Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 892927, 12 pages
doi:10.1155/2008/892927

Research Article

Bayesian Mixture Model Analysis for Detecting Differentially Expressed Genes

Zhenyu Jia[1] and Shizhong Xu[2]

[1] *Department of Pathology & Laboratory Medicine, University of California, Irvine, CA 92697, USA*
[2] *Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA*

Correspondence should be addressed to Zhenyu Jia, zhenyu.jia@gmail.com

Received 27 August 2007; Accepted 18 November 2007

Recommended by Nengjun Yi

Control-treatment design is widely used in microarray gene expression experiments. The purpose of such a design is to detect genes that express differentially between the control and the treatment. Many statistical procedures have been developed to detect differentially expressed genes, but all have pros and cons and room is still open for improvement. In this study, we propose a Bayesian mixture model approach to classifying genes into one of three clusters, corresponding to clusters of downregulated, neutral, and upregulated genes, respectively. The Bayesian method is implemented via the Markov chain Monte Carlo (MCMC) algorithm. The cluster means of down- and upregulated genes are sampled from truncated normal distributions whereas the cluster mean of the neutral genes is set to zero. Using simulated data as well as data from a real microarray experiment, we demonstrate that the new method outperforms all methods commonly used in differential expression analysis.

1. INTRODUCTION

Current microarray technology allows us to measure the expression of thousands of genes simultaneously in a small chip. Many advanced statistical methods have been developed to analyze data generated from this technology. For example, a Bayesian model has been proposed to normalize gene expression data and select genes that are differentially expressed [1]. Jörnsten et al. [2] developed a Bayesian method to impute missing values prior to any statistical analysis. Efron et al. [3], Broët et al. [4], Edwards et al. [5], Do et al. [6] used a Bayesian approach to identify differentially expressed genes. Methods of [7–9] are examples for Bayesian clustering analysis.

In this study, we focus on detecting differentially expressed genes in two-sampled designs, in which only two conditions (control and treatment) are examined and each condition is replicated several times. The data are assumed to have been properly imputed and normalized as needed. Much effort has been made on this study to detect genes whose expression levels respond to the treatment.

Numerous methods have been suggested based on the P-values of a test statistic. The P-values are obtained from separate tests and reported for all genes [10, 11]. Cui and Churchill [12] reviewed the test statistics for differential expression for microarray experiments. The gene-specific summary statistic provides a basis for the rank ordering of genes and the creation of a short list of genes declared as differentially expressed. However, genes are studied separately in the sense that calculating the test statistic for one gene does not use information from other genes. In limma (linear model for microarray data), Smyth [13] cleverly borrowed information from the ensemble of genes to make inference for individual gene based on the moderate t-statistic. Some other researchers also took advantages of shared information by examining data jointly. Efron et al. [3] proposed a mixture model methodology implemented via an empirical Bayes approach. For each gene, the expression levels were mapped to a single z-score, which is assumed to arise from two distributions (affected and unaffected). Pan [14] used a mixture model to cluster genes based on gene specific t-statistic. They are examples where the idea of clustering was first used in differential expression analysis (clustering methods are usually applied to microarray data generated from multiple conditions). Similarly, the methods of [4–6, 15] all used a mapped quantity, like the z-score or the t-statistic, to classify genes

into different groups. These methods are different from the proposed Bayesian mixture model in that genes were clustered based on a summary statistic for these methods while our Bayesian method clusters genes based on the pattern of expression. It is more desirable to use the expression profile than to use a summary score for cluster analysis because information is bound to be lost when mapping the expression profile into a single score. Recently, Newton et al. [16] developed a new method in which gene expression are directly used for classification. Newton et al. [16] used a hierarchical model that consists of two levels. The first level of the model describes the conditional distributions of the observed measurement of gene expression given the means of the control (μ_{1i}) and the treatment (μ_{2i}), where i indexes the gene. The second level describes the distribution of μ_{1i} and μ_{2i} jointly by a mixture of three multivariate distributions with constrained parameters of $\mu_1 > \mu_2$, $\mu_1 = \mu_2$, and $\mu_1 < \mu_2$. Genes become linked by virtue of having μ_{1i} and μ_{2i} drawn from a common distribution. Parameter estimation was accomplished via the Expectation-Maximization (EM) algorithm [17].

For the Bayesian clustering method presented in this study, the observed gene expression levels are described by a regression model as done by Gusnanto et al. [15]. For each gene, the irrelevant intercept is removed by a special normalization scheme. The slope of the regression represents the difference of expression under the two conditions. The Bayesian method is implemented via the Markov chain Monte Carlo (MCMC) algorithm. The regression coefficient of each gene is assumed to be sampled from a mixture of three normal distributions with constrained parameters. The three distributions are $N(\beta_k, \nu_k)$ for $k = 1, 2, 3$, where $\beta_1 < 0$, $\beta_2 = 0$, and $\beta_3 > 0$ are the constrained parameters. The proposed new method actually turns the problem of a complicated multivariate mixture distribution of Newton et al. [16] into that of a univariate mixture distribution.

The new Bayesian method proposed in this study (Method I) is compared to five different methods that are commonly used in differential expression analysis. These four methods, called Methods II, III, IV, V, and VI respectively, are described below. Method II is the regularized t-test (SAM, Tusher et al. [10]), in which a t-like score is calculated for each gene. Genes with scores greater than an adjusted threshold are said to have altered expressions under the treatment. Permutations are used to calculate the false discovery rate (FDR, Benjamini and Liu [18]). Method III is the model-based cluster analysis of Pan [14] where the variable used for clustering is the t-test statistic. Genes that are assigned into a cluster with a mean t value significantly different from zero are declared to be differentially expressed. Method IV is the hierarchical mixture model method of Newton et al. [16], where the expected expression values of the ith gene under the control (μ_{1i}) and the treatment (μ_{2i}) are assumed to arise from a mixture of three distributions. Clustering is actually conducted based on the latent parameters (μ_{1i} and μ_{2i}). The cluster of genes with $\mu_{1i} \neq \mu_{2i}$ are declared as differentially expression genes. A short list of significant genes is created based on FDR at 0.05. Method V is the linear model and empirical Bayes method (limma) of Smyth [13]. In limma, a hierarchical linear model is used to describe the gene expression levels. A moderated t-statistic, with extra information borrowed from other genes, is calculated for each gene. Adjusted P-values are calculated for genes based on the moderated t-statistics to achieve FDR control. Method VI is the mixture mixed model analysis of Gusnanto et al. [15], where a similar linear model is used to describe gene expressions. However, the variable used for clustering is the mean difference between the treatment and the control. The mathematical differences between the proposed method and the four methods compared will be evident after the new method is introduced. Advantages of the new method over the four methods will be demonstrated in Section 3 and further addressed in the Section 4.

