Genet. Res., Camb. (2005), **86**, *pp.* 193–207. *With 3 figures.* © 2005 Cambridge University Press
doi:10.1017/S0016672305007822 Printed in the United Kingdom

Clustering expressed genes on the basis of their association with a quantitative phenotype

ZHENYU JIA AND SHIZHONG XU*
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

(*Received 6 December 2004 and in revised form 13 May and 9 September 2005*)

Summary

Cluster analyses of gene expression data are usually conducted based on their associations with the phenotype of a particular disease. Many disease traits have a clearly defined binary phenotype (presence or absence), so that genes can be clustered based on the differences of expression levels between the two contrasting phenotypic groups. For example, cluster analysis based on binary phenotype has been successfully used in tumour research. Some complex diseases have phenotypes that vary in a continuous manner and the method developed for a binary trait is not immediately applicable to a continuous trait. However, understanding the role of gene expression in these complex traits is of fundamental importance. Therefore, it is necessary to develop a new statistical method to cluster expressed genes based on their association with a quantitative trait phenotype. We developed a model-based clustering method to classify genes based on their association with a continuous phenotype. We used a linear model to describe the relationship between gene expression and the phenotypic value. The model effects of the linear model (linear regression coefficients) represent the strength of the association. We assumed that the model effects of each gene follow a mixture of several multivariate Gaussian distributions. Parameter estimation and cluster assignment were accomplished via an Expectation-Maximization (EM) algorithm. The method was verified by analysing two simulated datasets, and further demonstrated using real data generated in a microarray experiment for the study of gene expression associated with Alzheimer's disease.

1. Introduction

With the advent of microarray technology, it is now possible to measure the expression levels of many genes simultaneously under various conditions. In the original microarray experiments, conditions were defined as control and treatment (Schena *et al*., 1995). Each control and treatment may be replicated several times to provide an assessment of the experimental error. The treatment–control design has been adopted in research with various organisms, including plants (Desprez *et al*., 1998), animals (Anholt & Mackay, 2004; Lazarov *et al*., 2005), humans (Zhang *et al*., 1997) and microbes (Han *et al*., 2004). The purpose of this kind of study is to detect genes whose expression levels respond to the treatment. Statistical methods for such data analysis include the simple *t*-test (Devore & Peck, 1997), the Bayesian method combined with the *t*-test (Baldi & Long, 2001), SAM (significance analysis of microarrays; Tusher *et al*., 2001), the regression approach (Thomas *et al*., 2001) and the model-based cluster analysis (Yeung *et al*., 2001). In many microarray experiments, gene expressions are examined in multiple (more than two) conditions. For example, the control may be represented by tissues sampled from affected persons with a certain disease before a special drug treatment and the treatment may be represented by tissues sampled from persons with the same disease but treated with various doses of the drug, each dose representing a level of the treatment. Because there are multiple levels of treatment, a traditional *t*-test is no longer sufficient and analysis of variance (ANOVA) seems to be more appropriate (Kerr *et al*., 2000; Wolfinger *et al*., 2001; Chu *et al*., 2002; Cui & Churchill, 2003). An alternative approach to

* Corresponding author. Tel: +1 (951) 8275898. e-mail: xu@genetics.ucr.edu

microarray data analysis with multiple conditions is the cluster analysis (Eisen *et al*., 1998), which aims to classify genes with similar expression patterns into the same cluster. Genes within the same cluster are studied as a whole group for their association with a certain disease.

Numerous clustering methods have been proposed for microarray data analysis. Commonly used ones include hierarchical clustering (Carr *et al*., 1997), k-means (Tavazoie *et al*., 1999), the graph-based CAST algorithm (Ben-Dor & Yakhini, 1999), support vector machines (Brown *et al*., 2000), self-organizing maps (Herrero *et al*., 2001) and multilayer perceptrons (Mateos *et al*., 2002). These algorithms are largely heuristically motivated and do not require any underlying statistical models. Recently, model-based methods of clustering have been paid much attention by many investigators because they are built on a solid mathematical and statistical foundation (Yeung *et al*., 2001; McLachlan *et al*., 2002; Ghosh & Chinnaiyan, 2002; Qu & Xu, 2004). The model-based clustering methods have been implemented in two ways: unsupervised (Yeung *et al*., 2001; McLachlan *et al*., 2002; Ghosh & Chinnaiyan, 2002) and supervised (Qu & Xu, 2004). The unsupervised clustering method simply classifies genes based on their expression patterns across treatments or conditions without resorting to any prior knowledge of gene function. The supervised clustering method, however, requires preclassified samples (arrays) and tries to classify unknown genes into these known clusters. If the functions of existing gene clusters are known, the supervised clustering method serves as a tool for identification of gene function.

If a microarray experiment is carried out across a large number of conditions (levels of treatment) and these levels change in a quantitative manner, we will expect that genes expressions will be highly correlated among levels that are quantitatively close. The model-based clustering analysis allows investigators to model and estimate the covariance structure. Use of an elegant covariance structure will significantly increase the efficiency of the cluster analysis. If the quantitative levels are the times after a certain drug injection, the experiment is called time-course microarray experiment (Saban *et al*., 2001). If the levels refer to different dosages of a certain drug treatment, the experiment is called dose-response microarray experiment (Peddada *et al*., 2003). Data collected from both time-course and dose-response experiments can be analysed using the mixed effects model developed by Luan & Li (2003). The mixed-model analysis is a model-based clustering method in which gene expression levels are described as a function of time or drug dosage. The functional relationship is approximated by the B-splines (Luan & Li, 2003). The parameters involved in the smooth function curve are partitioned into a vector of fixed effects and a vector of random effects. Conceptually, the cluster analysis is made based on these parameters rather than on the original data points. In other words, gene clustering is based on the shape of the expression profile because the shape is determined by the parameters.

There is another class of microarray experiments in which different conditions refer to different subjects (or individuals) selected based on their phenotypes. For example, Blalock *et al*. (2004) set up a microarray experiment to study the relationship of gene expression and the severity of Alzheimer's disease. The authors selected 31 subjects each with two quantitative measurement of disease severity (MMSE, Mini-Mental State Examination; NFT, Neurofibrillary Tangle count). The phenotypes of both traits vary more or less in a continuous manner. To study the association of gene expression with the two traits, Blalock *et al*. (2004) calculated the Pearson correlation of gene expression and the trait value. The genes are then sorted according to the magnitude of the correlation coefficients. Genes ranked at the top of the list are reported as being effective on the disease. We propose to use the magnitude of the regression coefficient of gene expression on phenotype to measure the strength of gene-trait association. This implies that a significant correlation is only biologically meaningful if the regression of the expression on phenotype is also high.

The simple regression analysis for gene expression data may be sensitive to outliers because the correlation coefficients are calculated one gene at a time, although the expression data are collected jointly. Joint analysis may extract more information from the data, and thus increase the efficiency of gene clustering. In addition, the threshold P value chosen to declare significance is somewhat arbitrary, making the simple regression analysis inconsistent. Instead of sorting genes based on the Pearson correlation, we propose to take a regression approach by clustering genes according to their regression coefficients. From the statistical test point of view, simple correlation analysis is equivalent to simple regression analysis. Therefore, we call the simple correlation or regression analysis SimpReg for short. In contrast to SimpReg, we call the proposed cluster regression analysis ClusReg. Although both methods are regression analyses, the result of ClusReg can be different from that of SimpReg. This is because gene expression data are analysed jointly in ClusReg so that information from other genes can be incorporated into the analysis of the current gene of interest. In addition, all linear models are directly expressed as linear functions of regression coefficients, which allow biologically uninterested effects, i.e., dye effect, to be explicitly taken into account when estimating the regression coefficients.

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

In ClusReg analysis, gene expression is the response variable and the phenotypic value is the independent variable. We further partition each regression coefficient into a fixed effect and a random effect. The Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977) is used to estimate parameters and to assign each gene to a cluster. The fixed effect for each cluster is estimated using all genes clustered in the same group and the random effect for each gene is estimated via the best linear unbiased prediction (BLUP) technique (see Robinson, 1991, for the theory of BLUP). Gene clustering and parameter estimation are conducted jointly by utilizing data of the entire microarray experiment.

