Biometrics 62, 1089–1098
December 2006
DOI: 10.1111/j.1541-0420.2006.00611.x

A Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification

Ming Yuan

School of Industrial and Systems Engineering, Georgia Institute of Technology,
765 Ferst Drive NW, Atlanta, Georgia 30332, U.S.A.
email: myuan@isye.gatech.edu

and

Christina Kendziorski

Department of Biostatistics and Medical Informatics, University of Wisconsin at Madison,
1300 University Avenue, Madison, Wisconsin 53706, U.S.A.

Summary. Although both clustering and identification of differentially expressed genes are equally essential in most microarray studies, the two tasks are often conducted without regard to each other. This is clearly not the most efficient way of extracting information. The main aim of this article is to develop a coherent statistical method that can simultaneously cluster and detect differentially expressed genes. Through information sharing between the two tasks, the proposed approach gives more sensible clustering among genes and is more sensitive in identifying differentially expressed genes. The improvement over existing methods is illustrated in both our simulation results and a case study.

Key words: Differential expression; Empirical Bayes; False discovery rate; Finite-mixture models; Model-based clustering.

1. Introduction

In contrast to traditional methods that analyze just tens of genes at any one time, microarrays can simultaneously measure the expression level for thousands, often the entire repertoire of a cell population or tissue under investigation. This innovation presents a powerful tool for studies of diverse biological systems.

With such a large number of genes monitored, clustering is one of the foremost tasks for microarray data analysis. It identifies groups of genes that have similar expression profiles across samples. Clustering can reduce the effort of studying individual genes and more importantly it can unmask the functional groups among genes. Since the seminal work by Eisen et al. (1998), various approaches have been developed to fulfill this task in the context of microarray experiments. To name a few, hierarchical clustering, K-means, and partitioning around medoids have all been applied in high throughput studies.

When gene expression measurements come from multiple biological conditions, a fundamental goal is to identify those genes that are differentially expressed under different conditions. This practice often helps investigators identify specific diagnostic, prognostic, and predictive factors for disease which can ultimately lead to the development of molecular-based therapies. The development of statistical methods to identify differentially expressed genes has received much attention, especially methods to identify genes that are differentially expressed between two conditions. For detailed discussion regarding this subject, the readers are referred to Parmigiani et al. (2003) and the references therein.

Although both clustering and differentially expressed gene identification are equally essential in most microarray studies, the two tasks are often conducted without regard to each other. This is clearly not the most efficient way of making inferences. Certainly, which cluster a gene belongs to has a great deal to do with whether or not the gene is differentially expressed. On the other hand, the knowledge of gene clusters provides valuable aids in determining a gene's differential expression pattern. It is the main aim of this article to develop a coherent statistical framework that can be used to simultaneously cluster and detect differentially expressed genes. To this end, we use a two-level mixture model to describe the way in which expression measurements arise. Comparing with the existing methods, the proposal here shares information between the tasks of clustering and detecting differentially expressed genes.

Like many other model-based clustering approaches, each cluster is represented by a mixture component in the first level of our statistical framework. Many advantages of the model-based clustering method, for example, those described in Yeung et al. (2001), are therefore inherited by our method. Different from the existing clustering methods, though, we

base our clustering decision not only on the average expression level and/or the sample variances of certain genes, but also on how likely a gene is to be differentially expressed. In the second level of the model, each cluster is further represented by a mixture model, each component representing the expression pattern of a gene. In this way, decision criteria of differential expression are allowed to vary among clusters; and, as a result, the approach introduced here is expected to outperform previously proposed methods that assume homogeneity among genes.

The article is organized as follows. The proposed statistical framework is detailed in the next section. Section 3 addresses issues of model fitting and posterior inferences. In Section 4, we conduct simulations to show advantages of the new method over existing approaches. A data set is analyzed in Section 5 for illustrative purposes, followed by a discussion in Section 6.

2. Unified Mixture Modeling Approach

Following the model-based clustering method proposed by Fraley and Raftery (2002), our modeling strategy is to capture the probability distribution of expression measurements taken on a set of genes $g = 1, \ldots, G$ by a finite mixture. Often, replicate measurements are obtained under different biological conditions. We assume that some preprocessing technique has been used to adequately normalize the data to remove systematic effects and provide a summary score of expression for each gene on each array. Appropriate normalization schemes exist for both single- and two-color arrays. In the latter case, a reference sample is often used for one of the colors to facilitate a normalization scheme that provides the required summary score of expression (Yang et al., 2002). However, for some two-color designs, obtaining scores of expression that are comparable across multiple conditions may not be possible, particularly if the experiment measures ratios of expression that are not connected by common samples.

For simplicity of presentation, we consider comparing two conditions, for example, control and treatment, with data $\mathbf{x}_g = (x_{g1}, \ldots, x_{gn_1})$ from the n_1 replicate measurements in the first condition and $\mathbf{y}_g = (y_{g1}, \ldots, y_{gn_2})$ from the second condition. This simplification is not required and is relaxed in the web-based Appendix available at `http://www.tibs.org/biometrics`.

2.1 *Model-Based Clustering*

For a moment, suppose we know a priori that there are C clusters among the genes. Expression measurements for genes from the same cluster are expected to have similar profiles, and therefore can be reasonably modeled as observations from the same distribution. More specifically, if gene g comes from the kth cluster, then

$$(\mathbf{x}_g, \mathbf{y}_g) \sim f_k(\mathbf{x}_g, \mathbf{y}_g). \tag{1}$$

Under this notion, measurements of a randomly picked gene g from the G genes we observed should follow

$$(\mathbf{x}_g, \mathbf{y}_g) \sim \pi_1 f_1(\mathbf{x}_g, \mathbf{y}_g) + \cdots + \pi_C f_C(\mathbf{x}_g, \mathbf{y}_g), \tag{2}$$

where π_k is the prior probability that g comes from the kth cluster ($\pi_1 + \cdots + \pi_C = 1$).

Different choices of the component distribution of (2) have been researched in the literature. The most common choices are variants of the multivariate normal. Yeung et al. (2001) systematically documented various options within this family. The structure of the multivariate normal distribution allows an efficient algorithm to compute the posterior probability that a gene belongs to a certain cluster. The main disadvantage of this specification, however, is that it lacks an intuitive way to model the information that the measures are taken under different biological conditions. In this article, we explore the flexibility of (2) further and propose a component distribution that not only accounts for the multiple biological conditions but also incorporates the likelihood for a gene to be differentially expressed among different conditions.