2. METHODS

2.1. Linear model

Let Y_{ij} ($i = 1, \ldots, m$ and $j = 1, \ldots, N$) be the log transformed expression level for gene i from chip j, where m is the total number of genes and N is the total number of chips. Let X_j be a dichotomous variable indicating the status of the chip with $X_j = 0$ for the control and $X_j = 1$ for the treatment. We use the following regression model to describe Y_{ij}:

$$Y_{ij} = \alpha_i + X_j \gamma_i + \varepsilon_{ij}, \tag{1}$$

where α_i is the intercept of the regression model, γ_i is the regression coefficient and ε_{ij} is the residual error. Note that the regression coefficient represents the difference in gene expression between the control and the treatment ($\mu_2 - \mu_1$). A large γ_i implies that gene i is differentially expressed. The intercept α_i represents the mean expression level of gene i in the control (μ_1). The intercept α_i is irrelevant to the differential expression analysis and thus should be removed from the model (see next section for detail). The regression coefficient γ_i is assumed to be sampled from one of three normal distributions, corresponding to the downregulated genes, the neutral genes, and the upregulated genes, respectively. The residual error is assumed to be $N(0, \sigma^2)$ distributed. In matrix notation, model (1) can be expressed as

$$Y_i = W\alpha_i + X\gamma_i + \varepsilon_i, \tag{2}$$

where $Y_i = \{Y_{ij}\}_{j=1}^{N}$ is a column vector for the expressions of gene i across all samples, $X = \{X_j\}_{j=1}^{N}$ is the incidence matrix (a column vector) for the linear model, $\varepsilon_i = \{\varepsilon_{ij}\}_{j=1}^{N}$ is a vector for the residual errors, and W is an $N \times 1$ unity vector, that is, a vector of 1 for all elements.

2.2. Normalization

As stated earlier, α_i is irrelevant to the differential expression analysis. We propose to remove the effect of α by using a linear contrast. This process, which is called normalization, is similar to the way of removing the fixed effects in the restricted maximum likelihood (REML) method for variance

component analysis [19]. For each gene, we subtract the observed expression level by the average level across all chips, that is,

$$Y_i^* = Y_i - W\overline{Y}_{i\cdot} \tag{3}$$

where $\overline{Y}_{i\cdot} = N^{-1}\sum_{j=1}^{N} Y_{ij}$ is the mean expression for gene i across all the chips. The mean expression can also be expressed in matrix notation as

$$\overline{Y}_{i\cdot} = (W^T W)^{-1} W^T Y_{i\cdot}. \tag{4}$$

Therefore, the normalized expression is described as a linear combination of the original expressions

$$Y_i^* = [I - W(W^T W)^{-1} W^T] Y_i = LY_i, \tag{5}$$

where

$$L = I - W(W^T W)^{-1} W^T. \tag{6}$$

We can prove that $LW = 0$. Therefore,

$$Y_i^* = LY_i = LX\gamma_i + L\varepsilon_{i\cdot} \tag{7}$$

Let $X^* = LX$ and $\varepsilon_i^* = L\varepsilon_i$, we have

$$Y_i^* = X^*\gamma_i + \varepsilon_i^*. \tag{8}$$

Now the intercept has been removed by using linear contrasts LY_i in place of the original Y_i. The residual variance-covariance matrix of the linear contrasts is $\mathrm{Var}(\varepsilon_i^*) = LL^T\sigma^2$. Note that the rank of matrix LL^T is $N - r(W) = N - 1$, indicating that the inverse of matrix LL^T does not exist. The model-based cluster analysis, however, requires the inverse. Therefore, we delete the last row of matrix L to form a new $(N-1)\times N$ matrix, called L^*, to build the linear contrasts. This treatment is equivalent to deleting the last element of Y_i^*, that is, the dimension of Y_i^* is $(N-1)\times 1$. The dimensions of matrix X^* and ε_i^* also change into $(N-1)\times 1$ accordingly. Let $R = L^*(L^*)^T$ be an $(N-1)\times(N-1)$ matrix of full rank. We now have $\mathrm{Var}(\varepsilon_i^*) = R\sigma^2$, which will be used in the model-based cluster analysis.

To simplify the notation, let us define $y_i = Y_i^*$, $x = X^*$ and $\epsilon_i = \varepsilon_i^*$, all with a dimension of $(N-1)\times 1$. The new model with the intercept removed becomes

$$y_i = x\gamma_i + \epsilon_i. \tag{9}$$

Note that the special way of normalization described above only applies to linear effects. Different methods should be used for adjusting nonlinear effects. In subsequent analysis, we assume that normalization via linear contrasts has been conducted and thus model (9) will be the basis for parameter estimation and statistical inference.

2.3. Mixture model and the Bayesian setup

Conditional on γ_i, the probability density of y_i is

$$p(y_i \mid \gamma_i, \sigma^2) = \phi_n(y_i; x\gamma_i, R\sigma^2), \tag{10}$$

where $n = N - 1$, $\phi_n(y_i; x\gamma_i, R\sigma^2)$ is the n-dimensional normal density of variable y_i with mean $x\gamma_i$ and variance-covariance matrix $R\sigma^2$. The gene-specific effect γ_i is assumed to be sampled from one of three normal distributions (clusters), $N(\beta_k, \nu_k)$ for $k = 1,2,3$. We define variable z_i as the cluster assignment of gene i, where $z_i = k$ if gene i comes from cluster k. For notational simplicity, we define $\eta(z_i,k)$ for $k = 1,2,3$ as a redundant variable to z_i, where $\eta(z_i,k) = 1$ if $z_i = k$ and $\eta(z_i,k) = 0$ otherwise. Let $\eta(z_i) = \{\eta(z_i,k)\}_{k=1}^3$ for $\eta(z_i,k) = 0,1$ and $\sum_{k=1}^3 \eta(z_i,k) = 1$. The new variable $\eta(z_i)$ will be used in place of z_i in subsequent derivation. Let $\beta = \{\beta_k\}_{k=1}^3$ and $\nu = \{\nu_k\}_{k=1}^3$. The density of the mixture distribution of γ_i is

$$p(\gamma_i \mid \eta(z_i), \beta, \nu) = \sum_{k=1}^{3} \eta(z_i,k)\phi_1(\gamma_i; \beta_k, \nu_k). \tag{11}$$

Note that variable $\eta(z_i,k)$ is unknown and it is treated as missing value. In fact, inferring the probability distribution of $\eta(z_i,k)$ is the main purpose of the proposed mixture model analysis. Define $\pi = \{\pi_k\}_{k=1}^3$ as the mixing proportions of the three components for $\pi_k \geq 0$ and $\sum_{k=1}^3 \pi_k = 1$. The distribution of $\eta(z_i)$ is multinomial with one observation,

$$p(\eta(z_i) \mid \pi) = \prod_{k=1}^{3} \pi_k^{\eta(z_i,k)}. \tag{12}$$

We have introduced the probability densities of the data, the regression coefficients, and the missing values for a single gene. We now combine the densities of individual genes to form the joint density of all genes. The probability density of the data $y = \{y_i\}_{i=1}^m$ is

$$p(y \mid \gamma, \sigma^2) = \prod_{i=1}^{m} p(y_i \mid \gamma_i, \sigma^2). \tag{13}$$

