Statistical significance analysis of longitudinal gene expression data

Xu Guo\textsuperscript{1}, Huilin Qi\textsuperscript{2}, Catherine M. Verfaillie\textsuperscript{2} and Wei Pan\textsuperscript{1,*}

\textsuperscript{1}Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455-0378, USA and \textsuperscript{2}Stem Cell Institute, Department of Medicine, University of Minnesota Medical School, 422 Delaware Street SE, MMC 716, Minneapolis, MN 55445, USA

Received on October 17, 2002; revised on November 2, 2002; accepted on March 11, 2003

ABSTRACT
Motivation: Time-course microarray experiments are designed to study biological processes in a temporal fashion. Longitudinal gene expression data arise when biological samples taken from the same subject at different time points are used to measure the gene expression levels. It has been observed that the gene expression patterns of samples of a given tumor measured at different time points are likely to be much more similar to each other than are the expression patterns of tumor samples of the same type taken from different subjects. In statistics, this phenomenon is called the within-subject correlation of repeated measurements on the same subject, and the resulting data are called longitudinal data. It is well known in other applications that valid statistical analyses have to appropriately take account of the possible within-subject correlation in longitudinal data.

Results: We apply estimating equation techniques to construct a robust statistic, which is a variant of the robust Wald statistic and accounts for the potential within-subject correlation of longitudinal gene expression data, to detect genes with temporal changes in expression. We associate significance levels to the proposed statistic by either incorporating the idea of the significance analysis of microarrays method or using the mixture model method to identify significant genes. The utility of the statistic is demonstrated by applying it to an important study of osteoblast lineage-specific differentiation. Using simulated data, we also show pitfalls in drawing statistical inference when the within-subject correlation in longitudinal gene expression data is ignored.

Contact: weip@biostat.umn.edu

INTRODUCTION
With the rapid development of microarray technologies (Brown and Botstein, 1999; Lander, 1999), we can simultaneously measure the expression levels of thousands of genes over a period of time to study biological processes in a temporal fashion. Most studies with time-course microarray data focus on clustering analysis to identify genes with similar expression patterns in an exploratory way. Statistical model-based clustering analysis approaches have appeared (Yeung et al., 2001; Ghosh and Chinnaiyan, 2002; Li et al., 2002 and references therein). Here, we consider significance analyses to detect differential gene expression with longitudinal array data, which arise from some time-course array experiments with repeated use of biological samples taken from the same subject over time. Consequently, the repeated measurements on the same subject correlate to each other. In this paper, the within-subject correlation specifically refers to the correlation among the expression levels of the same gene at different time points for the same subject, rather than that among the expression levels of different genes measured on the same subject at the same time point.

Many statistical methods have been proposed to detect differential gene expression between two groups using independent samples, such as the regression modelling method (Thomas et al., 2001; Zhao et al., 2001), the significance analysis of microarrays (SAM) method (Tusher et al., 2001), the Empirical Bayes method (Efron et al., 2001) and the mixture model method (MMM) (Pan et al., 2003, available at http://www.biostat.umn.edu/rrs.php). Pan (2002) provided a review and comparison of the above methods; more references can be found in Pan (2003). In particular, SAM has been generalized to independent data from experiments with more than two groups and applicable to a special case of longitudinal data, the paired data, such as for gene expression in tumors before and after therapy, which only involves two time points. At its present form, it cannot handle general longitudinal data with more than two time points, which is the subject of our study.

The goal of this paper is to propose a method to construct a test statistic that properly takes account possible within-subject correlation in longitudinal gene expression data. More details on a specific example are going to be given later. It is well known from other statistical applications that invalid statistical inferences may be obtained if we ignore the potential within-subject correlation in longitudinal data (e.g. Diggle

\*To whom correspondence should be addressed.
et al., 1994). We will show that this conclusion also holds in the context of longitudinal gene expression data. Since all the above methods assume independence and use the $t$-type statistics, none of them can be directly applied to longitudinal gene expression data. Here, we apply estimating equation techniques to construct a robust Wald statistic that is a variant of robust Wald statistic. We can incorporate the idea of SAM or MMM to associate significance levels to the proposed statistic.

To illustrate the utility of the proposed method, we apply it to an important study of osteoblast differentiation from human mesodermal progenitor cells.