2.2 *Differential Expression Pattern*

To account for the fact that the expression measurements come from different biological conditions and a gene can therefore have different expression patterns (defined below), we use a nested mixture model for each component density of (2).

$$f_k(\mathbf{x}_g, \mathbf{y}_g) = \sum_{j \in \mathcal{S}} p_{kj} f_{kj}(\mathbf{x}_g, \mathbf{y}_g), \tag{3}$$

where $\mathcal{S}$ is the collection of all possible differential expression patterns, and p_{kj} is the prior probability that a gene from the kth cluster has expression pattern $S_j (p_{k1} + \cdots + p_{k|\mathcal{S}|} = 1, |\mathcal{S}|$ represents the cardinality of $\mathcal{S}$).

Expression measurements for a gene can be regarded as noisy observations of a vector of latent expression levels for different biological conditions. In the current setup, the vector for g would be (μ_{gx}, μ_{gy}). Equality and inequality relationships among these expression levels (referred to as expression patterns) represent the biological differences and similarities among conditions.

One way of specifying $\mathcal{S}$ is as follows: Because we are only concerned with two conditions here, there could be three possibilities for a gene g. It is either equivalently expressed (EE), $\mu_{gx} = \mu_{gy}$; overexpressed, (OE) $\mu_{gx} > \mu_{gy}$; or underexpressed (UE), $\mu_{gx} < \mu_{gy}$. Usually, we call a gene differentially expressed (DE) if it is either OE or UE.

In each scenario, f_{kj} in (3) can be interpreted as the conditional distribution $f_k(\mathbf{x}_g, \mathbf{y}_g \mid S_j)$. As the biological information is completely contained in (μ_{gx}, μ_{gy}), we assume that $\mathbf{x}_g$ and $\mathbf{y}_g$ are conditionally independent given (μ_{gx}, μ_{gy}). Therefore, the component distribution f_{kj} can be rewritten as

$$\begin{aligned}
f_{kj}(\mathbf{x}_g, \mathbf{y}_g) &= f_k(\mathbf{x}_g, \mathbf{y}_g \mid S_j) \\
&= \int_{\mu_{gx}} \int_{\mu_{gy}} f_k(\mathbf{x}_g, \mathbf{y}_g, \mu_{gx}, \mu_{gy} \mid S_j)\, d\mu_{gx}\, d\mu_{gy} \\
&= \int_{\mu_{gx}} \int_{\mu_{gy}} f_k(\mathbf{x}_g, \mathbf{y}_g \mid \mu_{gx}, \mu_{gy}) \\
&\quad \times f_k(\mu_{gx}, \mu_{gy} \mid S_j)\, d\mu_{gx}\, d\mu_{gy} \\
&= \int_{\mu_{gx}} \int_{\mu_{gy}} f_k(\mathbf{x}_g \mid \mu_{gx}) f_k(\mathbf{y}_g \mid \mu_{gy}) \\
&\quad \times f_k(\mu_{gx}, \mu_{gy} \mid S_j)\, d\mu_{gx}\, d\mu_{gy} \\
&\equiv \int_{\mu_{gx}} \int_{\mu_{gy}} \prod_{i=1}^{n_1} g_{0k}(x_{gi} \mid \mu_{gx}) \prod_{i=1}^{n_2} g_{0k}(y_{gi} \mid \mu_{gy}) \\
&\quad \times f_k(\mu_{gx}, \mu_{gy} \mid S_j)\, d\mu_{gx}\, d\mu_{gy}.
\end{aligned} \tag{4}$$

The observational distribution g_{0k} represents how the expression measurement is observed for genes from the kth cluster. Furthermore, we assume that under EE, $\mu_{gx} = \mu_{gy}$ is sampled from a prior distribution h_k, that is, $f_k(\mu_{gx}, \mu_{gy} \,|\, \text{EE}) = h_k(\mu_{gx})I(\mu_{gx} = \mu_{gy})$; and under DE, μ_{gx} and μ_{gy} are independently sampled from the same prior distribution h_k with the additional constraint that $\mu_{gx} > (<)\mu_{gy}$ depending on whether the gene is OE (or UE), that is, $f_k(\mu_{gx}, \mu_{gy} \,|\, \text{EE}) = 2h_k(\mu_{gx})h_k(\mu_{gy})I(\mu_{gx} > (<)\mu_{gy})$. Following the discussion above, we can compute the marginal distribution of $(\mathbf{x}_g, \mathbf{y}_g)$ under different expression patterns. If a gene is EE, then

$$f_{k,\text{EE}}(\mathbf{x}_g, \mathbf{y}_g) = \int_{\mu_g} \prod_{i=1}^{n_1} g_{0k}(x_{gi} \,|\, \mu_g) \times \prod_{i=1}^{n_2} g_{0k}(y_{gi} \,|\, \mu_g) h_k(\mu_g)\, d\mu_g. \tag{5}$$

If we consider OE and UE instead, the marginal distributions are

$$f_{k,\text{OE}}(\mathbf{x}_g, \mathbf{y}_g) = 2 \iint_{\mu_{gx} > \mu_{gy}} \prod_{i=1}^{n_1} g_{0k}(x_{gi} \,|\, \mu_{gx}) \times \prod_{i=1}^{n_2} g_{0k}(y_{gi} \,|\, \mu_{gy}) h_k(\mu_{gx}) h_k(\mu_{gy})\, d\mu_{gx}\, d\mu_{gy}, \tag{6}$$

$$f_{k,\text{UE}}(\mathbf{x}_g, \mathbf{y}_g) = 2 \iint_{\mu_{gx} < \mu_{gy}} \prod_{i=1}^{n_1} g_{0k}(x_{gi} \,|\, \mu_{gx}) \times \prod_{i=1}^{n_2} g_{0k}(y_{gi} \,|\, \mu_{gy}) h_k(\mu_{gx}) h_k(\mu_{gy})\, d\mu_{gx}\, d\mu_{gy}. \tag{7}$$

In this article, we focus on two parametric specifications for g_{0k} and h_k. In the so-called gamma-gamma model (GG), we consider g_0 as a gamma distribution with mean value μ_{gx} or μ_{gy} and a common unknown shape parameter α_k. h is chosen so that the rate parameter of g_{0k} follows a gamma distribution with shape parameter α_{0k} and rate parameter ν_k. The lognormal-normal model (LNN) is an alternative specification, where g_{0k} is a lognormal distribution such that $\log x_{gi}$ and $\log y_{gi}$ have means μ_{gx} and μ_{gy}, respectively, and a common unknown variance σ_k^2, and h_k is another normal distribution $N(\mu_{0k}, \tau_k^2)$.