The density of $\gamma = \{\gamma_i\}_{i=1}^m$ is

$$p(\gamma \mid \eta, \beta, \nu) = \prod_{i=1}^{m} p(\gamma_i \mid \eta(z_i), \beta, \nu). \tag{14}$$

The density of $\eta = \{\eta(z_i)\}_{i=1}^m$ is

$$p(\eta \mid \pi) = \prod_{k=1}^{3} \pi_k^{m_k}, \tag{15}$$

where $m_k = \sum_{i=1}^m \eta(z_i,k)$ for $\sum_{k=1}^3 m_k = m$. The joint density of (y,γ,η) is

$$p(y,\gamma,\eta \mid \theta) = p(y \mid \gamma, \sigma^2) p(\gamma \mid \eta, \beta, \nu) p(\eta \mid \pi), \tag{16}$$

where $\theta = (\sigma^2, \beta, \pi, \nu)$ is the parameter vector.

The next step of the analysis is to find a suitable prior for θ. The following vague prior is chosen for each of the variance components, $\sigma^2 \sim \mathrm{Inv}-\chi^2(0,0)$ and $\nu_k \sim \mathrm{Inv}-\chi^2(0,0)$ for $k = 1,2,3$. They are called the scaled inverse chi-square

distributions with zero degrees of freedom and zero-scale parameter. The actual forms of the priors are $p(\sigma^2) = 1/\sigma^2$ and $p(\nu_k) = 1/\nu_k$ for $k = 1, 2, 3$. The scaled inverse chi-square distribution is conjugate, and thus the posterior distribution of the variance is also scaled inverse chi-square. The three clusters of genes are distinguished by the overall patterns of responses to the treatment. The first cluster consists of all the downregulated genes, the second cluster represents all the neutral genes, and the third cluster contains all the upregulated genes. The characteristics of the three clusters can be represented by enforcing the following constraints on the means of the three clusters: $\beta_1 < 0, \beta_2 = 0$, and $\beta_3 > 0$. Therefore, the prior distributions of the means of the three clusters are $p(\beta_k) = 1\ (\beta_k \in \Omega_k)$. Here, we adopt a special notation to indicate that $p(\beta_k) = 1$ if $(\beta_k \in \Omega_k)$ and $p(\beta_k) = 0$ if $(\beta_k \notin \Omega_k)$, where $\Omega_1 \equiv (\beta_1 < 0)$, $\Omega_2 \equiv (\beta_2 = 0)$, and $\Omega_3 \equiv (\beta_3 > 0)$. The prior for π is the multivariate generalization of the Beta distribution $p(\pi \mid \delta) = D(\pi \mid \delta, \delta, \delta)$, called the Dirichlet distribution, where $\delta = 1$ is used in this study. The joint prior for all the parameters is $p(\theta) = (1/\sigma^2)\prod_{k=1}^{3}(1/\nu_k)$. The posterior distribution of all the unknowns $p(\gamma, \eta, \theta \mid y)$ is proportional to the joint distribution of all variables,

$$\begin{aligned} p(\gamma, \eta, \theta \mid y) &\propto p(y, \gamma, \eta, \theta) \\ &= p(y \mid \gamma, \sigma^2)\, p(\gamma \mid \eta)\, p(\eta \mid \pi)\, p(\theta). \end{aligned} \tag{17}$$

Statistical inference of this distribution is the theme of the Bayesian analysis. There is no explicit form for the joint distribution (17). Therefore, we draw observations of the unknowns from the conditional distributions. Fortunately, with the current Bayesian setup, the conditional distribution of any single variable, given all other variables, has a known distribution. Therefore, the MCMC process can be proceeded exclusively using the Gibbs sampler.

2.4. Markov chain Monte Carlo

The detailed steps of the Markov chain Monte Carlo (MCMC) process are described as follows.

Step 1. Set $t = 0$ and initialize all variables $(\gamma^{(t)}, \eta^{(t)}, \beta^{(t)}, \sigma^{2(t)}, \nu^{(t)}, \pi^{(t)})$.

Step 2. Simulate $\gamma_i^{(t+1)}$ from $\gamma_i \sim N(\overline{\gamma}_i, s^2_{\gamma_i})$ conditional on $z_i = k$, where

$$\begin{aligned} s^2_{\gamma_i} &= \left[x^T R^{-1} x + \sigma^2/\nu_k\right]^{-1} \sigma^2 \\ \overline{\gamma}_i &= \left[x^T R^{-1} x + \sigma^2/\nu_k\right]^{-1} \left[x^T R^{-1} y + (\sigma^2/\nu_k)\beta_k\right], \end{aligned} \tag{18}$$

where all variables in the right-hand side take the most current values. In this step, the most current values of the variables in the right-hand side are the values at iteration t. This statement also holds for subsequent steps, except that the most current values of the variables in the right-hand side are values at iteration t or $t + 1$, depending on whether the variable has been updated or not in the current sweep.

Step 3. Simulate $\eta(z_i)^{(t+1)}$ from

$$\eta(z_i) \sim \text{Multinomial}\ (1, \pi_{i1}, \pi_{i2}, \pi_{i3}), \tag{19}$$

a multinomial distribution with one observation, where π_{ik} is

$$\pi_{ik} = \frac{\pi_k p(\gamma_i \mid \eta(z_i, k), \beta_k, \nu_k)}{\sum_{k'=1}^{3} \pi_{k'} p(\gamma_i \mid \eta(z_i, k'), \beta_{k'}, \nu_{k'})}. \tag{20}$$

Step 4. Simulate $\beta_k^{(t+1)}$ from a truncated normal distribution

$$\beta_k \sim N(\overline{\beta}_k, s^2_{\beta_k})\, 1(\beta_k \in \Omega_k). \tag{21}$$

The mean and the variance of the normal distribution are

$$\begin{aligned} \overline{\beta}_k &= \frac{1}{\sum_{i=1}^{m} \eta(z_i, k)} \sum_{i=1}^{m} \eta(z_i, k)\, \gamma_i, \\ s^2_{\beta_k} &= \frac{\nu_k}{\sum_{i=1}^{m} \eta(z_i, k)}, \end{aligned} \tag{22}$$

respectively. Note that $\beta_2 = 0$ is enforced and no sampling for β_2 is required. The general inverse transformation method [20] is used to sample β_1 and β_3 from the corresponding truncated normal distributions.

Step 5. Simulate $\sigma^{2(t+1)}$ from

$$\sigma^2 \sim \text{Inv} - \chi^2\left(mn, \sum (y_i - x\gamma_i)^T R^{-1} (y_i - x\gamma_i)\right), \tag{23}$$

where $m \times n$ is the degree of freedom, and the term with summation is the scale parameter of the inverse chi-square distribution.

