Journal of Biometrics & Biostatistics

Wang et al., J Biomet Biostat 2012, S8
DOI: 10.4172/2155-6180.S8-001

Research Article

Open Access

Statistical Analysis of Patient-Specific Pathway Activities via Mixed Models

Lily Wang[1]*, Xi Chen[2] and Bing Zhang[3]

[1]*Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA*
[2]*Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[3]*Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA*

Abstract

In the study of complex diseases, a major challenge is disease heterogeneity, where the dysregulation of different pathways often lead to similar disease phenotypes. As a result, a given pathway could be differentially expressed with respect to controls for some patients, but not for others. Therefore, to develop successful personalized treatment regime, in addition to identifying disease relevant pathways for the entire patient group, it's also important to test if a particular pathway is dysregulated for an individual patient. To this end, we compare pathway gene expression profile for a particular individual in the patient group to the "norm" (or standard) established by a group of control patients. We studied statistical analysis of patient-specific pathway activities under the mixed models framework. Using gene expression dataset with realistic correlation patterns, we showed the proposed hypothesis testing procedure had false positive rate (type I error) as expected. In addition, we illustrated the proposed methodology using a Type 2 Diabetes dataset. Our results showed a previously diabetes associated pathway was only differentially expressed (relative to the control group) in less than 30% of the diabetes patients, which provided an explanation for the moderate group level statistical significance seen in a previous study. This result also suggested targeting this particular pathway would likely be beneficial for only 30% of the patients. In addition to the case-control study we have illustrated, this model can be easily extended to handle more complex designs with additional covariates and multiple sources of variations. Moreover, the proposed model operates within a well-established statistical framework and can be implemented in common statistical packages.

Keywords: Microarray; Gene expression; Mixed models; Pathway analysis; Gene set analysis; Statistical significance

Introduction

In the study of complex diseases, a major challenge is disease heterogeneity, where the dysregulations of different pathways often lead to similar disease phenotypes. As a result, a given pathway can be differentially expressed (with respect to controls) for some patients, but not others. Therefore, in addition to identifying disease relevant pathways for the entire patient group, successful (personalized) treatment regimes will also depend upon knowing if a particular pathway is dysregulated for an individual patient. To this end, we compare pathway gene expression profile for a particular individual in the patient group to the "norm" (or standard) established by a group of control patients. The questions we are hoping to address are: for a particular patient, is his gene expression level for a given pathway normal (as compared to control subjects)? Will therapy targeted at a particular pathway likely be beneficial for the patient?

Over the past few years, many pathway analysis tools [1-6] have been developed. Typically, to identify differentially expressed pathways for a disease, the pathway expression profile for the patient group is compared to that for the control group. However, once a disease relevant pathway at the group level is identified, few if any of these methods can be used to assess the statistical significance of patient-specific pathway activities. Typically, the mean or principal component score are used to summarize pathway activity in each patient, with no significance assessment [7,8]. Hypothesis testing of pathway gene expression for individual patient is challenging, because when comparing a patient pathway gene expression to that of a group of control patients, different sources of variations - the between-sample variations as well as the within-sample variations (among different genes), need to be accounted.

In this paper, we study statistical analysis of patient-specific pathway activities under the mixed models framework [9]. Mixed effects models, which include fixed effects that model the mean structure in data and random effects that account for various sources of variations, is a flexible statistical modelling framework. Previously, mixed effects models have been successfully applied to the analysis of gene expression data, both at the single gene level [10,11] and at the pathway level [2,3,5,6].

In particular, Wang et al. [5] proposed a mixed effects random coefficient model for the analysis of time course experiments. Here we adapt this model to case-control studies, which is the most common design for human microarray studies. Our model includes both fixed effects that model mean gene expression profiles for the patient and control groups, and random effects that model how each subject's profile varies about the group mean, thus belongs to the general class of mixed effects models. To account for the complex correlation patterns between genes, we additionally including random effects based on eigenvectors and Eigen-values of the gene-gene covariance matrix. To assess the properties of the proposed test for patient-specific pathway activities, we conducted a simulation study using gene expression

***Corresponding author:** Lily Wang, Department of Biostatistics, S-2323 Medical Center North, Nashville, TN 37232, USA, Tel: (615) 343-3856; Fax:(615) 343-4924; E-mail: lily.wang@vanderbilt.edu

Received June 02, 2011; **Accepted** February 15, 2012; **Published** February 22, 2012

Citation: Wang L, Chen X, Zhang B (2012) Statistical Analysis of Patient-Specific Pathway Activities via Mixed Models. J Biomet Biostat S8:001. doi:10.4172/2155-6180.S8-001

data with realistic correlation patterns. In addition, we illustrate the proposed method for a real diabetes microarray dataset. Finally, we provide some discussions and concluding comments.

