Inference on differences between classes using cluster-specific contrasts of mixed effects

SHU KAY NG*
School of Medicine, Griffith Health Institute, Griffith University, Meadowbrook, QLD 4131, Australia
s.ng@griffith.edu.au

GEOFFREY J. MCLACHLAN, KUI WANG
Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia

ZOLTAN NAGYMANYOKI, SHUBAI LIU, SHU-WING NG
Laboratory of Gynecologic Oncology, Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women’s Hospital, Boston, MA 02115, USA

SUMMARY
The detection of differentially expressed (DE) genes, that is, genes whose expression levels vary between two or more classes representing different experimental conditions (say, diseases), is one of the most commonly studied problems in bioinformatics. For example, the identification of DE genes between distinct disease phenotypes is an important first step in understanding and developing treatment drugs for the disease. We present a novel approach to the problem of detecting DE genes that is based on a test statistic formed as a weighted (normalized) cluster-specific contrast in the mixed effects of the mixture model used in the first instance to cluster the gene profiles into a manageable number of clusters. The key factor in the formation of our test statistic is the use of gene-specific mixed effects in the cluster-specific contrast. It thus means that the (soft) assignment of a given gene to a cluster is not crucial. This is because in addition to class differences between the (estimated) fixed effects terms for a cluster, gene-specific class differences also contribute to the cluster-specific contributions to the final form of the test statistic. The proposed test statistic can be used where the primary aim is to rank the genes in order of evidence against the null hypothesis of no DE. We also show how a $P$-value can be calculated for each gene for use in multiple hypothesis testing where the intent is to control the false discovery rate (FDR) at some desired level. With the use of publicly available and simulated datasets, we show that the proposed contrast-based approach outperforms other methods commonly used for the detection of DE genes both in a ranking context with lower proportion of false discoveries and in a multiple hypothesis testing context with higher power for a specified level of the FDR.

Keywords: Contrast; Differential expression; Mixture model; Random effects modeling.

*To whom correspondence should be addressed.
1. Introduction

In the analysis of multivariate data from samples in \( m \) known classes of \( p \) objects, one of the major goals is the inference on differences in observations between classes. This leads to the identification of relevant features that differentiate the classes and the prediction of the class of origin for unclassified objects. An example of such an application is the analysis of gene-expression data, where the aim is to detect genes (features) that are differentially expressed (DE) in a known number \( m \) of classes (conditions), \( C_1, \ldots, C_m \). These \( m \) classes may correspond to tissues (cells) that are at different stages in some process or in distinct pathological states. For example, one can compare healthy cells to cancerous cells to learn which genes tend to be over- or underexpressed in the diseased cells. In this context, the intent is also to select a small subset of “marker” genes that characterize the different tissue classes and construct a classifier to predict the class of origin of an unclassified tissue sample with respect to one of a number of distinct disease phenotypes. Studies suggested that classifiers based on a small set of marker genes can classify more accurately tumor subtypes than can standard clinical criteria (McLachlan and others, 2004).

An obvious test statistic to employ in the detection of DE between two classes for a given gene is to form the usual two-sample (pooled) \( t \)-test statistic. However, as small class-sample sizes are often used for microarray experiments, there can be problems with a poor estimate of the variance in the denominator of the \( t \)-statistic if each gene is considered independently. One attempt “at borrowing strength” across the tests is to adopt a moderated form of the \( t \)-statistic whereby the gene-specific variance is weighted with a contribution from all the genes (Smyth, 2004). Another commonly used attempt to involve all the gene profiles in the formation of a test statistic for a gene is to partition the gene profiles into clusters (Dahl and Newton, 2007; Newton and others, 2004; Pan and others, 2002). Existing clustering-based methods have been implemented by assuming the existence of pure clusters of up- and downregulated genes, which are reflected by large class differences in the cluster-specific (estimated) mean expressions for the tissues (He and others, 2006; Qiu and others, 2008). In an ideal situation, there would be three clusters corresponding to null genes, upregulated DE genes, and downregulated DE genes. But in practice, there is a need for more than three clusters since not all the upregulated genes are assigned to the one cluster and, similarly, for the downregulated genes. With more than three clusters, there is the problem of how to identify the clusters corresponding null genes and up- and downregulated DE genes, assuming that the clusters are pure. One way of approaching this problem has been to adopt an hierarchical approach whereby each cluster is decomposed into three subclusters representing null, upregulated DE, and downregulated genes (Yuan and Kendziorski, 2006).

Given pure clusters, the question of whether a gene is DE is decided on the basis of its (estimated) posterior probabilities of membership with respect to the various clusters. However, the intent of obtaining pure clusters is not always possible or verifiable. For example, a gene might have a high estimated posterior probability of belonging to a cluster taken to correspond to DE genes. But it can still be quite atypical of the cluster to which it is assigned and, indeed, be a null gene as to be illustrated below. If no gene-specific differences are included in the contrast, the cluster-specific contrasts for tests on differences between the classes are the same for all genes in a cluster and can thus be misleading.