Step 6. Simulate $\nu_k^{(t+1)}$ from

$$\nu_k \sim \text{Inv} - \chi^2\left(m_k, \sum \eta(z_i, k)\, (\gamma_i - \beta_k)^2\right), \tag{24}$$

where $m_k = \sum_{i=1}^{m} \eta(z_i, k)$.

Step 7. Simulate $\pi^{(t+1)}$ from

$$\pi \sim \text{Dirichlet}\ (m_1 + 1, m_2 + 1, m_3 + 1). \tag{25}$$

So far, every variable has been updated. We then increment t by $1 (t = t + 1)$ and repeat Steps 3–7 until the chain gets sufficiently long to allow reliable post-MCMC inference about the parameters of interest. Gene i is assigned to group k if $\overline{\eta}(z_i, k) = \max\{\overline{\eta}(z_i, k')\}_{k'=1}^{3}$, where $\overline{\eta}(z_i, k)$ is the posterior mean of $\eta(z_i, k)$ calculated from the posterior sample. Schemes for sampling variables from the aforementioned distributions are discussed by Gelman et al. [21].

3. APPLICATIONS

3.1. Analysis of simulated data

We simulated the expression of $m = 1000$ genes from six different groups on $N = 16$ microarray chips. Chips 1–8 represent the control, and chips 9–16 represent the treatment.

Table 1: Parameters used in the simulation experiment and their estimates from the Bayesian mixture model analysis.

Parameter		α_k	β_k	π_k	ν_k	σ^2
True	Group 1	3.5	1.2	0.005	0.01	0.03
	Group 2	4	0	0.345	0.04	
	Group 3	4.5	−1	0.02	0.02	
	Group 4	0	0	0.615	0.04	
	Group 5	0	0.8	0.005	0.01	
	Group 6	0	0.9	0.01	0.01	
True (combined)	Cluster 1	—	−1	0.020	0.02	0.03
	Cluster 2	—	0	0.960	0.04	
	Cluster 3	—	0.95	0.020	0.03	
Estimated	Cluster 1	—	−0.984	0.023	0.03	0.03
	Cluster 2	—	0	0.937	0.04	
	Cluster 3	—	0.702	0.040	0.11	

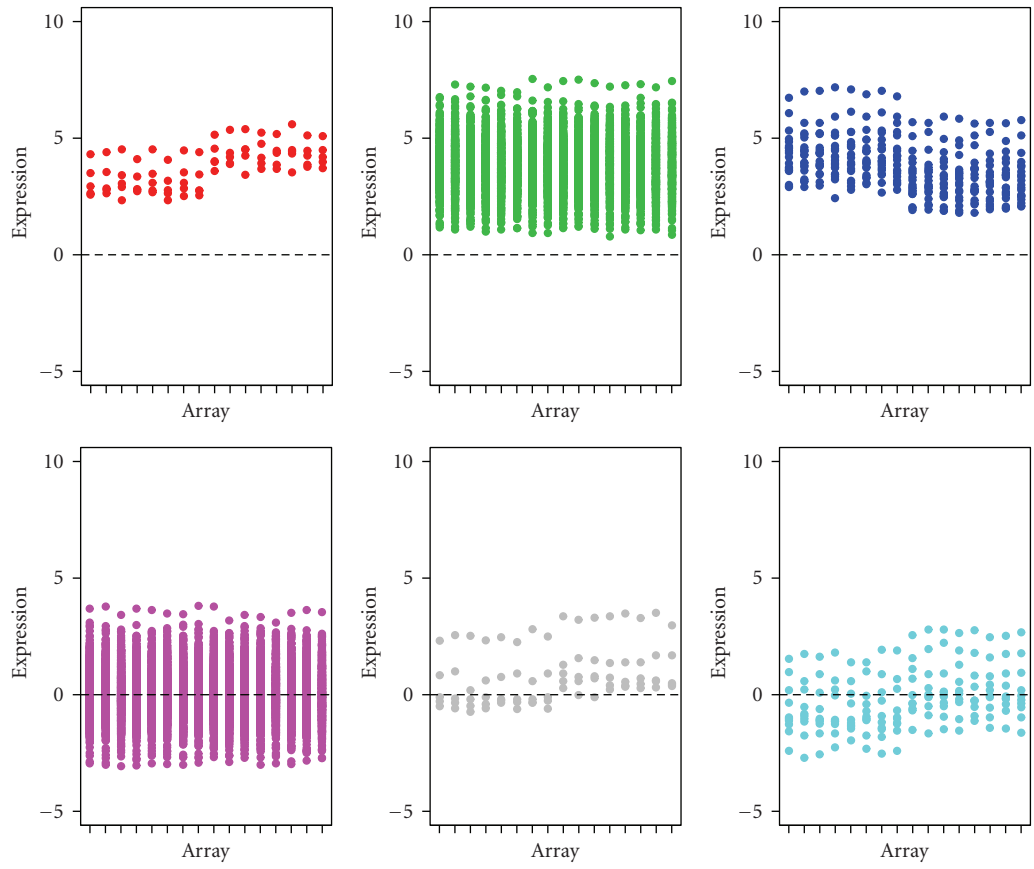

Figure 1: Original expression patterns for the six simulated groups of genes. In each plot, the first half represents the observed data from chips 1–8 and the second half represents the observed data from chips 9–16.

The true parameter values of the six groups of genes are shown in Table 1. If we ignore the intercepts, the six groups of genes really come from three functional clusters: Cluster 1 represents the ($m_1 = 20$) downregulated genes (Group 3); Cluster 2 contains the ($m_2 = 960$) neutral genes (Groups 2 and 4); and Cluster 3 consists of the ($m_3 = 20$) upregulated genes (Groups 1, 5, and 6). The true parameters of the original six groups and the parameters for the combined three clusters are given in Table 1. Note that there are actually three components (Group 1, 5, and 6) for the combined Cluster 3. The weighted β of three simulated components is treated as the true β for this cluster, that is, $1.2 \times 0.25 + 0.8 \times 0.25 + 0.9 \times 0.5 = 0.95$. The expression patterns are shown in Figure 1 for the original six groups and Figure 2 for the combined three clusters.

The data were analyzed using the Bayesian mixture model analysis reported in this study (Method I). We set the number of iterations equal to 12000 in the MCMC process.