Methods

Microarray pre-processing

Before fitting mixed models, there are several pre-processing steps. First, for each gene, to homogenize variances of all genes included in the mixed model, we standardize each (log transformed, normalized) gene expression value by subtracting its control group mean and dividing by its control group standard deviation. The standardized gene expression values then represent the number of standard deviations away from the "normal" gene expression values [5,6].

Mixed models for gene set analysis

Next, we link gene identifiers in the expression dataset with pre-defined gene sets such as those defined by Gene Ontology [12], so that genes are grouped by gene sets. For each gene set, we next construct the following mixed model for a case-control study:

$$Y_{ijk} = Group_j + Patient_k + r_{1i} + \ldots\ldots + r_{pi} + \varepsilon_{ijk} \qquad \text{(Model 1)}$$

where Y_{ijk} = standardized gene expression value for gene *i* from patient *k* in group *j* (*j* = 1 for patient group, *j* = 0 for control group) from the pre-processing step; $Group_0$, $Group_1$ are fixed effects that model the mean pathway gene expressions for the two groups; $Patient_1,\ldots\ldots,Patient_n$ are random effects that model patient variations, they model how pathway gene expression for a patient deviates from the group means. Since variations in patient samples may be different from (e.g. often more variable) than those in control samples, we assume separate variance components for the two groups:

$$Patient_1,\ldots,Patient_{n0} \sim independent\ N(0,\sigma_0^2)$$

and

$$Patient_{n0+1},\ldots,Patient_n \sim independent\ N(0,\sigma_1^2);$$

$$r_1,\ldots,r_p \sim independent\ N(0,\sigma_r^2)$$

are random effects included to account for the heterogeneous correlation patterns between genes, see details below; *p* is the rank of the gene-gene covariance matrix; $\varepsilon_{ijk} \sim N(0,\sigma^2)$ represent variations due to measurement error.

Note that while parameters for the fixed effects (e.g. $Group_0$, $Group_1$) are fixed unknown parameters to be estimated from data, random effects (e.g. $Patient_1$, $Patient_2,\ldots.Patient_n$)are random variables and the parameter associated with them $(e.g.\sigma_0^2\sigma_1^2)$ are called the variance components. Parameters from the mixed models are estimated using restricted maximum likelihood (REML) along with appropriate standard errors [9,13].

Modelling the heterogeneous correlation patterns between genes

In Model 1, the *Patient* random effects are constructed as indicator variables for each sample, that is, $Patient_k = I\{Patient\ k\}$, they account for the homogeneous covariance among all gene expression values from the same patient. On the other hand, the random effects $r_1,\ldots..,r_p$ model the heterogeneous correlation patterns between genes. Figure 1 illustrates the design matrix corresponding to $\{r_l; l=1,\ldots\ldots..,3\}$for a hypothetical gene set with 3 genes using SAS procedure PRINCOMP. Briefly, let $\hat{\boldsymbol{\Sigma}}$ be the sample gene-gene covariance matrix with dimension $p\times p$ (*p* = number of genes in the gene set), we specify the column in the design matrix corresponding to random effect r_l to be $\sqrt{\hat{\lambda}_l}\hat{\boldsymbol{\alpha}}_l$ where $\hat{\boldsymbol{\alpha}}_l$ = estimated *l*-th eigenvector of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\lambda}_l$ = estimated *l*-th eigenvalue of Σ, $l=1,..,p$. The eigenvectors and eigenvalues of a matrix $\boldsymbol{\Sigma}$ are defined as vectors α_1 and scalars λ_1 such that $\Sigma\alpha_l = \lambda_l\ \alpha_l$ $l=1,\ldots,p$. This design matrix for the random effects $r_1,\ldots.,r_p$ is motivated by the theorem on Spectral Decomposition [14], a short proof is given in the next section.