In our approach, we thus include gene-specific random effects terms in forming a cluster-specific contrast. Our initial clustering of the gene profiles into a number \( g \) of clusters is effected by the fitting of a mixture of linear mixed models (LMMs) that include random effects terms specific to the genes in addition to the class-specific fixed effects terms. This model extends earlier work of Ng and others (2006) by allowing correlated gene-specific random effects for the \( m \) tissue classes to follow a multivariate normal distribution, which gives a more flexible model with an additional parameter explicitly governing and accounting for the relationship between gene expressions among the tissue classes due to individual gene effects; see Section 2. The component LMMs also include random effects terms shared by all genes belonging to the same component of the mixture model, which implies that these genes are not assumed to be
independently distributed as usually assumed in other clustering or multiple hypothesis testing approaches; see also Desai and Storey (2012) and Qin and Self (2006). The choice of the number \( g \) of clusters is made by consideration of the likelihood via the BIC criterion (Ng and others, 2006). However, the choice of \( g \) is not crucial in that the subsequent use of them in our approach does not rely on the clusters being pure as to whether all cluster members are DE or not DE (null). In our approach, a (normalized) contrast for the test of no DE is formed based on class differences between the cluster-specific fixed effects terms for each class in addition to class differences between the gene-specific random effects terms. As discussed below, the inclusion of the latter terms in the cluster-specific contrast means that a gene not typical of the cluster to which has been assigned outright or in a soft manner by the initial clustering of the genes can modify the differences between the fixed effects terms for the cluster. The final form of our proposed test statistic is formed by weighting the cluster-specific (normalized) contrasts over the clusters. The test statistic can be used to rank the genes in order of evidence against the null hypothesis of no DE, which is often the primary aim of gene discovery experiments for which microarrays are designed. The distributional assumptions do not have to hold for the test statistic to perform well in ranking the genes. Our proposed test statistic can also be used to carry multiple hypothesis testing where the intent is to control the false discovery rate (FDR) at or below a specified level (Efron, 2010). In the case of calculating the \( P \)-value, we adopt a permutation approach to provide an approximation to the null distribution of the test statistic and hence to the calculation of the associated \( P \)-value.

2. Methods

We let \( p_h \) be the number of samples in the \( h \)th tissue class \( C_h \) \((h = 1, \ldots, m)\). The total number of measurements for each gene is \( p = \sum_{h=1}^{m} p_h \). Thus the gene-expression data can be represented by an \( n \times p \) matrix, where \( n \) is the number of genes. We let \( y_j = (y_{1j}, \ldots, y_{pj})^T \) contain the measurements on the \( j \)th gene, where superscript \( T \) denotes vector transpose. It is postulated that \( y_j \) has a \( g \)-component mixture distribution with probability \( \pi_i \) of belonging to the \( i \)th component \( G_i (i = 1, \ldots, g) \), where the \( \pi_i \) sum to one. We let the \( g \)-dimensional vector \( z_j \) denote the component membership of \( y_j \), where \( z_{ij} = (z_j)_i = 1 \) if \( y_j \) belongs to the \( i \)th component and zero otherwise \((i = 1, \ldots, g)\). We put \( y = (y_1^T, \ldots, y_n^T)^T \) and \( z = (z_1^T, \ldots, z_n^T)^T \).

2.1 The model

Conditional on its membership of the \( i \)th component \( G_i \), the distribution of \( y_j \) is specified by the LMM

\[
y_j = X\beta_i + U b_{ij} + V c_i + \epsilon_{ij},
\]

where \( X \) (a \( p \times m \) matrix), \( U \), and \( V \) denote the design matrices corresponding, respectively, to the fixed effects terms \( \beta_i \) and to the random effects terms \( b_{ij} \) and \( c_i \)(\( i = 1, \ldots, g \)).

The vector \( b_{ij} = (b_{1ij}, \ldots, b_{mij})^T \) contains the gene-specific random effects for each of the \( m \) tissue classes, and \( c_i = (c_{1i}, \ldots, c_{pi})^T \) contains the random effects common to all genes from the \( i \)th component. The latter induces a correlation between those genes from the same component and is an attempt to allow for the fact that in reality the gene profile vectors are not all independently distributed. The gene-specific random effects \( b_{hij} (h = 1, \ldots, m) \) allow for correlation between the gene expressions across the tissues both within a class and between classes for the same gene. The expressions on a given gene should be independent, but this may not hold in practice due to poor experimental conditions resulting in batch-effects.
The measurement error vector \( \varepsilon_{ij} \) is taken to be multivariate normal \( N_p(0, A_i) \), where \( A_i \) is a diagonal matrix. The vectors \( b_{ij} \) and \( c_i \) of random effects terms are taken to be multivariate normal \( N_m(0, B_i) \) and \( N_p(0, C_i) \), respectively, where \( C_i \) is assumed to be diagonal and \( B_i \) is a non-diagonal matrix given by

\[
B_i = \begin{pmatrix}
\sigma_{h1}^2 & \rho \sigma_{b1} \sigma_{b2} & \cdots & \rho \sigma_{b1} \sigma_{bmi} \\
\rho \sigma_{b2} \sigma_{b1} & \sigma_{b2}^2 & \cdots & \rho \sigma_{b2} \sigma_{bmi} \\
\vdots & \vdots & \ddots & \vdots \\
\rho \sigma_{bmi} \sigma_{b1} & \rho \sigma_{bmi} \sigma_{b2} & \cdots & \sigma_{bmi}^2
\end{pmatrix},
\]

where \( \rho \) accounts for the correlation between gene-specific random effects \( b_{hi} (h = 1, \ldots, m) \), which are shared, respectively, among the expressions on the \( i \)th gene in the \( h \)th tissue class (see Section 1 of the supplementary material available at *Biostatistics* online for the covariance structure of the unconditional distribution of the gene-expression profiles under (2.1) and (2.2)).