Under either the GG or the LNN model, the marginal distribution under EE has been obtained in closed form as documented in Kendziorski et al. (2003). Readily computable formulae for the marginal distributions under OE and UE can also be derived for both GG and LNN models. Write $\alpha_x = \int_{\mu_{gx}} g_{k0}(\mathbf{x}_g \,|\, \mu_{gx}) h_k(\mu_{gx})\, d\mu_{gx}$ and $\alpha_y = \int_{\mu_{gy}} g_{k0}(\mathbf{y}_g \,|\, \mu_{gy}) h_k(\mu_{gy})\, d\mu_{gy}$. Under the GG model

$$f_{k,\text{OE}}(\mathbf{x}_g, \mathbf{y}_g) = \alpha_x \alpha_y P(B > b_x/(b_x + b_y)), \tag{8}$$

$$f_{k,\text{UE}}(\mathbf{x}_g, \mathbf{y}_g) = \alpha_x \alpha_y P(B < b_x/(b_x + b_y)), \tag{9}$$

where $B \sim Be(a_x, a_y), a_x = n_1\alpha + \alpha_0, a_y = n_2\alpha + \alpha_0, b_x = \sum_j x_{gj} + \nu$, and $b_y = \sum_j y_{gj} + \nu$. Under the LNN:

$$f_{k,\text{OE}}(\mathbf{x}_g, \mathbf{y}_g) = \alpha_x \alpha_y \Phi\left(\frac{c_x - c_y}{\sqrt{d_x^2 + d_y^2}}\right), \tag{10}$$

$$f_{k,\text{UE}}(\mathbf{x}_g, \mathbf{y}_g) = \alpha_x \alpha_y \Phi\left(\frac{c_y - c_x}{\sqrt{d_x^2 + d_y^2}}\right), \tag{11}$$

where

$$c_x = \frac{\sigma^2/n_1\mu_0 + \tau^2 \bar{\mathbf{x}^*}_g}{\sigma^2/n_1 + \tau^2} \qquad d_x = \frac{\sigma^2\tau^2/n_1}{\sigma^2/n_1 + \tau^2} \tag{12}$$

$$c_y = \frac{\sigma^2/n_2\mu_0 + \tau^2 \bar{\mathbf{y}^*}_g}{\sigma^2/n_2 + \tau^2} \qquad d_y = \frac{\sigma^2\tau^2/n_2}{\sigma^2/n_2 + \tau^2} \tag{13}$$

and $\bar{\mathbf{x}^*}$ and $\bar{\mathbf{y}^*}$ are averaged log-transformed expression measures. Technical details of the derivation are available in the web-based Appendix.

3. Posterior Inferences

There are three different ways to view the unified model. Described by (2), we are able to make inference on which cluster a gene belongs to. An application of Bayes theorem gives us the posterior probability that gene g comes from a specific cluster:

$$P(g \in \ k\text{th cluster} \,|\, \mathbf{x}_g, \mathbf{y}_g) = \frac{\pi_k f_k(\mathbf{x}_g, \mathbf{y}_g)}{\pi_1 f_1(\mathbf{x}_g, \mathbf{y}_g) + \cdots + \pi_C f_C(\mathbf{x}_g, \mathbf{y}_g)}. \tag{14}$$

These posterior probabilities can guide us in separating cluster from cluster.

Alternatively, we can rewrite the unified model as

$$(\mathbf{x}_g, \mathbf{y}_g) \sim \sum_{j \in \mathcal{S}} p_j \left(\sum_{k=1}^{C} \pi_{kj}^* f_{kj}(\mathbf{x}_g, \mathbf{y}_g)\right), \tag{15}$$

where $p_j = \sum_k \pi_k p_{kj}$ and $\pi_{kj}^* = \pi_k p_{kj}/p_j$. Now we can also make inference on a gene's differential expression pattern according to the posterior probability:

$$P(g \text{ has pattern } S_j \,|\, \mathbf{x}_g, \mathbf{y}_g) = \frac{p_j \left(\sum_{k=1}^{C} \pi_{kj}^* f_{kj}(\mathbf{x}_g, \mathbf{y}_g)\right)}{\sum_{j' \in \mathcal{S}} p_{j'} \left(\sum_{k=1}^{C} \pi_{kj'}^* f_{kj'}(\mathbf{x}_g, \mathbf{y}_g)\right)}. \tag{16}$$

At last, we can also write the unified model as

$$(\mathbf{x}_g, \mathbf{y}_g) \sim \sum_{k=1}^{C} \sum_{j \in \mathcal{S}} \pi_{kj} f_{kj}(\mathbf{x}_g, \mathbf{y}_g), \tag{17}$$

where $\pi_{kj} = \pi_k p_{kj}$. Using this formulation, we are able to make joint inference on the cluster membership and the expression pattern for a gene. Similar to (14),

$$P(g \in \text{cluster } k, g \text{ has pattern } S_j \,|\, \mathbf{x}_g, \mathbf{y}_g) = \frac{\pi_{kj} f_{kj}(\mathbf{x}_g, \mathbf{y}_g)}{\sum_{k=1}^{C} \sum_{j \in \mathcal{S}} \pi_{kj} f_{kj}(\mathbf{x}_g, \mathbf{y}_g)}. \tag{18}$$

Once the posterior probabilities (14), (16), and (18) are obtained, inferences can be made based on these quantities.