Application of the spectral decomposition theorem

A general representation of the linear mixed model is

$$\boldsymbol{Y = X\beta + Zu + e}$$

$$\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G})$$

$$\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{R})$$

$$Cov[\boldsymbol{u}, \boldsymbol{e}] = \mathbf{0}$$

where ***X,Z*** are design matrices for the fixed and random effects, ***Xβ*** and ***Zu*** are the fixed effects and random effects components, and ***e*** is the error term. The marginal model for ***Y*** is then $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ}^t + \boldsymbol{R})$.

For Model 1 described above, the fixed effects consisted of $\boldsymbol{\beta} = [Group_0\ Group_1]^t$ and ***X*** is the corresponding design matrix. The random effects ***u*** has three parts:

$$Patient_1,\ldots,Patient_{n0} \sim independent\ N(0,\sigma_0^2), \qquad (1)$$

$$Patient_{n0+1},\ldots,Patient_n \sim independent\ N(0,\sigma_1^2), \qquad (2)$$

and

$$r_1,\ldots,r_p \sim independent\ N(0,\sigma_r^2).\ \text{We assume } \boldsymbol{R}=\sigma^2\boldsymbol{I} \qquad (3)$$

Let $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$ be the sub-matrices of the design matrix ***Z*** corresponding to the three sets of random effects, so that $\boldsymbol{Z} = [\boldsymbol{Z_1}, \boldsymbol{Z_2}, \boldsymbol{Z_3}]$. The design matrix $\boldsymbol{Z_3}$ corresponding to the random effects $r_1,\ldots.,r_p$ was motivated by the theorem on Spectral Decomposition [14] which states that under regularity conditions, for any symmetric matrix (with rank *p*), we have

$$\boldsymbol{\Sigma} = \lambda_1\alpha_1\alpha_1^t + \lambda_2\alpha_2\alpha_2^t + \ldots + \lambda_p\alpha_p\alpha_p^t,$$

where α_l and λ_l (l = 1, …, *p*) are *l*-th eigenvector and Eigen-value of Σ.

We next show the random effects $r_1,\ldots.,r_p$ model heterogeneous covariances between genes. Assume

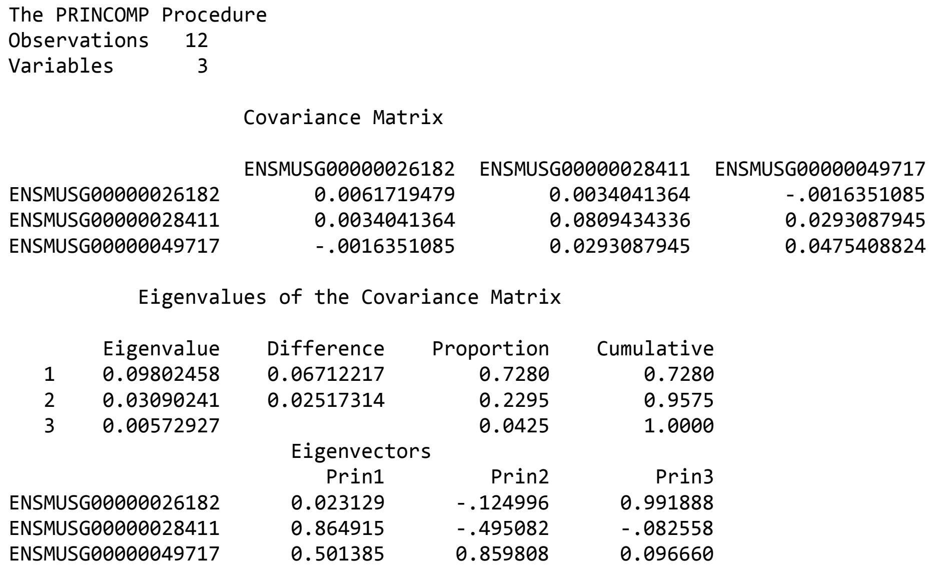
$$r_1,\ldots,r_p \sim \text{ independent } N(0,\sigma_r^2),$$