2. Theory and methods

(i) *Mixed model of gene expression*

Let X_j ($j=1, \ldots, n$) be the phenotypic value of a continuous variable for the jth individual in a population of size n (the number of subjects). Let Y_{ij} ($i=1, \ldots, m$ and $j=1, \ldots, n$) be the normalized (log transformed) expression level of the ith gene measured from the jth individual, where m is the total number of genes. We assume that these genes are from c different clusters, indexed by $k=1, \ldots, c$. Let Z_i be the cluster indicator for the ith gene, which takes one value from $\{1, \ldots, c\}$. For the ith gene in the kth cluster, we propose the following mixed-effects model for the observed gene expression of the jth individual:

$$Y_{ij}|_{Z_i=k} = (\beta_{k\mathbf{0}} + X_j\beta_{k\mathbf{1}}) + (\gamma_{i\mathbf{0}} + X_j\gamma_{i\mathbf{1}}) + \varepsilon_{ij}. \quad (1)$$

The first term of this model is used to describe the mean regression profile for the kth cluster. The second term in the above equation is used to model the random effect of the regression profile for the ith gene, where $\gamma_i = [\gamma_{i\mathbf{0}} \quad \gamma_{i\mathbf{1}}]^T$ is a vector of random regression coefficients with a joint $N(0, \Sigma)$ distribution. This term is used to model gene-specific deviation of the regression from the cluster mean. Here we assume that the random coefficients of all genes share a common covariance matrix, regardless of the clusters. The last term in (1) is used to model the uncorrelated measurement error ε_{ij} under the assumption of $\varepsilon_{ij} \sim N(0, \sigma^2)$, for $i=1, \ldots, m$ and $j=1, \ldots, n$.

The above model may be conveniently expressed in matrix notation. Let $Y_i = [Y_{i1} \ldots Y_{in}]^T$ be the vector of expression of the ith gene in all subjects. Let us further define $\beta_k = [\beta_{k\mathbf{0}} \quad \beta_{k\mathbf{1}}]^T$ as a 2×1 vector, $X = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix}^T$ as an $n \times 2$ matrix, and $\varepsilon_i = [\varepsilon_{i1} \ldots \varepsilon_{in}]^T$ as an $n \times 1$ vector. The matrix version of (1) is

$$Y_i|_{Z_i=k} = X\beta_k + X\gamma_i + \varepsilon_i. \quad (2)$$

The expectation and the covariance matrix of (2) are

$$\mu_k = E(Y_i|_{Z_i=k}) = X\beta_k \quad (3)$$

and

$$V = Var(Y_i|_{Z_i=k}) = X\Sigma X^{\mathrm{T}} + I\sigma^2, \quad (4)$$

respectively, where I is an identity matrix with dimension $n \times n$. Model (2) is a mixed-effects model (Laird & Ware, 1982; Robinson, 1991) where β_k is a vector of fixed effects and γ_i is a vector of random effects. The only difference between this model and a typical mixed-effects model is that the fixed effects and random effects share a common design matrix X.

(ii) *Likelihood function of Gaussian mixture*

Conditional on Z_i (the cluster label for gene i), Y_i is described by a mixed-effects model (2). However, Z_i is unknown and it is one of the important quantities that we want to infer in the cluster analysis. Let $\pi = \{\pi_1 \ldots \pi_c\}$ for $\sum_{k=1}^{c} \pi_k = 1$, be the proportions of genes contained by the c clusters. Before we observe Y_i, the probability that gene i belongs to the kth cluster is simply π_k, which is called the prior probability of $Z_i=k$ and denoted by $\pi_k = \Pr(Z_i=k)$. Because Z_i is missing, the probability density of Y_i is a mixture of c normal distributions. As a result, the log likelihood function has the following form:

$$L(\psi) = \sum_{i=1}^{m} \ln \left[\sum_{k=1}^{c} \pi_k p(Y_i|Z_i=k; \beta_k, \Sigma, \sigma^2) \right] \quad (5)$$

where ψ is a vector of parameters and

$$p(Y_i|Z_i=k; \beta_k, \Sigma, \sigma^2) \propto \frac{1}{|V|^{1/2}} \exp\left[-\frac{1}{2}(Y_i - X\beta_k)^{\mathrm{T}} V^{-1} \times (Y_i - X\beta_k) \right] \quad (6)$$

is a normal density (the kth component of the mixture distribution). There is no closed form for the maximum likelihood estimate (MLE) of ψ due to the mixture property of the likelihood (equation 5). However, explicit MLE of the parameters (except π) does exist if the Z_is and γ_is are known. We can take advantage of this property and use the EM algorithm (Dempster *et al*., 1977) to find the MLE of ψ, as described below.

(iii) *EM algorithm for cluster analysis*

The EM algorithm (Dempster *et al*., 1977) is a specific numerical algorithm for solving the MLE of parameters. It is particularly suitable for the mixed-model analysis because such a model can be formulated as a missing value problem. The missing values are the cluster labels (Z_i) and the random regression coefficients (γ_i).

The target likelihood function to be maximized in the M-step and the derivation of it are

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

given in Appendix A. Here we only describe the EM-steps:

Step 0: Initializing parameter, $\psi = \psi^{(0)}$.
Step 1 (E1): Updating the cluster indicator variable:

$$\pi_{ik} = E[\delta(Z_i, k)] = \frac{\pi_k p(Y_i | Z_i = k; \beta_k, \Sigma)}{\sum_{k'=1}^{c} \pi_{k'} p(Y_i | Z_i = k'; \beta_{k'}, \Sigma)}. \quad (7)$$

Step 2 (E2): Updating the random effect:

$$\hat{\gamma}_i = E(\gamma_i | Y_i, Z_i = k) = \Sigma X^{\mathrm{T}} (X \Sigma X^{\mathrm{T}} + I\sigma^2)^{-1} (Y_i - X\beta_k). \quad (8)$$

The conditional covariance matrix of γ_i should also be calculated in this step:

$$\hat{S}_i = V(\gamma_i | Y_i, Z_i = k) = \Sigma - \Sigma X^{\mathrm{T}} (X \Sigma X^{\mathrm{T}} + I\sigma^2)^{-1} X \Sigma. \quad (9)$$

Such a matrix will be used later in the M-steps.

Step 3 (M1): Updating the mixing proportions:

$$\pi_k = \frac{1}{m} \sum_{i=1}^{m} \pi_{ik}. \quad (10)$$

Step 4 (M2): Updating cluster means:

$$\beta_k = (\pi_k m X^{\mathrm{T}} X)^{-1} \sum_{i=1}^{m} \pi_{ik} X^{\mathrm{T}} [Y_i - X E(\gamma_i | Y_i, Z_i = k)]. \quad (11)$$

Step 5 (M3): Updating the covariance matrix of the random effects:

$$\begin{aligned} \Sigma &= \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{c} \pi_{ik} E(\gamma_i \gamma_i^{\mathrm{T}} | Y_i, Z_i = k) \\ &= \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{c} \pi_{ik} (\hat{\gamma}_i \hat{\gamma}_i^{\mathrm{T}} + \hat{S}_i). \end{aligned} \quad (12)$$

Step 6 (M4): Updating the residual variance:

$$\begin{aligned} \sigma^2 &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{k=1}^{c} \pi_{ik} E(||Y_i - X\beta_k - X\gamma_i||^2) \\ &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{k=1}^{c} \pi_{ik} Y_i^T (Y_i - X\beta_k - X\hat{\gamma}_i). \end{aligned} \quad (13)$$

Steps 1 to 6 are repeated until a certain criterion of convergence is reached. Like any other numerical algorithms of optimization, the EM algorithm only provides a local solution. Usually, several different initial values of ψ should be tried to increase the probability of finding the global solution.

(iv) *Number of clusters and hypothesis tests of cluster means*

The above EM algorithm applies to situations where the number of clusters, c, is fixed. In reality, c should be estimated from the data. A commonly used method for inferring c is to calculate the Bayesian information criterion (BIC) and choose c that minimizes the BIC value. BIC is considered as an approximation to the Bayesian factor, and is defined as

$$BIC_c = -2L(\hat{\psi}_c) + \dim(\psi_c) \ln(m) \quad (14)$$

for c clusters, where $\hat{\psi}_c$ is the MLE of the parameter vector ψ_c under c clusters and $\dim(\psi_c)$ is the dimension of ψ_c, i.e., the number of parameters in the model.