Recently, a rare bone marrow cell was identified and called mesodermal progenitor cell (MPC) (Reyes et al., 2001). It can differentiate at single-cell level into mesenchymal cell types such as osteoblasts, chondroblasts and adipocytes, and also into cells of visceral mesodermal origin. Thus, MPC can be an ideal source of cells to generate osteoblasts to treat bone diseases such as osteoporosis or non-healing fractures, and osteogenesis imperfecta (Horwitz et al., 1999). Before utilizing MPC to generate osteoblasts specifically, we need to understand the differentiation process of MPC into osteoblasts first, which involves gene regulations of specific signaling proteins and transcription factors by numerous extrinsic factors, including hormones and growth factors (Yamaguchi et al., 2000; Ducy et al., 2000). To identify genes involved in the osteoblast lineage-specific differentiation, the expressed gene profiles of undifferentiated MPC (at day 0) and MPC induced osteoblast lineage-specific differentiation, the expressed gene were used to measure gene expression across the seven days. Thus, a longitudinal data set with four different time points was generated. We propose a test statistic that properly accounts for the possible within-subject correlation in the longitudinal data, then apply the test statistic to SAM and MMM to identify genes differentially expressed over time, which may play important roles in the process of osteoblast lineage-specific differentiation.

**METHODS**

**Notation**

Suppose we have a longitudinal data set with $J$ subjects, there is a response vector with $n_j$ responses, $Y_j = (y_{j1}, \ldots, y_{jn_j})^T$ and an $n_j \times p$ covariance matrix $X_j$ for the $j$th ($j = 1, \ldots, J$) subject. $y_{jk}$'s ($k = 1, \ldots, n_j$) are assumed correlated within each subject and independent across different subjects. The marginal expectation of the response, $E(y_{jk}) = \mu_{jk}$, is modelled by the equation $g(\mu_{jk}) = X_{jk}\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional vector of unknown regression parameters and $g(\cdot)$ is a known link function. The marginal variance is $\text{Var}(y_{jk}) = v(\mu_{jk})\phi$, where $v(\cdot)$ is a known variance function and $\phi$ is a scale parameter. The within-subject correlation matrix $R_0 = \text{corr}(Y_j)$ is generally unknown. We can consistently estimate $\beta$ by specifying a working correlation matrix $R_W(\alpha)$, which may depend on some parameter $\alpha$, and solving the following generalized estimating equations:

$$U(\beta, \phi) = \sum_{j=1}^{J} D_j^T V_j^{-1} S_j = 0,$$  \hspace{1cm} (1)

where $D_j = \partial \mu_j/\partial \beta^T$, $\mu_j = (\mu_{j1}, \ldots, \mu_{jn_j})^T$, $V_j = \phi A_j^{1/2} R_W(\alpha) A_j^{1/2}$, $A_j = \text{diag}(v(\mu_{j1}), \ldots, v(\mu_{jn_j}))$ and $S_j = Y_j - \mu_j$.

Under mild regularity conditions, $\hat{\beta}$ is consistent and asymptotically normal when the number of subjects is sufficiently large (Liang and Zeger, 1986). $\text{Cov}(\hat{\beta})$ can be consistently estimated by the so-called sandwich estimator,

$$V_S = \left( \sum_{j=1}^{J} D_j^T V_j^{-1} D_j \right)^{-1} \left( \sum_{j=1}^{J} D_j^T V_j^{-1} S_j S_j^T V_j^{-1} D_j \right) \times \left( \sum_{j=1}^{J} D_j^T V_j^{-1} D_j \right)^{-1}. \hspace{1cm} (2)$$

We can plug in $\hat{\beta}$ and $\hat{\alpha}$ to get $\hat{V}_S$. If $V_j$ correctly specifies $\text{Cov}(Y_j)$, the covariance matrix of $\hat{\beta}$ is consistently estimated by,

$$V_M = \left( \sum_{j=1}^{J} D_j^T V_j^{-1} D_j \right)^{-1}, \hspace{1cm} (3)$$

which is called the model-based or naive variance estimator.

**Robust Wald statistic**

A composite null hypothesis is specified as $H_0 : L\beta = 0$, where $L$ is an $r \times p$ matrix with full row rank $r$. The robust Wald statistic is

$$W = (L\hat{\beta})^T (L\hat{\beta} S_L L^T)^{-1} (L\hat{\beta}) \hspace{1cm} (4)$$

(Kent, 1982; White, 1982). When the number of subjects is sufficiently large, the robust Wald statistic $W$ has an asymptotic chi-squared null distribution $\chi^2_r$. However, this approximation is not adequate when the number of subjects is small, and we propose to use permutation methods as in SAM and MMM to draw inference.