For example, under 0–1 loss, we shall assign gene g to a cluster and/or an expression pattern with the highest posterior probability. Certainly, in practice, other thresholds might also be used to give more conservative lists of potential differentially expressed genes. A natural question is how we measure the effectiveness of a cutoff probability τ. The false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) is a common criterion in the multiple testing setup. In the context of determining whether a gene is differentially expressed, it can be interpreted as P(a gene is EE | its posterior probability of DE $> \tau$). Simple mathematical derivation leads to the following estimate of the FDR (Newton et al., 2004):

$$\widehat{\text{FDR}} = \frac{\sum_{g:P(\text{DE}\,|\,\mathbf{x}_g,\mathbf{y}_g)>\tau} P(\text{EE}\,|\,\mathbf{x}_g,\mathbf{y}_g)}{\text{card}\{g : P(\text{DE}\,|\,\mathbf{x}_g,\mathbf{y}_g) > \tau\}}. \tag{19}$$

Using (19), we can estimate FDR for a specific cutoff τ, that is, $\tau = 0.5$. Alternatively, for a given FDR level α, that is, $\alpha = 0.05$, we can also identify a cutoff τ, which leads to the most powerful list of genes with FDR controlled at the given level.

In order to carry out the inferences formulated above, one needs to first know the parameters associated with the unified model: the number of clusters C; the prior probabilities for clusters $\pi_1, \ldots, \pi_C$; prior probabilities for different expression patterns; and parameters associated with f_{kj}. Ideally, these parameters should be set based on scientific knowledge. In practice, such prior information is oftentimes not available. In these situations, we suggest these parameters be estimated in an empirical Bayes fashion. An expectation-maximization (EM) algorithm is described in the web-based Appendix.

4. Simulations

Accounting for the heterogeneity among the genes, the unified approach can potentially increase the sensitivity in detecting those genes that are differentially expressed. To investigate the advantage of respecting the heterogeneity in terms of identifying differentially expressed genes, we generated 4500 genes under two conditions from two clusters. One cluster contains 3000 genes following the LNN model with parameters $\mu_0 = 8$, $\tau = 1.39$, and $\sigma = 0.3$. Another group contains 1500 genes following the LNN model with parameters $\mu_0 = 5.7$, $\tau = 0.8$, and $\sigma = 0.9$. These parameters are chosen to be similar to those obtained from the data set discussed in the next section. Among the first cluster of genes, a randomly selected 5% were chosen to be differentially expressed; for the second cluster, the proportion of differentially expressed genes was varied from 5% to 50%. Varying the second proportion allowed us to see how the difference between the two clusters affected the performance of different methods. We considered four different implementations of the proposed approach (for details on these, see the web-based Appendix).

(1) AIC: LNN model with number of clusters selected using the Akaike information criterion;
(2) BIC: LNN model with number of clusters selected using the Bayesian information criterion;
(3) HQ: LNN model with number of clusters selected using the criterion proposed by Hannan and Quinn (1979);
(4) TC: LNN model with number of clusters fixed at the true value, in this case, 2.

For each implementation, we consider the number of clusters from 1 to 20 and control the false discovery rate at 5% as discussed in the last section. We compared these implementations with several others in the literature. The methods compared include,

(1) EBarrays: The method given in Kendziorski et al. (2003) with false discovery controlled in the same fashion as the proposed method.
(2) Qval: Two-sample t-test with p value adjustment made by the q-value to control the overall FDR at 0.05. This approach is proposed in a series of articles by Storey and coauthors (see Storey, 2002 and references therein).
(3) LIMMA: The approach proposed by Smyth (2004) with FDR controlled at 5%. The FDR is calibrated in the same fashion as Qval.

Figure 1 reports the FDR, sensitivity, and specificity averaged over 100 simulated data sets. From the figure, we can see that in this example, all four implementations of the proposed method perform essentially the same. All four implementations of the proposed methods, as well as Qval and LIMMA, successfully controlled the FDR. However, the proposed method is much more sensitive than Qval and LIMMA. This could be due to the fact that the simulation favors the proposed approach in that the model assumptions are satisfied. It is worth investigating whether the advantage of the proposed method persists if the model assumptions do not hold.

For this reason, we conducted another set of simulations, which were motivated by the example used in Newton et al. (2004). The data set is a synthesis of three sets of gene expressions. In each cluster, we have $N = 2000$ genes, $n_1 = n_2 = 3$ replicates per condition, and a gamma observation component with shape parameters $a_1 = a_2 = 20$ that are common to all genes. Each cluster differs in the status of the underlying mixing components in f:

(1) Inverse gamma, shape parameter $a_0 = 2$, location $x_0 = 10$;
(2) Uniform on $5 \le A \equiv \log((\mu_{g,1}\mu_{g,2})^{\frac{1}{2}}) \le 11$ and $-1 \le M \equiv \log(\mu_{g,1}/\mu_{g,2}) \le 1$; and $M = 0$ if $\mu_{g,1} = \mu_{g,2}$;
(3) Uniform on $5 \le A \equiv \log((\mu_{g,1}\mu_{g,2})^{\frac{1}{2}}) \le 11$ and $-2 \le M \equiv \log(\mu_{g,1}/\mu_{g,2}) \le 2$; and $M = 0$ if $\mu_{g,1} = \mu_{g,2}$.

The proportions of differential expression are 0.05, 0.1, and 0.2, respectively, for the three clusters. Table 1 documents the operating characteristics of each of the above methods based on 100 runs. The numbers in the brackets are the standard errors. Except for the number of DE calls, all other standard errors are less than 0.001, and are therefore not reported here. Again, we see that the proposed method is much more sensitive than the other methods. Slightly elevated FDRs are observed for AIC and HQ. A more careful examination reveals that the reason is that they tend to select too many clusters to overcome the model misspecification. The more conservative BIC protected against this problem.