Once the number of clusters is determined, we can concentrate on each of the clusters and test the significance of the cluster means. There are numerous methods for the significance test. For illustration purposes, we use the general Wald test statistic (Fahrmeir & Tutz, 1994) to test the cluster means; in principle any appropriate statistics can be used here. The estimated mean of cluster k is $\hat{\beta}_k = [\hat{\beta}_{k\mathbf{0}} \quad \hat{\beta}_{k\mathbf{1}}]^T$ and the variance matrix of the estimate is approximated by

$$\mathrm{Var}(\hat{\beta}_k) = V_{\beta_k} = [m\pi_k X^{\mathrm{T}} (X \hat{\Sigma} X^{\mathrm{T}} + I\hat{\sigma}^2)^{-1} X]^{-1}. \quad (15)$$

To test the hypothesis that $\hat{\beta}_k = 0$, the following test statistic may be used:

$$w_k = \hat{\beta}_k^{\mathrm{T}} V_{\beta_k}^{-1} \hat{\beta}_k. \quad (16)$$

In microarray data analysis, investigators may be interested in the hypothesis that $\hat{\beta}_{k\mathbf{1}} = 0$ and the value of the intercept is irrelevant. The test statistic can be derived using a general method for testing a linear contrast. Let $L^T \hat{\beta}_k$ be a linear contrast of the cluster means. To test the hypothesis that $L^T \hat{\beta}_k = 0$, we use

$$w_k = \hat{\beta}_k^{\mathrm{T}} L (L^{\mathrm{T}} V_{\beta_k} L)^{-1} L^{\mathrm{T}} \hat{\beta}_k. \quad (17)$$

It is now obvious that $\hat{\beta}_{k1} = 0$ can be formulated as $L^T \hat{\beta}_k = 0$ where $L^T = [0 \quad 1]$. The Wald statistic in (17) is then compared with the threshold $\chi^2_{\mathbf{0{\cdot}95,1}} = 3{\cdot}82$ (95th percentile of the chi-square distribution with one degree of freedom). $\hat{\beta}_{k\mathbf{1}}$ is said to be significantly different from zero if w_k is greater than this quantity. We can also compare two clusters using the Wald test statistic. Assume that we want to test the hypothesis that $\hat{\beta}_k - \hat{\beta}_{k'} = 0$. The test statistic appears to be

$$w_{kk'} = (\hat{\beta}_k - \hat{\beta}_{k'})^{\mathrm{T}} (V_{\beta_k} + V_{\beta_{k'}})^{-1} (\hat{\beta}_k - \hat{\beta}_{k'}). \quad (18)$$

(v) *BLUP of the gene expression profile*

After the EM algorithm converges and the number of clusters is determined, we obtain all the estimated parameters. We also have the posterior probability

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

that the ith gene belongs to the kth cluster, $\pi_{ik}=\Pr(Z_i=k|Y_i)$, for $i=1, \ldots, m$ & $k=1, \ldots, c$. Based on these probabilities, we can assign the ith gene to the kth cluster if $\pi_{ik}=\max(\pi_{i1}, \ldots, \pi_{ic})$. We can also try to cluster only those genes with the maximum π_{ik} greater than a predetermined cut-off value, and declare other genes as unclassified.

After gene clustering, we can obtain the estimate of the gene expression profile using gene expression data from the same cluster. For the ith gene in the kth cluster, the best linear unbiased predictor (BLUP) of the random regression coefficients γ_i is

$$\hat{\gamma}_i=\hat{\Sigma}X^{\mathrm{T}}(X\hat{\Sigma}X^{\mathrm{T}}+I\hat{\sigma}^2)^{-1}(Y_i-X\hat{\beta}_k). \tag{19}$$

The corresponding estimate of the individual gene expression profile for the ith gene in the kth cluster is

$$\hat{Y}_i|_{Z_i=k}=X\hat{\beta}_k+X\hat{\gamma}_i \tag{20}$$

The estimated gene expression profiles can be plotted against the phenotypic value.

(vi) *Extension to the heteroscedastic covariance matrix*

The basic assumption of the EM analysis presented earlier is that the covariance matrix of the random regression coefficients is constant across clusters, the so-called homoscedastic covariance matrix. In this section, we try to extend the algorithm to handle situations where the covariance matrix is not constant but varies across clusters. In addition, we may relax the assumption of independent residual errors for genes within the same cluster. The model for the ith gene in the kth cluster is

$$Y_i|_{Z_i=k}=X\beta_k+X\gamma_i|_{Z_i=k}+\varepsilon_i \tag{21}$$

where $\gamma_i|_{Z_i=k}\sim N(0, \Sigma_k)$ is the random regression coefficient with a different covariance matrix for a different cluster. The residual error is assumed to be $\varepsilon_i\sim N(0, D)$, where D is a diagonal matrix. The expectation and covariance matrix of model (21) are

$$\mu_k=E(Y_i|Z_i=k)=X\beta_k \tag{22}$$

and

$$V_k=Var(Y_i|Z_i=k)=X\Sigma_kX^{\mathrm{T}}+D, \tag{23}$$

respectively.

The EM algorithm is largely the same as that of the homoscedastic covariance matrix model except that the step of updating Σ is replaced by c steps of updating Σ_k for $k=1, \ldots, c$, and updating σ^2 is replaced by updating D, as shown below:

$$\Sigma_k=\frac{1}{m\pi_k}\sum_{i=1}^{m}\pi_{ik}E(\gamma_i\gamma_i^{\mathrm{T}}|Y_i, Z_i=k) \tag{24}$$

and

$$D=\frac{1}{m}\sum_{i=1}^{m}\mathrm{diag}\left\{\sum_{k=1}^{c}\pi_{ik}E[(Y_i-X\beta_k-X\gamma_i)\times(Y_i-X\beta_k-X\gamma_i)^{\mathrm{T}}]\right\}. \tag{25}$$

3. Application

(i) *Simulation studies under the homoscedastic covariance matrix model*

In the first simulation experiment, we chose $c=5$ and simulated expression of 1000 genes on 50 subjects (microarray chips), each of which has a phenotypic value normally distributed with mean 20 and standard deviation 9, which were estimated from the phenotypes of a real microarray experiment (Blalock *et al*., 2004). The parameters used in the simulation are given in Table 1 (dataset 1). These parameters are similar or comparable to the estimated parameters obtained from the Alzheimer's disease microarray experiment (Blalock *et al*., 2004). We intentionally made the differences among clusters 2, 3 and 4 very small to evaluate the efficiency of our method in dealing with the difficult situation. The initial values of parameters for the EM iterations were chosen as follows. First, we fitted each gene to a simple linear regression model on the phenotypic value and estimated b_{i0} (intercept) and b_{i1} (regression coefficient) for each gene, indexed by i for the ith gene. We then sorted the genes by the estimated b_{i1} and divided the 1000 genes into five clusters each with 200 genes based on their ranks in the sorted dataset. The mean intercept and the mean regression coefficient for group k were treated as the initial value of vector β_k. The average value of all the 1000 estimated residual variance was used as the initial value of σ^2. The initial value of the covariance matrix for the random regression coefficients was chosen as $\Sigma^{(0)}=I\sigma^2$ where I is an identity matrix with dimension 2×2. Finally, the prior for the membership probability of each gene was chosen from a uniform distribution. The EM iterations were stopped when the difference between successive parameter estimates was less than 0·0001 for each parameter. Similar criteria and rules were used in all subsequent data analyses.

Table 1 also gives the results of ClusReg analysis at $c=5$. We can see that the estimated parameters agree well with the true parameters. The last column of Table 1 gives the Wald test statistic (w_k) of each cluster for testing the null hypothesis of $\beta_{k1}=0$. The test statistic is significant for each of the clusters when $\chi^2_{0\cdot 95,1}=3\cdot 82$ (95th percentile of the chi-square distribution with one degree of freedom) was used as the critical value of the test statistic. Even if we chose $\chi^2_{0\cdot 995,1}=7\cdot 8$ (correction for multiple tests) as the

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

Table 1. *True parameters and the estimated parameter values in the simulation experiment (dataset 1): (a) parameters used in the simulation experiment and their estimated values when the number of clusters was set at five ($c=5$); (b) estimated parameters from the same simulated data but with the number of clusters set at three ($c=3$)*

	Cluster		β_{k0}	β_{k1}	π_k	Σ	σ^2	w_k
a	1	True	6·48	−0·1	0·05	True		
		Estimate	6·608	−0·101	0·049	$\Sigma=\begin{bmatrix}1 & 0{\cdot}006\\ 0{\cdot}006 & 0{\cdot}001\end{bmatrix}$	True	503·70
	2	True	6·25	−0·001	0·20		0·36	
		Estimate	5·757	−0·008	0·200			14·27
	3	True	6·18	0·0001	0·26			
		Estimate	6·368	0·013	0·245			11·92
	4	True	5·99	0·001	0·44	Estimate	Estimate	
		Estimate	6·164	−0·005	0·458		0·358	44·09
	5	True	5·25	0·1	0·05	$\hat{\Sigma}=\begin{bmatrix}1{\cdot}0065 & 0{\cdot}0055\\ 0{\cdot}0055 & 0{\cdot}0009\end{bmatrix}$		
		Estimate	5·007	0·108	0·049			573·90
b	I	Estimate	6·633	−0·099	0·05	Estimate		466·50
	II	Estimate	6·128	−0·0008	0·9		0·358	0·57
	III	Estimate	4·997	0·107	0·05	$\hat{\Sigma}=\begin{bmatrix}1{\cdot}047 & 0{\cdot}007\\ 0{\cdot}007 & 0{\cdot}001\end{bmatrix}$		533·30

Table 2. *Number of genes assigned to clusters for the simulated data (dataset 1) when $c=5$ was chosen. The sum of each column represents the true number of genes simulated from that cluster and the sum of each row represents the number of genes assigned to that cluster (similar table structure was also used in other tables for the purpose of comparison)*

	True					
Estimate	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Sum
Cluster 1	40	2	1	3	0	46
Cluster 2	0	0	8	5	1	14
Cluster 3	0	28	32	48	4	112
Cluster 4	10	170	218	381	1	780
Cluster 5	0	0	1	3	44	48
Sum	50	200	260	440	50	

critical value, the test statistic is still significant. Table 2 gives the error rates of cluster assignment of the genes. Clusters 1 and 5, which are far removed from the remaining clusters, have quite small error rates. Clusters 2, 3 and 4, however, are subject to high error rates due to the small differences among the means of the three clusters.

