introducing a prior that allows us to detect differentially expressed genes in multiple-sample experiments, where the number of samples can be greater than two. The prior is written as a mixture of singular Gaussian distributions. We show how one can use MCMC to estimate the parameters, even though the model contains a component that is a mixture of singular distributions. Inference is based on the posterior probabilities of differential expression calculated from our model. We call our method BRIDGE (Bayesian Robust Inference for Differential Gene Expression).

The article is organized as follows. Section 2 introduces the data structure and the notation. In Section 3 we present the Bayesian hierarchical model, and in Section 4 we show how it is used to test for differential expression. Section 4 also reviews six other baseline and commonly used methods to test for differentially expressed genes in the two-sample case. In Section 5, we apply the methods to experimental data and compare the results. In Section 6, we show how one can extend our model to multiple-sample experiments and use it to detect differentially expressed genes in a three-sample experiment. Finally, in Section 7 we discuss our results and possible extensions.

2. Data

We used two data sets that are fairly typical of data in this area. We use the word "sample" to describe different experimental or biological conditions.

The HIV data: This data set, described by Wout et al. (2003), consists of four experiments using the same RNA preparation on four different slides. The expression levels of 7680 cellular RNA transcripts were assessed in CD4-T-cell lines at time $t = 24$ hours after infection with HIV virus type 1. This data set contains two samples, one of which corresponds to the HIV-infected cells and the other to noninfected cells. It also contains 12 HIV-1 genes used as positive controls, i.e., genes known in advance to be differentially expressed. This data set is the result of a balanced dye-swap experiment. Two of the four (technical) replicates were hybridized with the green dye (Cy3) for the control and the red dye (Cy5) for the treatment, then the dyes were reversed on the other two replicates. After the image analysis, the data take the form y_{isr}, $i = 1, \ldots, I$; $s = 1, 2$; $r = 1, \ldots, R$, where y_{isr} is the log transformed estimated intensities of gene i in sample s from replicate r.

The BRCA data: Hedenfalk et al. (2001) conducted a study to examine breast cancer tissues from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2, or from patients not expected to carry a hereditary mutation. Hedenfalk et al. (2001) examined 22 breast cancer tumor samples: 7 tumors with BRCA1, 8 tumors with BRCA2, and 7 sporadic tumors, i.e., with neither mutation. In these data, "samples" refer to tissue sample types and there is no color swap. A set of 3226 genes was pre-selected by Hedenfalk et al. (2001) by filtering the raw images. The data take the form $y_{isr} \equiv \log_2(x_{isr}/\text{ref}_{ir})$, $i = 1, \ldots, I$; $s = 1, 2, 3$; $r = 1, \ldots, R_s$, where x_{isr} is the intensity from gene i of the rth (biological) replicate in sample s, and ref_{ir} is the intensity from a common reference sample. Note that here, Hedenfalk et al. (2001) used a reference sample because there are three samples of interest: BRCA1, BRCA2, and sporadic.

Each data set was normalized so that the mean of the log expression values across genes in each experiment is zero. There are more elaborate normalization techniques (Tseng et al., 2001; Yang et al., 2002) but we do not address normalization issues in this article, and we assume that it was done as a preprocessing step.

Both data sets are available at `www.stat.washington.edu/raph/data`.

3. Differential Expression with Two Samples

In this section, we introduce the Bayesian hierarchical model used to test for differentially expressed genes in the two-sample case. Our model extends the one used by Gottardo et al. (2003b) to allow it to detect differentially expressed genes with two samples. We consider two types of design: direct comparison using cDNA microarrays, and indirect comparison using cDNA microarrays or oligonucleotide arrays (e.g., Affymetrix). The difference lies in how the errors are modeled. In the case of direct comparison with cDNA microarrays, we use a bivariate distribution with nonzero correlation, because measurements come in pairs. In the case of indirect comparison designs and oligonucleotide arrays, there is no need for this additional feature. In each situation, we model the measurements from each gene as the sum of a sample effect and an error term. The sample effects and error variances are assumed to arise from a genomewide distribution with hyperparameters specific to each sample.

3.1 *The Models*

3.1.1 *Direct comparison using cDNA microarrays.* Two-color microarrays can be used to compare two samples of

interest directly. Each measurement consists of a pair of observations (y_{i1r}, y_{i2r}) from the two samples. The model is as follows:

$$\begin{aligned} y_{isr} &= \gamma_{is} + \frac{\epsilon_{isr}}{\sqrt{w_{ir}}}, \\ (\epsilon_{i1r}, \epsilon_{i2r})' \,|\, \mathbf{V}_i &\sim \mathrm{N}_2(\mathbf{0}, \mathbf{V}_i), \\ (w_{ir} \,|\, \nu_r) &\sim \mathcal{G}a(\nu_r/2, \nu_r/2), \end{aligned} \tag{1}$$

where w_{ir} and $(\epsilon_{i1r}, \epsilon_{i2r})'$ are independent. Since the w's are independent of the ϵ's, we have $\epsilon_{isr}/(w_{ir})^{1/2} \sim \mathcal{T}_{(\nu_r, \mathbf{0}, \mathbf{V}_i)}$, i.e., the (bivariate) errors have a bivariate t-distribution with ν_r degrees of freedom and covariance matrix $\mathbf{V}_i$.