**Gene-specific score**

The robust Wald statistic for the $i$th gene in a longitudinal microarray data set is

$$W(i) = [L\hat{\beta}(i)]^T [L\hat{\beta}(i) S_L L^T]^{-1} [L\hat{\beta}(i)]. \hspace{1cm} (5)$$

We define the gene-specific score as

$$w(i) = [L\hat{\beta}(i)]^T [L\hat{\beta}(i) L^T + \lambda_w I_{rxr}]^{-1} [L\hat{\beta}(i)]. \hspace{1cm} (6)$$

where $\lambda_w$ is a positive scalar.
The algorithm is summarized in the following four steps:

1. The determinant value of $L\hat{V}_S(i)L^T$ is denoted by $\lambda_i$. For $\alpha \in (0, 0.05, 0.10, \ldots, 1.0)$, let $\lambda^a$ be the $\alpha$ quantile of $\lambda_i$'s and $w^a(i) = [L\hat{b}(i)]^T[L\hat{V}_S(i)L^T + \lambda^a I_{nxr}]^{-1}[L\hat{b}(i)]$.

2. Compute the $100$ quantiles of the $\lambda_i$ values, denoted by $q_1 < q_2 < \cdots < q_{100}$.

3. For each value of $\alpha$,
   (a) Compute $v_n = mad(w^a(i)|\lambda_i \in [q_n, q_{n+1}]), n = 1, 2, \ldots, 99$, where $v_n$ is a function of $\alpha$, $\lambda_i$'s, $q_n$, $q_{n+1}$ and etc. $mad$ is the median absolute deviation from the median of $w^a(i)$'s for $\{i|\lambda_i \in [q_n, q_{n+1}]\}$ divided by $0.64$.
   (b) Compute $cv(\alpha) = \text{coefficient of the variation of the } v_n \text{ values.}$

4. Choose $\hat{\alpha}$ to minimize $cv(\alpha)$ and compute $\lambda_w = \lambda_{\hat{\alpha}}$, $\lambda_w$ is fixed at the value $\lambda_w$.

**RESULTS**

**Data**

There are three donors in the study of osteoblast differentiation. For each donor, RNA samples were harvested from undifferentiated MPC (baseline) and MPC induced to the osteoblast at day 1, day 2 and day 7. Radiolabeled cDNA probes were generated and hybridized to Human GeneFilters microarrays. The resulting longitudinal microarray data contain expression levels of 4132 genes on each of the twelve arrays. The raw data were normalized to assure that mean intensities were the same on each array using software Pathways 3. We subtract the baseline expression levels of each gene of each donor to obtain the relative expression levels. For the $i$th gene of the $j$th donor, it is denoted as $Y_{j}(i) = (y_{j,1}(i), y_{j,2}(i), y_{j,7}(i))^T$, $i = 1, \ldots, 4132$, $j = 1, \ldots, 3$, which is a vector containing relative gene expression levels at three time points, day 1, day 2 and day 7 after the induction, with the mean $\mu(i) = (\mu_1(i), \mu_2(i), \mu_7(i))^T$.

**Gene-Specific Score**

The gene expression levels for the $i$th gene from the $j$th donor is modelled as

$$E[Y_{j}(i)] = X_{j}(i)\beta(i),$$

where $X_{j}(i) = I_{3 \times 3}$ and $\beta(i) = \mu(i)$. Since we only have three donors in the study, we do not have enough degrees of freedom to test whether any gene expression change occurs during the entire time course of seven days, which is specified by the null hypothesis $H_0 : \mu_1(i) = \mu_2(i) = \mu_7(i) = 0$, or in shortened notation, $H_0 : \mu(i) = 0$. Here, we only try to identify genes activated or suppressed at day 1 or/and day 2 of the differentiation process relative to baseline with $H_0 : \mu_1(i) = \mu_2(i) = 0$ or $H_0 : L\mu(i) = 0$, where $L$ is a $2 \times 3$ matrix with the two rows as $(1, 0, 0)$ and $(0, 1, 0)$. Significant genes are more likely to be those responsible for the initial commitment from MPC to osteoblast lineage. Note that this $H_0$ cannot be tested using the usual paired $t$-test or its analog in SAM, for which the null hypothesis would be $\mu_1(i) = \mu_2(i)$.

Since our primary interest is about regression parameters, we treat covariance parameters as nuisance parameters. We assume that the working covariance matrix $V_j(i)$ is the same for the three donors. Since the variance of gene expression levels may be related with mean expression levels, we assume $V_j(i) = \text{diag}(\sigma_1^2(i), \sigma_2^2(i), \sigma_7^2(i))$ with different variances at different time points. For simplicity, we use independence working correlation structure, i.e., $R_{W} = I_{3 \times 3}$. Thus, $V_j(i) = \text{diag}(\sigma_1^2(i), \sigma_2^2(i), \sigma_7^2(i))$.