as described above (and illustrated in Figure 1), we then have

$$Z_3 = \left[\sqrt{\hat{\lambda}_1}\hat{\alpha}_1\ \ \sqrt{\hat{\lambda}_2}\hat{\alpha}_2\ \ \ldots\ \ \sqrt{\hat{\lambda}_p}\hat{\alpha}_p\right]$$

$$\boldsymbol{G}_3 = \begin{bmatrix} \sigma_r^2 & 0 & \ldots & 0 \\ 0 & \sigma_r^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \sigma_r^2 \end{bmatrix}$$

where $\boldsymbol{G}_3$ is diagonal matrix corresponding to $\boldsymbol{Z}_3$. Let $\boldsymbol{r} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots\ldots \boldsymbol{r}_p]^t$,



```
The PRINCOMP Procedure
Observations   12
Variables       3

                     Covariance Matrix

                     ENSMUSG00000026182  ENSMUSG00000028411  ENSMUSG00000049717
ENSMUSG00000026182         0.0061719479        0.0034041364        -.0016351085
ENSMUSG00000028411         0.0034041364        0.0809434336        0.0293087945
ENSMUSG00000049717         -.0016351085        0.0293087945        0.0475408824

             Eigenvalues of the Covariance Matrix

        Eigenvalue    Difference    Proportion    Cumulative
   1    0.09802458    0.06712217        0.7280        0.7280
   2    0.03090241    0.02517314        0.2295        0.9575
   3    0.00572927                      0.0425        1.0000
                         Eigenvectors
                            Prin1         Prin2         Prin3
ENSMUSG00000026182       0.023129      -.124996      0.991888
ENSMUSG00000028411       0.864915      -.495082      -.082558
ENSMUSG00000049717       0.501385      0.859808      0.096660
```

Figure 1: An illustration of computation for random effects $\{r_l; l=1,...,p\}$ in Model 1, using a hypothetical gene set with 3 genes (variables) and 12 samples (observations). Covariance Matrix = estimated gene-gene covariance matrix $\hat{\boldsymbol{\Sigma}}$. Under "Eigenvalues of the Covariance Matrix", $\hat{\lambda}_1$ = 0.09802458 is the estimated first eigenvalue of $\hat{\boldsymbol{\Sigma}}$. Under "Eigenvectors", Prin 1 = $\hat{\boldsymbol{\alpha}}_1$ is the estimated first eigenvector of $\hat{\boldsymbol{\Sigma}}$. r_1 is computed as a scaled product of the first eigenvalue and eigenvector, or $\sqrt{0.098}\hat{\boldsymbol{\alpha}}_1$, note that they vary according to genes, so the random effects have sub-index *i* in Model 1.

the contribution of random effects $r_1,...,r_p$ to the covariance matrix of $\boldsymbol{Y}$ in the marginal model would then be

$$\text{var}(\boldsymbol{Z}_3\boldsymbol{r}) = \boldsymbol{Z}_3\boldsymbol{G}_3\boldsymbol{Z}_3{}' = \sigma_r^2\hat{\lambda}_1\hat{\boldsymbol{\alpha}}_1\hat{\boldsymbol{\alpha}}_1^t + ... + \sigma_r^2\hat{\lambda}_p\hat{\boldsymbol{\alpha}}_p\hat{\boldsymbol{\alpha}}_p^t$$

Next, we show the approximation of gene-gene covariance matrix $\boldsymbol{\Sigma}$ using $\boldsymbol{Z}_3\boldsymbol{G}_3\boldsymbol{Z}_3^t$ based on the estimated eigenvectors and Eigen-values is asymptotically unbiased. To see this, note that

$$\begin{aligned}
&E\left(\hat{\lambda}_l\hat{\boldsymbol{\alpha}}_l\hat{\boldsymbol{\alpha}}_l^t\right) \quad l=1,...,p \\
&= E(\hat{\lambda}_l)\,E\left(\hat{\boldsymbol{\alpha}}_l\hat{\boldsymbol{\alpha}}_l^t\right) \quad \hat{\lambda}_l \text{ and } \hat{\boldsymbol{\alpha}}_l \text{ are independent (Jolliffe 2002, p48)} \\
&= \lambda_l\left\{E\left(\hat{\boldsymbol{\alpha}}_l\right)E\left(\hat{\boldsymbol{\alpha}}_l\right) + Var\left(\hat{\boldsymbol{\alpha}}_l\right)\right\} \\
&= \lambda_l\left\{\boldsymbol{\alpha}_l\boldsymbol{\alpha}_l^t + O\left(1/n\right)\right\} \quad \text{(Jolliffe 2002, p48)} \\
&\to \lambda_l\boldsymbol{\alpha}_l\boldsymbol{\alpha}_l^t \quad \text{as } n\to\infty
\end{aligned}$$