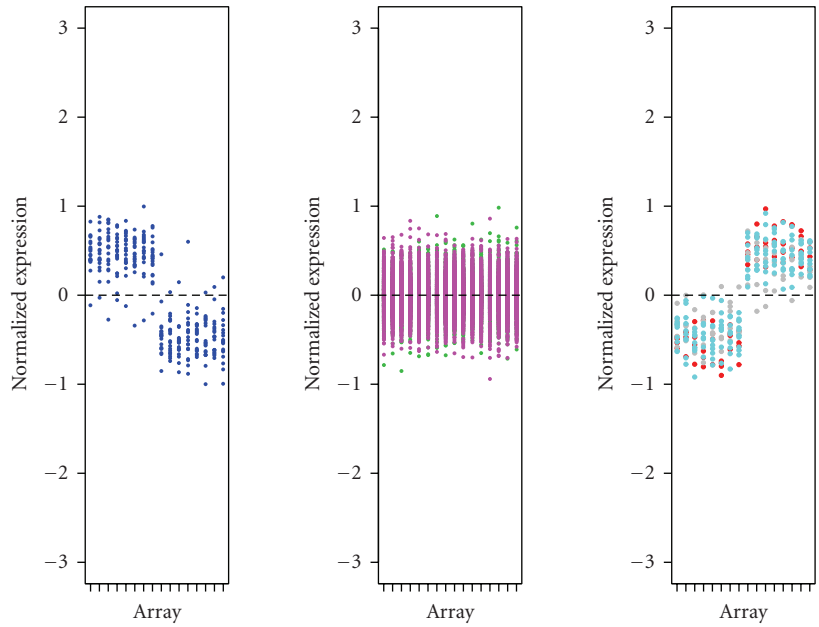

Figure 2: Expression patterns for the three combined clusters after normalization. See Figure 1 for the legends.

The results generated from the first 2000 iterations were discarded for the burn-in period. For the remaining 10000 iterations, we saved one observation in every 20 iterations to reduce series correlation between consecutive observations. Therefore, the posterior sample contained 500 observations. The posterior mean of each variable is considered as the Bayesian estimate. Hypothesis tests were only performed on the cluster means β rather than on individual genes. Genes assigned into a significant cluster were simultaneously declared to be differentially expressed. From Table 1, we can see that the estimated parameters agree well with the true parameters. Genes are correctly assigned into three clusters except that two genes in Cluster 2 (the neutral cluster) are incorrectly assigned into the other two clusters (down- and upregulated clusters) (see Table 2). Upon enforcing constraints on the cluster means, we are able to fix the number of clusters to three, rather than finding the optimal number of clusters using the Bayes factor or its BIC [22] approximation.

An interesting question is what would happen if the cluster means are not constrained? To answer this question, we also analyzed the simulated data with unconstrained β. We varied the number of clusters from 3 to 6 and found out that 4 was the optimal number of clusters (maximum value of Bayes factor). One cluster mean was close to zero. Another cluster had a negative mean. The third and fourth clusters all had positive means but the two means were almost identical. Therefore, Clusters 3 and 4 may be well considered as a single cluster. When treated as three clusters, the results were similar to those reported earlier when the cluster means were restricted (data now shown).

For comparison, the data were also analyzed using the five methods (Methods II–VI) described in Section 1. The numbers of genes classified into each of the three clusters are given in Table 2 for all the six methods, including Method I (the new method). Cutoff value 0.05 was used for gene detection in Method II and Method V to achieve FDR control. All methods have successfully found the 40 truly differentially expressed genes. However, Methods II, III, IV, V, and VI detected more false-positive genes than Method I. The model-based cluster analysis of Pan (Method III) required BIC to decide the optimal number of clusters. For simplicity of comparison, we only used the result of Method III under three clusters. We noticed that Method III always put the down- and upregulated genes into a cluster with a large variance, but placed the neutral genes into the other two cluster with small variances. The two clusters with small variances can be combined as a single cluster. Because all the methods have successfully detected the 40 nonneutral genes, none of them have suffered the Type II error. However, different methods tended to have different numbers of genes from the neutral cluster misplaced into the functional clusters. This means that different methods have different Type I errors.

Let $N(k,k')$ be the number of genes from cluster k assigned into cluster k' for $k,k' = 1,2,3$. Then $N(k) = \sum_{k'=1}^{3} N(k,k')$ is the number of genes in cluster k. Recall that Clusters 1 and 3 contain differentially expressed genes and Cluster 2 is reserved for the neutral genes. The empirical Type I error is then defined as

$$\text{Type I error} = \frac{N(2,1)+N(2,3)}{N(2)}. \tag{26}$$

The empirical Type II error is defined as

$$\text{Type II error} = \frac{N(1,2)+N(3,2)}{N(1)+N(3)}. \tag{27}$$

The empirical power is defined as

$$\text{Power} = 1 - \text{Type II error}. \tag{28}$$

Table 2: Numbers of genes assigned into each of the three clusters for six different methods of differential expression analysis. (The sum of each column within a method represents the true number of genes simulated from that cluster and the sum of each row represents the number of genes assigned into that cluster.)

Method	Estimate	True			Sum	Type I error
		Cluster 1	Cluster 2	Cluster 3		
I	Cluster 1	20	1	0	21	0.002
	Cluster 2	0	958	0	958	
	Cluster 3	0	1	20	21	
II	Cluster 1	20	205	0	225	0.459
	Cluster 2	0	519	0	519	
	Cluster 3	0	236	20	256	
III	Cluster 1	0	0	0	0	0.033
	Cluster 2	0	928	0	928	
	Cluster 3	20	32	20	72	
IV	Cluster 1	20	37	0	57	0.073
	Cluster 2	0	890	0	890	
	Cluster 3	0	33	20	53	
V	Cluster 1	20	153	0	173	0.325
	Cluster 2	0	648	0	648	
	Cluster 3	0	159	20	179	
VI	Cluster 1	20	5	0	25	0.011
	Cluster 2	0	949	0	949	
	Cluster 3	0	6	20	26	
	Sum (method)	20	960	20	1000	

The empirical Type I errors for all the five methods are listed in Table 2. Method I (the new method) have the smallest Type I error. Note that this method of validation is valid only for simulation study where the true gene assignment is known.

3.2. Analysis of mice data

We analyzed a cDNA microarray dataset published by Dudoit et al. [11]. The data are publicly available on the website (http://www.stat.berkeley.edu/users/terry). The data consist of expression measurements of 6342 genes from a study of lipid metabolism in mice [23]. This experiment was set up to find genes that are differentially expressed in the livers of mice with very low HDL cholesterol levels (treatment) in contrast to a group of normal inbred mice (control). The treatment group consists of eight scavenger receptor BI (SR-BI) transgenic mice and the control group consists of eight normal inbred mice.

We first analyzed the data with the Bayesian mixture model method proposed in this study (Method I). We used exactly the same setup of the MCMC as that used in the simulation experiment. The estimated parameters are given in Table 3. Only β_3 (cluster mean of up-regulated genes) is significantly different from zero. The mixing proportions of the three clusters show that majority of the genes (82.46%) are neutral. Note that the estimated mixing proportions ($\hat{\pi}_1 = 0.1724$, $\hat{\pi}_2 = 0.8246$, $\hat{\pi}_3 = 0.0030$) are some nuisance parameters, which do not necessarily match the actual numbers of genes assigned to the three clusters ($m_1 = 0$, $m_2 = 6332$, $m_3 = 10$). Report should be made based on the actual numbers of genes assigned to the clusters.

Table 3: Parameters estimated for the mice data using the Bayesian mixture model analysis.