For a given gene, the correlation matrix, $\mathbf{V}_i$, allows the measurements from the two samples to be correlated and each gene to have its own variance. The precision matrix (i.e., the inverse of the covariance matrix) is given by

$$\left(\mathbf{V}_i^{-1} \,\middle|\, \rho, \lambda_{\epsilon_{i1}}, \lambda_{\epsilon_{i2}}\right) = \frac{1}{(1-\rho^2)} \begin{pmatrix} \lambda_{\epsilon_{i1}} & -\sqrt{\lambda_{\epsilon_{i1}}\lambda_{\epsilon_{i2}}}\rho \\ -\sqrt{\lambda_{\epsilon_{i1}}\lambda_{\epsilon_{i2}}}\rho & \lambda_{\epsilon_{i2}} \end{pmatrix},$$

where ρ is the correlation between samples and $\lambda_{\epsilon_{is}}$ is the precision of gene i in sample s.

3.1.2 *Indirect comparison and oligonucleotide arrays.* In this case, the experimental design can be unbalanced and there is no physical reason to model the observation as bivariate. We modify the model as follows:

$$\begin{aligned} y_{isr} &= \gamma_{is} + \frac{\epsilon_{isr}}{\sqrt{w_{isr}}}, \\ \left(\epsilon_{isr} \,|\, \lambda_{\epsilon_{is}}\right) &\sim \mathrm{N}\left(0, \lambda_{\epsilon_{is}}^{-1}\right), \\ (w_{isr} \,|\, \nu_r) &\sim \mathcal{G}a(\nu_r/2, \nu_r/2), \end{aligned} \tag{2}$$

where w_{isr} and ϵ_{isr} are independent. It follows that $\epsilon_{isr}/(w_{isr})^{1/2}$ has a t-distribution with ν_r degrees of freedom and scale parameter $\lambda_{\epsilon_{is}}^{-1}$.

In both of models (1) and (2), we model γ_{is}, the effect of sample s on gene i, as a random effect with a mixture of two singular Gaussian distributions, i.e.,

$$\begin{aligned} (\boldsymbol{\gamma_i} \,|\, \boldsymbol{\lambda_\gamma}, p) \sim{}& (1-p)\mathrm{N}\left(\gamma_{i1}; 0, \lambda_{\gamma_{12}}^{-1}\right)\mathbf{1}_{[\gamma_{i1}=\gamma_{i2}]} \\ &+ p\mathrm{N}\left(\gamma_{i1}; 0, \lambda_{\gamma_1}^{-1}\right)\mathrm{N}\left(\gamma_{i2}; 0, \lambda_{\gamma_2}^{-1}\right)\mathbf{1}_{[\gamma_{i1}\neq\gamma_{i2}]}, \end{aligned} \tag{3}$$

where $\boldsymbol{\gamma_i} = (\gamma_{i1}, \gamma_{i2})'$ and $\boldsymbol{\lambda_\gamma} = (\lambda_{\gamma_1}, \lambda_{\gamma_2}, \lambda_{\gamma_{12}})$. The first component corresponds to the genes that are not differentially expressed ($\gamma_{i1} = \gamma_{i2}$), while the second component corresponds to the genes that are differentially expressed ($\gamma_{i1} \neq \gamma_{i2}$). Note that the formulation is not standard as it is not absolutely continuous with respect to two-dimensional Lebesgue measure. However, it defines a proper distribution with respect to a more general dominating measure, namely, the sum of a one-dimensional Lebesgue measure on the line $\gamma_{i1} = \gamma_{i2}$ and the two-dimensional Lebesgue measure (Gottardo and Raftery, 2004).

Finally, we use an exchangeable prior for the precisions, so that information is shared between genes, namely, $\lambda_{\epsilon_{is}} \sim \mathcal{G}a(a_{\epsilon_s}^2/b_{\epsilon_s}, a_{\epsilon_s}/b_{\epsilon_s})$, i.e., a gamma distribution for each sample with mean a_{ϵ_s} and variance b_{ϵ_s}.

3.2 *Priors*

We use a vague but proper prior for the parameters $\lambda_{\gamma_{12}}$, λ_{γ_1}, λ_{γ_2} of the distributions of the sample effects γ_{is} in (3). This is exponential with mean 200, so that $\lambda_\gamma \sim \mathcal{G}a(1, 0.005)$.

We also use vague but proper priors for the error precisions, specified by $a_{\epsilon_s} \sim \mathcal{U}_{[0,1000]}$ and $b_{\epsilon_s} \sim \mathcal{U}_{[0,1000]}$. The prior for the correlation between the two samples is given by $\rho \sim \mathcal{U}_{[-1,1]}$. The prior for the mixing parameter, p, is uniform over [0, 1]. The prior for the degrees of freedom, ν_r, is uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$.

3.3 *Parameter Estimation*

Realizations were generated from the posterior distribution using MCMC algorithms (Gelfand and Smith, 1990). All updates are straightforward except for $\boldsymbol{\gamma}$, for which the update is nonstandard since the distribution is formed by two singular components. However, there is a common dominating measure and so the Metropolis–Hastings algorithm can be used (Gottardo and Raftery, 2004). In the case of model (1), if the errors were taken to be independent, i.e., $\rho = 0$, the full conditional of $\boldsymbol{\gamma}$ would be given by