Here, the solution to the estimating equation,

$$\sum_{j=1}^{J} V_j(i)^{-1} [Y_j(i) - \mu(i)] = 0,$$

is

$$\hat{\mu}(i) = (\hat{y}_1(i), \hat{y}_2(i), \hat{y}_7(i))^T,$$

where

$$\hat{y}_1(i) = \sum_{j=1}^{J} y_{j,1}(i)/J, \hat{y}_2(i) = \sum_{j=1}^{J} y_{j,2}(i)/J$$

and

$$\hat{y}_7(i) = \sum_{j=1}^{J} y_{j,7}(i)/J.$$
The sandwich variance estimator is
\[ \hat{V}_2(i) = \sum_{j=1}^{J} (Y_j(i) - \hat{\mu}(i))(Y_j(i) - \hat{\mu}(i))^T / J^2. \]

Thus,
\[ L\hat{\mu}(i) = (\hat{y}_1(i), \hat{y}_2(i))^T \]
and its variance is
\[ \hat{V}_w(i) = \frac{1}{J^2} \left( \sum_{j=1}^{J} (y_j(i) - \hat{\mu}(i))^2 \sum_{j=1}^{J} (y_j(i) - \hat{\mu}(i))(y_j(i) - \hat{\mu}(i)) - \sum_{j=1}^{J} (y_j(i) - \hat{\mu}(i))^2 / J \right). \]

The robust Wald statistic for the \( i \)th gene is
\[ W(i) = (\hat{y}_1(i), \hat{y}_2(i))\hat{V}_w^{-1}(i)(\hat{y}_1(i), \hat{y}_2(i))^T \]
and the gene-specific score is
\[ w(i) = (\hat{y}_1(i), \hat{y}_2(i))\hat{V}_w^{-1}(i)(\hat{y}_1(i), \hat{y}_2(i))^T, \]
where \( \lambda_w = 0.045 \), calculated from the osteoblast differentiation data.

We have applied both SAM (Tusher et al., 2001) and MMM (Pan et al., 2003) to associate statistical significance levels to the gene-specific scores to detect differential gene expression during the process of osteoblast lineage-specific differentiation.

**SAM**

Genes are ranked by the magnitude of \( w(i) \) values, so that \( w(1) \) is the largest score and \( w(i) \) is the \( i \)th largest score. Under \( H_0: \mu_1(i) = \mu_2(i) = 0 \), we permute the relative expression values for the \( i \)th gene at day 1 and day 2 after the induction by multiplying a randomly selected +1 or -1 with an equal probability of 0.5. With \( B \) permutations, we have \( B \) longitudinal data sets and obtain \( B \) sets of ordered scores. For permutation \( b (b = 1, \ldots, B) \), the ordered scores are denoted as \( w_0(b) = (w_0(1)(b), \ldots, w_0(4132)(b)) \). \( w_0(i)(b) \) represents the \( i \)th largest score within permutation \( b \). To identify potentially significant changes in expression, we use a scatter plot of the observed scores \( w(i) \) vs. the expected null scores \( w_E(i) = \frac{\sum_{b=1}^{B} w_0(i)(b)}{B} \). For most of the genes, \( w(i) \approx w_E(i) \), but some genes are represented by points deviated from the equality line \( w(i) = w_E(i) \) by a distance greater than a threshold \( \Delta \), and may correspond to significant genes. We can adjust the threshold \( \Delta \) to control the false discovery rate (FDR), which is estimated using the null scores. The FDR is the ratio of the number of false positives (FP) over the total number of positives (TP). Note that as always, use of FDR (or any other statistics) cannot tell which genes are actually false positives. With different values of \( \Delta \), we identify different numbers of significant genes (TP) with corresponding FDRs. As the value of \( \Delta \) decreases, the total number of positives increases and the number of false positives also increases.

**MMM**

We can also use MMM (Pan et al., 2003) to obtain a cut-off value for observed scores \( w \) from the osteoblast differentiation data to declare statistical significance while controlling the number of false positives. Since the ordered null scores \( w_0 \) from the \( B \) permutations are positive and skewed, we take a natural logarithm transformation to obtain \( \log(w_0) \), whose distribution is more symmetric about zero. We fit a Normal mixture model to \( \log(w_0) \) from multiple permutations to estimate the null distribution of \( \log(w) \) by using the expectation–maximization (EM) algorithm (Dempster et al., 1977). For any given false positive rate or Type I error \( \alpha \), we can use the bisection method (Press et al., 1992) to solve the equation
\[ \alpha = P(\log(w_0) > s \mid H_0) \]
to obtain the corresponding cut-off value \( s \) for \( \log(w) \) to define statistical significance. Then we can get the cut-off value for \( w \) as \( \exp(s) \).

