Published in final edited form as:

Stat Biosci. 2015 October 1; 7(2): 167–186. doi:10.1007/s12561-014-9109-1.

Random Effects Model for Multiple Pathway Analysis with Applications to Type II Diabetes Microarray Data

Herbert Pang,

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina 27705, U.S.A. Tel.: +919-681-5011, Fax: +919-668-5888

Inyoung Kim, and

Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, U.S.A. Tel.: +540-231-5366, Fax: +540-231-3863

Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, and Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, U.S.A. Tel.: +203-785-6271, Fax: +203-785-6912

Herbert Pang: herbert.pang@duke.edu; Inyoung Kim: inyoungk@vt.edu; Hongyu Zhao: hongyu.zhao@yale.edu

Abstract

Close to three percent of the world's population suffer from diabetes. Despite the range of treatment options available for diabetes patients, not all patients benefit from them. Investigating how different pathways correlate with phenotype of interest may help unravel novel drug targets and discover a possible cure. Many pathway-based methods have been developed to incorporate biological knowledge into the study of microarray data. Most of these methods can only analyze individual pathways but cannot deal with two or more pathways in a model based framework. This represents a serious limitation because, like genes, individual pathways do not work in isolation, and joint modeling may enable researchers to uncover patterns not seen in individual pathway-based analysis. In this paper, we propose a random effects model to analyze two or more pathways. We also derive score test statistics for significance of pathway effects. We apply our method to a microarray study of Type II diabetes. Our method may elucidate how pathways crosstalk with each other and facilitate the investigation of pathway crosstalks. Further hypothesis on the biological mechanisms underlying the disease and traits of interest may be generated and tested based on this method.

Keywords

Diabetes; Gene expression analysis; Microarray; Pathway tests; Random pathway effects; Score test

Correspondence to: Herbert Pang, herbert.pang@duke.edu.

Present address: School of Public Health, Li Ka Shing Faculty of Medicine, Hong Kong, China

Supplementary Materials

Detailed derivations and additional tables are given in the Supplementary Materials, is available online.

1 Introduction

1.1 Background on Diabetes

At least 171 million people, or 2.8% of the population, around the world suffer from diabetes, and this number may double in 20 years from now [42]. In the United States, diabetes was the 7th leading cause of death in 2007. The Centers for Disease Control and Prevention (CDC) gave a national estimate of a total of 25.8 million people, or 8.3% of the population, having diabetes [7]. About three-quarters of them are diagnosed and the remaining are undiagnosed. There are three main types of diabetes: Type I, Type II, and Gestational. Type II diabetes, also called noninsulin-dependent diabetes mellitus is the most common of all. In adults, it accounts for about 90% to 95% of all diagnosed cases of diabetes [7]. Type II diabetes usually begins with a disorder in which the malfunctioning cells cannot use insulin properly. This is called insulin resistant. If this is not resolved, it causes the pancreas to gradually loses its ability to produce insulin as its demand rises. Exposure to high levels of glucose over years can result in serious damage to major organs. Thus, diabetes is well known to cause several serious complications, such as blindness, heart disease, kidney damage, and limb amputations. It can also result in high medical cost burden with the total direct and indirect costs per year estimated to be 245 billion [1].

Recent years have seen great research efforts for diabetes studies. From the treatment perspective, Glycemic Control and Complications in Diabetes Mellitus Type II, a VA Administration study researches the effect of intensive glucose control on patients with Type II diabetes [13]. There are also prevention trials like the HEALTHY trial [5]. Some ongoing studies include the Action to Control Cardiovascular Risk in Diabetes study that seeks to identify comorbidity related deaths in adults with Type II diabetes using intensive management [14], and the Treatment Options for Type II Diabetes in Adolescents and Youth trial aims to identify the best treatment of Type II diabetes in children and teens [44]. Several drugs have been developed to treat Type II diabetes in the past decade, including Exenatide [4], Pramlintide [36], Liraglutide [11], and Sitagliptin phosphate [34]. However, not all patients benefit from these treatments. Therefore, scientists have continued to work on understanding the molecular and genetic basis of diabetes to advance the development of novel treatments. One promising approach is to study biological pathways. For example, [27] characterized the behavior of apoptotic signal transduction pathways in diabetes, and [37] identified a possible link between cancer and diabetes by studying the LKB1-AMPK pathway.

1.2 Pathway-based analysis

Numerous statistical methods have been developed to associate clinical outcomes with pathways based on microarray gene expression data. Most pathways considered in this context are predefined gene sets available from external databases such as KEGG [19]. In 2003, Mootha et al. published a paper that illustrates the advantages of taking the pathway-based approach over single gene based analysis in studying patients with Type II diabetes. Their data set consists of 22,283 genes measured in 17 patients with normal glucose tolerance and 18 samples with Type II diabetes mellitus. Combining microarray data with prior biological knowledge may help researchers better understand the biological process of

diseases and generate further hypotheses for testing. Mootha's research motivated us to develop our methodology to jointly examine multiple pathways.

Pathway-based methods involving a continuous measure as the outcome of interest include the score-based global test [15], multivariate linear model ANCOVA [28], tree-based regression random forests regression [31], least-squares kernel machines and linear mixed models [25], and semi-parametric Bayesian approach [21]. However, as far as we know, none of the existing methods looks at how two or more pathways jointly affect the clinical outcome of interest.

Understanding how multiple pathways relate to the phenotype of interest can help investigators discover possible crosstalks between pathways. These crosstalks provide additional information on the underlying mechanism of diseases. For example, pathways that give similar signals with respect to the phenotype of interest can be considered to have similar behavior, and genes that link these pathways may serve as a bridge between two different biological functions. Recently, Pang and Zhao [32] investigated how to build pathway clusters and uncover crosstalks between pathways using a tree-based classification algorithm. In this paper, we propose a random effects model based approach for continuous outcome data that jointly considers multiple pathways.