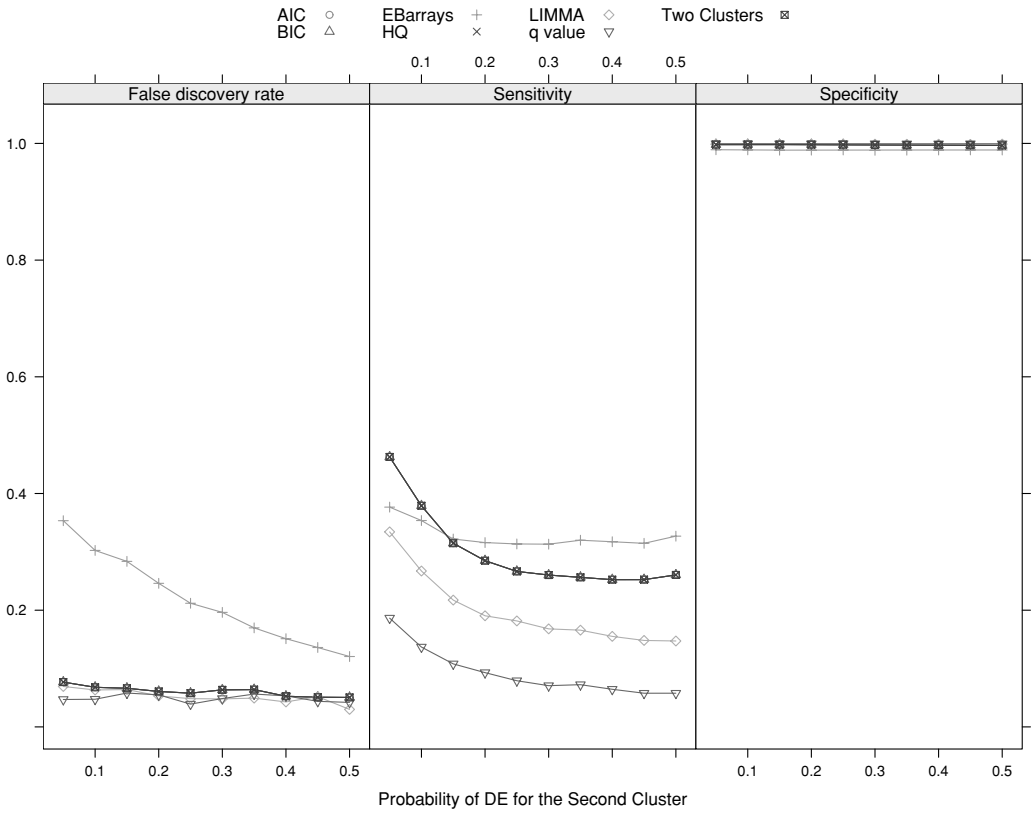

Figure 1. Operating characteristics for simulation I.

The unified approach is also capable of identifying coregulated clusters among genes. To demonstrate this ability, we generated genes from 10 clusters. The cluster sizes were uniformly sampled from 300 to 600. The expression data were then simulated from the LNN model with parameters $\mu_0 = 1, \ldots, 10$. Parameters τ and σ for each cluster are randomly sampled from $0.5\delta(1) + 0.5\delta(1.39)$ and $0.5\delta(1) + 0.5\delta(0.3)$, respectively. The proportions of differential expression for clusters were uniformly sampled from 5% to 45%. Figure 2 shows a typical simulated gene expression data set and the clustered version of the same expression data.

5. Application

To further investigate the utility of this approach, we here consider a real data set obtained from an experiment designed to study the genetic basis for differences between two inbred mouse populations (B6 and BTBR) that show diverse response to a mutation in the leptin gene. Leptin is a protein hormone with important effects in regulating body weight, metabolism, and reproductive function (Zhang et al., 1994). A mutation in the leptin gene causes only mild and transient type 2 diabetes in B6 mice (Coleman and Hummel, 1973), but severe diabetes in BTBR mice (Stoehr et al., 2000). To gain insight into the genetic basis for these differences, Affymetrix MGU74Av2 microarrays were used to probe liver tissues in two pools of two mice in each condition; the data were processed using the DNA-Chip Analyzer (Li and Wong, 2001). Further details can be found in Lan et al. (2003).

An analysis of these data using EBarrays identifies 185 genes to be differentially expressed when FDR is controlled at 5%; the unified approach finds 294. Interestingly, the q-value calculations implemented as in the Qval and LIMMA approaches ((2) and (3) of Section 4) estimate the proportion of differentially expressed genes at 31.5% and 35%, respectively; but no individual genes are called differentially expressed when FDR is controlled at 5%.

Table 1
Operating characteristics for simulation II

	EBarrays	BIC	AIC	HQ	Qval	LIMMA
Number of DE calls	361.05 (0.2055)	494.15 (0.3705)	535.20 (0.2770)	528.40 (0.3422)	367.50 (0.19445)	453.90 (0.2134)
Sensitivity	0.2860	0.4110	0.4410	0.4360	0.3090	0.3740
Specificity	0.9900	0.9930	0.9910	0.9910	0.9960	0.9902
FDR	0.0970	0.0520	0.0620	0.0600	0.0430	0.0600

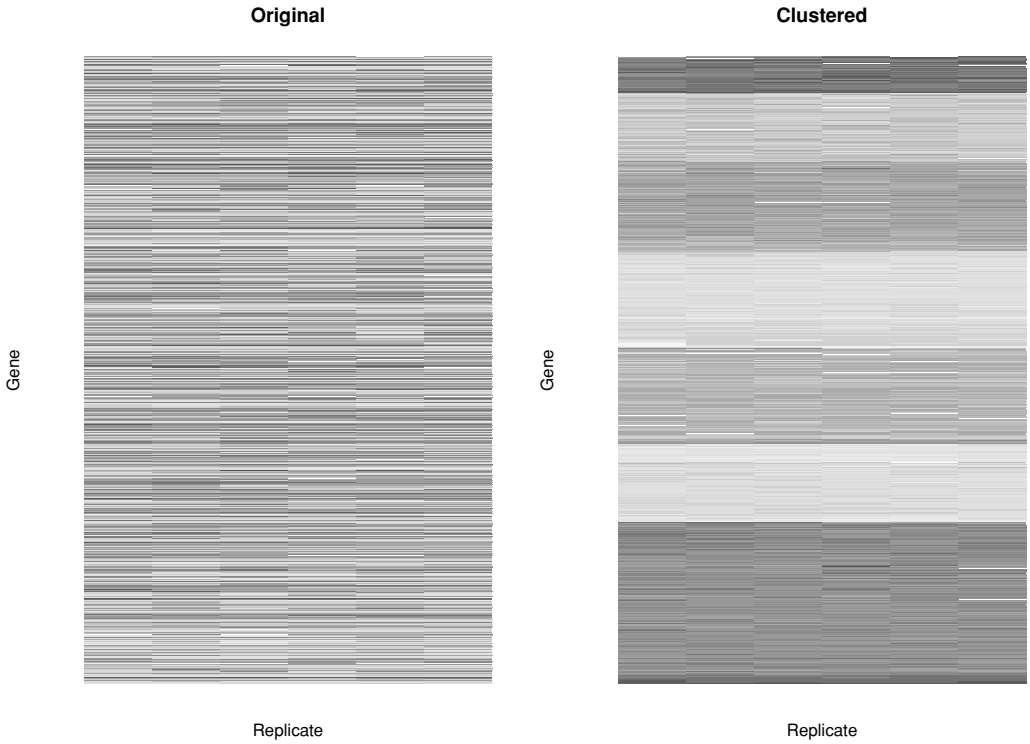

Figure 2. Clustering results for simulation III.