### 2.2 The test statistic

We now define our test statistic for the ranking of the genes in the order of their significance of being DE. We proceed initially under the assumption that each gene profile is classified with respect to the \( g \) components \( G_1, \ldots, G_g \) in its mixture distribution with components specified by (1); that is, \( z \) is known. We also assume initially that the vector \( \xi_i \) containing the distinct elements in the component-covariance matrices \( A_i, B_i, \) and \( C_i \) is known.

We let \( b_{Gi} \) be the \( (mn_i) \)-dimensional vector containing the random effects terms for the \( n_i \) genes belonging to the \( i \)th component of the mixture model \( (i = 1, \ldots, g) \). We can write \( b_{Gi} \) as

\[
b_{Gi} = (b_{i1}^T, \ldots, b_{in_i}^T)^T,
\]

where \( i_1, \ldots, i_{n_i} \) denote the labels of the \( n_i \) genes belonging to the \( i \)th component \( G_i (i = 1, \ldots, g) \) of the mixture model. We let

\[
r_i = (\beta_i^T, b_{Gi}^T, c_i^T)^T \quad (i = 1, \ldots, g)
\]

be the vector containing the fixed and random effects for the \( n_i \) genes belonging to the \( i \)th component \( G_i (i = 1, \ldots, g) \). For an individual gene \( j \) belonging to the \( i \)th component \( G_i \), we can form the cluster-specific normalized contrast \( S_{ij} \) given by

\[
S_{ij} = d_j^T r_i / \lambda_{ij} \quad (i = 1, \ldots, g),
\]

where \( d_j \) is a vector whose elements sum to zero and \( \lambda_{ij} \) is the normalizing term. The choice of \( d_j \) has direct implication on the inference space of the contrast \( S_{ij} \). For the case of \( m = 2 \) classes of tissue samples, a typical form for \( d_j \) is

\[
d_j^T = (1 -1 : 0 0, \ldots, 0 0, 1 -1, 0 0, \ldots : 0 \ldots 0),
\]

where only one pair of \( (1 -1) \) exists in the second partition corresponding to the gene-specific random effects \( b_{ij} \) for a gene in the \( i \)th cluster. The contrast (2.6) represents an “intermediate inference space” in that the inference is “narrow” to gene-specific random effects \( b_{ij} \) but “broad” to tissue-specific random effects \( c_i \) (McLean and others, 1991). This means that a contrast of differential expressions between two classes of tissues is being considered and the inference applies to the specific genes studied in the experiment (narrow) and to the entire population from which biological tissue samples were obtained (broad).
Our proposed test statistic is based on an estimate \( \hat{S}_{ij} \) of \( S_{ij} \) obtained by replacing \( r_i \) with an estimate \( \hat{r}_i \) in the right-hand side of (2.5) and now taking the normalizing term \( \lambda_{ij} \) to be the standard error of \( d_j^T \hat{r}_i \) (conditional on membership of the \( j \)th gene to the \( i \)th component \( G_i \)). That is,

\[
\hat{S}_{ij} = d_j^T \hat{r}_i / \lambda_{ij},
\]

where \( \hat{r}_i \) is an estimate of the vector \( r_i \) of fixed and random effects terms and

\[
\lambda_{ij} = \sqrt{d_j^T \text{cov}(\hat{r}_i)d_j}.
\]

In the case where \( \hat{r}_i \) is the BLUP estimator of \( r_i \),

\[
\text{cov}(\hat{r}_i) = \Omega_i(\zeta_i; z),
\]

where \( \Omega_i \) is defined as follows.

The matrix \( \Omega_i \) (and \( d_j \)) is partitioned conformally corresponding to \( \beta_i | \theta_{G_i} | c_i \) with dimensions \( m, mn_i \), and \( p \), respectively. That is,

\[
\Omega_i = \begin{bmatrix} \Omega_{i\beta} & \Omega_{i\beta b} & \Omega_{i\beta c} \\ \Omega_{i\beta b}^T & \Omega_{i b} & \Omega_{i bc} \\ \Omega_{i\beta c}^T & \Omega_{i bc}^T & \Omega_{i c} \end{bmatrix}^{-1},
\]

and

\[
\Omega_{i\beta} = n_i X_i^T A_i^{-1} X, \quad \Omega_{i\beta b} = 1_{n_i} \otimes X_i^T A_i^{-1} U, \quad \Omega_{i\beta c} = n_i X_i^T A_i^{-1} V, \\
\Omega_{i b} = I_{n_i} \otimes (U_i^T A_i^{-1} U + B_i^{-1}), \quad \Omega_{i bc} = 1_{n_i} \otimes U_i^T A_i^{-1} V, \quad \Omega_{i c} = n_i (V_i^T A_i^{-1} V + C_i^{-1}),
\]

where \( 1_{n_i} \) is a \( n_i \)-dimensional vector of ones, \( I_{n_i} \) is an identity matrix with dimension \( n_i \), and the sign \( \otimes \) denotes the Kronecker product of two matrices.