Linear mixed effects models have a long history, Henderson et al. in 1959 [17] published a paper applying this approach in the agricultural context. Liu et al. [25] was one of the first applications of this approach to genetic pathways. In this context, our model can be considered as a random effects model without the fixed effects which was described in [35]. We extend this modeling framework to two or more random effects. Since we are modeling pathways, we call these effects as pathway effects. We assume that the pathway effects have a multivariate normal distribution with zero mean and covariance structure having a special case of the polynomial kernel. This kernel was used in [15] to model the dependencies among genes within a gene set.

Our paper is organized as follows. In section 2, we define our model, discuss parameter estimation and statistical tests of pathway effects. In section 3, we generalize our model to incorporate multiple pathways. In section 4, we present a set of simulation results. In section 5, we apply our method to a microarray data. We conclude our work in the final section.

2 Model

2.1 Model

Suppose we have microarray data from n subjects. For each pathway k , let i denote the i -th subject, $\mathbf{y} = (y_1, \dots, y_n)$ be the continuous outcome vector for the n subjects, such as glucose level in the case of the diabetes study in section 4, and $\mathbf{x}_{ik}$ be the $g_k \times 1$ vector of gene expression with continuous values, where g_k is the number of genes for pathway k . Let $X_k = (\mathbf{x}_{1k}, \mathbf{x}_{2k}, \dots, \mathbf{x}_{nk})^T$ be the matrix consisting of gene expression for a pathway k for all the individuals.

[25] considered a linear mixed effects model which consists of both the fixed effects and the random effects. Because the fixed effects are relatively easy to handle, we will focus on

modeling pathways solely based on random effects in this paper. When there is one pathway, the model can be written as

$$\mathbf{y}=\mathbf{r}+\mathbf{e}$$

where $\mathbf{r}$ is an $n \times 1$ vector of random effects, which are called the pathway effects for this pathway, $\mathbf{r} \sim N_n(0, \tau_1 K_1)$, $\mathbf{e} \sim N_n(0, R = \sigma^2 I)$, τ_1 is the variance parameter for this pathway, and K_1 is a kernel of interest of dimension n by n . We will consider the kernel $K_1 = X_1 X_1^T$ used in [15], which is a special case of the polynomial kernel.

Under the above setup, the BLUP estimates of the pathway effects $\mathbf{r}$ are given by $\hat{\mathbf{r}} = \{\sigma^{-2}I + (\tau_1 K_1)^{-1}\}^{-1} \sigma^{-2} I \mathbf{y}$. Now we first extend this model to the two pathway case,

$$\mathbf{y}=\mathbf{r}_1+\mathbf{r}_2+\mathbf{e} \quad (1)$$

where both $\mathbf{r}_1$ and $\mathbf{r}_2$ are $n \times 1$ pathway effects vectors following $\mathbf{r}_1 \sim N_n(0, \tau_1 K_1)$, $\mathbf{r}_2 \sim N_n(0, \tau_2 K_2)$, $\mathbf{e} \sim N_n(0, R = \sigma^2 I)$, and τ_1, τ_2 are the pathway-specific variance parameters for pathways one and two, respectively. And let $K_1 = X_1 X_1^T$ and $K_2 = X_2 X_2^T$.

Assuming independence between the pathway effects $\mathbf{r}_1$ and $\mathbf{r}_2$, i.e. additive pathway effect, the joint density of $(\mathbf{y}, \mathbf{r}_1, \mathbf{r}_2)$ under model (1) is given by the following formula,

$$\frac{\exp\{-[(\mathbf{y} - \mathbf{r}_1 - \mathbf{r}_2)^T R^{-1}(\mathbf{y} - \mathbf{r}_1 - \mathbf{r}_2)]/2\} \exp\{-[\mathbf{r}_1^T (\tau_1 K_1)^{-1} \mathbf{r}_1]/2\} \exp\{-[\mathbf{r}_2^T (\tau_2 K_2)^{-1} \mathbf{r}_2]/2\}}{(2\pi)^{n/2+n/2+n/2} |\tau_1 K_1|^{1/2} |\tau_2 K_2|^{1/2} |R|^{1/2}}.$$

The independence assumption may have biological context. In the literature, there is evidence to support that independent pathways may play a part in biological development, biological transport, and biological regulation ([3], [46], [26]). Even when there is one gene overlapping between two pathways, the independence assumption may not be too severely violated when the 2nd order polynomial kernel is used to model the covariance structure in the multivariate normal pathway effect if the main effects from the second pathway are from genes other than the overlapping gene. There are also different forms of pathways such as metabolic, signaling and regulatory pathways. They serve different biological functions, some are cellular component, molecular functions, biological process to name a few.

Our goal is to maximize the above joint density with respect to $\mathbf{r}_1$ and $\mathbf{r}_2$.

This is equivalent to minimizing the following:

$$J=(\mathbf{y} - \mathbf{r}_1 - \mathbf{r}_2)^T R^{-1}(\mathbf{y} - \mathbf{r}_1 - \mathbf{r}_2) + \mathbf{r}_1^T (\tau_1 K_1)^{-1} \mathbf{r}_1 + \mathbf{r}_2^T (\tau_2 K_2)^{-1} \mathbf{r}_2$$

Differentiating with respect to $\mathbf{r}_1$ and $\mathbf{r}_2$, we get

$$\frac{\partial J}{\partial \mathbf{r}_1} = 2(\tau_1 K_1)^{-1} \mathbf{r}_1 - 2R^{-1} \mathbf{y} + 2R^{-1} \mathbf{r}_2 + 2R^{-1} \mathbf{r}_1 \quad \frac{\partial J}{\partial \mathbf{r}_2} = 2(\tau_2 K_2)^{-1} \mathbf{r}_2 - 2R^{-1} \mathbf{y} + 2R^{-1} \mathbf{r}_1 + 2R^{-1} \mathbf{r}_2$$

Setting these derivatives to zero, we obtain the following paired pathway effects model equations,

$$\begin{bmatrix} (\tau_1 K_1)^{-1} + R^{-1} & R^{-1} \\ R^{-1} & (\tau_2 K_2)^{-1} + R^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} R^{-1} \mathbf{y} \\ R^{-1} \mathbf{y} \end{bmatrix}, \quad (2)$$

where the first n equations are from the first derivative $J/\mathbf{r}_1$ and the second n equations are from the second derivative $J/\mathbf{r}_2$. Estimation of pathway effects are discussed in supplementary materials.