	β_k	π_k	ν_k	σ^2
Cluster 1	−0.0065	0.1724	0.0616	0.0904
Cluster 2	0	0.8246	0.0006	
Cluster 3	0.6418	0.0030	2.0266	

All six methods that were used to analyze the simulated data were applied here to analyze the mice data. The numbers of genes assigned to the three clusters are given in Table 4 for all the five methods. Clearly, Methods I and VI were similar to each other and Methods II, III, IV, and V placed much more genes into the nonneutral clusters, a phenomenon which has been observed early in the simulation experiment.

In the original study [23], individual P-values were calculated based on the t-test statistics for all the genes and then the adjusted P-values were calculated using the method described in Westfall and Young [24]. Five genes (represented by eight array elements) were detected by Callow et al. [23] as differentially expressed between the treatment and the control (see Table 5). Note that several array elements may represent the same gene. We found that gene EST AI746730 is not included in the dataset and thus was not analyzed with any of the methods (marked as NA in Table 5).

Table 4: Numbers of genes assigned to the three clusters for the mice data for six different methods.

Method	Cluster 1 (down)	Cluster 2 (neutral)	Cluster 3 (up)
I	0	6332	10
II	40	6182	120
III	66	6276	0
IV	13	6300	29
V	12	6293	37
VI	0	6329	13

The Bayesian mixture model analysis detected 10 genes (or spots), four of which were reported in Callow et al. [23]. The expressions of the 10 genes are plotted in Figure 3 (the first two rows). Significant differences between the control and treatment groups can be easily seen. Note that element 2374 (the first plot of the second row of Figure 3) was a "blank" spot on the chip. However, it was detected by all methods due to the remarkable difference between the control and the treatment. The third row of the plots represents five randomly selected genes which were not detected by any of the methods. Genes of this kind have the same expression levels under the control and treatment. The last row of the plots (in Figure 3) shows the expression patterns of five genes that were not detected by the Bayesian mixture model analysis (Method I), but detected by the other five methods (Methods II, III, IV, V, and VI). When examining these five genes, we found that the differences between the treatment and the control are not obvious. The statistical significance may be caused entirely by the outliers. Therefore, the five methods (II, III, IV, V, and VI) are perhaps too sensitive to outliers. For example, the first plot (5203) of the last row in Figure 3 is the spot that represents β-globin in the dataset. We cannot tell much difference between the control and the treatment. Genes with numbers 2375, 2377, 2379, and 2384 are all "blank" spots. The observed differences between the control and the treatment are purely caused by chance, yet these blank spots were detected by Methods II, III, IV, V, or VI. Finally, Method III [14] and Method V [13] missed gene SR-BI at array element number 3, which is known to have altered the expression between the control and the SR-BI transgenic mice.

When we examined the plots of the ten genes detected by the Bayesian mixture model analysis, we found that seven of them have increased the expression level by the treatment while three of them have decreased the expression level by the treatment. Interestingly, our method assigned the three genes with negative regression coefficients into the cluster with positive regression coefficients. If we look at the estimated parameters (Table 3) again, we realized that both Clusters 1 and 2 may be considered as the neutral cluster (mean close to zero and variance very small). The third cluster contains both the up-and downregulated genes because it has a much larger variance than the other two clusters. The same phenomenon also occurred for Method VI [15], but the issue was not discussed by Gusnanto et al. [15]. A possible explanation is that there are too few genes with negative regression coefficients, which made them hard to form a single cluster by themselves. However, why were these three genes assigned into Cluster 3 ($\beta > 0$) instead of Cluster 2 ($\beta = 0$), which is closer to the three genes in terms of the cluster mean? The reason may be that Cluster 3 has a larger variance than Cluster 2.

A separate simulation experiment has been conducted to verify the notion that ν plays a more important role than β in calculating the posterior probability of cluster assignment π_{ik} (data not shown).

4. DISCUSSION

Similar to Method IV [16], our method is based on a hierarchical model in which the parameters of interest (gene effects) are further described by a mixture distribution. Clustering is made based on the parameter rather than on a summary statistic such as the t-like statistical score. This allows us to incorporate the error structure of the gene expression profile into the linear model (see (10)), and thus capture the most information from the data. This explains why our method is different from (or even better than) the other four methods (Methods II, III, V, and VI). But why is our method better than (or different from) Method IV? This may be contributed by the two differences between Methods I and IV. One difference is that the incorporated normalization scheme in our model allows us to deal with only the effect of differential expression (regression coefficient) whereas Method IV deals with effects of gene expressions in both the control and the treatment. In other words, we are dealing with a single variable (regression coefficient, the only parameter of interest) but Method IV deals with two variables (intercept and regression coefficients). The dimension reduction from two to one and the simplified Gaussian mixture distribution of our model may largely contribute to the higher efficiency of our method. The second difference between our method and Method IV is that we used the probabilities of cluster assignment to classify genes into three clusters, and genes assigned into the neutral cluster are excluded from the list of differentially expressed genes, but Method IV goes one step further by employing an FDR criterion to select a list of differentially expressed genes. It is obvious that the FDR generated list of genes depends largely on the subjective cutting rule set by the investigator. We consider that the extra step of FDR analysis after gene clustering is not only redundant but also leads to subjective decision for differential expression analysis.

Method II [10] sorts genes based on the P-value for a regularized t-statistic calculated from each gene. Information from other genes plays no role in calculating the P-value for the current gene of interest. This, combined with the subjectiveness of the cutting rule for gene selection, may largely explains the difference between Method II and the proposed new method.

Method III [14] uses a t-statistic as the data point for cluster analysis. The original gene expression profile is converted into the t score. Some information may be lost during the conversion because the method fails to incorporate the error for calculating the t score. In addition, the t score is the

Table 5: Some differentially expressed genes for the mice data detected by six different methods.

	Gene or element	Method					
		I	II	III	IV	V	VI
a	SR-BI (ID: 3)	√	√		√		√
	SR-BI (ID: 783)	√	√	√	√	√	√
	SR-BI (ID: 1581)	√	√	√	√	√	√
	Glutathione-S-transferase	√	√	√	√	√	√
	β-globin		√	√	√	√	
	Cytochrome P450 2B10						
	EST AI746730	NA	NA	NA	NA	NA	NA
b	Growth factor receptor bound pr2	√	√	√	√	√	√
	Creatine Kinase, muscle	√	√		√	√	√
	Ubiquitin-conjugating enzyme E2H	√	√	√	√	√	√
	Capping protein alpha 2	√	√	√	√	√	√
	Diff. Ass. Prot. 13 kD	√	√	√	√	√	√
	Blank (ID: 2374)	√	√	√	√	√	√
	Blank (ID: 2375)				√		
	Blank (ID: 2377)		√		√		
	Blank (ID: 2379)		√		√		√
	Blank (ID: 2384)				√		

a: genes reported in Callow et al. [23]; b: a subset of genes not reported in Callow et al. [23].

differential gene expression divided by the standard deviation of the difference. However, the calculated standard deviation is subject to large error when the number of replicates per condition is small, which is usually the case in microarray experiments.