The computation of the inverse matrix (2.10) is not straightforward as the partition corresponding to \( b_{G_i} \) involves a large dimensional block matrix; see Section 2 of the supplementary material available at Biostatistics online for the derivation of (2.10) in terms of (2.11). In evaluating \( \hat{\Omega}_i \) in (2.10), it is noted that the empirical approximation, using the maximum likelihood (ML) estimates \( (A_i, \hat{B}_i, \hat{C}_i) \), tends to underestimate the true sampling variability of the estimated contrast \( \hat{S}_{ij} \). It is because the uncertainty in estimating the variance components is not accounted for. Based on a nonparametric bootstrap method, we found that the variances of the variance component estimates are generally small (contributing less than 5% of variance component estimates). Thus, the use of estimated variance components does not appear to introduce a large bias on \( \hat{\Omega}_i \); see McLean and others (1991).

Now in practice we do not know the classification \( z \) of the gene profiles with respect to the \( g \) components \( G_1, \ldots, G_g \) nor the vector \( \zeta_i \) of variances/covariances. We therefore fit a mixture of \( g \)-component LMM distributions as defined by (2.1) with the value of \( g \) is chosen according to the BIC criterion. The vector \( \Psi \) containing the unknown mixing proportions \( \pi_i \), the fixed effects \( \beta_i \), and the vector \( \zeta_i \) of variances/covariances \( i = 1, \ldots, g \) is estimated by ML. The random effects terms \( b_{G_i} \) and \( c_i \) are estimated by \( \hat{b}_{G_i} = E_{\hat{\Psi}}[b_{G_i} | y_j, z_{ij} = 1] \) and \( \hat{c}_i = E_{\hat{\Psi}}[c_i | y] \), respectively, where \( E_{\hat{\Psi}} \) denotes expectation using \( \hat{\Psi} \) for \( \Psi \). For unknown \( \zeta_i \) and \( z \), we make the approximation (see Section 3 of the supplementary material...
The weighted contrast

\[ \text{cov}(\hat{r}_i) \approx \Omega_i(\hat{\xi}_i, \hat{\zeta}). \]  

(2.12)

Conditional on membership of the \( j \)th gene from component \( G_i \), this leads to the estimated (normalized) contrast,

\[ \hat{S}_{ij} = d_i^T \hat{r}_i / \sqrt{d_i^T \Omega_i(\hat{\xi}_i, \hat{\zeta}) d_i}, \]  

(2.13)

where

\[ \hat{r}_i = (\hat{\beta}_i^T, b_{G_i}^T, c_i^T)^T. \]  

(2.14)

On weighting now the estimated (normalized) contrast \( \hat{S}_{ij} \) over the \( g \) components in the mixture model, we obtain as our test statistic for the \( j \)th gene,

\[ W_j = \sum_{i=1}^{g} \tau_i(y_j; \hat{\Psi}, \hat{c}_i) \hat{S}_{ij}, \]  

(2.15)

where \( \tau_i(y_j; \hat{\Psi}, c_i) \) is the posterior probability that the \( j \)th gene belongs to the \( i \)th component \( G_i \) conditional on \( y_j \) and \( c_i \).

### 2.3 Null distribution of \( W_j \)

The weighted contrast \( W_j \) can be formed to test the null hypothesis of \( H_j : j \)th gene is not DE \((j = 1, \ldots, n)\). The null distribution of \( W_j \) can be assessed using a permutation method described as follows. The steps are:

1. Let \( H \) be the \( n \times p \) data matrix, where \( H = (y_1, \ldots, y_n)^T \). Permute the class labels \( B \) times (corresponding to the \( p \) columns of \( H \)) and let \( y_{(b)}^j \) be the gene profile for the \( j \)th gene after the \( b \)th permutation \((b = 1, \ldots, B; j = 1, \ldots, n)\). Then after the \( b \)th permutation we have corresponding to \( H \), the data matrix \( H^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)})^T \).

2. Then for each gene \( j \) in turn, compute the \( B \) replications \( W_j^{(b)} \) of \( W_j \) \((b = 1, \ldots, B)\), where \( W_j^{(b)} \) is calculated in the same way that \( W_j \) is calculated except that \( y_j \) is replaced by \( y_{j}^{(b)} \). That is, at the end of the estimation process for the original data \( H \), we compute the \( W_j \). Now to compute the \( W_j^{(b)} \) we repeat the last process but with \( H \) replaced by \( H^{(b)} \).

As stated above, it follows from (2.13) to (2.15) that \( W_j^{(b)} \) is given by

\[ W_j^{(b)} = \sum_{i=1}^{g} \tau_i(y_j^{(b)}; \hat{\Psi}, \hat{c}_i) \hat{S}_{ij}^{(b)}, \]  

(2.16)

where

\[ \hat{S}_{ij}^{(b)} = d_i^T \hat{r}_i^{(b)} / \sqrt{d_i^T \Omega_i(\hat{\xi}_i, \hat{\zeta}) d_i} \]  

(2.17)

and where

\[ \hat{r}_i^{(b)} = (\hat{\beta}_i^T, b_{G_i}^{(b)}^T, c_i^T)^T. \]  

(2.18)

In (2.18), we have

\[ b_{G_i}^{(b)} = E_{\hat{\Psi}}[b_{G_i} | y_{j}^{(b)}, z_{ij} = 1]. \]  

(2.19)
2.4 Data preprocessing

In practice, it is common to standardize each column of the data matrix to have zero mean and unit standard deviation, followed by the row standardization to shift expression profiles to the same baseline (zero mean, unit standard deviation) for comparison. However, row standardization will transform the data into directional data that lie on a unit hypersphere (Dortet-Bernadet and Wicker, 2008). It implies that multivariate normal distributions may not be appropriate for modeling row-standardized gene-expression profiles of high dimensions (Banerjee and others, 2005). In this paper, data preprocessing is implemented by column standardization only. Without row standardization, genes with different mean expressions shall be clustered into different clusters, even though their expression profiles may exhibit similar differentiation across the tissue classes. In other words, genes within the same cluster may have different patterns of differential expression. A cluster of genes may therefore contain both DE and non-DE genes; see Section 3.