We fit a Normal mixture model to \( \log(w_0) \),
\[ f_0(\log(w_0); \Phi_{B_0}) = \sum_{g=1}^{g_0} \pi_g \Phi(\log(w_0) ; \mu_g, V_g), \]
where \( \Phi(\log(w_0) ; \mu_g, V_g) \) denotes the normal density function with mean \( \mu_g \) and variance \( V_g \), and \( \pi_g \)'s are mixing proportions. The number of components \( g_0 \) can be determined by the Akaike Information Criterion (AIC) (Akaike, 1973) and

### Table 1. Identifying temporal gene expression change by SAM and MMM for osteoblast differentiation data

<table>
<thead>
<tr>
<th>SAM</th>
<th>w(i)</th>
<th>Threshold</th>
<th>FP %</th>
<th>TP %</th>
<th>FDR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>w(i)</td>
<td>Threshold</td>
<td>FP %</td>
<td>TP %</td>
<td>FDR %</td>
</tr>
<tr>
<td>Delta 1</td>
<td>Constant</td>
<td>4.2</td>
<td>12</td>
<td>35</td>
<td>4.9</td>
</tr>
<tr>
<td>Delta 1</td>
<td>Quantile</td>
<td>4.2</td>
<td>12</td>
<td>35</td>
<td>4.9</td>
</tr>
<tr>
<td>Delta 1</td>
<td>Permuted</td>
<td>4.2</td>
<td>12</td>
<td>35</td>
<td>4.9</td>
</tr>
<tr>
<td>Delta 1</td>
<td>Parameter</td>
<td>4.2</td>
<td>12</td>
<td>35</td>
<td>4.9</td>
</tr>
<tr>
<td>Delta 1</td>
<td>Optimal</td>
<td>4.2</td>
<td>12</td>
<td>35</td>
<td>4.9</td>
</tr>
</tbody>
</table>

| Delta 1 | Constant | 4.2 | 12 | 35 | 4.9 | 33 | 15 |
| Delta 1 | Quantile | 4.2 | 12 | 35 | 4.9 | 33 | 15 |
| Delta 1 | Permuted | 4.2 | 12 | 35 | 4.9 | 33 | 15 |
| Delta 1 | Parameter | 4.2 | 12 | 35 | 4.9 | 33 | 15 |
| Delta 1 | Optimal | 4.2 | 12 | 35 | 4.9 | 33 | 15 |
To select the number of components $g_0$, a series of models with various values of $g_0$ are fitted and the $g_0$ corresponding to the first local minimum of AIC or BIC is selected (Fraley and Raftery, 1998). We use BIC here because it is preferred based on previous studies (Fraley and Raftery, 1998).

We use the EMMIX package (McLachlan et al., 1999, available at http://www.jstatsoft.org/v04/i02/) to fit five Normal mixture models to $\log(w_0)$ with $g_0 = 1, \ldots, 5$. According to BIC, we select the model with $g_0 = 3$ and obtain the estimates $\hat{\pi} = (0.540, 0.405, 0.054)$, $\hat{\mu} = (0.533, 1.297, -1.162)$ and $\hat{\Sigma} = (1.774, 0.663, 4.473)$. Compared to the empirical distribution (i.e. histogram) of $\log(w_0)$, the fitted Normal mixture model with three components (smooth line) fits very well. From the fitted Normal mixture model, we get the cut-off values for $\log(w)$ for different Type I errors. After taking the inverse of natural logarithm transformation, we get corresponding cut-off values for $w$. With different cut-off values, different numbers of significant genes are identified (Table 1).

**Significant genes**

The top twelve significant genes identified by both SAM and MMM using gene-specific score $w(i)$ are listed in Table 2. HSPCB is identified twice. In the microarray experiment, HSPCB were spotted to two different locations of the same array to monitor the homogeneity of the hybridization. Here, the gene-specific scores for the two locations are close to each other and both of them have been identified. As we know, some transcription factors required for cell proliferation are to be suppressed as cells differentiate. For example, MYC is down-regulated after the induction of differentiation. In addition, some other genes involved in cell proliferation, such as UBC110 and PSMD2, are also down-regulated. Both of them can help destroy cyclins, which is required for cells to complete mitosis and enter anaphase of the next cycle (Townsley et al., 1997). ODC1 is involved in polyamine biosynthesis, can also act as mitochondrial carrier protein and regulate cellular level of polyamine. These results are consistent with the notion that differentiation is negatively associated with cell cycle progression. The remaining genes in the list may have roles in the differentiation process previously not known, or may be among the set of false positive genes.

**Ignoring within-subject correlation**

We also analyze the osteoblast differentiation data by ignoring the within-subject correlation. When the data are treated as independent, the sandwich variance estimator is reduced to the model-based variance estimator. The resulting robust Wald statistic for the $i$th gene is

$$W_{\text{ind}}(i) = (\bar{y}_1(i), \bar{y}_2(i)) \hat{V}^{-1}_{\text{ind}}(i)(\bar{y}_1(i), \bar{y}_2(i))^T,$$

where

$$\hat{V}_{\text{ind}} = \frac{1}{J} \left[ \sum_{j=1}^{J} (y_{1,ij} - \bar{y}_1(i))^2 & 0 \\ 0 & \sum_{j=1}^{J} (y_{2,ij} - \bar{y}_2(i))^2 \right].$$

Then we define the gene-specific score $w_{\text{ind}}(i)$ based on $W_{\text{ind}}(i)$ and associate significance levels using either SAM or MMM (Table 1).