Now letting $\sigma_r^2 = 1$, we then have

$$E(\boldsymbol{Z}_3\boldsymbol{G}_3\boldsymbol{Z}_3{}') \to \lambda_1\alpha_1\alpha_1^t + ... + \lambda_p\alpha_p\alpha_p^t = \boldsymbol{\Sigma} \text{ as } n\to\infty$$

In practice, to increase the goodness-of-fit of Model 1, we include σ_r^2 as an unknown parameter and use restricted maximum likelihood to obtain its estimate. In addition, since $\boldsymbol{\Sigma}$ is not known, we replace $\boldsymbol{\Sigma}$ with its unbiased estimate $\hat{\boldsymbol{\Sigma}}$.

Significance testing of patient-specific pathway activities

Once estimates for fixed effects $\{Group_j;\ j=0,1\}$ and random effects $\{Patient_k;\ k=1,..,K\}$in Model 1 are obtained, to assess pathway significance for a particular patient (*Patient k*) from the patient group, we test the null hypothesis

$$H_0{:}Group_1{+}Patient_k - Group_0 = 0$$

Here, $Group_1{+}Patient_k$ represents average pathway expression (over all genes in the pathway) for $Patient_k$ and $Group_0$ represents average pathway expression for control patients.

To perform hypothesis testing for $Group_1{+}Patient_k - Group_0$, an approximate *t*-statistics can be computed using PROC MIXED in SAS software (Version 9.1, SAS Institute, Inc., Cary, North Carolina) based on a formula described in details in SAS 9.1.3 help documentation (p2740-2741) and [15,16]. More specifically, for the general linear mixed model describe above in the section "Application of the Spectral Decomposition Theorem", to test the null hypothesis $H_0 : L\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix} = 0$ where L is a single row of parameter contrast, we can construct an approximate t-statistic $t = L\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{u}}\end{pmatrix} / \sqrt{L\hat{C}L'}$ where $\hat{C} = \begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1}\end{bmatrix}^{-}$ is the variance-covariance of $\left(\hat{a} - \hat{a}, \hat{u} - u\right)$ and - denotes a generalized inverse. This approximation is based on asymptotic argument and $\hat{\boldsymbol{C}}$ tends to underestimate the true sampling variability of $(\hat{\boldsymbol{a}},\hat{\boldsymbol{a}})$ because no account is made for the uncertainty in estimating **G** and **R**. To account for the downward bias, we included DDFM=KR option in the MODEL statement, which prompted PROC MIXED to compute a specific inflation factor along with Satterthwaite-based degrees of freedom based on Kenward and Roger [17]. In the statistics literature, the estimates for the random *patient* effects are called Empirical Best Linear Unbiased Predictors (EBLUPs), and they have been shown to have several desirable properties [18] that make them especially attractive for estimating pathway activities for individual patients: (1) they are "best" in the sense that they are linear functions of the gene expression data Y that minimizes mean squared error between the predictor (estimates for random effects) and the true values of the random effects; (2) they are "unbiased" that their expectation is equal to the expectation of the random effects; (3) they are shrinkage estimates that borrow information across all subjects in the study, by accounting for the underlying variability between and within patients; (4) in the Bayesian literature, they can be formulated as the mean of posterior distribution of the random effects given observed data $\boldsymbol{Y}$, and are called the "empirical bayes" estimates. We illustrate significance testing of pathway expression for individual patients in the Results section.