2.2 Estimation of Pathway-specific Parameters

The inference of the pathway effects $\mathbf{r}_1$ and $\mathbf{r}_2$ assumes that the pathway-specific parameters, τ_1 and τ_2 , are known. However, in practice, they are unknown and need to be estimated. By treating them as variance components in the random effects model, we can estimate them using maximum likelihood.

Specifically, the marginal log-likelihood function for model (1) can be written as

$$l_{ML}(\tau_1, \tau_2) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma(\tau_1, \tau_2)| - \frac{1}{2} \mathbf{y}^T \Sigma^{-1}(\tau_1, \tau_2) \mathbf{y},$$

where $\Sigma = \tau_1 K_1 + \tau_2 K_2 + \sigma^2 I$.

The gradient of τ_1 , τ_2 and σ^2 and the Hessian of them are derived and given in supplementary materials.

2.3 Score test

The scientific question of interest is: Given pathway A in the model, does pathway B remain significant? Pairs of pathways with strong individual pathway-specific effects can be put into the pair-pathway model to assess the significance of the effects of both pathways on the outcome of interest.

We can test for $H_0 : \tau_1 = 0$ using the score test, which is given by

$$U_{\tau_1}(\tau_2, \sigma^2) = -\frac{1}{2} \text{trace}(\Sigma_0^{-1}(\tau_2, \sigma^2) K_1) + \frac{1}{2} \mathbf{y}^T \Sigma_0^{-1}(\tau_2, \sigma^2) K_1 \Sigma_0^{-1}(\tau_2, \sigma^2) \mathbf{y},$$

where $\Sigma_0(\tau_2, \sigma^2) = \tau_2 K_2 + \sigma^2 I$.

The information matrix is given by

$$I_{\tau_1}(\tau_2, \sigma^2) = -\frac{1}{2} \text{trace}(\Sigma_0^{-1}(\tau_2, \sigma^2) K_1 \Sigma_0^{-1}(\tau_2, \sigma^2) K_1) + \mathbf{y}^T \Sigma_0^{-1}(\tau_2, \sigma^2) K_1 \Sigma_0^{-1}(\tau_2, \sigma^2) K_1 \Sigma_0^{-1}(\tau_2, \sigma^2) \mathbf{y}.$$

Therefore, the test statistic can be written as

$$S_{\tau_1}(\tau_2, \sigma^2) = \frac{[U_{\tau_1}(\tau_2, \sigma^2)]^2}{I_{\tau_1}(\tau_2, \sigma^2)}.$$

The mean and variance of $U_{\tau_1}(\tau_2, \sigma^2)$ are $\zeta = \text{trace}(\Sigma_0^{-1} K_1)/2$ and

$I_{\tau_1 \tau_1} = \text{trace}((\Sigma_0^{-1} K_1)^2)/2$, respectively. But since τ_2 and σ^2 are unknown, we estimate its maximum likelihood estimator by fitting the null model. The test statistics then becomes

$$S_{\tau_1}(\hat{\tau}_2, \hat{\sigma}^2) = \frac{2\zeta [U_{\tau_1}(\hat{\tau}_2, \hat{\sigma}^2)]^2}{I_{\tau_1 \tau_1}}.$$

We replace $I_{\tau_1 \tau_1}$ by the efficient information $\tilde{I}_{\tau_1 \tau_1} = I_{\tau_1 \tau_1} - I_{\tau_1 \tau_2} I_{\tau_2 \tau_2}^{-1} I_{\tau_2 \tau_1}^T$ to account for the fact that $\hat{\tau}_2$ and $\hat{\sigma}^2$ are maximum likelihood estimates.

Similarly, for $H_0 : \tau_2 = 0$, we can replace τ_1 by τ_2 and vice versa to obtain $U_{\tau_2}(\tau_1, \hat{\sigma}^2)$.

Intuitively, this may suggest a χ^2 distribution with one degree of freedom. However, as [45] pointed out, in contrast to the variance component score statistic of [23], the above score test can be shown to follow a mixture of chi-squares under H_0 instead. Specifically, following the Appendix of [45], assuming that the null hypothesis $H_0 : \tau_1 = 0$ holds, which implies that

$\mathbf{y} = \mathbf{r}_2 + \mathbf{e}$ is true. Let $M = \frac{1}{2}(\Sigma_0^{-1} K_1 \Sigma_0^{-1})$, then U_{τ_1} can be written as

$$U_{\tau_1} = \mathbf{y}^T M \mathbf{y} - \text{trace}(\Sigma_0^{1/2} M \Sigma_0^{1/2}) = \tilde{\mathbf{y}}^T \Sigma_0^{1/2} M \Sigma_0^{1/2} \tilde{\mathbf{y}} - \text{trace}(\Sigma_0^{1/2} M \Sigma_0^{1/2}),$$

where $\tilde{\mathbf{y}} = \Sigma_0^{-1/2} \mathbf{y} \sim N(0, I)$.

Let $\lambda_1 \dots \lambda_p > 0$ be the ordered non-zero eigenvalues of $\Sigma_0^{1/2} M \Sigma_0^{1/2}$ and let E be a p by n matrix consisting of the corresponding eigenvectors of λ_i such that E is orthonormal. It follows that,

$$U_{\tau_1} = \tilde{\mathbf{y}}^T E^T \Lambda E \tilde{\mathbf{y}} - \text{trace}(\Lambda) = \sum_{i=1}^p \lambda_i (Z_i^2 - 1),$$

where $\Lambda = \text{diag}(\lambda_i)$, $\mathbf{Z} = (Z_1, \dots, Z_p)^T = E \tilde{\mathbf{y}}$ and $Z_i \sim N(0, 1)$.