To further investigate the genes identified using EBarrays and the unified approach, we tested for enrichment in functional categories recorded in the Gene Ontology database (`http://www.geneontology.org/`). In GO, transcripts are categorized at varying levels of biological detail (the three broadest levels are molecular function, cellular component, and biological process—there are many subcategories within each). For the two sets of transcripts (those identified by EBarrays and those identified using the unified approach), we tested for enrichment of a common biological process using *GOHyperG* in Bioconductor (Gentleman, 2005). For each biological process considered, the proportion of transcripts on the array labeled with that process was compared with the proportion on the list of genes identified by EBarrays (or the unified approach) labeled with the process. *GOHyperG* carries out a hypergeometric calculation to determine whether there is significant overrepresentation of the biological process among the identified transcripts. Interpretation of the resulting p values is not straightforward due to the many dependent hypotheses tested. Furthermore, the hypergeometric calculation for a particular biological process will tend to result in a small p value when few transcripts on the array are labeled with that process. For these reasons, it has been suggested that one only consider interesting small p values obtained for processes with a relatively large number of transcripts across the array (>10) (Gentleman, 2005). For similar reasons, we further restrict to cases where the number of identified transcripts labeled with the process is relatively large (>5).

Considering processes with at least 10 labeled transcripts across the array, at least 5 labeled transcripts on the identified list, and $p < 10^{-4}$, we found that the genes identified using EBarrays were most enriched for response to pest, pathogen, or parasite ($p = 8.4 \times 10^{-5}$). It is not clear how these genes might be involved in diabetes or obesity. Lowering our thresholds slightly did not improve the results. Considering the genes identified using the unified approach, we found highest enrichment for carbohydrate metabolism ($p = 8.5 \times 10^{-5}$). Reducing our thresholds slightly, we found enrichment for glucose metabolism at $p = 1.5 \times 10^{-4}$. These results make sense as both carbohydrate and glucose metabolism are clearly involved in diabetes and obesity. This suggests improved specificity for the list of genes identified using the unified approach. Improvements in sensitivity are also suggested if we consider the set of genes identified by the unified approach, but not by EBarrays. This list is enriched for only one process—carbohydrate metabolism ($p = 1.1 \times 10^{-3}$—the slightly elevated p value here is due to the fact that carbohydrate metabolism genes found by both methods are removed prior to analysis). As argued above, these improvements are likely due, at least in part, to the fact that the unified approach accounts for clusters inherent in the data thereby improving DE inferences. Figure 3 shows clear clusters in this data set.

6. Discussion

Two of the most important tasks in microarray data analysis are clustering and identifying differentially expressed genes. Although related, each task is most often addressed without regard to the other. In this article, we propose a unified approach that can simultaneously cluster and identify differentially expressed genes. Results can be used to make inferences

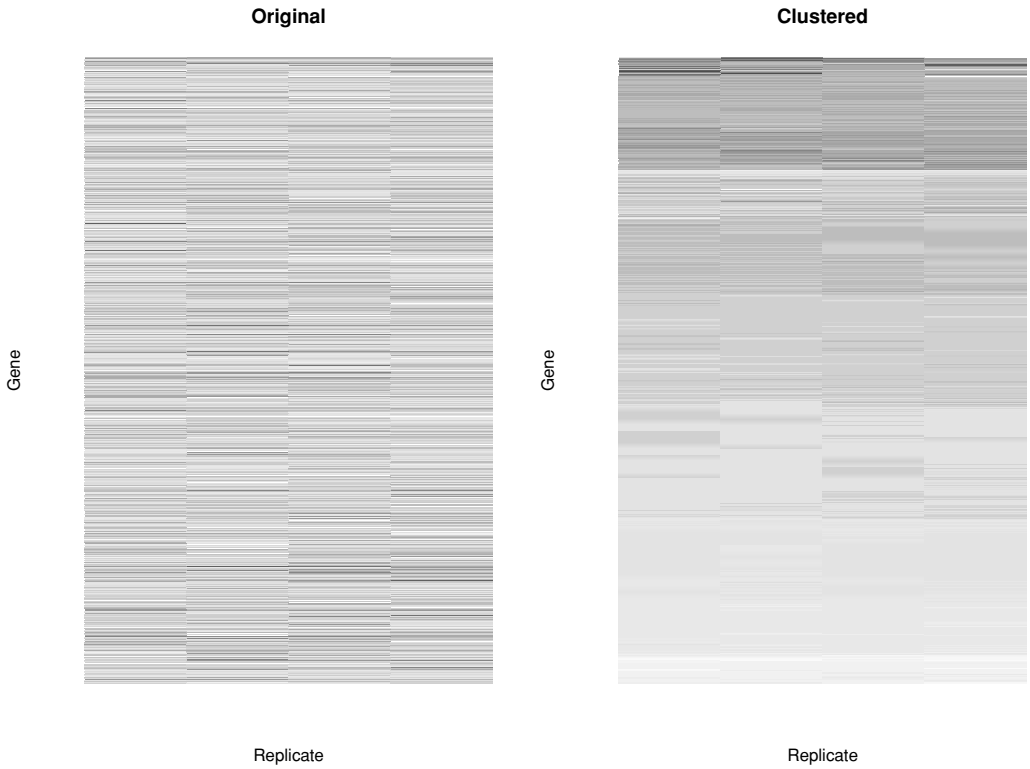

Figure 3. Clustering results for the B6/BTBR data set.

on clusters only, on differentially expressed genes only, or on both.