Summary of the proposed procedure

In summary, the steps for the mixed model analysis are as follows:

Pre-processing: to homogenize variances of all genes, standardize each (log transformed, normalized) gene expression value by subtracting its control group mean and dividing by its control group standard deviation. Next, group genes by gene sets and for each gene set:

For a gene set with p genes, specify design matrix (i.e. values) for the random effects $r_1,\dots,r_{p.}$

a. For each gene, to remove the mean effects, we fit the linear model $Y_{ijk} = Group_j + Patient_k + \in_{ijk}$ where Y_{ijk} denotes standardized expression value for gene i, patient k in group j obtained from step 1). From this model, the studentized residuals ([13] p415), which are residuals divided by an estimate of its standard deviation are then computed.

b. Let $\hat{\boldsymbol{\Sigma}}$ be the sample gene-gene covariance matrix (an unbiased estimate of Σ), calculated based on the studentized residuals from the gene-wise linear models in step 2) a. Specify the column in design matrix corresponding to random effect r_1 to be $\sqrt{\hat{\lambda}_l}\hat{\boldsymbol{\alpha}}_l$ where $\hat{\boldsymbol{\alpha}}_l$ = estimated l-th eigenvector of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\lambda}_l$ = estimated l-th eigenvalue of $\hat{\boldsymbol{\Sigma}}$ (see Figure 1 for an example).

Fit mixed model $Y_{ijk} = Group_j + Patient_k + r_{1i} + \dots + r_{pi} + \varepsilon_{ijk}$ described in the section "Mixed Models for Gene Set Analysis", obtain estimates for model parameters. To test for significance of the patient specific pathway activities, we test H_0: $Group_1 + Patient_k - Group_0 = 0$. Here, $Group_1 + Patient_k$ represents average pathway expression (over all genes in the pathway) for $Patient_k$ and $Group_0$ represents average pathway expression for control patients.

Supplementary file 1 shows example SAS program for implementing the proposed mixed model analysis.

Results

A simulation study

To ensure false positive rate based on the mixed effects model is as expected, we simulated null gene sets, for which disease status for each sample was generated randomly from a Bernoulli distribution and estimated the type I error rate of patient specific tests described in section "Significance testing of patient-specific pathway activities". To obtain realistic correlation patterns between genes for this simulation study, we used gene expression data from a real microarray experiment along with simulated disease outcomes (i.e. case-control statuses for each patient). More specifically, we used the dataset from Mootha et al. [19] where gene expression of skeletal muscle biopsy samples from 18 diabetes patients (DMT group) were compared to those from 17 control patients with normal glucose tolerance (NGT group).

Some pre-processing steps are in order: first, we grouped genes based on the biological process categories in Gene Ontology, by using the C5BP collection of gene sets from the MSigDB database [4] (http://www.broadinstitute.org/gsea/msigdb/) with sizes (the number of genes in the gene set) ranging from 5 to 200, this resulted in a total of 744 gene sets. Next, for each gene set, fixing the gene expression data, we generated a set of 35 case control status for the samples, randomly from the Bernoulli distribution with parameter (success probability) 0.5. Therefore, by design of experiment, for each gene set, the gene expression data were not related to disease outcome and the null hypothesis H_0: *gene expression profile for a patient sample is the same as the gene expression profiles in control samples* is true. This process was then repeated 10 times. For each repetition, for the 18 diabetes patients across all 744 gene sets, we obtained a total of 13392 patient-specific P-values (744 gene sets 18 patients). We then estimated type I error rate by the proportion of these patient-specific pathway P-values less than 0.05 among the total of 13392 P-values.

For comparison, we also included another simple and tempting approach using Wilcoxon signed-rank test. Briefly, to estimate pathway significance for a patient sample, we compared the gene expression profile for the patient sample with the average gene expression profile for control samples using the Wilcoxon signed-rank test for paired data. The signed-rank test is used here since for each gene, we have a pair of values, one from the patient sample, and another from the average of the control samples. Note that since the gene expression levels in control samples were averaged, this method doesn't account for between-patient variations in the control samples. In addition, the within-sample correlations (between different genes) were also ignored since the signed-rank test assumes independent units (i.e. genes).

Table 1 shows the estimated type I error rate for the proposed mixed model and Wilcoxon signed-rank test for each of the 10 repetitions. To estimate type I error rate, we pooled results for the 18 patient samples, so that for each repetition, each error estimate was calculated based on a total of 13392 tests (744 gene sets ×18 patient samples). Since under H_0, we expect the P-values to follow a uniform distribution, a method with type I error rate roughly equal to or less than the significance cutoff 0.05 is desirable. The results showed that while the type I error rate for mixed Model 1 was preserved (overall type I error = 0.044), it was excessive for the simpler method using Wilcoxon signed-rank test (overall type I error = 0.137).