Therefore, under H_0 , we obtain a mixture of chi-square distribution for the distribution of the first term of U_{τ_1} given the true value of τ_2 and σ with one degree of freedom. But since the computation is intensive for the above procedure, the Satterthwaite method can be used to approximate the distribution of the score test by a scaled chi-square distribution $\kappa \chi_{\nu}^2$,

where $\kappa = I_{\tau_1 \tau_1}/2\zeta$, $\nu = 2\zeta^2/I_{\tau_1 \tau_1}$, $I_{\tau_1 \tau_1} = \text{trace}[(\Sigma_0^{-1} K_1)^2]/2$, $\zeta = \text{trace}(\Sigma_0^{-1} K_1)/2$.

3 Extension to More Than Two Pathways

We can generalize the above method to multiple pathways. For a total of q pathways, we can use the following model for the joint effects from these pathways,

$$\mathbf{y} = \mathbf{r}_1 + \mathbf{r}_2 + \dots + \mathbf{r}_q + \mathbf{e}$$

Based on the joint likelihood, say for q pathways, we can obtain for $\mathbf{r}_1, \dots, \mathbf{r}_q$ the pathway effects model equations as follows:

$$\begin{bmatrix} (\tau_1 K_1)^{-1} + R^{-1} & R^{-1} & \dots & R^{-1} \\ R^{-1} & (\tau_2 K_2)^{-1} + R^{-1} & \dots & R^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ R^{-1} & R^{-1} & \dots & (\tau_q K_q)^{-1} + R^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_q \end{bmatrix} = \begin{bmatrix} R^{-1} \mathbf{y} \\ R^{-1} \mathbf{y} \\ \vdots \\ R^{-1} \mathbf{y} \end{bmatrix},$$

where $\tau_1, \dots, \tau_q$ are the pathway-specific parameters for their respective q pathways, $K_1, \dots, K_q$ are the kernels for their respective q pathways, and R remains as the covariance for the error term in the model.

We can estimate the pathway effects after some tedious algebra to solve the above equations to obtain the following generalized form for the q pathway effects,

$$\begin{aligned} \widehat{\mathbf{r}}_1 &= [(\tau_1 K_1)^{-1} + (R + \tau_2 K_2 + \dots + \tau_q K_q)^{-1}]^{-1} (R + \tau_2 K_2 + \dots + \tau_q K_q)^{-1} \mathbf{y}, \\ \widehat{\mathbf{r}}_2 &= [(\tau_2 K_2)^{-1} + (R + \tau_1 K_1 + \tau_3 K_3 + \dots + \tau_q K_q)^{-1}]^{-1} (R + \tau_1 K_1 + \tau_3 K_3 + \dots + \tau_q K_q)^{-1} \mathbf{y}, \\ &\vdots \\ \widehat{\mathbf{r}}_q &= [(\tau_q K_q)^{-1} + (R + \tau_1 K_1 + \dots + \tau_{q-1} K_{q-1})^{-1}]^{-1} (R + \tau_1 K_1 + \dots + \tau_{q-1} K_{q-1})^{-1} \mathbf{y}. \end{aligned}$$

The pathway-specific parameters, $\tau_1, \dots, \tau_q$, can be obtained for the pathway effects model equations using the following generalized form of the gradient and Hessian.

For $\boldsymbol{\theta} = (\tau_1, \tau_2, \dots, \tau_q, \sigma^2)$, the gradient is as follows,

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = -\frac{1}{2} \text{trace}(\Sigma^{-1} \Sigma_{\boldsymbol{\theta}}) + \frac{1}{2} \mathbf{y}^T \Sigma^{-1} \Sigma_{\boldsymbol{\theta}} \Sigma^{-1} \mathbf{y},$$

where $\Sigma = \tau_1 K_1 + \dots + \tau_q K_q + \sigma^2 I$, and $\Sigma_{\boldsymbol{\theta}}$ is the first derivative of Σ with respect to $\boldsymbol{\theta}$.

The Hessian of $\theta_1, \theta_2 = (\tau_1, \tau_2, \dots, \tau_q, \sigma^2)$ is then,

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_2} = \frac{1}{2} \text{trace}(\Sigma^{-1} \Sigma_{\boldsymbol{\theta}_1} \Sigma^{-1} \Sigma_{\boldsymbol{\theta}_2}) - \frac{1}{2} \mathbf{y}^T (\Sigma^{-1} \Sigma_{\boldsymbol{\theta}_1} \Sigma^{-1} \Sigma_{\boldsymbol{\theta}_2} \Sigma^{-1} + \Sigma^{-1} \Sigma_{\boldsymbol{\theta}_2} \Sigma^{-1} \Sigma_{\boldsymbol{\theta}_1} \Sigma^{-1}) \mathbf{y}.$$

3.1 Score test for individual null in multiple pathways

The scientific question of interest is: Given other pathways in the model, does a particular pathway remain significant? In the multiple pathways setting, one may test for the absence of the individual random effects, such as $H_0 : \tau_j = 0$. The score statistics for testing the composite null hypothesis of $H_0 : \tau_j = 0$ against the one-sided alternative hypothesis, $H_a : \tau_j > 0$ was considered by [23]. In our case, the score for testing $H_0 : \tau_j = 0$ is

$$U_{\tau_j}(\hat{\tau}_{-j}, \hat{\sigma}) = \frac{\partial l(\tau, \sigma)}{\partial \tau_j} \bigg|_{\tau_j=0, \hat{\tau}_{-j}, \hat{\sigma}} = \left[-\frac{1}{2} \text{trace}(\Sigma_{-j}^{-1} \Sigma_{\tau_j}) + \frac{1}{2} y^T \Sigma_{-j}^{-1} \Sigma_{\tau_j} \Sigma_{-j}^{-1} y \right] \bigg|_{\tau_j=0, \hat{\tau}_{-j}, \hat{\sigma}}$$

where $\Sigma_{\tau_j} = K_j$, and $\Sigma_{-j} = \tau_1 K_1 + \dots + \tau_{j-1} K_{j-1} + \tau_{j+1} K_{j+1} + \dots + \tau_q K_q + \sigma^2 I$.