When clustering is of primary interest, posterior probabilities of cluster membership can be used for gene cluster assignment. In addition, they can be used to help address one of the hardest questions regarding clustering—that of evaluating and interpreting a clustering result. In practice, a unique correct clustering does not exist and cluster validity depends on whether the cluster provides useful information in visualizing and further analyzing the data. Recently, Tseng and Wong (2005) proposed the concept of tight clustering, which is motivated by a similar argument. Instead of forcing every gene into one of the clusters, it might be more reasonable in many applications to identify the "cores" of clusters, namely, those genes which are believed to form the centers of clusters. Our method can achieve the same goal naturally. For example, the core of a cluster can be defined by those genes whose posterior probability of being in the cluster is among the highest or greater than $1 - \alpha$ with a prespecified level α. To illustrate this utility, for each of five clusters identified in the B6/BTBR data (Figure 3), Figure 4 provides 20 genes that have the highest posterior probabilities. Clear coexpressions are observed for genes forming the same cluster.

In addition to providing interpretable cluster assignments, derived posterior probabilities of differential expression improve upon those obtained from existing empirical Bayes methods. The particular hierarchical empirical Bayes method we focused on is similar to that introduced by Newton et al. (2001) and further developed by Kendziorski et al. (2003). Their approach, EBarrays, is useful as it allows for information sharing across genes and provides an adjustment for multiple tests. A disadvantage of their approach, however, is that the model assumptions do not always hold. We demonstrated the price paid in increased FDR when the model is misspecified. The proposed approach was much less sensitive to model misspecification, largely because the flexible cluster structure can appropriately accommodate both parametric and nonparametric distributions.

We note that in practice, model misspecification can often be identified and other methods can be used. For example, a key assumption made by the LNN model for EBarrays is the constant coefficient of variation (CCV) across genes. Figure 5 (left panel) shows a plot of the sample standard deviation versus the sample mean for a typical simulated data set of simulation I. The line represents the lowess fit (Cleveland, 1979) indicating that the assumption is not met. As a result, the inflated levels of FDR observed are not surprising. Similar structure is observed in the B6/BTBR data set (Figure 5, right panel). If diagnostics such as these were checked in practice, EBarrays would not be recommended and an alternative approach would be required. Although our model is not directly motivated to specifically address cases of model misspecification, it does inherit much flexibility in allowing for cluster-specific hyperparameters, thereby relaxing the CCV assumption. Improved model fit is perhaps responsible for the increase in sensitivity and specificity observed for the gene lists identified by the unified approach applied to the B6/BTBR data.

An approach proposed by Newton et al. (2004) specifically addresses cases where the parametric assumptions of

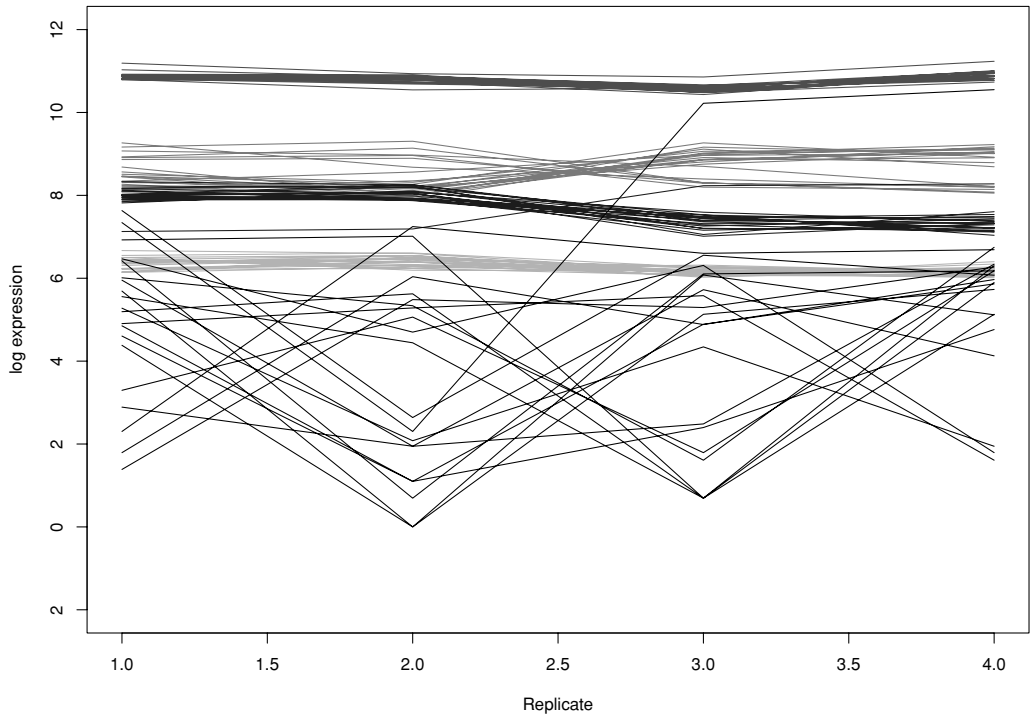

Figure 4. Core genes for clusters.

EBarrays are not met. The approach is a semiparametric extension of EBarrays (SPfit) that models h nonparametrically. The modification certainly robustifies EBarrays, but it is computationally demanding. Furthermore, it fails to address the relationship between which cluster a gene belongs to and how likely it is to be differentially expressed. We note that our approach inherits much of the flexibility provided by SPfit. Some of the advantage gained is demonstrated using a data set