When we detect a similar number of significant genes by using $w(i)$ and $w_{\text{ind}}(i)$ in SAM respectively, the estimated FDR by using $w_{\text{ind}}(i)$ is smaller than that by using $w(i)$ instead (Table 1). For example, we detect 287 significant genes with 148.4 estimated FP by using $w(i)$, whereas we detect 285 significant genes with 84.8 FP by using $w_{\text{ind}}(i)$. For MMM, the estimated FDR by using $w_{\text{ind}}(i)$ is also smaller than that by using $w(i)$ at a given Type I error rate. As shown later in simulations, use of $w_{\text{ind}}(i)$ may lead to under-estimation of FDR.

We also consider the top twelve significant genes identified by both SAM and MMM using the gene-specific score

<table>
<thead>
<tr>
<th>Rank</th>
<th>Accession</th>
<th>$w(i)$</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R60343</td>
<td>278.08</td>
<td>NT5: S' nucleotidase</td>
<td>DNA/RNA</td>
</tr>
<tr>
<td>2</td>
<td>AA430504</td>
<td>205.76</td>
<td>UBC110: ubiquitin carrier protein E2-C</td>
<td>Metabolism</td>
</tr>
<tr>
<td>3</td>
<td>AA478543</td>
<td>174.38</td>
<td>AKAP12: A kinase (PRKA) anchor protein (gravin) 12</td>
<td>Signal transduction</td>
</tr>
<tr>
<td>4</td>
<td>N54794</td>
<td>155.90</td>
<td>PAK1: plasminogen activator inhibitor, type 1</td>
<td>Protein synthesis/degradation</td>
</tr>
<tr>
<td>5</td>
<td>AA461467</td>
<td>152.65</td>
<td>ODC1: ornithine decarboxylase 1</td>
<td>Metabolism</td>
</tr>
<tr>
<td>6</td>
<td>R44334</td>
<td>147.70</td>
<td>HSPCB: heat shock 90kD protein 1, beta</td>
<td>Secretion pathway</td>
</tr>
<tr>
<td>7</td>
<td>R44334</td>
<td>144.77</td>
<td>HSPCB: heat shock 90kD protein 1, beta</td>
<td>Secretion pathway</td>
</tr>
<tr>
<td>8</td>
<td>AA464600</td>
<td>136.95</td>
<td>MYC: v-myc avian myelocytomatosis viral oncogene homolog</td>
<td>Transcription pathway</td>
</tr>
<tr>
<td>9</td>
<td>R42894</td>
<td>114.94</td>
<td>CADPS: Ca$^{2+}$-dependent activator protein for secretion</td>
<td>Signal transduction</td>
</tr>
<tr>
<td>10</td>
<td>R93124</td>
<td>110.04</td>
<td>AKR1C1: aldo-keto reductase family 1, member C1</td>
<td>Metabolism</td>
</tr>
<tr>
<td>11</td>
<td>AA461467</td>
<td>101.39</td>
<td>TPI1: triosephosphate isomerase 1</td>
<td>Metabolism</td>
</tr>
<tr>
<td>12</td>
<td>H05893</td>
<td>100.27</td>
<td>PSMD2: proteasome (prosome, macropain) 26S subunit, non-ATPase, 2</td>
<td>Protein synthesis/degradation</td>
</tr>
</tbody>
</table>
4.2 from MMM. This is also the case for Table 2, the number of false positives is 2.8 from SAM and as explained earlier are likely to have differential expressions during the osteoblast differentiation process based on previous biological studies.

Comparison between SAM and MMM

The results from SAM and MMM are consistent with each other. When they identify a similar number of significant genes, the number of false positives estimated from SAM is similar to that from MMM (Table 1). For instance, when they both identify 12 significant genes using \( w(i) \), listed in Table 2, the number of false positives is 2.8 from SAM and 4.2 from MMM. This is also the case for \( w_{ind}(i) \). When both SAM and MMM identify 33 significant genes using \( w_{ind}(i) \), the number of false positives is 4.2 from SAM and 4.9 from MMM.

SIMULATION

In order to evaluate the performance of the proposed statistic and compare SAM to MMM for longitudinal microarray data, we have done a small simulation study. Two data sets with 4132 genes for two simulation set-ups respectively were generated to mimic the osteoblast differentiation data.