$$\begin{aligned} (\boldsymbol{\gamma}_i \,|\, \ldots) \propto{}& (1-p)c_i\mathrm{N}\left(\gamma_{i1}; \mu_i^*, \lambda_i^{*-1}\right)\mathbf{1}_{[\gamma_{i1}=\gamma_{i2}]} \\ &+ pc_{i1}c_{i2}\mathrm{N}\left(\gamma_{i1}; \mu_{i1}^*, \lambda_{i1}^{*-1}\right)\mathrm{N}\left(\gamma_{i2}; \mu_{i2}^*, \lambda_{i2}^{*-1}\right)\mathbf{1}_{[\gamma_{i1}\neq\gamma_{i2}]}, \end{aligned}$$

where

$$\lambda_i^* = \sum_{r,s} w_{ir}\lambda_{\epsilon_{is}} + \lambda_{\gamma_{12}}, \quad \mu_i^* = \lambda_i^{*-1}\sum_{r,s} w_{ir}\lambda_{\epsilon_{is}}y_{isr},$$

and

$$\lambda_{is}^* = \lambda_{\epsilon_{is}}\sum_{r} w_{ir} + \lambda_{\gamma_s}, \quad \mu_{is}^* = \lambda_{is}^{*-1}\sum_{r} w_{ir}\lambda_{\epsilon_{is}}y_{isr}.$$

The constants c_i, c_{i1}, and c_{i2} are given by

$$c_i = \sqrt{\frac{\lambda_{\gamma_{12}}}{\lambda_i^*}}\exp\left\{-0.5\sum_{r,s} w_{ir}\lambda_{\epsilon_{is}}y_{isr}^2 + 0.5\lambda_i^{*-1}\left(\sum_{r,s} w_{ir}\lambda_{\epsilon_{is}}y_{isr}\right)^2\right\},$$

and

$$c_{is} = \sqrt{\frac{\lambda_{\gamma_s}}{\lambda_{i1}^*}}\exp\left\{-0.5\lambda_{\epsilon_{is}}\sum_{r} w_{ir}y_{isr}^2 + 0.5\lambda_{is}^{*-1}\left(\sum_{r} w_{ir}\lambda_{\epsilon_{is}}y_{isr}\right)^2\right\}.$$

To update $\boldsymbol{\gamma}$, one draws new pairs (γ_{i1}, γ_{i2}) from the null component of the full conditional with probability $(1-p^*) \equiv (1-p)c/[pc_1c_2 + (1-p)c]$, or from the other component with probability p^*. When $\rho \neq 0$, we used the full conditional given above as proposal and corrected the acceptance probability with a Hastings correction factor. This gave a high acceptance rate, about 90% in the example presented here. In the case of model (2), the errors are independent and the Gibbs sampler as described above can be used, replacing w_{is} by w_{isr}.

Trace plots and autocorrelation plots were used as convergence diagnostic tools. For each data set presented here, we

found that a sample of 50,000 iterations with 1000 burn-in iterations and storing every tenth iteration was sufficient to produce reliable results. Guided by this, and leaving some margin, we used 100,000 iterations with 5000 burn-in iterations, and stored every tenth iteration after the burn-in period. This took about 9 hours for the HIV data and 6 hours for the BRCA data (BRCA1 and BRCA2 samples only) on a Linux workstation with an Intel Xeon processor at 3.06 GHz. An `R` software package called `bridge` implementing the method is available from Bioconductor at `www.bioconductor.org`.

4. Methods to Be Compared

In this section, we describe seven methods for detecting differentially expressed genes with cDNA microarrays. These seven methods will be compared in Section 5.

BRIDGE: From models (1) and (2) we can compute the marginal posterior probability of differential expression of gene i, namely, $\Pr(\gamma_{i1} \neq \gamma_{i2} \,|\, \mathbf{y})$. For each gene i, the marginal posterior probability of differential expression corresponds to the posterior probability that $\gamma_{i1} \neq \gamma_{i2}$ given the data. For a given posterior sample S of size B, we estimate the posterior probabilities by $\frac{1}{B}\sum_{k\in S} \mathbf{1}_{[\gamma_{i1}^{(k)} \neq \gamma_{i2}^{(k)}]}$, where $\gamma_{i1}^{(k)}$ and $\gamma_{i2}^{(k)}$ are the values generated at the kth MCMC iteration and $\mathbf{1}_{[\gamma_{i1}^{(k)} \neq \gamma_{i2}^{(k)}]}$ is the indicator function, equal to one if $\gamma_{i1}^{(k)} \neq \gamma_{i2}^{(k)}$. When the Gibbs sampler is available for updating $\boldsymbol{\gamma}$, which is the case for model (2), Rao–Blackwellization could be used to improve the estimates. One would need to replace the indicator function by the conditional probability $P(\gamma_{i1}^{(k)} \neq \gamma_{i2}^{(k)} \,|\, \cdots)$, i.e., the probability that $\gamma_{i1}^{(k)} \neq \gamma_{i2}^{(k)}$ given all the other parameters. Inference is based on the posterior probabilities and we discuss possible thresholds at the end of this section.

Raw and Bonferroni-adjusted t-tests: A classical procedure for testing a null hypothesis about the mean of a distribution or the equality of two means is the t-test. Here we apply one-sample t-tests to log-ratios (i.e., HIV data), or two-sample t-tests on the log measurements in the case of oligonucleotide arrays or designs with a reference sample, such as the BRCA data. Because of the large number of hypotheses tested we also report the results of adjusting the p-values using the Bonferroni adjustment. The Bonferroni adjustment method controls the FWER, which is the probability of yielding one or more false positives. If P is the raw p-value, then the Bonferroni-adjusted p-value is $\min\{IP, 1\}$. We declare a gene to be differentially expressed if its raw or adjusted p-value is less than 0.05.