Similar to Method II, Method V [13] creates a list of significant genes based on the moderated t-statistic. Extra information is borrowed, on the basis of the hierarchical model, from the ensemble of genes which can assist in inference about each gene individually. However, users need to subjectively specify the mixing proportions before the algorithm is applied. The authors pointed out that the estimations of mixing proportions are somewhat unstable and suggested using 0.01 as a universal prior. In real mice data, the proportion of differentially expressed genes is about 0.002, which is overestimated by limma. This may explain why limma detected too many false positives both in simulation study and mice data analysis.

Method VI [15] is quite similar to the proposed method except that the observed differential expression between the control and the treatment is used for cluster analysis instead of using the expected differential expression (parameter) as the basis for clustering. Again, the error structure of the expression profile is not properly incorporated into the model when the normalization-like procedure is used. This explains why the proposed new method outperforms Method VI. In addition, Method VI is based on the EM algorithm, while the proposed method is based on the MCMC algorithm. The EM algorithm sometimes may stuck in the so-called local optimal, while the MCMC has reasonable chance to jump out of the "trap." This may explain why Method VI failed in the situation where the borders of clusters are fuzzy with the effects of differentially expressed genes varying within a wide span, rather than a narrow span (data not shown).

In a quantitative trait-associated microarray data analysis, Jia and Xu [25] classified genes into several clusters based on the association of gene expression and the phenotypic value of a quantitative trait. The Bayesian method developed here can be used for such a quantitative trait-associated microarray data analysis. Recall that, in the differential expression analysis, the design matrix X is a variable indicating whether a gene comes from the control or the treatment. To make such an extension, the design matrix X in the differential expression analysis is simply replaced by the phenotypic values of the trait in the quantitative trait-associated microarray analysis. The method developed here has the following extra features compared to that of Jia and Xu [25]: (a) Bayesian method implemented via the MCMC algorithm, (b) constraints on the cluster means, (c) an imbedded normalization step, and (d) a fixed number of clusters.

In differential gene expression analysis, we usually deal with two conditions, control and treatment, with a purpose of identifying genes whose expression levels have altered between the two conditions. In time-course and dose-response [26] microarray experiments, however, gene expression are measured from multiple conditions. The Bayesian mixture model analysis developed here may be extended to detect differentially expressed genes across multiple conditions. To make such an extension, the dimensions of X and γ_i in model (2) need to be changed to reflect the multiple conditions. Let $d+1$ be the number of conditions. The modified dimensions of X and γ_i should be $N \times d$ and $d \times 1$, respectively. The change of γ_i from a scaler into a vector leads to the following consequences: (1) the clusters of down- and upregulated

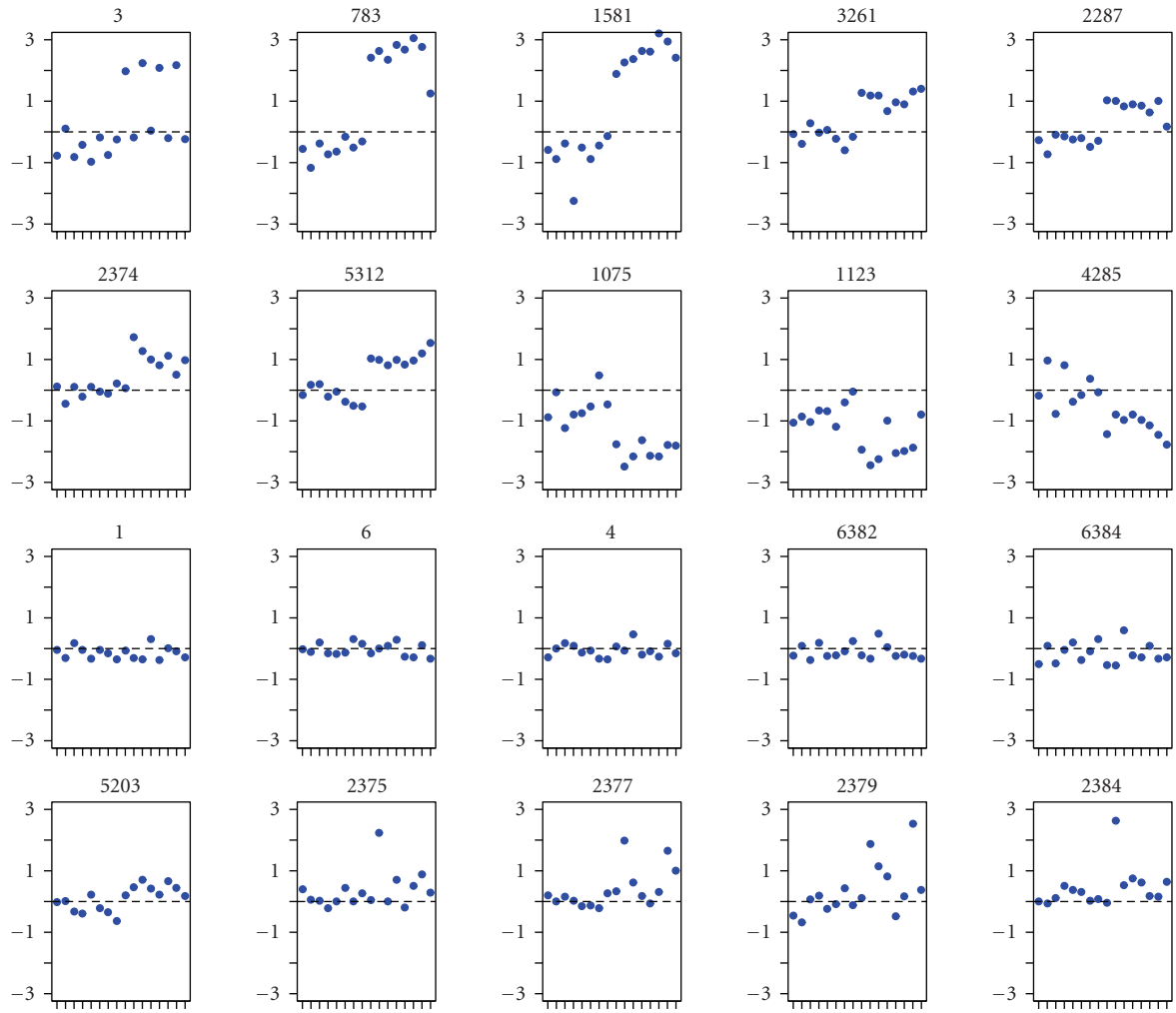

Figure 3: Expression patterns of some genes from the mice data. The first two rows (representing ten genes) are genes detected by all methods. The third row (five genes) are genes detected by none of the methods. The last row (five genes) are those detected by Methods II–VI but not by Methods I.