We also analysed the same dataset (dataset 1) by varying the number of clusters (c changes from 2 to 9) and found that the optimal BIC occurs at $c=3$ (see Fig. 1a for the BIC profile across the number of clusters). This is expected because clusters 2, 3 and 4 in the simulation were indeed very close to each other and could be treated as a single large group. Eventually, we evaluated the results with $c=3$ by combining the three closely related clusters as cluster II whereas the original clusters 1 and 5 were renamed as clusters I and III, respectively. The results with $c=3$ are listed in Tables 1 and 3. The test statistic for the mean of cluster II is no longer significant. The error rates are also improved.

To test efficiency of various methods, empirical Type I (α) and Type II (β) error rates were calculated in this simulation study. Let N be the total number of true neutral genes (cluster II) and Ne be the number of neutral genes that were incorrectly assigned into cluster I or III. Let S be the total number of true significant genes (cluster I + cluster III) and Se be the number of significant genes that were claimed to be neutral by mistake. We defined $\alpha=Ne/N$, $\beta=Se/S$ and $1-\beta$ as the empirical Type I error, Type II error and statistical power, respectively.

For comparison, we also reanalysed the data with simple regression (SimpReg) in the following steps: (1) we calculated the P values for individual genes based on the simple regression (Pearson's correlation) analysis, (2) we selected q which lies between 0 and 1, and this is the maximum FDR (false discovery rate;

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

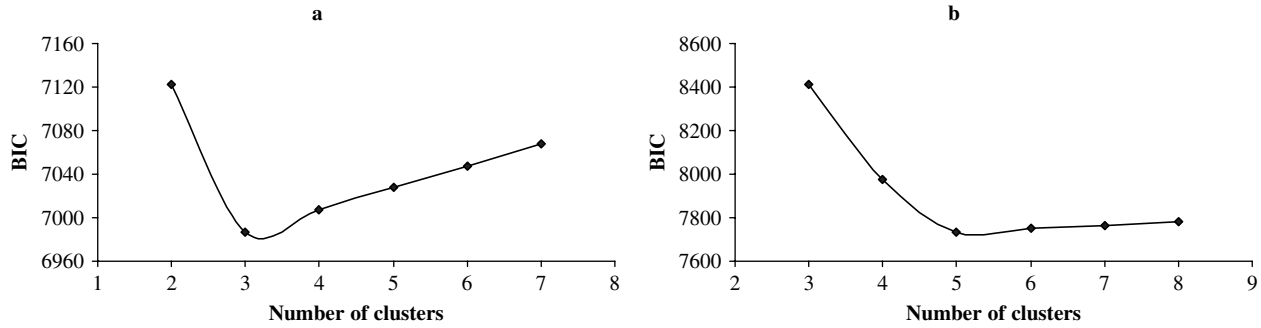

Fig. 1. BIC profiles across the number of clusters for the simulated data: (*a*) a plot of dataset 1 using the homoscedastic covariance matrix model and (*b*) a plot of dataset 2 using the heteroscedastic covariance model.

Benjamini & Liu, 1999) that we are willing to tolerate on average, (3) we sorted genes based on the P values in ascending order ($p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)}$), and let $H_{(i)}$ be the null hypothesis corresponding to P value $p_{(i)}$, (4) we let r be the largest i for which $p_{(i)} \leqslant \frac{i}{m} \frac{q}{c(m)}$, where $c(m) = 1$ under the assumption of independence of the P values, (5) we reject the null hypotheses, $H_{(1)}, \ldots, H_{(r)}$, which means that these genes are significantly related to the phenotype, and (6) for significant genes, we assigned those with positive correlation coefficients to one cluster and those with negative correlation coefficients to another cluster; the remaining non-significant (neutral) genes were classified into the third cluster.

We set the FDR at $q = 5\%$ and the cut-off P value in this situation was 0·0297. The result is given in Table 3, showing a $1-\beta = 0{\cdot}98$ power with an $\alpha = 0{\cdot}5589$ Type I error. This indicates that the simple regression analysis detected more significant genes than the actual number of genes. Many neutral genes (in cluster II) were incorrectly assigned into groups I and III. This is not efficient in gene identification as demonstrated by Wayne & Mclntyre (2002). We then set the number of significant genes detected by SimpReg analysis to the same number as we detected with the ClusReg analysis. Overall, the two methods generated similar results (see Table 3 for comparison). However, the cut-off P value to declare significance in the SimpReg analysis is about 3·6E−13 and the expected FDR is as small as 4·3E−12. This cut-off P value and the corresponding FDR may be too stringent in reality. We also see that both the Type I and Type II errors are greater than those in the ClusReg analysis.

Table 3. *Numbers of genes assigned to clusters for the new method and comparisons with simple regression analysis (dataset 1): (a) numbers of genes assigned to clusters for the simulated data using ClusReg at $c = 3$, leading to $\alpha = 0{\cdot}0056$ and $1-\beta = 0{\cdot}82$; (b) numbers of genes assigned to groups for the simulated data using SimpReg with the cut-off P value equal to 0·0297 and the expected FDR equal to 0·05, leading to $\alpha = 0{\cdot}5589$ and $1-\beta = 0{\cdot}98$; (c) numbers of genes assigned to groups for the simulated data using SimpReg with the cut-off P value equal to 3·62E−13 and the expected FDR equal to 4·3E−12, leading to $\alpha = 0.0122$ and $1-\beta = 0{\cdot}76$*

		True			
	Estimate	Cluster I	Cluster II	Cluster III	Sum
a	Cluster I	39	4	0	43
	Cluster II	11	895	7	913
	Cluster III	0	1	43	44
b	Group I	50	255	0	305
	Group II	0	397	2	399
	Group III	0	248	48	296
c	Group I	37	4	0	41
	Group II	13	889	11	912
	Group III	0	7	39	46
	Sum	50	900	50	

(ii) *Simulation studies under the heteroscedastic covariance matrix model*

In the second simulation experiment, we simulated a heteroscedastic covariance matrix for the random regression coefficients. Again, we chose $c = 5$ and the true parameters are given in Table 4 (dataset 2). We simulated 1000 genes on 10 subjects (microarrays), each of which has a phenotypic value evenly distributed between 1 and 5. The data were analysed using the heteroscedastic covariance matrix model. The BIC profile shows that $c = 5$ leads to the optimal BIC score (see Fig. 1b for the BIC profile). The estimated parameters are also given in Table 4, showing excellent agreement with the true parameter values. Table 5 lists the error rates of the cluster assignment of the genes.

We also examined the sensitivity of the method under the homoscedastic covariance matrix model to the assumption of heteroscedasticity. We reanalysed the same data (dataset 2) under the homoscedastic covariance matrix model. The BIC profile indicates

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

Table 4. *Parameters used in the simulation experiment and estimated parameters from the simulated experiment (dataset 2): (a) parameters used in the simulation experiment and their estimated values under the heteroscedastic covariance matrix model when the number of clusters was set at five ($c=5$); (b) estimated parameters from the same simulated data but under the homoscedastic covariance matrix model with the number of clusters set at six ($c=6$)*

Cluster			β_{k0}	β_{k1}	π_k	Σ_{11}	Σ_{12}	Σ_{22}	σ^2	w_k
a	1	True	−2	−4	0·2	0·6	0·2	0·6		
		Estimate	−1·91	−3·94	0·20	0·561	0·128	0·541	True	5451·32
	2	True	5	−2	0·2	1	0·4	1	0·291	
		Estimate	5·10	−1·97	0·20	1·274	0·504	0·997		756·17
	3	True	0	2	0·2	0·5	0	0·5		
		Estimate	−0·06	1·81	0·20	0·569	0·052	0·441		1405·75
	4	True	−6	3	0·2	1	0·5	0·8	Estimate	
		Estimate	−5·96	3·05	0·20	1·313	0·331	0·535	0·294	3309·63
	5	True	3	4	0·2	0·4	0·1	0·6		
		Estimate	3·09	4·01	0·20	0·403	−0·04	0·553		5542·66
b	1	Estimate	−1·9050	−3·9360	0·2	Estimate			Estimate	5452·71
	2	Estimate	4·4295	−2·6089	0·089				0·2942	1060·02
	3	Estimate	5·6812	−1·4015	0·111	$\hat{\Sigma}=\begin{bmatrix}0.7593 & 0.1241\\ 0.1241 & 0.5420\end{bmatrix}$				383·28
	4	Estimate	−0·0673	1·8079	0·2					1146·03
	5	Estimate	−5·9849	3·0359	0·2					3233·92
	6	Estimate	3·0670	4·0094	0·2					5671·23