Significance analysis of microarrays (SAM): This is a statistical technique for finding significant genes in a set of microarray experiments, proposed by Tusher et al. (2001). SAM uses regularized t-tests where the estimate of the standard deviation is regularized with a common estimate of the standard deviation and controls an estimate of the FDR that is the proportion of falsely identified genes among the genes declared to be differentially expressed. Using the SAM software (Chu et al., 2002), we select the largest rejection region with estimated FDR less than 0.1. In the literature, FDR values between 5% and 10% are commonly used (Tusher et al., 2001; Efron, 2004). In our experiments, the results were not very sensitive to the choice of the FDR value and we used 10%.

Empirical Bayes (EBarrays) lognormal–normal and gamma–gamma models: Newton et al. (2001) developed a method for detecting changes in gene expression in a single two-channel cDNA slide using a hierarchical gamma–gamma model. Kendziorski et al. (2003) extended this to replicate chips with multiple conditions, and provided the option of using a hierarchical lognormal–normal (LNN) model. For the gamma–gamma (GG) model, the observation component is a gamma distribution with mean γ and scale α/γ. The coefficient of variation α is taken to be constant across genes, and α/γ is assumed to have an inverse gamma distribution. For the LNN model, the observation component is a lognormal distribution with mean parameter γ and variance parameter σ^2, taken to be the same for all the genes. The conjugate prior for γ is normal with mean γ_0 and variance τ_0^2. For both models, the prior can be integrated out and the EM algorithm can be used to estimate the unknown parameters. Inference is based on the posterior probabilities of differential expression (Kendziorski et al., 2003) and we discuss possible thresholds at the end of this section.

Efron's Empirical Bayes: Efron (2004) used an empirical Bayes approach combined with a local version of the FDR to test for differential expression. He uses a two-component mixture to model z-scores (gene specific summary statistics) and detect differentially expressed genes. In the two-sample case, the z-scores are obtained from two-sample t-statistics; see Efron (2004). For each gene, inference is based on the local false discovery rate $\text{fdr}(z) \equiv f_0(z)/f(z)$ where f_0 is the empirical null density and f is the mixture density estimated from the observed data. Efron (2004) used a 10% FDR to call a gene differentially expressed.

Using posterior probabilities, a common rule is to declare a gene to be differentially expressed if its posterior probability is greater than 0.5, which corresponds to the usual 0-1 loss. To compare EBarrays and BRIDGE to the different cutoffs used in SAM and Efron's method, we also used two other posterior probability thresholds for declaring genes, one comparable to SAM and the other comparable to Efron's method. The first threshold controls the FDR at 10%, which corresponds to selecting the largest list having average null posterior probability less than 0.1 (Genovese and Wasserman, 2002). The second threshold is 0.9.

5. Results

5.1 *HIV Data*

We fitted model (1), described in Section 3.1, to the HIV data. The posterior modes of the degrees of freedom of the t-distribution, ν_r, ranged from 4 to 100, indicating that the sampling errors can be heavier-tailed than the Gaussian distribution and that the proportion of outliers varies from array to array. There is substantial between-sample correlation, estimated as 0.56, even after removing the sample effects.

The proportion of differentially expressed genes is estimated to be 0.007. Figure 1 is a plot of the posterior probabilities against the posterior means of the log-ratios computed from our model. A relatively small number of genes seem to be differentially expressed. To evaluate the effect of the t-distribution, we fitted the model given by (1) replacing the t-errors with Gaussian errors. The t-based model clearly leads to a more powerful test than the Gaussian as the

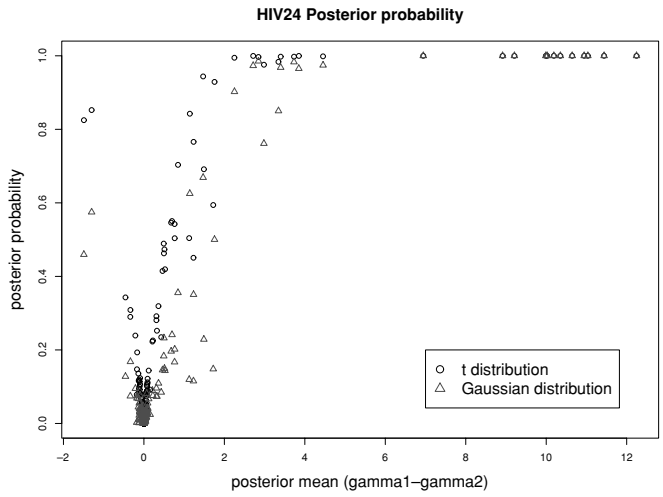

Figure 1. Posterior probabilities from the BRIDGE method with both Gaussian and t-errors plotted against the posterior differences between γ_1 and γ_2 (estimated log-ratios) from the model with t-distribution for the HIV24 data. Most of the log-ratios are shrunk close to zero and have very low posterior probabilities of differential expression. The use of the t-distribution increases the posterior probabilities of expression for several of the genes.

corresponding posterior probabilities are higher (Figure 1). The model with the t-distribution detects more genes, 35, as against 25 at the 0.5 posterior threshold. Table 1 illustrates this for two genes containing outliers. The posterior probabilities of expression from the model with the t-distribution are more than twice as large as those from the Gaussian model. This is because the downweighting of extreme values leads to smaller estimated variances and hence a higher posterior probability of differential expression for genes such as those in Table 1.