3. Results

3.1 Identification of differentially expressed genes (breast cancer data, Hedenfalk and others, 2001)

The data comprised the measurement of 3226 genes with \( p_1 = 7 \) BRCA1-mutation-positive tumors and \( p_2 = 8 \) BRCA2-mutation-positive tumors. The aim is to identify DE genes between the tumors associated with the two mutations. There is the matter that the dataset does not include healthy women samples. Thus the DE genes identified may not be distinguishable when compared with healthy women. In cases where control samples from healthy individuals were available or there were three or more tissue classes of samples, specific sets of DE genes can be identified with the proposed method by using different contrast \( d_j \) in (2.6); see Section 3.2.

Here we column normalized the logged expressions and fitted the proposed extended random-effects model to the data with \( g = 3 \) to \( g = 15 \) clusters. Using BIC, we identified five clusters of genes (see Section 4 of the supplementary material available at Biostatistics online for details of the clustering results). The difference \( \hat{\beta}_{1i} - \hat{\beta}_{2i} \) between the estimated mean expressions in Class 1 and Class 2 for the \( i \)th component are \(-0.351, -0.131, 0.303, 0.047, \) and \(-0.147 \) for \( i = 1, \ldots, 5 \), respectively. In the clusters with large absolute differences \( \hat{\beta}_{1i} - \hat{\beta}_{2i} \) (such as for \( i = 1 \) or 3), it could happen that some genes are DE and some are not. Similarly, a gene \( j \) in Cluster 4 could have a large weighted contrast (a potential DE gene) due to a large difference \( |\hat{b}_{1ij} - \hat{b}_{2ij}| \) between the (estimated) gene-specific random effects terms for \( i = 4 \).

As an illustration, we depict in Figure 1 three gene profiles. Two of them are among the top 100 ranked genes, but one is from Cluster 3 (large \( |\hat{\beta}_{13} - \hat{\beta}_{23}| \)) and the other is from Cluster 4 (small \( |\hat{\beta}_{14} - \hat{\beta}_{24}| \)). The remaining gene profile may correspond to a non-DE gene which, however, was assigned to Cluster 1 with large \( |\hat{\beta}_{11} - \hat{\beta}_{21}| \); see Table S2 of the supplementary material available at Biostatistics online.

The ranking of DE genes is implemented on the basis of the weighted estimates of contrast of mixed effects (2.15). The top 50 ranked upregulated genes (corresponding to large negative weighted contrast) and the top 50 downregulated genes (corresponding to large positive weighted contrast) in BRCA2-mutation-positive tumors relative to tumors with BRCA1 mutations are presented for functional annotation and enrichment analyses, using DAVID Functional Annotation Tool Version 6.7 (Dennis and others, 2003) and GeneGo MetaCore pathway analysis program.

The results of the functional annotation and enrichment analyses using DAVID are presented in Table 1, with comparison to previous results of Hedenfalk and others. Besides the same annotated functions with higher significance (enrichment with more genes), we also identified additional annotated functions with the proposed method which may provide singular insight into the biology of the disease. In the analysis by Hedenfalk and others (2001), coordinated transcription activation of genes involved in DNA repair was found in tumors with BRCA1 mutations, and genes involved in suppression of apoptosis in tumors with...
Inference on differences between classes using contrasts

Fig. 1. Three selected gene profiles (seven BRCA1- and eight BRCA2-mutation-positive tumor tissue samples): Solid line (from Cluster 3, $\hat{\beta}_{13} - \hat{\beta}_{23} = 0.303$, difference in gene-specific random effects $\hat{b}_{13j} - \hat{b}_{23j} = 0.646$); Dashed line (from Cluster 4, $\hat{\beta}_{14} - \hat{\beta}_{24} = 0.047$, $\hat{b}_{14j} - \hat{b}_{24j} = 0.487$); Dotted line (from Cluster 1, $\hat{\beta}_{11} - \hat{\beta}_{21} = -0.351$, $\hat{b}_{11j} - \hat{b}_{21j} = 0.387$).

BRCA2 mutations. Our contrast method not only has identified the same characteristics, but has also identified upregulated genes involved in new significant pathways represented by Gene Ontology (GO) terms such as “response to drug” and “response to wounding” in tumors with BRCA2 mutations, suggesting the differences between BRCA1 and BRCA2 tumors to chemotherapy and the potential involvement of TGFbeta pathway in BRCA2 tumors. To this end, there are some recent studies that support the different response rates of BRCA patients, such as Yang and others (2011), and the evidence of TGFbeta downstream target SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers (Walker and others, 2010). We have also used GeneGo MetaCore pathway analysis tool to compare network enrichment of genes identified by the two methods (see Table S3 of the supplementary material available at Biostatistics online). The genes identified by the contrast method show significant enrichment in genes that regulate cell cycle and cell adhesion functions. All together these new findings provide further insight into the differences between BRCA1- and BRCA2-mutated tumors. As far as we know, there are no other reports of using mathematical methods to uncover the same results as ours.