Let $v^T = (\tau_{-j}^T, (\sigma^2)^T)^T$, and $\hat{v}^T = (\hat{\tau}_{-j}^T, (\hat{\sigma}^2)^T)^T$ is the maximum likelihood estimates under the null. To test the null hypothesis of $H_0 : \tau_j = 0$ against the one-sided alternative hypothesis, $H_a : \tau_j > 0$, we can use the score statistic

$$S_{\tau_j}(\hat{v}^T) = U_{\tau_j}(\hat{v})^T \tilde{\vartheta}_{jj}^{-1}(\hat{v}) U_{\tau_j}(\hat{v})$$

where

$$\tilde{\vartheta}_{jj} = \vartheta_{\tau_j \tau_j} - \vartheta_{v \tau_j}^T \vartheta_{vv}^{-1} \vartheta_{v \tau_j}$$

is the efficient information for τ_j . Here,

$$\vartheta_{\tau_j \tau_j} = E \left(\frac{\partial l}{\partial \tau_j} \frac{\partial l}{\partial \tau_j} \right), \vartheta_{v \tau_j} = E \left(\frac{\partial l}{\partial v} \frac{\partial l}{\partial \tau_j} \right), \vartheta_{vv} = E \left(\frac{\partial l}{\partial v} \frac{\partial l}{\partial v^T} \right)$$

where the scores $\frac{\partial l}{\partial \tau_j}$ and $\frac{\partial l}{\partial v}$ and their expectations are calculated under $\tau_j = 0$.

Let μ_1^T, μ_2^T be the first and second moments of y , respectively, v_l be the l -th term of $v \setminus \{\sigma^2\} = (\tau_1, \dots, \tau_{j-1}, \tau_{j+1}, \dots, \tau_q)$, and $R(s, t) = \Sigma_{-j}^{-1} K_s \Sigma_{-j}^{-1} K_t \Sigma_{-j}^{-1} + \Sigma_{-j}^{-1} K_t \Sigma_{-j}^{-1} K_s \Sigma_{-j}^{-1}$, the elements of the efficient information matrix for this test are given by

$$\vartheta_{\tau_j \tau_j} = \frac{1}{2} \text{trace}(\Sigma_{-j}^{-1} K_j \Sigma_{-j}^{-1} K_j) - \frac{1}{2} \sum_i R_{ii}(j, j) \mu_{2i}^T - \sum_{i \neq j} R_{ij}(j, j) \mu_{1i}^T \mu_{1j}^T$$

The l -th element of vector v_{τ_j} is:

$$\vartheta_{v_l \tau_j} = \frac{1}{2} \text{trace}(\Sigma_{-j}^{-1} K_l \Sigma_{-j}^{-1} K_j) - \frac{1}{2} \sum_i R_{ii}(j, l) \mu_{2i}^T - \sum_{i \neq j} R_{ij}(j, l) \mu_{1i}^T \mu_{1j}^T$$

The (l, l') -th element of matrix ϑ_{vv} is:

$$\vartheta_{v_l v_{l'}} = \frac{1}{2} \text{trace}(\Sigma_{-j}^{-1} K_l \Sigma_{-j}^{-1} K_{l'}) - \frac{1}{2} \sum_i R_{ii}(l, l') \mu_{2i}^T - \sum_{i \neq j} R_{ij}(l, l') \mu_{1i}^T \mu_{1j}^T$$

where $\Sigma_{-j} = \tau_1 \hat{K}_1 + \dots + \tau_{j-1} \hat{K}_{j-1} + \tau_{j+1} \hat{K}_{j+1} + \dots + \tau_q \hat{K}_q + \sigma^2 I$.

Under some regularity conditions discussed in the Appendix, one can show that $S_{\tau_j}(\hat{v}^T) - E(U_{\tau_j})(\hat{v}^T)$ follows an asymptotically χ^2 distribution with q degrees of freedom.

Alternatively, the Satterthwaite method can be used to approximate the distribution of the score test by a scaled chi-square distribution similar to the previous section, $\kappa \chi_\nu^2$, where $\kappa =$

$$I_{\tau_j \tau_j} / 2\zeta, \nu = 2\zeta^2 / I_{\tau_j \tau_j}, I_{\tau_j \tau_j} = \text{trace}[(\Sigma_{-j}^{-1} K_j)^2] / 2, \zeta = \text{trace}(\Sigma_{-j}^{-1} K_j) / 2.$$

4 Simulation Studies

4.1 Two pathways

We carried out simulation study to assess the accuracies of the estimates, where 500 runs were performed for each of the simulation scenarios. Model 1 was considered in our simulations. Let g_1 denote the number of genes in pathway 1, g_2 the number of genes in pathway 2, and n the total number of samples.

Each cell of the expression data matrices X_1 and X_2 were simulated using $N(-3, 1)$ for the normal group and $N(3, 1)$ for the disease group. This simulation set-up resembles the real diabetes data set described in the next section, where there are two groups of patients with different glucose tolerance with the interest in predicting a continuous outcome of interest. We first fixed the sample size and the pathway sizes varied the values of the pathway-specific parameters, τ_1 and τ_2 , to assess the estimation accuracies under the following four set-ups.