Generating simulated data

We assume a general statistical model for the osteoblast differentiation data, which models the gene expression level for the \( i \)th gene at \( j \)th donor at time \( k \) (\( k = 0, 1, 2, 7 \)) as

\[
y_{ijk} = \mu_i + \tau_{ik} + b_{ij} + \epsilon_{ijk}. \tag{15}
\]

where \( \mu_i \) is the mean expression level for the \( i \)th gene at baseline, \( \tau_{ik} \) is the main effect of time (\( \tau_{i0} = 0 \)), \( b_{ij} \) is the random subject effect and \( \epsilon_{ijk} \) is random error. After adjusting for the baseline \( y_{ij0} \), we get

\[
y_{ij1} = y_{ij1} - y_{ij0} = \tau_1 + (-\epsilon_{ij0}) + \epsilon_{ij1}, \tag{16}
\]

\[
y_{ij2} = y_{ij2} - y_{ij0} = \tau_2 + (-\epsilon_{ij0}) + \epsilon_{ij2}. \tag{17}
\]

The observations from the \( i \)th donor are correlated through the term \( \epsilon_{ij0} \). We assume \( \epsilon_{ij0} \sim N(0, \sigma_{ij0}^2), \epsilon_{ij1} \sim N(0, \sigma_{ij1}^2), \epsilon_{ij2} \sim N(0, \sigma_{ij2}^2) \) and they are mutually independent. So, we have \( \text{Var}(y_{ij1}) = \sigma_{ij0}^2 + \sigma_{ij1}^2, \text{Var}(y_{ij2}) = \sigma_{ij1}^2 + \sigma_{ij2}^2 \) and \( \text{Var}(y_{ij1} - y_{ij2}) = \sigma_{ij1}^2 + \sigma_{ij2}^2 \).

When no change in gene expression occurs at day 1 and day 2 under \( H_0 : \tau_1 = \tau_2 = 0 \), it is reasonable to assume that \( \text{Var}(y_{ij1}) = \text{Var}(y_{ij2}) \) for the purpose of data generation, i.e., \( \sigma_{ij1}^2 = \sigma_{ij2}^2 \). We estimate \( \sigma_{ij0}^2, \sigma_{ij1}^2 \) and \( \sigma_{ij2}^2 \) from the osteoblast differentiation data and generate a data set with 4132 genes in it from the above model under \( H_0 \).

Table 3. False discovery rate estimated by SAM and MMM for the simulated data set under \( H_0 \) with the use of \( w(i) \) or \( w_{ind}(i) \) (FDR is taken as 100% when FP is greater than TP)

<table>
<thead>
<tr>
<th>SAM</th>
<th>( \Delta = 2 )</th>
<th>FP</th>
<th>TP</th>
<th>FDR (%)</th>
<th>( \Delta = 3.0 )</th>
<th>FP</th>
<th>TP</th>
<th>FDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w(i) )</td>
<td>9.0</td>
<td>10</td>
<td>90</td>
<td>0.01</td>
<td>3.6</td>
<td>8</td>
<td>45</td>
<td>0.001</td>
</tr>
<tr>
<td>( w_{ind}(i) )</td>
<td>3.8</td>
<td>4</td>
<td>95</td>
<td>0.005</td>
<td>1.6</td>
<td>4</td>
<td>40</td>
<td>0.005</td>
</tr>
<tr>
<td>( w_{ind}(i) )</td>
<td>0.6</td>
<td>1</td>
<td>60</td>
<td>0.01</td>
<td>0.7</td>
<td>2</td>
<td>30</td>
<td>0.01</td>
</tr>
</tbody>
</table>

| Generating simulated data with 4132 genes for two simulation set-ups respectively were generated to mimic the osteoblast differentiation data.

Another data set with 4132 genes in it is generated similarly, in which 200 genes have differential expression and the rest of the genes have no expression changes.

Ignoring within-subject correlation

We apply both gene-specific scores, \( w(i) \) and \( w_{ind}(i) \), to the simulated data set under \( H_0 \). When we detect a similar number of significant genes by using \( w(i) \) and \( w_{ind}(i) \) in SAM respectively, the estimated FDR by using \( w_{ind}(i) \) is much smaller than that by using \( w(i) \) instead (Table 3). For example, we detect 4 significant genes with 3.8 false positives using \( w(i) \), whereas we detect 4 significant genes with only 1.6 false positives using \( w_{ind}(i) \). By the nature of how the simulated data set was generated under \( H_0 \), we know that all the identified significant genes are false positives, i.e., the true FDR = 1. Thus, use of \( w(i) \) performs better than use of \( w_{ind}(i) \) because the estimated FDR from the former is closer to the true value 1 than that from the latter. The same conclusion can be drawn when MMM is used: FDR can be dramatically underestimated if \( w_{ind}(i) \), rather than \( w(i) \), is used (Table 3).