In comparison, the unadjusted t-test yields many p-values less than 0.05. More than 700 genes have raw p-values less than 0.05 whereas only two have adjusted p-values less than 0.05. The adjustment is clearly too conservative since we know for sure from external information that at least 12 genes are

Table 1
Log-ratios of two genes of the HIV data sets. The BRIDGE posterior probabilities were computed using both t-distributed errors and Gaussian errors. The use of the t-distribution downweights the potential outliers and increases the posterior probability of expression.

	Replicates				Post. prob.	
	1	2	3	4	Gauss.	t
Gene 1	1.29	0.75	2.39	1.82	0.20	0.51
Weights	0.99	0.94	0.60	0.88		
Gene 2	1.35	1.67	2.77	1.18	0.35	0.77
Weights	1.00	1.14	0.50	0.98		

Note: The weights are the posterior means of the w's computed from our model.

differentially expressed. SAM, EBarrays GG and LNN (at the 0.5 threshold), and Efron's method report 125, 83, 75, and 122 differentially expressed genes, respectively, while BRIDGE (at the 0.5 threshold) reports 35 differentially expressed genes. BRIDGE and EBarrays GG and LNN controlling the FDR at 10% detect 33, 86, and 81 genes, respectively. BRIDGE and EBarrays GG and LNN, with a cutoff posterior probability of 0.9 (comparable to Efron's method with 10% local FDR), detect 23, 63, and 60 genes, respectively. All the methods except the t-test with adjusted p-values correctly detect the 12 positive control genes, i.e., those known to be differentially expressed.

In order to evaluate the performance of each method, we divided the four replicates of the HIV data into groups of two replicates. We did this division in both the possible ways that preserve the dye-swap design. We applied each method to each group of replicates and looked at the agreement and disagreement between the genes declared to be differentially expressed. All the methods except the t-test (both raw and adjusted p-values) detected the 12 positive controls. Overall, the number of genes declared to be differentially expressed was smaller when two replicates were used than when four replicates were used, for all the methods except EBarrays and Efron's method.

Table 2 shows the number of agreements and disagreements between the two groups of replicates, for each of the seven methods we consider. The raw and adjusted t-tests both performed poorly: the raw t-test recorded a very high level of disagreement, presumably corresponding to a large number of false positives, while the adjusted t-test was clearly too conservative. Among the other methods, the numbers of genes on which there was agreement were comparable, but EBarrays and Efron's method recorded much larger numbers of genes on which there was disagreement than BRIDGE or SAM. Comparing BRIDGE with SAM, BRIDGE recorded agreement on more genes than SAM, with lower levels of disagreement for each group. It thus seems that for this particular data set, using the criterion of agreement between replicates, BRIDGE did better than the other methods at identifying differentially expressed genes.

5.2 *BRCA Data*

This time, we fitted model (2) described in Section 3.1 to the BRCA data. The posterior modes of the degrees of freedom of the t-distribution, ν_r, again ranged from 4 to 100, indicating that the sampling errors can be heavier-tailed than the Gaussian distribution and that the proportion of outliers varies from array to array. The proportion of differentially expressed genes is estimated to be 0.36, which is much larger than for the HIV data and indicates that more genes are differentially expressed. This is consistent with the results of Hedenfalk et al. (2001) where the author observed that the BRCA1 and BRCA2 mutations differed significantly in their global patterns of gene expression.

We compare the numbers of genes declared to be differentially expressed by each method. This time we only used the LNN model, because if the intensity measurements arise from a lognormal distribution, then so should the ratio. However this is not the case for a gamma distribution. Efron's method and the t-test with adjusted p-values seem very conservative

Table 2

Agreement and disagreement about the differential expression of genes in the HIV data when the four replicates are divided into two sets of two. For each method, "Agreement" denotes the number of genes declared to be differentially expressed based on both sets of two replicates, while "Disagreement" refers to the number of genes that were declared to be differentially expressed based on one set of two replicates, and not to be differentially expressed based on the other set of two replicates. Using BRIDGE and EBarrays, we report three numbers, the first two corresponding to posterior probability thresholds of 0.5 *and* 0.9*, while the third controls the FDR at* 10%*.*

		Rep. 1&3 vs. 2&4		Rep. 1&4 vs. 2&3	
		Agreement	Disagreement	Agreement	Disagreement
BRIDGE	PP > 0.5	18	11	18	6
	PP > 0.9	14	7	15	4
	FDR < 0.1	18	12	20	5
SAM	FDR < 0.1	17	16	16	6
EBarrays GG	PP > 0.5	22	78	22	46
	PP > 0.9	20	54	22	28
	FDR < 0.1	23	82	22	50
EBarrays LNN	PP > 0.5	30	133	26	99
	PP > 0.9	26	100	24	59
	FDR < 0.1	31	143	26	103
Efron's method	Local FDR < 0.1	22	308	44	399
t-test	Raw $p < 0.05$	25	467	21	427
	Adj. $p < 0.05$	0	0	0	0

with only zero and two genes declared differentially expressed, respectively. BRIDGE, SAM, and EBarrays controlling the FDR at 10% detect about the same numbers of genes: 291, 374, and 375, respectively. BRIDGE and EBarrays with a cutoff posterior probability of 0.9, which is comparable to Efron's method with 10% local FDR, still detect 157 and 153 genes. BRIDGE with posterior probabilities greater than 0.5 detect more genes than EBarrays with the same threshold, 880 against 415. The different methods are also compared in the next section when using the full BRCA data.