Table 5. *Numbers of genes assigned to clusters and comparisons between the homoscedastic and the heteroscedastic models (dataset 2): (a) numbers of genes assigned to clusters for the simulated data under the heteroscedastic covariance matrix model when $c=5$; (b) numbers of genes assigned to clusters for the simulated data under the homoscedastic covariance matrix model at $c=6$*

		True					
	Estimate	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Sum
a	Cluster 1	200	0	0	0	0	200
	Cluster 2	0	200	0	0	0	200
	Cluster 3	0	0	195	0	3	198
	Cluster 4	0	0	0	200	0	200
	Cluster 5	0	0	5	0	197	202
b	Cluster I	200	0	0	0	0	200
	Cluster II	0	200	0	0	0	200
	Cluster III	0	0	195	1	5	201
	Cluster IV	0	0	0	199	0	199
	Cluster V	0	0	5	0	195	200
	Sum	200	200	200	200	200	

that the optimal BIC occurs when $c=6$, although the data were simulated at $c=5$. This observation further verified that more components of the mixture model are required when a simple covariance structure (homoscedastic) is used than when a complex covariance structure (heteroscedastic) is used (Fraley & Raftery, 2002). We noted that when the homoscedastic covariance model is used for the heteroscedastic covariance data, cluster 2 was divided into two subclusters. The estimated parameters for the six-cluster analysis under the homoscedastic covariance model are given in Table 4. We also reported the results of five-cluster analysis under the homoscedastic covariance model (although six is the optimal number of clusters under this model for this particular dataset). In this analysis, cluster I corresponds to the original cluster 1, cluster II combines the original clusters 2 and 3, and clusters III, IV and V correspond to the original clusters 4, 5 and 6, respectively. The error rates are given in Table 5, showing good agreement between the two models, and thus the robustness of the homoscedastic covariance model.

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

Table 6. *Estimated parameters of gene expression associated with Alzheimer's disease with the model-based ClusReg analysis at $c=3$*

Cluster	$\hat{\beta}_{k0}$	$\hat{\beta}_{k1}$	$\hat{\pi}_k$	$\hat{\Sigma}$	$\hat{\sigma}^2$	w_k
1	6·0012	−0·0340	0·0047	$\hat{\Sigma}=\begin{bmatrix}1\cdot16840 & 0\cdot00180\\ 0\cdot00180 & 0\cdot00005\end{bmatrix}$	0·17368	770·05
2	6·3442	−0·0014	0·9602			145·56
3	5·2263	0·0303	0·0351			3071·78

(iii) *Application to Alzheimer's disease*

Blalock *et al*. (2004) reported results of simple Pearson correlation analysis of 9921 expressed genes associated with some quantitative measurements for the severity of Alzheimer's disease with 31 subjects. The data collected from the Alzheimer's disease microarray experiment were Affymetrix data. Each data point represents the average difference of 20 PM − MM pairs of intensity differences. Therefore, the technology that our expression data best mirrors is the Affymetrix approach. Blalock *et al*. (2004) examined two traits (MMSE and NFT) in the analysis and divided the genes into three groups based on their correlations with MMSE and NFT. A total of 1977 genes were declared as up-regulated (either negatively correlated with MMSE, positively correlated with NFT, or both) and 1436 genes were down-regulated (positively correlated with MMSE, negatively correlated with NFT, or both). The majority of the genes were neutral (correlated with neither MMSE nor NFT). The original data are available through the Gene Expression Omnibus (GEO) website with an accession number GSE1297.

According to E. M. Blalock (personal communication), Alzheimer's is a complicated disease and cannot be unequivocally determined by either MMSE or NFT measurement. Blalock *et al*. (2004) analysed two traits separately and classified genes based on both traits. The results are hard to compare with our analysis in which only one phenotype is considered. The purpose of our analysis for the real data is to demonstrate the method and verify the computer program. Therefore, we only analysed one trait, MMSE, as a working example. According to the selection criterion given by Blalock *et al*. (2004), we selected 9754 genes for analysis (note the difference between this number and the number of genes selected by Blalock *et al*.). Expression values of the 9754 genes (subset of the original data) and the disease phenotypic values (MMSE) for the 31 subjects are available on request. We recalculated the Pearson correlations of all the 9754 genes in the way we described in the simulation experiment.

We first analysed gene expression data with the model-based method of ClusReg under the homoscedastic covariance model and then analysed the data with SimpReg for comparison. In the model-based ClusReg analysis, we examined the BIC scores under $c=2, \ldots, 10$ and found that $c=3$ is the optimal number of clusters. The estimated parameters at $c=3$ are listed in Table 6. The BLUP estimates of gene expression for the three clusters are depicted in Fig. 2 (the upper panels). The lower panels of Fig. 2 represent the BLUP estimates and the original observed data points for a typical gene picked from each cluster. Overall, 46 genes (0·47%) were negatively associated with MMSE, 342 genes (3·51%) were positively associated with MMSE, and the remaining 9366 genes (96·02%) were neutral. The test statistic of the regression coefficient of the neutral cluster was still significant, but the magnitude was negligible compared with the other two clusters.

In the simple regression analysis, we first set FDR at 0·05 to select the cut off P value of 0·0006. Using this cutoff P value, only 46 genes (0·47%) were negatively associated and 79 genes (0·81%) were positively associated with MMSE (see Table 7 for the comparison of the two methods). The number of significant genes was much less than that detected with the ClusReg analysis. We then chose 0·0032 as the cut-off P value so that the SimpReg analysis detected exactly the same number of genes (388) as that with the ClusReg analysis. The expected FDR value with this cutoff P value was approximately 0·08. The SimpReg analysis detected 161 genes (1·65%) with negative association and 227 genes (2·33%) with positive association. Table 7 also shows the comparison of the two methods with regard to the numbers of genes detected in each cluster. Among the 195 genes with negative association (46 genes by the ClusReg analysis and 161 genes by the SimpReg analysis), only 12 were detected by both methods. Among the 472 positively associated genes (342 by the ClusReg analysis and 227 by the SimpReg analysis), only 97 were detected by both. Many genes detected by the ClusReg analysis failed to show significance in the simple regression analysis and vice versa.

We examined all genes detected by one method and missed by the other method through scatter plots. The plots of four typical genes are shown in Fig. 3. Genes 7874 and 6205 were detected by the ClusReg analysis but missed by the SimpReg analysis (designated as class I genes), whereas genes 3261 and 3476 were

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

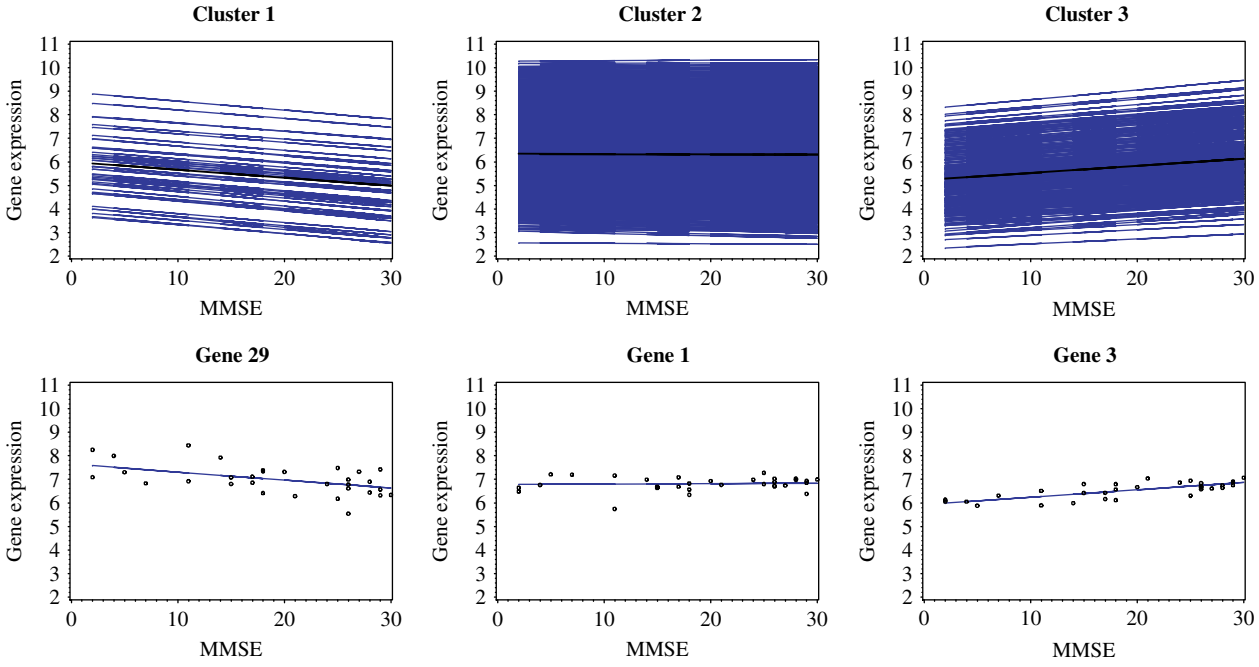

Fig. 2. BLUP estimates of gene expression of the three detected gene clusters (upper panels) and the BLUP estimates along with the original observed data points for a typical gene picked from each cluster (lower panels) in the MMSE analysis using the homoscedastic covariance model.