Application to a diabetes dataset

We next applied the mixed model analysis to the diabetes dataset

Repetition	*Mixed Model*	*Wilcoxon Signed-Rank Test*
1	0.041	0.139
2	0.043	0.138
3	0.049	0.137
4	0.044	0.136
5	0.045	0.135
6	0.040	0.135
7	0.041	0.134
8	0.044	0.143
9	0.043	0.134
10	0.054	0.138
Overall	0.044	0.137

For each repetition, the simulation dataset consisted of 744 gene sets and *P*-values for patient specific pathway activities were calculated for each of the 18 patient samples. Therefore the type I error rates were calculated based on a total of 13392 tests (744 gene sets ×18 patient samples) for each repetition.

Table 1: Type I error rate for Mixed Model 1 and the Wilcoxon Signed-Rank test ($\alpha = 0.05$).

Patient ID	P-value	Adjusted P-value
S1	0.028	0.505
S2	0.743	1.000
S3	0.019	0.334
S4	0.001	0.019
S5	0.200	1.000
S6	0.757	1.000
S7	0.059	1.000
S8	2.899E-04	0.005
S9	0.003	0.059
S10	0.048	0.866
S11	0.879	1.000
S12	0.090	1.614
S13	8.905E-08	1.603E-06
S14	3.312E-05	0.001
S15	0.013	0.228
S16	4.444E-04	0.008
S17	0.006	0.113
S18	0.020	0.352

S1 = Patient sample 1
P-value = nominal P-value for testing differential expression of patient-specific pathway activities (against mean gene expression of control group)
Adjusted P-value = Bonferroni corrected P-value.

Table 2: Significance testing of patient-specific pathway activities for the Oxidative Phosphorylation Pathway.

[19] with real disease outcomes. In particular, we studied the *Oxidative Phosphorylation Pathway* from the KEGG database [20]. This pathway was initially studied by Mootha et al. [19] and was shown to be associated with type 2 diabetes (T2D) based on multiple computational and experimental evidences. When the entire patient group was compared to controls, the *P*-value based on mixed model (by testing *Ho: $Group_1$- $Group_0$*) was 0.0231, indicating significant association between *Oxidative Phosphorylation Pathway* gene expression with T2D. Next, we computed statistical significance of the pathway activity for each diabetes patient compared to the controls. To correct for multiple comparisons, we performed Bonferroni corrections to *P*-values of the patient specific tests. The results showed there were 5 out of the 18 patients with differentially expressed pathway expression compared to controls (adjusted *P*-value < 0.05, Table 2). In Mootha et al. [19], the authors noted a typical gene in this pathway was only moderately decreased by about 20% in diabetes patients. As table 2 shows, one possible explanation for the moderate group level estimate could be that this pathway was only differentially expressed for some patients (with respect to controls), but not all patients.

Discussion

In this paper, we were mainly concerned with hypothesis testing for individual patients, which differed from the prediction analysis framework: a prediction model trains a classifier based on a group of patients and a group of controls and assumes everyone in the patient group have dysregulated pathway activities; on the other hand, in the mixed models, we assume as a result of disease heterogeneity, a given pathway can be differentially expressed (with respect to controls) for some patients, but not others.

In the proposed mixed model, the mean for control group served the purpose of establishing a "norm" or standard for gene expression level of a particular pathway. It is worth noting that although pathway gene expression for each patient was compared to the mean pathway gene expression of the controls, our approach did account for between-subjects variability in the control group, by modeling the mean as well as its standard error in the test statistic.

In the course of this study, we also considered resampling-based procedures for estimating patient specific *P*-values. However, this was difficult for two reasons: first, when re-sampling genes, the underlying assumption is that the group of genes in each re-sample is exchangeable (e.g. have the same correlation patterns) with genes in other re-samples, which may not be reasonable considering the complexities in gene expression datasets. Second, when permuting samples for pathway expressions, as discussed in details previously [5], the hypothesis being tested is a *global* hypothesis that gene expression level for a patient is the same as the control group gene expression for *all the genes* in the gene set. On the other hand, here we aimed at testing a *central* hypothesis that the *average gene expression* activity for pathway genes in a patient is the same as that in controls. In our experience, testing global hypothesis may be difficult for some gene expression datasets with strong signals by generating excessive number of significant results.