6. The Multiple Sample Case

Even though the models introduced in Section 3.1 are intended to test for differential expression between two samples, they can be extended to situations where there are more than two samples and differences in expression of the same gene between any two samples may be of interest. Here the alternative hypothesis is not as simply defined as before, because there are many possible patterns of differential gene expression. In order to account for all possible patterns, the prior for the γ's in (3) needs to be modified. Because most microarray technologies allow the direct comparison of at most two samples, we need only to introduce a single model, which is similar to (2). We need to show how the prior for $\boldsymbol{\gamma}$ can be modified for the case where we have three samples. The generalization to more than three samples should be straightforward.

6.1 *The Model*

The general model is the same as (2) with three samples, i.e., $s = 1, 2, 3$. We need only to modify the distribution of $\boldsymbol{\gamma}$, the vector of sample effects in each sample, to allow each possible pattern of differential expression to have positive probability. We still model it as a random effect, but this time with a mixture of five singular Gaussian distributions, as follows:

$$
\begin{aligned}
&(\boldsymbol{\gamma} \,|\, \boldsymbol{\lambda_\gamma}, \boldsymbol{p}) \\
&\quad \sim\ p_1 \mathrm{N}\big(\gamma_1; 0, \lambda_{\gamma_1}^{-1}\big)\mathbf{1}_{[\gamma_1=\gamma_2=\gamma_3]} \\
&\qquad + p_2 \mathrm{N}\big(\gamma_1; 0, \lambda_{\gamma_1}^{-1}\big)\mathrm{N}\big(\gamma_2; 0, \lambda_{\gamma_{23}}^{-1}\big)\mathbf{1}_{[\gamma_1\neq\gamma_2=\gamma_3]} \\
&\qquad + p_3 \mathrm{N}\big(\gamma_2; 0, \lambda_{\gamma_2}\big)\mathrm{N}\big(\gamma_1; 0, \lambda_{\gamma_{13}}^{-1}\big)\mathbf{1}_{[\gamma_1=\gamma_3\neq\gamma_2]} \\
&\qquad + p_4 \mathrm{N}\big(\gamma_3; 0, \lambda_{\gamma_3}^{-1}\big)\mathrm{N}\big(\gamma_1; 0, \lambda_{\gamma_{12}}^{-1}\big)\mathbf{1}_{[\gamma_1=\gamma_2\neq\gamma_3]} \\
&\qquad + p_5 \mathrm{N}\big(\gamma_1; 0, \lambda_{\gamma_1}^{-1}\big)\mathrm{N}\big(\gamma_2; 0, \lambda_{\gamma_2}^{-1}\big)\mathrm{N}\big(\gamma_3; 0, \lambda_{\gamma_3}^{-1}\big)\mathbf{1}_{[\gamma_1\neq\gamma_2\neq\gamma_3]}, \quad (4)
\end{aligned}
$$

where $\boldsymbol{\lambda_\gamma} = (\lambda_{\gamma_1}, \lambda_{\gamma_2}, \lambda_{\gamma_3}, \lambda_{\gamma_{12}}, \lambda_{\gamma_{13}}, \lambda_{\gamma_{23}}, \lambda_{\gamma_{123}})$ is the vector of precisions, and $\boldsymbol{p}$ is the vector of probabilities for the five patterns, constrained to sum to one. The five components correspond to all five possible patterns of expression. As before, the formulation is not standard since it is not absolutely continuous with respect to the three-dimensional Lebesgue measure, but it does define a proper distribution with respect to a more general dominating measure (Gottardo and Raftery, 2004).

We kept the priors described in Section 3.2. All the λ_γ's have exponential prior distributions with mean 200. The mixing probabilities have a prior distribution that is uniform over the simplex $S = \{\mathbf{p} : \sum_i p_i = 1\}$, i.e., the prior is $\mathcal{D}(1, 1, 1, 1, 1)$, Dirichlet with common hyperparameters 1. Due to the independence of the errors, the full conditional for $\boldsymbol{\gamma}$ is available and so Gibbs sampling can be used as described in Section 3.3.

6.2 *Results*

We fitted the model given by (2) with three samples to the BRCA data. The posterior modes of the degrees of freedom of

Table 3

Estimates of the mixing probabilities for the five patterns of expression, and the numbers of genes classified into each pattern on the BRCA data from model (2) *with three samples, and from EBarrays. A gene was classified into a pattern if the corresponding posterior probability of its conforming to that pattern was greater than* 0.5. *The null pattern has the highest average posterior probability according to both methods.*

		$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$
BRIDGE	Mix. prob.	0.66	0.11	0.22	0.01	0.0006
	No. of genes	2415	89	356	17	0
EBarrays	Mix. prob.	0.79	0.067	0.11	0.019	0.00025
(LNN)	No. of genes	2721	147	272	36	3

the t-distribution, ν_r, ranged as low as 4, indicating that the sampling errors can be more heavy-tailed than the Gaussian distribution. The posterior mixing probabilities, $p_1, \ldots, p_5$, of the five patterns of differential expression, are summarized in Table 3 by their posterior means. The mixing probabilities clearly favor the null pattern of no differential expression, suggesting that most of the genes are not differentially expressed. They also suggest that there is more difference between the BRCA1 and the BRCA2 tumors than any BRCA with sporadic tumors. This confirms results from Hedenfalk et al. (2001).