Setup 1: $n = 100, g_1 = 50, g_2 = 50, \tau_1 = \sqrt{2} \approx 1.414, \tau_2 = \sqrt{2} \approx 1.414, \sigma^2 = 1$

Setup 2: $n = 100, g_1 = 50, g_2 = 50, \tau_1 = \sqrt{9}, \tau_2 = \sqrt{9}, \sigma^2 = 1$

Setup 3: $n = 100, g_1 = 50, g_2 = 50, \tau_1 = \sqrt{2} \approx 1.414, \tau_2 = \sqrt{9}, \sigma^2 = 1$

Setup 4: $n = 100, g_1 = 50, g_2 = 50, \tau_1 = \sqrt{9}, \tau_2 = \sqrt{30} \approx 5.477, \sigma^2 = 4$

The results are presented in Table 1. We can see that the means of the estimates of all the parameters τ_1, τ_2 and σ through the 500 simulations are very close to the true values in these setups. As expected, when the parameter values increase, the variances of the parameter estimates also increase.

We then varied the sample size and the number of genes in each pathway. The results are given in the bottom half of Table 1.

Setup 5: $n = 70, g_1 = 35, g_2 = 35, \tau_1 = \sqrt{5} \approx 2.236, \tau_2 = \sqrt{5} \approx 2.236, \sigma^2 = 1$

Setup 6: $n = 200$, $g_1 = 50$, $g_2 = 50$, $\tau_1 = \sqrt{5} \approx 2.236$, $\tau_2 = \sqrt{5} \approx 2.236$, $\sigma^2 = 1$

Setup 7: $n = 100$, $g_1 = 30$, $g_2 = 30$, $\tau_1 = \sqrt{5} \approx 2.236$, $\tau_2 = \sqrt{5} \approx 2.236$, $\sigma^2 = 1$

Additionally, as expected, the variance of the parameter estimates increases as the number of samples decreases. The variance increase is also noticeable when there are fewer genes in the pathway. Compared with one pathway, the variation of the τ estimates are similar. However, the σ estimate is tighter for the one pathway compared with the two pathway, (Table 7 in supplementary materials).

Simulations were carried out to assess the type I error and power of the score test proposed in section 2.5. As we can see from Table 2, under the true null, the score test had a type I error rate below the nominal level of 0.05. In some instances, it is on the conservative side. Table 3 summarizes the results on power and the test was able to achieve more than 75% power under the alternatives for various settings using 1000 simulations.

To address how robust the pathway effects estimates are if two pathways are not independent of each, we performed the following simulations. In Tables 5 and 6 in supplementary materials, we present the results for the setups above with one slight variation. The only difference between these tables and Table 1 is that the data generated is 50% and 30% correlated for Tables 5 and 6 in supplementary materials, respectively. We found that the bias of the estimates remains low, however, the variance of these estimates increases as the proportion of correlated genes increases. For example, for case 2, the variance for the pathway effects goes from 0.37–0.43 to 0.47–0.52 to 0.62–0.68, when the correlation increases from 0% to 30% to 50%. Similarly for other cases.

We provided additional simulation results with varying fraction of the genes in a pathway that have an effect on the outcome to provide a more realistic demonstration. Similar to the setups above, we vary the number of genes that have an effect from 10%, 30%, to 50%. The results of these simulations can be found in Figure 4 of supplementary materials. For sample size larger than 100, we see that the type I errors are all below 0.03, which is more conservative than the desired level of 0.05. For smaller sample size, such as 70, it hovers around 0.02 to 0.05. However, for sample size of 36, we can see that the Type I error is slightly inflated, but not too severely. In terms of power, as with the simulations for all genes changed, we see that the power is reasonable for all settings when $\tau_2 = 0$, see Figure 4 in supplementary materials.

Furthermore, we performed more simulations to investigate the impact on the score test when we have non-normality of trait data. Instead of simulating the expression data and pathway effects with a multivariate normal, the expression data were simulated using X_1, X_2 that are drawn from independent identically distributed $U(-1, 0.75)$ for the normal group and X_1, X_2 that are drawn from independent identically distributed $U(-0.75, 1)$ for the disease group. The outcome/trait data was then based on the principal component(s) of the individual pathways. For type I error, we simulated the trait data with just one pathway and the other pathway had zero pathway effect. The results were shown in Figure 5 of supplementary materials. For this simulation, again we see a similar pattern as before where

for sample size of 70 or above, the type I error is between 0.02 to 0.05. But for smaller sample size of 36, we see that the type I error is slightly inflated. For power, we simulated the trait data was based on the principal components of the first pathway and the second pathway. In terms of power, all of the settings showed power of 0.8 or greater, see Figure 5 in supplementary materials.

In order to understand the performance of single pathway versus joint pathway analysis, we also performed simulations with one pathway, Table 7 in supplementary materials. Compared with Table 1, we see the results are fairly similar with the exception that the estimation of σ is now more precise. In Table 8 in supplementary materials, we assess the power for the single pathway case and found that even with small sample size, a power of more than 98% can be achieved. This observation is most likely due to the fact that we have more precise estimation when we are modeling only one pathway.

Additionally, to simulate more realistic non-linear effects, we have added simulations based on the correlation structure from real microarray data [12]. In the two pathway case, Figure 6 in supplementary materials, we used the real correlation structure from two pathways, one with 49 genes and another with 36 genes. We set $\tau_1 = 2$ and $\sigma = 1$ to assess the type I error and power of the score test proposed. τ_2 is set to 2 for the power assessment and 0 for the type I error assessment. Under the non-linear effects scenarios, the score test appears to control type I error, but leaning towards the conservative side. Power remains above 70% power under the alternatives for various settings, see Figure 6 in supplementary materials.