Both \( w(i) \) and \( w_{ind}(i) \) are applied to the simulated data set with 200 genes differentially expressed to see which of them is more powerful to detect genes truly differentially expressed (Table 4). Among the genes identified as differentially expressed by the proposed methods (TP), some of them are in the set of 200 genes known to be differentially expressed and the number of these genes are indicated as ‘True’ in Table 4. When a similar number of genes are identified by \( w(i) \) and \( w_{ind}(i) \) in SAM, \( w(i) \) gives a larger number of the genes truly differentially expressed. This is also true for MMM at different Type I error rates. As we have noticed, even though \( w(i) \) is more powerful to detect genes that are truly differentially expressed than \( w_{ind}(i) \), it still misses quite
some true positives and identifies false positives unless the microarray experiments are performed with more replications.

Both $w(i)$ and $w_{\text{ind}}(i)$ are constructed based on estimating equation techniques. The only difference between them is whether to consider the potential within-subject correlation for longitudinal microarray data. If we ignore the within-subject correlation and use the corresponding gene specific score $w_{\text{ind}}(i)$, we tend to underestimate FDR and lose power. Hence, we should use a statistic like $w(i)$ to take account of the within-subject correlation to make valid statistical inferences.

### Comparison between SAM and MMM

For the simulated data set under $H_0 : \tau_{11} = \tau_{12} = 0$, the number of significant genes should be close to the number of false positives; that is, FDR should be close to 1. The results from SAM and MMM are consistent with each other (Table 3). When we use $w(i)$ to identify significant genes, both SAM and MMM provide good FDR estimates. For example, when both of them identify 4 significant genes, the estimated FP is 3.8 from SAM and 3.5 from MMM.

When we apply both SAM and MMM to the simulated data set with 200 genes differentially expressed, we find that SAM and MMM have similar power to detect genes truly differentially expressed (Table 4). When the two methods both identify 69 significant genes using $w(i)$, 64 of them are truly differentially expressed. This is also true when we use $w_{\text{ind}}(i)$. When SAM identifies 113 significant genes and MMM identifies 116 significant genes, the number of genes truly differentially expressed is 88 for SAM and 89 for MMM.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>True TP (%)</th>
<th>True/TP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>0.005</td>
<td>54</td>
<td>66</td>
</tr>
<tr>
<td>0.01</td>
<td>89</td>
<td>116</td>
</tr>
<tr>
<td>0.05</td>
<td>165</td>
<td>325</td>
</tr>
</tbody>
</table>

Table 4. Identifying genes truly differentially expressed by SAM and MMM for the simulated data, in which 200 genes have differential expression, with the use of $w(i)$ or $w_{\text{ind}}(i)$.

### CONCLUSIONS

It has been observed that the gene expression patterns of samples of a given tumor measured at different times are much more similar to each other than are the expression patterns of tumor samples of the same type taken from different subjects (Perou et al., 2000). In statistics, this phenomenon is called the possible within-subject correlation of repeated measurements on the same subject, and the resulting data are called longitudinal data. It is well known in other applications that valid statistical analyses have to appropriately take account of possible within-subject correlation in longitudinal data. To our knowledge, there are few studies considering significance tests with longitudinal gene expression data. In addition to the current study, the only other exception is Zhao et al. (2002) who proposed a parametric method, in contrast to our nonparametric method. In particular, though Xu et al. (2002) discussed regression analysis for time-course gene expression data, due to the way the experiment was done, the authors (correctly) assumed the independence between a gene’s expression levels measured at different time points; that is, there is no within-subject correlation. However, for our example of osteoblast differentiation data, it is verified that the within-subject correlation exists, due to the nature of repeated measurements arising from the use of multiple samples taken on the same subject. In this paper, we have presented a robust statistic to detect genes with temporal changes in expression for longitudinal gene expression data. Built on estimating equation techniques, this test statistic appropriately accounts for possible within-subject correlation present in longitudinal gene expression data. Following the basic idea in SAM, we have also proposed a method to stabilize the covariance matrix and thus the test statistic. Then the proposed statistic can be applied in a similar way to that in some existing statistical methods, such as SAM or MMM, to identify significant genes. Using simulated data, we have shown the importance of properly taking account of the within-subject correlation in longitudinal gene expression data. Ignoring the correlation may lead to invalid statistical inference. Although we only showed its application to test two mean parameters in our example using a mesodermal progenitor cell data set, this test statistic is general and in principle can be applied to detect more complex expression patterns.

### ACKNOWLEDGEMENTS

We thank the reviewers and the editor for many helpful and constructive comments. X.G. and W.P. were supported in part by an NIH grant (R01-HL65462) and a Minnesota Medical Foundation grant.

### REFERENCES