In addition to matching patients with the most appropriate treatments, the estimated patient-specific pathway activities *P*-values can be applied to a number of other settings. For example, in analysis that integrates different types of omics datasets such as the expression QTL study, the patient specific *P*-values can be used to summarize information for each patient and help reduce dimensionality. Furthermore, these patient specific summaries on pathway activities can be applied to model disease progression [21] over time or compare diseases at a systems level [22].

In summary, we have proposed a new strategy for significance testing of patient-specific pathway activities using a mixed model. This model compares pathway gene expression for a patient to that of a group of controls, while modelling between-patient and within-patient (among different genes) variations at the same time. Using gene expression dataset with realistic correlation patterns, we have shown the proposed model had false positive rate (type I error) as expected. In addition, our results on a type 2 diabetes dataset showed that a previously diabetes associated pathway was only differentially expressed (relative to the control group) in less than 30% of the diabetes patients, which provided an explanation for the moderate group level statistical significance. These results also suggested targeting this particular pathway would likely be beneficial for only 30% of the patients. In addition to the case-control study we have illustrated, the mixed model can be easily extended to handle more complex designs with additional covariates and multiple sources of variations. Moreover, the proposed model operates within a well-established statistical framework and can be implemented in common statistical packages.

Acknowledgements

L.W. was partially supported by NICHD grant 5P30-HD015052-25 and NIH grant 1P50-MH078028-01A1. X.C. was partially supported by 5P30CA068485-13. B.Z. was partially supported by NIH grant 1P50-MH078028-01A1 and R01-GM088822.

References

1. Chen X, Wang L, Smith JD, Zhang B (2008) Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics 24: 2474-2481.

2. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC (2005) Testing association of a pathway with survival using gene expression data. Bioinformatics 21: 1950-1957.

3. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93-99.

4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545-15550.

5. Wang L, Chen X, Wolfinger RD, Franklin JL, Coffey RJ, et al. (2009) A unified mixed effects model for gene set analysis of time course microarray experiments. Stat Appl Genet Mol Biol 8: Article 47.

6. Wang L, Zhang B, Wolfinger RD, Chen X (2008) An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet 4: e1000115.

7. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353-357.

8. Guo Z, Zhang T, Li X, Wang Q, Xu J, et al. (2005) Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics 6: 58.

9. McCulloch CE, Neuhaus JM (2001) Generalized, Linear and Mixed Models. John Wiley & Sons, Inc.

10. Chu TM, Weir B, Wolfinger R (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci 176: 35-51.

11. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, et al. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8: 625-637.

12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

13. Littell RC, Miliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006) SAS for Mixed Models. SAS Institute Inc, Cary, NC, USA.

14. Jolliffe IT (2002) Principal Component Analysis. Wiley Online Library, New York, NY.

15. SAS 9.1.3 Help and Documentation. SAS Institute Inc, Cary, NC, USA, Page no: 2740-2741.

16. McLean RA, Sanders WL (1988) Approximating degrees of freedom for standard errors in mixed linear models. Am Stat Assoc 50-59.

17. Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53: 983-997.

18. Robinson GK (1991) That BLUP is a Good Thing: The Estimation of Random Effects. Statist Sci 6: 15-32.

19. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267-273.

20. Kanehisa M (2002) The KEGG database. Novartis Found Symp 247: 91-101.

21. Edelman EJ, Guinney J, Chi JT, Febbo PG, Mukherjee S (2008) Modeling cancer progression via pathway dependencies. PLoS Comput Biol 4: e28.

22. Li Y, Agarwal P (2009) A pathway-based view of human diseases and disease relationships. PLoS One 4: e4346.

This article was originally published in a special issue, **Biostatistics in systems biology** handled by Editor(s). Dr. Jimmy Efird, East Carolina University, USA; Dr. Rongling Wu, The Pennsylvania State University, USA; Dr. Chengjie Xiong, Washington University School of Medicine, USA