detected by the SimpReg analysis but failed to show up in the ClusReg analysis (designated as class II genes). A common feature of class I genes is that the observed points are spread widely around the regression line, which is clearly in contrast to the opposite feature shared by class II genes whose observed points are all concentrated around the regression line. In addition, class I genes tend to have a steeper slope than class II genes. The conclusion is that the two methods detect different sets of genes because they use different criteria of gene detection. The ClusReg method classifies genes based on the magnitudes of the regression coefficients while the SimpReg method classifies genes based on the magnitudes of the correlation coefficients. A gene whose observed data points are widely spread (i.e. with a large error) is unlikely to be detected with the SimpReg analysis because the correlation coefficient tends to be low even if the regression coefficient may be large. On the other hand, a gene whose observed data points are highly concentrated (i.e. with a small error) is likely to be detected with the SimpReg analysis even though the regression coefficient may be small. These genes are likely to be excluded by the ClusReg analysis.

4. Discussion

The proposed study of gene expression associated with a quantitative trait phenotype differs from quantitative trait locus (QTL) mapping in several respects. In QTL mapping, the response variable is the phenotype of a quantitative trait and the independent variables are (discrete) genotype indicator variables of QTL (pieces of DNA on the genome). In the phenotype-associated gene expression study, the discrete genotype indicator variables of QTL are replaced by the continuously distributed gene expression variables. Because the number of microarrayed genes can be extremely large and the number of microarrayed individuals is usually small, we flipped the roles of phenotype and gene expression in the linear model by treating gene expression as response variables and the phenotype as the independent variable. Since there are multiple gene expression variables involved in a microarray experiment, the problem becomes a multivariate linear model problem. However, traditional multivariate analysis is incapable of handling such a high dimensionality of the multivariate model. As a result, we proposed the model-based clustering method and treated it as a special dimension reduction approach. The phenotype-associated gene expression analysis also differs from expression QTL (eQTL) analysis (Schadt *et al*., 2003) in that the expression levels of genes are treated as the response variables whereas the genotype indicators of marker loci are treated as independent variables. The quantitative phenotype of a trait plays no role in an eQTL analysis other than being used to select important markers for inclusion in the analysis. Theoretically, one can analyse gene expression, quantitative phenotype and markers jointly in a single step.

For the first time, we developed a clustering method to classify expressed genes based on their association with a continuously distributed disease phenotype.

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

Table 7. *Comparison of ClusReg analysis with SimpReg analyses: (a) comparison of the numbers of genes detected by ClusReg and SimpReg analyses for association with Alzheimer's disease when the cut-off P value was 0·00064 (the expected FDR was 0·05); (b) comparison of the numbers of genes detected by ClusReg and SimpReg analyses for association with Alzheimer's disease when the cut-off P value was 0·0032 (the expected FDR was 0·08)*

		Clustering			
	Correlation	Cluster 1	Cluster 2	Cluster 3	Sum
a	Group 1 (negative association)	6	40	0	46
	Group 2 (neutral)	40	9286	303	9629
	Group 3 (positive association)	0	40	39	79
b	Group 1 (negative association)	12	149	0	161
	Group 2 (neutral)	34	9087	245	9366
	Group 3 (positive association)	0	130	97	227
	Sum	46	9366	342	

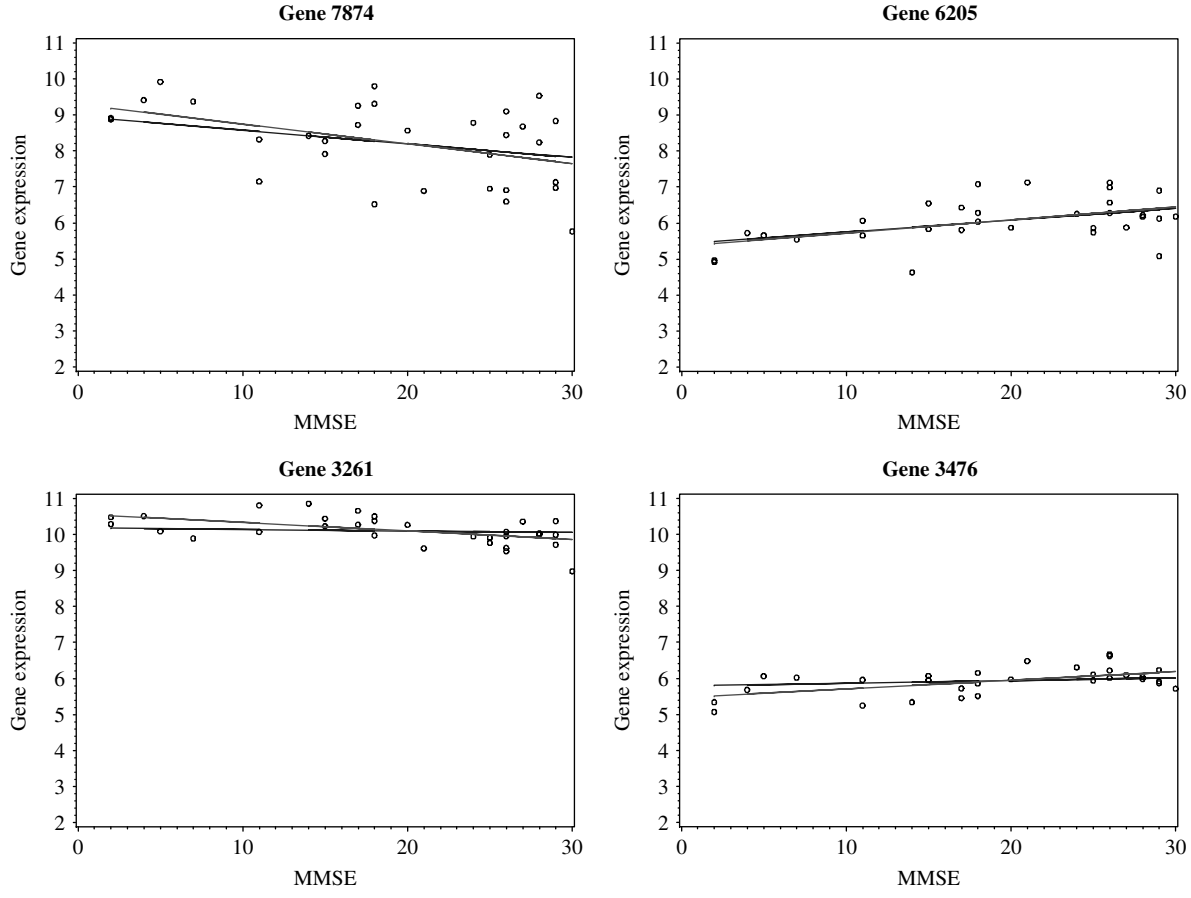

Fig. 3. Scatter plots of the expression levels of four genes against MMSE phenotype along with the regression lines. Genes 7874 and 6205 (upper panels) were detected by the ClusReg analysis but failed to show statistical significance in the SimpReg analysis, whereas genes 3261 and 3476 (lower panels) were detected by the SimpReg analysis but missed by the ClusReg analysis. The dark and light continuous lines represent the regression lines of ClusReg analysis and SimpReg analysis, respectively.

Simple regression analysis has been applied to this type of microarray analysis (Blalock *et al*., 2004). However, genes were analysed separately in the SimpReg analysis. We strongly believe that joint analysis implemented via ClusReg is more meaningful than SimpReg in revealing the connection between genes and phenotype. In the SimpReg analysis, choosing the appropriate significance level for the

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

correlation coefficients is somewhat arbitrary, which makes SimpReg analysis inconsistent (see Table 3). The model-based ClusReg analysis classifies genes based on the regression coefficients, not based on the individual *P* values. In the ClusReg analysis, all genes are analysed simultaneously in a single model, which may avoid all potential problems associated with the separate analysis. Because genes clustered in the same group are combined and reported as a group, information has been greatly increased.