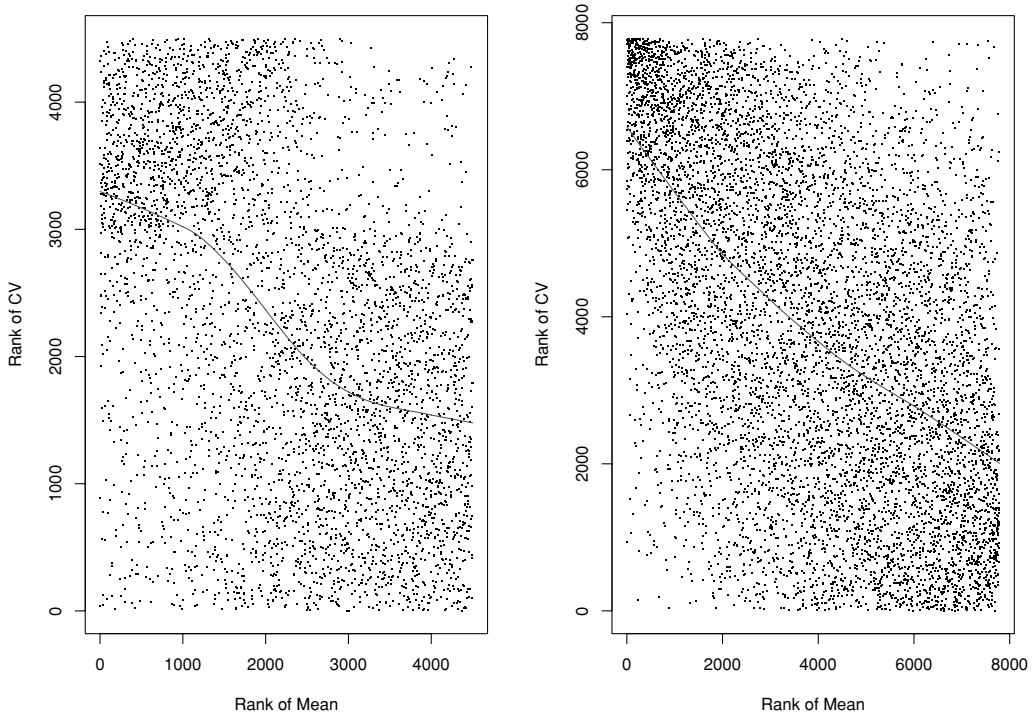

Figure 5. Coefficient of variation as a function of the mean for the simulation I data set (left panel) and the B6/BTBR data set (right panel).

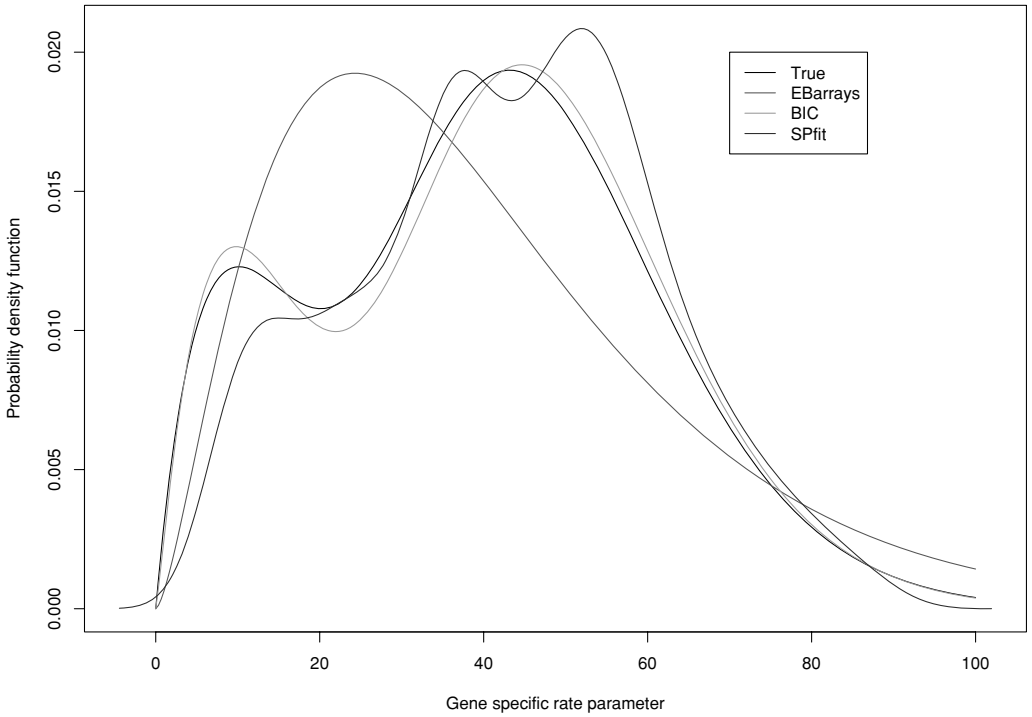

Figure 6. Mixing distribution estimates obtained from a simulated data set with 4500 genes. The method of Newton et al. (2004) completed in 138.77 seconds; the proposed method completed in 16.98 seconds on the same machine. True, bimodal black; EBarrays, unimodal black; BIC, bimodal gray; SPfit, trimodal black.

simulated according to the model assumptions made in SPfit. Consider a data set with observational distribution f following a gamma distribution with shape parameter $\alpha = 20$ and rate parameter sampled from $\frac{1}{3}Ga(2, 0.1) + \frac{2}{3}Ga(10, 0.2)$ ($Ga(a, b)$ represents a gamma distribution with shape parameter a and rate parameter b). Figure 6 presents the estimated mixing distribution for the latent rate parameter h using the proposed method and SPfit. As shown, the proposed method provides a more accurate estimate.

In summary, our proposed approach can be used to cluster genes, to identify differentially expressed genes, or to make inference on both cluster membership and differential expression status simultaneously. The approach preserves the computational efficiency of parametric empirical Bayes methods while at the same time allows for increased flexibility in model assumptions. Improved performance was observed for both simulated and case study data compared with methods that treat these questions separately. As a result, we expect the proposed approach will increase the utility of currently used empirical Bayes methods for clustering and important gene identification. Further work is required to develop diagnostics and identify the conditions under which the proposed approach is most useful.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74,** 829–836.

Coleman, D. L. and Hummel, K. P. (1973). The influence of genetic background on the expression of the obese (Ob) gene in the mouse. *Diabetologia* **9,** 287–293.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 14863–14868.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97,** 611–631.

Gentleman, R. (2005). *Using GO for Statistical Analysis.* Bioconductor vignette. Available at: http://bioconductor.org.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* **41,** 190–195.

Kendziorski, C. M., Newton, M. A., 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.

Lan, H., Rabaglia, M. E., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Zou, F., Yandell, B. S., and Attie, A. D. (2003). Gene expression profiles of non-diabetic and

diabetic obese mice suggest a role of hepatic lipogenic capacity in diabetes susceptibility. *Diabetes* **52,** 688–700.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 31–36.

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

Parmigiani, G., Garrett, E. S., Irizarry, R., and Scott, S. L. (eds). (2003). *The Analysis of Gene Expression Data: Methods and Software.* New York: Springer.

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

Stoehr, J. P., Nadler, S. T., Schueler, K. L., Rabaglia, M. E., Yandell, B. S., Metz, S. A., and Attie, A. D. (2000). Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. *Diabetes* **49,** 1946–1954.

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

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61,** 10–16.

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.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17,** 977–987.

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L., and Friedman, J. M. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature* **372,** 425–431.

Received June 2005. *Revised March* 2006.
Accepted March 2006.