3.2 Additional illustrations and comparisons

The proposed method is applied to the breast cancer dataset of van’t Veer and others (2002). An aim of the original study was to search for prognostic signatures that could discriminate between patients of
| Table 1. Comparison of GO term association of the top 50 ranked up- and downregulated genes in tumors with BRCA2 mutations relative to tumors with BRCA1 mutations (breast cancer data of Hedenfalk and others, 2001) |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
|                                  | Our method                       | Hedenfalk method                 |                                  |
|                                  | Count (fold)                     | P-value                          | Count (fold)                     | P-value                          |
| (a) GO terms associated with upregulated genes |                                  |                                  |                                  |                                  |
| Response to hormone stimulus     | 9 (7.9)                          | 1.2E−05                          | 6 (5.5)                          | 3.9E−03                          |
| Cell motion                      | 8 (5.4)                          | 5.0E−04                          | 3 (2.1)                          | 4.0E−01                          |
| Enzyme-linked receptor protein signal pathway | 7 (6.6)                          | 5.3E−04                          | 6 (5.9)                          | 2.9E−03                          |
| Regulation of cell proliferation | 9 (3.7)                          | 2.2E−03                          | 8 (3.4)                          | 6.6E−03                          |
| Regulation of phosphorylation    | 7 (4.8)                          | 2.6E−03                          | 5 (3.6)                          | 4.4E−02                          |
| Transcription factor binding     | 7 (4.5)                          | 3.5E−03                          | 4 (2.7)                          | 1.7E−01                          |
| Negative regulation of cell differentiation | 5 (7.5)                          | 4.0E−03                          | 4 (6.3)                          | 2.4E−02                          |
| Cell-matrix adhesion             | 3 (11)                           | 3.0E−02                          | 4 (15)                           | 2.1E−03                          |
| Response to drug                 | 7 (10)                           | 4.4E−05                          |                                  |                                  |
| Negative regulation of apoptosis | 7 (6.4)                           | 6.4E−04                          |                                  |                                  |
| Blood vessel morphogenesis       | 5 (7.6)                           | 3.7E−03                          |                                  |                                  |
| Response to wounding             | 7 (4.3)                           | 4.9E−03                          |                                  |                                  |
| Cellular response to stress      |                                  |                                  | 6 (3.7)                          | 2.2E−02                          |
| (b) GO terms associated with downregulated genes |                                  |                                  |                                  |                                  |
| Interphase of mitotic cell cycle | 6 (18)                           | 1.6E−05                          | 6 (17)                           | 2.2E−05                          |
| Cyclin-dependent protein kinase activity | 5 (29)                           | 2.3E−05                          | 4 (22)                           | 7.6E−04                          |
| Nuclear lumen                    | 13 (4)                           | 1.6E−04                          | 13 (3)                           | 7.0E−04                          |
| Posttranscriptional regulation—gene expression | 6 (9.0)                           | 4.7E−04                          | 5 (7.0)                           | 5.2E−03                          |
| Regulation to DNA damage stimulus | 6 (5.1)                           | 5.7E−03                          | 9 (7.1)                           | 2.7E−05                          |
| Induction of apoptosis           | 3 (3.0)                           | 2.6E−01                          | 5 (4.6)                           | 2.1E−02                          |
| Regulation of miRNA stability    | 3 (43)                           | 2.1E−03                          |                                  |                                  |
| Cellular protein metabolic process | 5 (8.7)                           | 2.3E−03                          |                                  |                                  |
| Single-stranded DNA binding      |                                  |                                  | 5 (28)                           | 2.6E−05                          |

Note: “count” is the number of genes belonging to an annotation term, “fold” is the fold-enrichment of involved genes over total genes in the list relative to the number of genes with the same term category in the human genome background, P-value is the modified Fisher exact P-value for assessing the gene-enrichment; see Dennis and others (2003).
with P276 (a novel CDK inhibitor) alone, $p_2 = 6$ tissue samples treated with gemcitabine alone, and $p_3 = 3$ tissue samples treated with combination of P276 and gemcitabine (P276-Gem). The aim of the original study was to study whether the combination of P276-Gem downregulated Akt-mTOR signaling pathway. With $m = 3$ tissue classes, we adopt

$$d_j = (0.5 \ 0.5 \ -1: \ 0 \ 0 \ 0, \ \ldots, \ 0 \ 0 \ 0, \ 0.5 \ 0.5 \ -1, \ 0 \ 0 \ 0, \ \ldots: 0 \ \ldots \ 0),$$

(3.1)

where both Classes 1 and 2 contribute half to the contrast relative to Class 3 (P276-Gem). The results are presented in Section 6 of the supplementary material available at Biostatistics online. We compare the gene lists obtained by our method and existing approaches: significance analysis of microarray (SAM) of Tusher and others (2001), linear models for microarray data (LIMMA) of Smyth (2004), optimal discovery procedure (ODP) of Storey (2007), and the $t$-test. With SAM, ODP, and the $t$-test, Classes 1 and 2 were combined and then was compared with Class 3. With LIMMA, a linear contrast of $(0.5 \ 0.5 \ -1)$ was adopted in multiple testing for all genes. From the results presented in Table S10 of the supplementary material available at Biostatistics online, different gene lists were obtained by the methods. The results of enrichment analysis using DAVID is shown in Table S11 of the supplementary material available at Biostatistics online. In addition to the interesting genes identified in Subramaniam and others (2012), our method has also uncovered new pathways and cellular processes that show enrichment of downregulated genes. The new cellular processes such as “DNA replication” and “Pyrimidine metabolism” may be related to the effects of combined drug treatment to pancreatic cancer.