Replication is always needed in various microarray experimental designs for specific purposes (Churchill, 2002). Replication in the phenotype-associated microarray experiment means that the same individual should be microarrayed more than once. With such a replication, we are able to partition the residual error into biological error (between individuals) and technical error (within individuals). The model developed in this study is suitable for non-replicated experiments, such as the Alzheimer's disease microarray experiment (Blalock *et al*., 2004). However, it can be modified with little extra effort to handle data from replicated experiments. For replicated data, we need to modify the design matrix, X, and the dimensions of all other matrices. Technical details of the modification are given in Appendix B for interested readers.

We searched the NCBI gene bank and found that 117 genes have been reported to be related to Alzheimer's disease. Among the 117 genes, 64 of them appeared in the list of genes analysed in this study. Our ClusReg analysis detected 14 of the 64 genes whereas the SimpReg analysis detected only eight. Six genes were detected by both methods. We examined all 50 genes that failed to be detected by ClusReg ($64-14=50$) and plotted the expression profiles. All the profiles were very flat (with no particular trends; data not shown). These genes did not show any association with the phenotype we analysed. There are several explanations for this: (1) these genes may affect the disease through other phenotypes, (2) their expression levels do not change across any phenotype of interest (like housekeeping genes) but their gene products are essential to the function of other genes that are directly related to the disease, (3) microarray expression experiments may have some limitation for detecting these genes (microarray is not a technology applicable to all biological problems), (4) there might be some subtle non-linear relationships between the expression and the phenotypes, which failed to be captured by the linear regression analysis, and (5) the small sample size (31 subjects is not a large group). Although incorporating multiple traits in the model may increase power, we believe that the ClusReg analysis based on one phenotype is already very informative. It detected more indexed genes than the SimpReg analysis. The majority of the genes detected by the ClusReg analysis (highly associated with the disease phenotype) have not been reported or studied, leaving a tremendous gap for molecular biologists to fill.

Clustering genes based on their association with a phenotype is a true functional genomics study. If the phenotype is an important disease in humans or an economically important trait in agricultural species, the technique will provide a way to identify genes in the metabolic pathways. In the Alzheimer's disease analysis (Blalock *et al*., 2004), the subjects are a group of genetically heterogeneous individuals. Individuals vary due to genotypic differences as well as environmental variants. The functional genes identified are confounded with genotypic differences. This kind of confounding can be eliminated by choosing genetically identical individuals (sampled from an inbred line) as material for the microarray experiment. However, many microarray experiments using inbred laboratory animals have been designed for different purposes, such as detecting differential gene expression responding to a drug treatment, which is not a functional genomics study. If there is a continuous phenotype associated with each individual from the inbred line, genes expressions can be analysed using the model proposed in this study. The expressed genes identified in such a study would truly reflect the difference in expression of the same allele because all individuals carry exactly the same genotype.

Cluster analysis aims at classifying genes into different groups. Model-based ClusReg analysis, however, also facilitates statistical tests, not for individual genes but for gene clusters, although statistical test has never been the focus of cluster analysis. We proposed the Wald test statistic for testing significance of cluster means from zero. If the test statistic is significant, all genes in the cluster are declared as significant. We have successfully applied this test statistic to identify significant clusters. In the real data analysis, we reported genes in the extreme clusters (with a high test statistic value). The proposed Wald test statistic would allow researchers to study the statistical power of the cluster analysis for gene identification via simulation experiments.

We proposed two structures of the covariance matrix for the random regression coefficients: homoscedastic and heteroscedastic. Both structures have been tested and worked satisfactorily. For the simulated data, clustering results were almost identical, showing that the homoscedastic model is robust. A general suggestion on selection between the two models is to use the homoscedastic model when the dimension of the covariance matrix is high and the number of clusters is large. In the opposite situation, the data are considered to be sufficiently rich to allow the use of the heteroscedastic model. Theoretically, it is possible to use the BIC score to select the appropriate model. The homoscedastic model is nested

within the heteroscedastic model so that the likelihood value of the latter is always greater than the former. Incorporating the penalty of the large number of parameters in the heteroscedastic model will eventually lead to an optimal model choice.

Computer note. Data were analysed using programs written in SAS (SAS Institute, 1989). The SAS code of the programs and the data (including both the simulated data and the real data collected from the Alzheimer's disease microarray experiment) are available on request.

We are grateful to Drs E. M. Blalock and P. W. Landfield of the University of Kentucky for providing the microarray data and the Alzheimer's disease phenotypes. Dr Blalock also provided comments on the first draft of the manuscript for improvement. We thank the Editor Trudy Mackay and two anonymous reviewers for their constructive comments on an early version of the manuscript. The research was supported by the National Institute of Health Grant R01-GM55321 to S.X.

Appendix A. Expectation of the complete-data log likelihood

The missing values are the cluster labels (Z_i) and the random regression coefficients (γ_i). Conditional on these missing values, the density of Y_i is

$$p(Y_i|Z_i,\gamma_i;\beta,\sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{c}\delta(Z_i,k)(Y_i-X\beta_k-X\gamma_i)^{\mathrm{T}} \times (Y_i-X\beta_k-X\gamma_i)\right] \quad (A1)$$

where $\beta=\{\beta_k\}$ is the array of mean regression coefficients for all clusters and

$$\delta(Z_i,k)=\begin{cases}1 & \text{if } Z_i=k\\ 0 & \text{otherwise}\end{cases} \quad (A2)$$

is an indicator variable for the cluster label, which has a multinomial distribution with probability

$$p(Z_i;\pi)\propto\prod_{k=1}^{c}\pi_k^{\delta(Z_i,k)}. \quad (A3)$$

The random regression coefficient has a normal distribution with probability density

$$p(\gamma_i;\Sigma)\propto\frac{1}{|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}\gamma_i^{\mathrm{T}}\Sigma^{-1}\gamma_i\right]. \quad (A4)$$

Note that equation (6) in the main text is derived from $p(Y_i|Z_i=k,\gamma_i;\beta,\sigma^2)$ using

$$p(Y_i|Z_i=k;\beta_k,\Sigma,\sigma^2)=\int p(Y_i|Z_i=k,\gamma_i;\beta,\sigma^2)\times p(\gamma_i;\Sigma)d\gamma_i. \quad (A5)$$

The joint density of $\{Y_i, Z_i, \gamma_i\}$ is

$$p(Y_i,Z_i,\gamma_i;\pi,\beta,\Sigma,\sigma^2)\propto p(Y_i|Z_i,\gamma_i;\beta,\sigma^2)p(Z_i;\pi)p(\gamma_i;\Sigma). \quad (A6)$$

The joint distribution of the data and the missing values for all genes simply takes the product of all gene specific densities,

$$p(Y,Z,\gamma;\pi,\beta,\Sigma,\sigma^2)\propto\prod_{i=1}^{m}p(Y_i|Z_i,\gamma_i;\beta,\sigma^2)p(Z_i;\pi)p(\gamma_i;\Sigma). \quad (A7)$$

Let $\psi=\{\pi,\beta,\Sigma,\sigma^2\}$ be the array of parameters. Given the missing values, the log likelihood function is

$$\begin{aligned}L_{Z,\gamma}(\psi)&=\ln p(Y,Z,\gamma;\pi,\beta,\Sigma,\sigma^2)\\&=\sum_{i=1}^{m}[\ln p(Y_i|Z_i,\gamma_i;\beta,\sigma^2)\\&\quad+\ln p(Z_i;\pi)+\ln p(\gamma_i;\Sigma)].\end{aligned} \quad (A8)$$

This likelihood function is called the complete-data log likelihood. In contrast to this likelihood, the likelihood function defined in the main text (equation 5) is called the observed (or incomplete-data) log likelihood. The EM algorithm requires many steps of iteration to achieve the final MLE of ψ. In each step, however, the target function that is maximized is not the observed log likelihood but the so-called expected complete-data log likelihood function, i.e.

$$eL(\psi)=E[L_{Z,\gamma}(\psi)]=T_1(\psi)+T_2(\psi)+T_3(\psi) \quad (A9)$$

where the expectation is taken with respect to the missing values Z (in the form of δ) and γ conditional on ψ and Y. The three components of (A9) are

$$\begin{aligned}T_1(\psi)&=\sum_{i=1}^{m}E[\ln p(Y_i|Z_i,\gamma_i;\beta,\sigma^2)]\\&=-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left\{\sum_{k=1}^{c}E[\delta(Z_i,k)(Y_i-X\beta_k-X\gamma_i)^{\mathrm{T}}\times(Y_i-X\beta_k-X\gamma_i)]\right\}-\frac{mn}{2}\ln(\sigma^2),\end{aligned} \quad (A10)$$

$$T_2(\psi)=\sum_{i=1}^{m}E[\ln p(Z_i;\pi)]=\sum_{i=1}^{m}\sum_{k=1}^{c}E[\delta(Z_i,k)]\ln\pi_k \quad (A11)$$

and

$$\begin{aligned}T_3(\psi)&=\sum_{i=1}^{m}E[\ln p(\gamma_i;\Sigma)]\\&=-\frac{1}{2}\sum_{i=1}^{m}E[\gamma_i^{T}\Sigma^{-1}\gamma_i]-\frac{m}{2}\ln(\Sigma).\end{aligned} \quad (A12)$$

The E-steps of the EM algorithm involve calculating the expectation of terms containing the missing

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

values (δ and γ). In the M-steps, we take

$$\frac{\partial}{\partial\psi}eL(\psi)=\frac{\partial}{\partial\psi}T_1(\psi)+\frac{\partial}{\partial\psi}T_2(\psi)+\frac{\partial}{\partial\psi}T_3(\psi)=0 \tag{A13}$$

and solve for ψ. Given the expectations of the terms involving the missing values, the solutions of ψ in the M-steps have closed forms, which are given in the main text.