EBarrays is similar in this case to what it was for the two-sample case described in Section 4; see Kendziorski et al. (2003) for further details. Efron's method and SAM can also be used in a three-sample context. In Efron's method, one can simply replace the t-test by an F-test, and in SAM the modified t-statistic is replaced by a modified F-statistic (Chu et al., 2002).

For each gene, we wish to compare the five different patterns of differential expression given by each component of (4), which we denote by $\mathcal{P}_i$. For each gene and pattern, we can compute the posterior probability that the gene conforms to that pattern. We computed the BRIDGE posterior probabilities from our model and the posterior probabilities using EBarrays. For comparison purposes, we followed Kendziorski et al. (2003) in classifying a gene as conforming to a pattern if the corresponding posterior probability was greater than 0.5. The numbers of genes classified into each pattern are quite different; see Table 3.

This may be due to two main differences between the two methods. First, EBarrays models within-gene variances as constant, while BRIDGE allows for them to vary between genes. If within-gene variances do indeed vary between genes, then a method such as EBarrays that assumes constant variance may be more likely to incorrectly classify genes with high variances. Second, in its lognormal–normal version, EBarrays assumes measurement errors to be normally distributed, while BRIDGE allows their distribution to have heavier tails. If the tails are indeed heavier, for example if there are outliers, then a method such as EBarrays that assumes normality may be more likely to (incorrectly) declare genes with outlying measurements in some replicates to be differentially expressed.

To illustrate the latter point, consider Figure 2 which shows the expression levels for the three genes classified into pattern $\mathcal{P}_5$ by EBarrays. The posterior probabilities of being in pattern $\mathcal{P}_5$ reported by EBarrays are 0.99, 0.75, and 0.96.

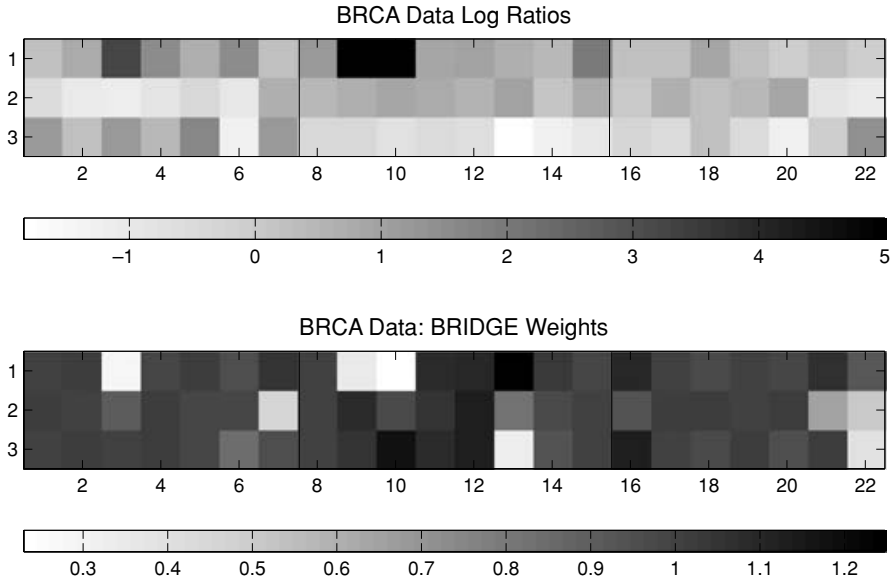

Figure 2. The three genes with the greatest EBarrays posterior probabilities of conforming to pattern $\mathcal{P}_5$ (all three samples are different) based on the LNN model (top). The corresponding posterior weights from our model (bottom). The first seven samples (1–7) correspond to BRCA1 tumors, the next eight samples (8–15) correspond to BRCA2 tumors, and the last seven samples (16–22) correspond to sporadic tumors. Several possibly outlying samples were heavily downweighted by our model. The two dark outliers, measurements 9 and 10 from gene 1, were truncated to 5 for clarity. The actual data values are 6.64 and 7.13, respectively.

The same three genes are not classified into the $\mathcal{P}_5$ pattern by our method but into patterns $\mathcal{P}_1, \mathcal{P}_3$, and $\mathcal{P}_3$, respectively, with BRIDGE posterior probabilities 0.74, 0.58, and 0.58. The first gene in Figure 2 seems to contain several possible outliers, e.g., measurements 9 and 10. The posterior mean of the weights, $\boldsymbol{w}$, from our model shows that these values have been downweighted. As a result, the gene is not declared to be differentially expressed in any way by our method. Similarly, the sporadic tumor samples of the two other genes (samples 16–22) contain several possible outliers that are downweighted by our model. The three tumor types no longer seem differentially expressed and our method classifies the two genes into pattern $\mathcal{P}_3$, though the posterior probabilities of being in model $\mathcal{P}_2$ were also high at 0.41 and 0.31, respectively. Visual inspection suggests that the choices made by EBarrays were quite influenced by a few extreme values, which our method downweights.

Looking at the numbers of genes that are significantly different from the null pattern, Efron's method is still conservative with no genes declared differentially expressed. BRIDGE, SAM, and EBarrays controlling the FDR at 10% detect about the same numbers of genes: 252, 324, and 447, respectively. BRIDGE and EBarrays with a cutoff posterior probability of 0.9, which is comparable to Efron's method with 10% local FDR, detect 133 and 295 genes. BRIDGE with posterior probabilities greater than 0.5 detects more genes than EBarrays with the same threshold, 806 against 505. These results are consistent with the two-sample case.