genes are not explicitly defined, although the neutral cluster is still defined as that with $\beta_k = 0$, and (2) the variance of γ_i in the experiment with two conditions becomes a variance-covariance matrix in the experiment with multiple conditions. The first problem may be solved via the following approaches. Let C be the total number of clusters from which these genes are sampled. Let the first cluster be the one consisting of all the neutral genes and thus $\beta_1 = 0$ is enforced for this cluster. The means of the remaining clusters $\beta_2 - \beta_C$ are not restricted, that is, they are sampled freely from their posterior distributions (multivariate normal). The number of clusters C are found based on the value of Bayesian factor or BIC. Genes not classified into the neutral cluster are claimed to be differentially expressed. The actual number of neutral genes may not be sensitive to the choice of the total number of clusters. Therefore, one may simply choose any reasonable number of clusters for the analysis. This number may be a function of d, say $C = 3^d$. Note that $C = 3$ for a single dimension (one regression coefficient) as done in this study. For two regression coefficients, $C = 3 \times 3 = 9$. Therefore, for d regression coefficients, the number of clusters becomes $C = 3^d$. The second problem may be tackled as follows. The scaled inverse chi-square distribution chosen for the variance may be replaced by the generalization of the scaled inverse chi-square distribution for the variance-covariance matrix, called the inverse-Wishart distribution [21]. This prior distribution for the variance-covariance matrix is conjugate, and thus the standard sampling algorithm for a random vector from an inverse-Wishart distribution applies.

The intercept α_i of gene i in model (2) is irrelevant to the differential expression analysis and thus it is removed from the model via a special normalization process, called linear contrasting. In fact, all effects not related to differential expression can be removed via such a normalization process. Deleting these irrelevant effects (e.g., dye effect, block effect, etc.) may avoid tedious estimating procedure for unspecified parameters used in Zhang et al. [1] and save substantial time in calculation. To remove these effects, matrices W and α_i need to be customized to reflect the general nature of the normalization. Let h be the number of irrelevant effects.

The dimensions of W and α_i should be $N \times h$ and $h \times 1$, respectively. The coefficient matrix for the linear contrasts L remains the same as that defined by model (6). However, matrix L^* takes the first $N - r(W)$ eigenvectors of matrix L. This generalized approach of normalization is conceptually similar to the ANOVA analysis proposed by Kerr et al. [27], where two steps are involved in the ANOVA. In the first step, α_i is estimated under model $Y_i = W\alpha_i + \epsilon_i$ and then $Y_i^* = Y_i - W\hat{\alpha}_i$, the residuals adjusted by the irrelevant effects, is used as the raw data for further differential expression analysis. In the second step, Y_i^* is described by model $Y_i^* = X\gamma_i + \epsilon_i$ and then γ_i is estimated and tested to make an inference about the significance of gene i. Three issues need to be emphasized for the comparison: (1) the generalized approach of normalization proposed in this study removes α_i using no explicit estimate of α_i, (2) the reduced degrees of freedom after adjusting for the irrelevant effects are used for the new method, and (3) appropriate covariance structure for the residuals is used for the new method.

5. CONCLUSIONS

In this paper, we propose a Bayesian mixture model approach to detect genes that differentially expressed in control-treatment microarray experiments. Genes are assigned into one of three constrained clusters, corresponding to clusters of downregulated, neutral, and upregulated genes, respectively. The Bayes method is implemented via the Markov chain Monte Carlo (MCMC) algorithm. Genes that have been assigned into nonneutral clusters are the target genes which we would like to disclose. Using simulated data as well as data from real microarray experiments, we demonstrate that the new method outperforms the methods commonly used in differential expression analysis. Although the new method was demonstrated using data generated from laboratory animals, it is able to generalize to genome studies for plants.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Health Grant R01-GM55321, and the National Science Foundation Grant DBI-0345205 to SX.

REFERENCES

[1] D. Zhang, M. T. Wells, C. D. Smart, and W. E. Fry, "Bayesian normalization and identification for differential gene expression data," *Journal of Computational Biology*, vol. 12, no. 4, pp. 391–406, 2005.

[2] R. Jörnsten, H.-Y. Wang, W. J. Welsh, and M. Ouyang, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, no. 22, pp. 4155–4161, 2005.

[3] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1151–1160, 2001.

[4] P. Broët, S. Richardson, and F. Radvanyi, "Bayesian hierarchical model for identifying changes in gene expression from microarray experiments," *Journal of Computational Biology*, vol. 9, no. 4, pp. 671–683, 2002.

[5] J. W. Edwards, G. P. Page, G. Gadbury, et al., "Empirical Bayes estimation of gene-specific effects in micro-array research," *Functional and Integrative Genomics*, vol. 5, no. 1, pp. 32–39, 2005.

[6] K.-A. Do, P. Müller, and F. Tang, "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society. Series C*, vol. 54, no. 3, pp. 627–644, 2005.

[7] M. Medvedovic and S. Sivaganesan, "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206, 2002.

[8] C. Vogl, F. Sanchez-Cabo, G. Stocker, S. Hubbard, O. Wolkenhauer, and Z. Trajanoski, "A fully bayesian model to cluster gene-expression profiles," *Bioinformatics*, vol. 21, supplement 2, pp. 130–136, 2005.

[9] A. E. Teschendorff, Y. Wang, N. L. Barbosa-Morais, J. D. Brenton, and C. Caldas, "A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data," *Bioinformatics*, vol. 21, no. 13, pp. 3025–3033, 2005.

[10] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[11] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[12] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, vol. 4, no. 4, p. 210, 2003.

[13] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[14] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[15] A. Gusnanto, A. Ploner, and Y. Pawitan, "Fold-change estimation of differentially expressed genes using mixture mixed-model," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, article 26, pp. 1–22, 2005.

[16] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist, "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, vol. 5, no. 2, pp. 155–176, 2004.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977.

[18] Y. Benjamini and W. Liu, "A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence," *Journal of Statistical Planning and Inference*, vol. 82, no. 1-2, pp. 163–170, 1999.

[19] H. D. Patterson and R. Thompson, "Recovery of inter-block information when block sizes are unequal," *Biometrika*, vol. 58, pp. 545–554, 1971.

[20] L. Devroye, *Non-Uniform Random Variate Generation*, Springer, New York, NY, USA, 1986.

[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, New York, NY, USA, 1995.

[22] G. Schwartz, "Estimating the dimensions of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[23] M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin, "Microarray expression profiling identifies genes with altered expression in HDL-deficient mice," *Genome Research*, vol. 10, pp. 2022–2029, 2000.

[24] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, Wiley, New York, NY, USA, 1993.

[25] Z. Jia and S. Xu, "Clustering expressed genes on the basis of their association with a quantitative phenotype," *Genetical Research*, vol. 86, no. 3, pp. 193–207, 2005.

[26] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach, "Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference," *Bioinformatics*, vol. 19, no. 7, pp. 834–841, 2003.

[27] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of Computational Biology*, vol. 7, no. 6, pp. 819–837, 2001.