4.2 Three pathways

Similar to the results presented in Table 1. We performed simulations for three pathways, see Table 9 in supplementary materials. All the parameters τ_1 , τ_2 , τ_3 , and σ through the 500 simulations are very close to the true values in these setups. As expected, when the parameter values increase, the variances of the parameter estimates also increase. Again, the variance of the parameter estimates increases as the number of samples decreases. In general, the variances of the estimates are higher in the three pathways than the two pathways scenarios. Similar simulations as those described in the previous paragraph was performed for the three pathway case, i.e. correlation structure from real microarray data, see Figure 7 in supplementary materials, the third pathway has 41 genes. The setup is the same, except that we have the addition of τ_3 in both the type I error and power assessments. The results are similar with the two pathway case except for the fact that we see an inflated type I error for the case of small sample size.

Finally, we presented simulation results for the score test for three pathway case. We used the non-normality trait data setup which better reflects real data. The expression data were simulated using $X_1, X_2, X_3 \sim U(-1, 0.75)$ for the normal group and $X_1, X_2, X_3 \sim U(-0.75, 1)$ for the disease group. The outcome/trait data was then based on the principal components of the individual pathways. For type I error, we simulated the trait data with just pathway one and three, and pathway two had zero pathway effect. In the three pathway case, we see that the type I error is below the nominal value of 0.05 for sample size of 100 or larger. The type I error, however, is inflated in both 36 and 70 sample sizes, see Figure 8 of supplementary materials. In terms of power, the trait data was simulated with the principal components of

all three pathways, see Figure 8 of supplementary materials. Across all settings, the power is above 85%. The type I error is inflated for $n=70$ in addition to the smallest sample size which was seen in the two-pathway case. In terms of power, it is also showing less power in the three-pathway case compared to the two-pathway simulations. Intuitively, it does make sense since having three-pathway makes the modeling a little more complex than just two with an additional pathway effect parameter involved.

We provided additional simulation results with varying fraction of the genes in a pathway that have an effect on the outcome for a more realistic demonstration. Similar to the two pathway setups above, we vary the number of genes that have an effect from 10%, 30%, to 50%. The results of these simulations can be found in Figure 9 of supplementary materials. For sample size larger than 100, we see that the type I errors are all below 0.03, which is more conservative than the desired level of 0.05. For smaller sample size, such as 70, it hovers around 0.04 to 0.06. However, for sample size of 30, we can see that the Type I error is inflated, but not too severely. In terms of power, as with the simulations for all genes changed, we see that the power is very high for all settings, except for sample size of 36, it is only around 60%, see Figure 9 in supplementary materials.

5 Application to microarray data

5.1 Data set

We analyzed a diabetes data set from [29]. They utilized the HGU-133a Affymetrix genechip with 22,283 genes to study 17 normal glucose tolerance individuals versus 18 Type II diabetes mellitus patients.

Let X_k ($n \times g_k$) be the gene expression levels for pathway k , where $n = 35$ is the sample size and g_k is the number of genes in pathway k . The outcome is the glucose level, which is in our model. This variable is centralized for analysis. The goal of our study is to identify pathways that affect the glucose level related to diabetes based on the random effects with kernel $X_k X_k^T$, i.e. to identify pathways with strong pathway effects, and then to study joint effects from pathway pairs among the these top pathways.

We considered 171 pathways including 95 KEGG pathways, 30 BioCarta pathways (<http://www.biocarta.com>) and 46 curated pathways, constructed from biological experiments performed by the Broad Institute, [29]. These pathways contained between 35 and 543 genes. The KEGG pathway database [19] is a collection of curated pathways. It mainly consists of metabolic pathways including molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases. Most of the pathways from BioCarta are related to signal transduction for human and a smaller set of metabolic pathways. The remaining pathways were curated by Mootha and colleagues from biological experiments.

5.2 Diabetes Data Set Results

5.2.1 One pathway—Our results for the one pathway case were consistent with previous findings from this diabetes data set. The results are given in Table 4 ranked by the pathway-specific effects.

Many of the pathways with high τ values are known to be related to diabetes. These include “Oxidative phosphorylation”, “ATP synthesis”, and “OXPHOS HG-U133A probes” ranked 1, 3, and 7, respectively. “Oxidative phosphorylation” was found in [29] analysis on binary phenotype of interest. Researchers have found that oxidative phosphorylation expression is coordinately decreased in human diabetic muscle. It has been hypothesized that PGC-1 α , a cold-inducible regulator of mitochondrial biogenesis, thermogenesis and skeletal muscle fiber-type switching, induces the oxidative phosphorylation pathway [24]. It is not surprising to see that “ATP synthesis” follows closely behind because it is in fact a subset of “Oxidative phosphorylation”. Similarly, “OXPHOS HG-U133A probes”, is a Broad Institute curated version of “Oxidative phosphorylation”.

“Actions of Nitric Oxide in the Heart” is ranked as one of the top pathways as it has been found by researchers that nitric oxide synthesis plays a role in reduction of glucose uptake for individuals with Type II diabetes compared with control groups [22]. JAK-Stat Signaling Pathway is also related to glucose level as the inhibition of this pathway was found to prevent glucose-induced increase in glomerular mesangial cell growth, a finding published in *Diabetes* [40].