7. Discussion

We have developed a framework for testing for differential expression in gene expression arrays, in a way that is robust to outlying measurements and powerful even with a small number of replicates. Our Bayesian hierarchical model is based on a model used by Gottardo et al. (2003b) to estimate microarray intensities in a robust way. We modified the model by using a novel form of prior that allows us to detect differentially expressed genes in multiple-sample experiments. When there are three or more samples, the model allows us to detect the differentially expressed genes, and also to classify them into the different patterns of differential expression. In an experiment with HIV two-sample data, we compared our method with six other baseline and commonly used methods, and it performed better, at least in terms of agreement and disagreement between groups of replicates. Our model requires more computing than some other methods because it involves MCMC, and users would need to decide whether the improved results are worth the additional computing time.

We assume that normalization was done as a preprocessing step but it is possible to include normalization effects in the model (Gottardo et al., 2003b). We also assumed that the genes were independent and that measurements for a given gene were independent too. While it would be hard to incorporate dependence among the genes, it would be possible to introduce some genewise correlation for the measurement errors. This was not done here but could be done as in Kendziorski et al. (2003). As we have shown in the three-sample case, our model can easily be generalized to test for differential expression among an arbitrary number of samples. However, the number of components needed and the associated number of parameters grow very rapidly. It will be hard to fit the model when the number of samples is too large. One solution would be to consider only certain patterns of expression to reduce the number of components in the mixture (Kendziorski et al., 2003).

In order to compare our method with others, we identified a gene as differentially expressed if the posterior probability of its being so was greater than a given threshold. In practice, though, we would often not use a cutoff, but instead would report the posterior probabilities themselves. A biologist could then choose to do further research on a number of the most likely genes, taking account of resource constraints, or to study genes whose posterior probability exceeds a pre-specified threshold. The posterior probabilities from our model are well calibrated and easy to interpret.

In this article, we have compared our model with six alternatives, but there are many other methods for detecting differentially expressed genes with gene expression data. We chose these six because they are either obvious baseline methods or widely used; they are also representative of other methods. For example, there are several other empirical Bayes methods that we could have used. These include the lognormal–normal models of Lönnstedt and Speed (2002) and Gottardo et al. (2003a) and the less parametric approaches of Efron et al. (2001) and Newton et al. (2004). More comparisons between statistical tests can be found in Cui and Churchill (2003). Among explicit adjustments for multiple testing, we considered only the Bonferroni adjustment and the FDR control methods given by SAM and Efron's method; these are widely used and easy to understand. But other adjustments for multiple comparisons have been proposed, and we refer the reader to Dudoit, Shaffer, and Boldrick (2003) for a recent review.

Acknowledgements

The authors thank Nema Dean for careful reading of the manuscript, Angelique van't Wout for providing us with some of the data, John Storey for helpful discussions about FDR, and four anonymous referees and the associate editor for suggestions that clearly improved an earlier draft of the article. This research was supported by NIH grant 8 R01 EB002137-02, and Raftery's research was also partially supported by ONR grant N00014-01-10745, Yeung's research was funded by NIH-NCI grant 1K25CA106988-01, and Bumgarner's research was funded by NIH-NIAID grants 5P01 AI052106-02 and 1U54AI057141-01, NIH-NIEHA grant 1U19ES011387-02, NIH-NHLBI grants 5R01HL072370-02 and 1P50HL073996-01, and NIH-NCRR grant 1S10RR019423-01.

References

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

Besag, J. E. and Higdon, D. M. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B* **61,** 691–746.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA

microarray images. *Journal of Biomedical Optics* **2,** 364–374.

Chu, G., Narasimham, B., Tibshirani, R., and Tusher, V. (2002). *SAM "significance analysis of microarrays," users guide and technical document.* Stanford University.

Cui, X. and Churchill, G. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4,** 210.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12,** 111–139.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18,** 71–103.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

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.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Genovese, C. and Wasserman, L. (2002). *Bayesian and frequentist multiple testing.* Technical Report, Carnegie Mellon University.

Gottardo, R. and Raftery, A. (2004). *Markov chain Monte Carlo computations with mixture of singular distributions.* Technical Report 470, Statistics Department, University of Washington, Seattle.

Gottardo, R., Pannucci, J. A., Kuske, C. R., and Brettin, T. (2003a). Statistical analysis of microarray data: A Bayesian approach. *Biostatistics* **4,** 597–620.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. (2003b). *Robust estimation of cDNA microarray intensities.* Technical Report 438, Statistics Department, University of Washington, Seattle.

Hedenfalk, I., Duggan, D., Chen, Y., et al. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344,** 539–548.

Ibrahim, J., Chen, M. H., and Gray, R. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97,** 88–99.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22,** 3899–3914.

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

Newton, M. C., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8,** 37–52.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* **270,** 467–470.

Tadesse, M., Ibrahim, J., and Mutter, G. (2003). Identification of differentially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometrics* **59,** 542–554.

Tseng, G. C., Oh, M., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29,** 2549–2557.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98,** 5116–5121.

van't Wout, A. B., Lehrma, G. K., Mikheeva, S. A., O'Keeffe, G. C., Katze, M. G., Bumgarner, R. E., Geiss, G. K., and Mullins, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of $CD4^+$-T-cell lines. *Journal of Virology* **77,** 1392–1402.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30,** e15.

Received July 2004. *Revised March* 2005.
Accepted April 2005.