3.3 Golden Spike benchmark dataset

Golden Spike is a dataset generated to provide a benchmark for comparing different approaches for the analysis of Affymetrix GeneChips (Choe and others, 2005). The experiment compared two classes of tissue samples: $p_1 = 3$ (control) and $p_2 = 3$ (spike-in). By design, the dataset has 1331 (9.5%) DE (nonnull) genes (nominal spike-in to control ratio $> 1$) and 12 666 (90.5%) true nulls (nominal spike-in to control ratio $= 1$ or do not matched to any cRNA). In this study, the log-transformed data set “10a” is used, which is available in the website of Choe and others (2005). We fitted the extended random-effects model to the data and identified four clusters of genes based on BIC for model selection. Figure 2 presents the proportion of true nulls among the top-ranked genes at various cut-off rankings. We compare our method with SAM, LIMMA, ODP, and the $t$-test. From Figure 2, it can be seen that the proposed method outperforms the other approaches by having the smallest proportion of true nulls when the cut-off ranking is larger than 700. When the cut-off ranking is small, the proposed method performs equally well as ODP. Among the top 1000 ranked genes (declared to be DE), our method gives the least of 79 true nulls. The numbers of true nulls for the other approaches are, respectively, 311 (SAM), 141 (SAM with its fudge factor $s_0 = 0$), 290 (LIMMA), 227 ($t$-test), and 140 (ODP).

3.4 Simulation experiment 1: ranking of correlated genes

We consider the simulation procedure in Allison and others (2002) to study the relative performance of our method when gene expressions are correlated. For $m = 2$ classes of tissues with $p_1 = p_2 = 10$, we generated 3000 gene expressions independently for each of the 20 tissues in 6 blocks of size 500 from multivariate normal distributions with mean vector $10 \times 1_{500}$ and covariance matrix $\sigma^2(1_{500}1_{500}^T \rho + (1 - \rho)I_{500}) \otimes I_6$. Here, $\sigma^2 = 4$ is the variance and $\rho$ is the correlation between gene expressions. The above covariance structure implies that genes in the same block are correlated, while genes in different blocks are independent. Finally, for 20% of the genes (300 randomly selected genes each for up- and downregulated
gene groups), a true mean difference in expression $\pm \delta$ between the two tissue classes was added to the expression levels for the last 10 tissues.

We consider $\rho = 0.0$ (independence), $\rho = 0.4$ (moderate), $\rho = 0.6$ (moderately strong), $\rho = 0.8$ (strong dependence), and $\delta = 2$ (moderate differential expression). For each set of parameter values, 100 independent simulation experiments were conducted. We fitted the extended random-effects model to the column-normalized data with $g = 3$ components. Given that there are 600 DE genes from a total of $n = 3000$ genes in each simulated dataset, we obtained the top ranked 600 genes and noted the proportion of null genes among them; that is, the false discovery proportion (FDP), and the proportion of the 600 DE genes among them (the power). The results presented in Figure 3 show that our contrast method again outperforms the $t$-test, SAM, LIMMA, and ODP, with higher power and a markedly smaller error rate in FDP.

### 3.5 Simulation experiment 2: controlling the FDR

In this simulation experiment, we consider $m = 2$ two classes of tissues and the same setting as in Simulation experiment 1 above. Four sets of parameter values of $(\delta, \rho)$ are considered: $(\delta = 2, \rho = 0.0)$, $(\delta = 2, \rho = 0.4)$, $(\delta = 3, \rho = 0.0)$, and $(\delta = 3, \rho = 0.4)$.

In this experiment we illustrate our contrast method, where the intent is to control the FDR. One way to achieve this is to use the Benjamini and Hochberg (1995) procedure applied to the $P$-values. In order to be able to compute a $P$-value for the test statistic $W_j$ of our contrast methods, we need to know its null distribution under the null hypothesis that the $j$th gene is not DE. A rough approximation would be to take $W_j$
as having a standard normal distribution. We use an improved approximation by fitting a three-component mixture of $t$-distributions to some replicated values of $W_j^{(b)}$ obtained by permuting the column labels as detailed in Section 2.3. The component corresponding to the central portion of $W_j^{(b)}$ can be considered as representing the null distribution of $W_j$ and be adopted to estimate the $P$-value for each original observed value of $W_j$. Having obtained $P$-values for each gene, we can proceed as in McLachlan and others (2006) and convert them to $z$-scores to which a mixture of two normals is fitted with the first component corresponding to the null genes and the second component corresponding to the DE genes. The (estimated) posterior probability $\hat{\tau}_0(z_j)$ that the $j$th genes belongs to the first (null) component is the estimate of the local FDR. With each method, genes having an estimated local FDR less than some threshold $c_0$ (that is, $\hat{\tau}_0(z_j) < c_0$) can be taken to be DE. The implied FDR and power can be estimated as outlined in Section 2.

In Table 2, the true FDR and power are presented. It can be seen that the contrast method has high power when there is moderate differential expression or moderate dependence. The other methods have less power (see Simulation experiment 1) and their results are not included in Table 2.