Appendix B. Algorithm for replicated microarray experiments

For simplicity, we assume that each individual is microarrayed r times (equal replication). Let $N=nr$ and Y_i becomes an $N\times 1$ vector, which is described by the following linear model:

$$Y_i|_{Z_i=k}=WX\beta_k+WX\gamma_i+W\eta_i+\varepsilon_i \tag{B1}$$

where

$$W=\begin{bmatrix} J & 0 & \cdots & 0\\ 0 & J & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & J\end{bmatrix} \tag{B2}$$

is a block diagonal matrix and J is an $r\times 1$ column vector of unity. If the number of replicates r varies across individuals, the unity vector J should have a variable dimension across individuals. Here, we introduce an additional vector of errors, $\eta_i=[\eta_{i1}\ \cdots\ \eta_{in}]^T$, called the between-individual errors or biological errors. In contrast to η_i, the residual errors ε_i are called the within-individual errors or simply technical errors. We assume $\eta_i\sim N(0, I\sigma_\eta^2)$, where I is an identity matrix with dimension n. Given $Z_i=k$, the expectation and covariance matrices of Y_i are

$$\mu_k=E(Y_i|Z_i=k)=WX\beta_k \tag{B3}$$

and

$$V=Var(Y_i|Z_i=k)=WX\Sigma X^{\mathrm{T}}W^{\mathrm{T}}+WW^{\mathrm{T}}\sigma_\eta^2+I\sigma^2, \tag{B4}$$

respectively.

The EM algorithm remains largely the same as that of the non-replicated experimental data analysis. Here, we only emphasize the differences. The parameter vector now includes an additional parameter σ_η^2. The list of missing values now include an extra vector η_i.

Let us define the conditional expectations and variances of the random effects as

$$\hat{\gamma}_i=E(\gamma_i|Y_i,Z_i=k)=\Sigma X^{\mathrm{T}}W^{\mathrm{T}}(WX\Sigma X^{\mathrm{T}}W^{\mathrm{T}}+WW^{\mathrm{T}}\sigma_\eta^2+I\sigma^2)^{-1}(Y_i-WX\beta_k), \tag{B5}$$

$$\hat{S}_i=Var(\gamma_i|Y_i,Z_i=k)=\Sigma-\Sigma X^{\mathrm{T}}W^{\mathrm{T}}(WX\Sigma X^{\mathrm{T}}W^{\mathrm{T}}+WW^{\mathrm{T}}\sigma_\eta^2+I\sigma^2)^{-1}WX\Sigma, \tag{B6}$$

$$\hat{\eta}_i=E(\eta_i|Y_i,Z_i=k)=\sigma_\eta^2W^{\mathrm{T}}(WX\Sigma X^{\mathrm{T}}W^{\mathrm{T}}+WW^{\mathrm{T}}\sigma_\eta^2+I\sigma^2)^{-1}(Y_i-WX\beta_k), \tag{B7}$$

$$\hat{R}_i=Var(\eta_i|Y_i,Z_i=k)=I\sigma_\eta^2-\sigma_\eta^2W^{\mathrm{T}}(WX\Sigma X^{\mathrm{T}}W^{\mathrm{T}}+WW^{\mathrm{T}}\sigma_\eta^2+I\sigma^2)^{-1}W\sigma_\eta^2. \tag{B8}$$

Finding these quantities represents the E-step. The M-step includes computing the following terms:

$$\hat{\beta}_k=(\pi_k mX^{\mathrm{T}}W^{\mathrm{T}}WX)^{-1}\sum_{i=1}^{m}\pi_{ik}X^{\mathrm{T}}W^{\mathrm{T}}(Y_i-WX\hat{\gamma}_i-W\hat{\eta}_i), \tag{B9}$$

$$\hat{\Sigma}=\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{c}\pi_{ik}(\hat{\gamma}_i\hat{\gamma}_i^{\mathrm{T}}+\hat{S}_i), \tag{B10}$$

$$\hat{\sigma}_\eta^2=\frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{c}\pi_{ik}(\hat{\eta}_i^{\mathrm{T}}\hat{\eta}_i+\hat{R}_i), \tag{B11}$$

$$\hat{\sigma}^2=\frac{1}{mnr}\sum_{i=1}^{m}\sum_{k=1}^{c}\pi_{ik}Y_i^T(Y_i-WX\hat{\beta}_k-WX\hat{\gamma}_i-W\hat{\eta}_i). \tag{B12}$$

The detailed EM-steps follow those of the non-replicated microarray data analysis described in the main text.

References

Anholt, R. R. H. & Mackay, T. F. C. (2004). Quantitative genetic analyses of complex behaviours in *Drosophila*. *Nature Reviews Genetics* **5**, 838–849.

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

Ben-Dor, A. & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology* **6**, 281–298.

Benjamini, Y. & Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* **82**, 163–170.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R. & Landfield, P. W. (2004). Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumour suppressor responses. *Proceedings of the National Academy of Sciences of the USA* **101**, 2173–2178.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines.

https://doi.org/10.1017/S0016672305007822 Published online by Cambridge University Press

Proceedings of the National Academy of Sciences of the USA **97**, 262–267.
Carr, D. B., Somogyi, R. & Michaels, G. (1997). Templates for looking at the gene expression clustering. *Statistical Computing and Statistical Graphics Newsletter* **8**, 20–29.
Chu, T. M., Weir, B. & Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* **176**, 35–51.
Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495.
Cui, X. Q. & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.1–210.10.
Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
Desprez, T., Amselem, J., Caboche, M. & Hofte, H. (1998). Differential gene expression in *Arabidopsis* monitored using cDNA arrays. *Plant Journal* **14**, 643–652.
Devore, J. & Peck, R. (1997). *Statistics: The Exploration and Analysis of Data*, 3rd edn. Pacific Grove, CA: Duxbery Press.
Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* **95**, 14863–14868.
Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Berlin: Springer.
Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.
Ghosh, D. & Chinnaiyan, A. M. (2002). Mixture modeling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275–286.
Han, Y. P., Zhou, D. S., Pang, X., Song, Y. J., Zhang, L., Bao, J. Y., Tong, Z. Z., Wang, J., Guo, Z. B., Zhai, J. H., Du, Z. M., Wang, X. Y., Zhang, X. Q., Wang, J., Huang, P. T. & Yang, R. F. (2004). Microarray analysis of temperature-induced transcriptome of *Yersinia pestis*. *Microbiology and Immunology* **48**, 791–805.
Herrero, J., Valencia, A. & Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **17**, 126–136.
Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.
Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.
Lazarov, O., Robinson, J., Tang, Y. P., Hairston, I. S., Korade-Mirnics, Z., Lee, V. M. Y., Hersh, L. B., Sapolsky, R. M., Mirnics, K. & Sisodia, S. S. (2005). Environmental enrichment reduces Aβ levels and amyloid deposition in transgenic mice. *Cell* **120**, 701–713.
Luan, Y. & Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19**, 474–482.
Mateos, A., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M. & Stolovizky, G. (2002). Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Research* **12**, 1703–1715.
McLachlan, G. J., Bean, R. W. & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.
Peddada, S. D., Lobenhofer, E. K., Li, L. P., Afshari, C. A., Weinberg, C. R. & Umbach, D. M. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834–841.
Qu, Y. & Xu, S. (2004). Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics* **20**, 1905–1913.
Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–32.
Saban, M. R., Hellmich, H., Nguyen, N. B., Winston, J., Hammond, T. G. & Saban, R. (2001). Time course of LPS-induced gene expression in a mouse model of genitourinary inflammation. *Physiological Genomics* **5**, 147–160.
SAS Institute (1989). *SAS/IML Software: Usage and Reference*, version 6, 1st edn. Cary, NC: SAS Institute.
Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B. & Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.
Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* **270**, 467–470.
Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 218–285.
Thomas, J. G., Olson, J. M., Tapscott, S. J. & Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **11**, 1227–1236.
Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA* **98**, 5116–5121.
Wayne, M. L. & Mclntyre, L. M. (2002). Combining mapping and arraying: an approach to candidate gene identification. *Proceedings of the National Academy of Sciences of the USA* **99**, 14903–14906.
Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Rushel, P., Afshari, C. & Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.
Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.
Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272.