4. Discussion

We have presented a clustering-based contrast approach to draw inference on differences between classes using full gene-expression profiles. An extended random-effects model is adopted for the clustering of gene-expression data, which allows for a direct modeling of correlation among genes and within genes via
correlated gene-specific random effects. Our approach thus enables the partition of overall variation into correlated random components and independent random noise. More importantly, the predicted random effects have a meaningful interpretation and, together with the fixed effects, they can be adopted to form a weighted contrast for assessing directly the differential expression between tissue classes for each gene. Unlike the extended random-effects model, the mixture of factor analyzers aims to cluster multivariate data based on a reduced dimension via factor analysis. It is therefore more appropriate to cluster tissue samples (McLachlan and others, 2004). Alternatively, biclustering is becoming a popular approach for two-way clustering of microarray data, such as Martella and others (2011). This method involves the clustering of both tissue samples and genes. Similar to existing clustering-based approaches, this approach produces clustering of genes that may not correspond to DE genes. The applicability of the proposed method has been demonstrated using real and simulated datasets.

The use of BIC for model selection has been successfully applied in many Bioinformatics applications (He and others, 2006; McLachlan and others, 2004; Pan and others, 2002). As described on Section 1, the gene ranking is robust to the choice of $g$ that is determined by BIC. Here we illustrate this robustness using a sensitivity analysis (Section 7 of the supplementary material available at Biostatistics online). With the Hedenfalk data, the next smallest BIC value indicates a six-cluster solution. The top 50 up- and downregulated genes from this solution are compared with the five-cluster solution presented in Section 3.1. From the results given in Table S12 of the supplementary material available at Biostatistics online, it can be seen that 82% of the top ranked genes remain, while 17% are still within the top 70 up- or downregulated genes of the six-cluster solution.

The proposed approach could be modified for RNA sequencing (RNA-seq) data that quantify gene expression by providing counts of transcripts. One way is to use the voom function of the LIMMA package to convert sequence read counts into measurements on the real line which can then be analyzed using the proposed method as for microarray data. Alternatively, a Poisson mixture model may be adopted to cluster RNA-seq data, where gene-specific and tissue-specific random effects are added to the conditional mean rate of Poisson components. To estimate the random components, an ML or a residual ML approaches may be used (McLachlan and others, 2004); see also Li and others (2012). The work for handling RNA-seq data will be pursued in future research.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at [http://biostatistics.oxfordjournals.org](http://biostatistics.oxfordjournals.org).

**ACKNOWLEDGMENTS**

We thank the Associate Editor and two referees for helpful comments. **Conflict of Interest:** None declared.
Inference on differences between classes using contrasts

Funding

This work was supported by the Australian Research Council.

References

Allison, D., Gadbury, G., Heo, M. and others. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis* 39, 1–20.

Banerjee, A., Dhillon, I., Ghosh, J. and Sra, S. (2005). Clustering on the unit hypersphere using Von-Mises-Fisher distributions. *Journal of Machine Learning Research* 6, 1345–1382.

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, 259–300.

Choe, S., Boutros, M., Michelson, A. and others. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* 6, R16.

Dahl, D. B. and Newton, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association* 102, 517–526.

Dennis, Jr, G., Sherman, B. T., Hosack, D. A. and others. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology* 4, P3.

Desai, K. H. and Storey, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association* 107, 135–151.

Dortet-Bernadet, J. and Wicker, N. (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics* 9, 66–80.

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge University Press.

He, Y., Pan, W. and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics and Data Analysis* 51, 641–658.

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

Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13, 523–538.

Martella, F., Alfo, M. and Vichi, M. (2011). Hierarchical mixture models for biclustering in microarray data. *Statistical Modelling* 11, 489–505.

Mclachlan, G. J., Bean, R. W. and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarray. *Bioinformatics* 22, 1608–1615.

Mclachlan, G. J., Do, K. A. and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. New York: Wiley.

McLean, R. A., Sanders, W. L. and Stroup, W. W. (1991). A unified approach to mixed linear models. *American Statistician* 45, 54–64.

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.

Ng, S. K., Mclachlan, G. J., Wang, K. and others. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* 22, 1745–1752.
PAN, W., LIN, J. AND LE, C. T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biology* 3, 0009.1–0009.8.

QIN, L. X. AND SELF, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* 62, 526–533.

QIU, W., HE, W., WANG, X. AND LAZARUS, R. (2008). A marginal mixture model for selecting differentially expressed genes across two types of tissue samples. *International Journal of Biostatistics* 4, Article 20.

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

STOREY, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B* 69, 347–368.

SUBRAMANIAM, D., PERIYASAMY, G., PONNURANGAM, S. and others. (2012). CDK-4 inhibitor P276 sensitizes pancreatic cancer cells to Gemcitabine-induced apoptosis. *Molecular Cancer Therapeutics* 11, 1598–1608.

TUSHER, V. G., TIBSHIRANI, R. AND 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.

VAN’T VEER, L. J., DAI, H., VAN DE VIJVER, M. J. and others. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.

WALKER, L. C., FREDERICKSEN, Z. S., WANG, X. and others. (2010). Evidence for SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers. *Breast Cancer Research* 12, R102.

YANG, D., KHAN, S., SUN, Y. and others. (2011). Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *Journal of the American Medical Association* 306, 1557–1565.

YUAN, M. AND KENDZIORSKI, C. (2006). A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics* 62, 1089–1098.

[Received August 13, 2013; revised February 18, 2014; accepted for publication May 14, 2014]