5.2.2 Two Pathways—To investigate whether pathways affect the outcome additively we analyzed the top 10 highest τ values in the pair-pathway model. Only one pair of pathways, “c25 U133 probes” and “Oxidative phosphorylation”, with at most one overlapping gene between the two passed multiple testing correction $p < 0.0006$ with a p-value of 0.0005. For the pair pathway tests, the number of tests performed is approximately the ‘number of pathways choose two’. Each pathway of a pathway pair with at most one gene overlapping is tested using the score test. To address the multiple testing issues, we have chosen to use the conservative Bonferroni multiple testing correction strategy. The table displays two pathways, both of these pathways are significant individually. When “Oxidative phosphorylation” is added to the model, Pathway 1 is no longer significant while Pathway 2 remains significant with the p-values given in the table. It is interesting to note that “Oxidative phosphorylation” was found to be related to the patients in the original analysis of the diabetes data set [29]. Given a sample size of $n=35$, permutations are also performed to verify the significance of these score tests. We found the significant score test has permutation p-values of less than 0.01. This pathway is very interesting, as there are no overlaps between “c25 U133 probes” and “Oxidative phosphorylation”, see Figure 1 in supplementary materials. There could potentially be a crosstalk, or the two pathways are functionally related with regards to the outcome of interest; if both pathways have significant pathway effects when modeled individually, but only one pathway effect remains significant when they are modeled together. There is literature evidence that there can be potential crosstalks between these two pathways. For example, CD40, a gene in the gene set “c25 U133 probes”, is known to induce oxidative stress [9] and has been found to cause mild impairment of oxidative metabolism [20]. In addition, CD40 enhances phosphorylation of AKT [8]. In turn, AKT activates oxidative phosphorylation [38] and AKT translocation has been investigated in relation to oxidative phosphorylation in diabetic cardiac muscle [43]. Another gene REL of “c25 U133 probes” is a family consisting of Nuclear Factor kappa B, a well known oxidative stress transcription factor ([6], [2], and [30]). One source

of oxidative stress is the leakage of activated oxygen from mitochondria during oxidative phosphorylation. Another method to help tease out biological relevance is to use binding and protein-protein interaction information. In Figure 2 in supplementary materials, NDUFA6 is an enzyme which binds to both BAT3 in “c25 U133 probes” and ATP5J2 in “Oxidative phosphorylation” [39]. The above findings can help scientists identify potential drug targets and generate further biological hypothesis for testing. We have investigated whether gene expression correlation is driving the results that are found for the pair pathway tests. We discovered that there are other pathways that have higher proportion of genes that have $\rho > 0.7$ or $\rho > 0.8$, but these were not significant in our tests. Using Akaike information criterion (AIC), see supplementary materials, we compared the model fit of one pathway and two pathways analyses. In close to half of the instances, the two pathways analyses have smaller AIC than each of the one pathway analyses.

5.2.3 Three Pathways—To investigate whether pathways are associated with the outcome additively we analyzed the same set of pathways in the pair-pathway in three pathways models. Table 10 in supplementary materials displays the results for the score test for three pathways and testing whether the pathway effect of pathway 2 is equal to 0 under the null. The table displays three pathways on each row, these pathways are significant individually. Quite a few pathways meet the Bonferroni multiple testing corrected p-value of 0.0002. Among these trios, the Ubiquitin mediated proteolysis is no longer significant after the inclusion of “JAK-Stat Signaling Pathway” and 1) “ATP synthesis”, 2) “c25 U133 probes”, 3) “Oxidative phosphorylation”, 4) “OXPHOS HG-U133A”, and 5) “Actions of Nitric Oxide in the Heart”. Therefore, it seems that the pathways “Ubiquitin mediated proteolysis” and “JAK-Stat Signaling Pathway” have strong crosstalk potential, or the two pathways are functionally related with regards to the outcome of interest. These two pathways do not have any overlapping genes. However, very interestingly, the origin of this pathway from KEGG illustrates a crosstalk between these two pathways, see Figure 3 in supplementary materials. On the “JAK-Stat Signaling Pathway” diagram, the orange star indicates the linkage of this pathway with “Ubiquitin mediated proteolysis” [19].

6 Discussion

The development of pathway-based approaches is motivated by the fact that genes do not work independently. Because pathways also work together, it is important to investigate how pathways are related to each other. These pathway crosstalks can help further understand the biological mechanisms underlying diseases and facilitate the discovery of potential biomarkers. However, the existing approaches are not tailored towards the joint analysis of multiple pathways.

In this paper, we have developed a random effects model for analyzing two and more pathways with a continuous clinical outcome and derived score test statistic for significance of pathway effects. Our method helps answer the scientific question of given other pathways in the model, does a particular pathway remains significant. We illustrated an application of our method by performing pathway analysis on a diabetes microarray data. This method allows us to rank pathways according to pathway specific parameters and test the impact on one pathway after the inclusion of another pathway. By identifying potential links among

pathways, bioinformaticians and biologists can make use of the proposed methods to discover pathway and find genes that bridge between pathways. This allows researchers to obtain results that are more closely tied to the biological mechanism of diseases and to examine pathway crosstalk.

A number of issues should be investigated further. Although we have focused on the polynomial kernel, other kernels should also be explored. The properties of score tests for testing random effects of covariance matrix of the polynomial kernel deserve further investigations. Pathways have overlapping genes, and we should derive specific models for investigating pathways with shared genes. In some instances, K_1 and K_2 are close to singular. Adding a regularization term to K can help with this issue. The choice of the tuning constant can be investigated further in the future.

Possible extensions to our method include a modification to allow for the modeling of survival outcomes, such as time to death. Survival outcome has been studied by several researchers ([16], [41], [33]). In addition, our approach is suitable for modeling continuous outcome, but there are data with binary phenotype of interest. In those cases, a logistic random effects model can be used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was partially supported by National Institutes of Health (NIH) grant GM59507, CA142538, a pilot grant from the Yale Pepper Center, the National Science Foundation (NSF) grant DMS 0714817, and start-up funds from Duke University School of Medicine. We would also like to thank 'Yale University Biomedical High Performance Computing Center' NIH grant RR19895, which funded the instrumentation.

Appendix

Asymptotic distribution of score test for individual null

Regularity conditions:

1. $\Sigma_{-j}^{-1} \Sigma_{\tau_j}$ is of full rank.
2. As $n \rightarrow \infty$, the number of observations at any level of any random effect is bounded by a constant which follows from that fact that each pathway effect is a vector of size n .
3. The sequences $\{\xi_i \psi_i^{-1} y_i\}$ and $\{\xi_i\}$ are uniformly bounded $\forall i = 1, \dots, n$.
4. There exists a positive definite matrix I^0 such that $\lim_{n \rightarrow \infty} \frac{I}{n} = I^0$, which is reasonable given conditions 1 and 2.

5. For any given q by 1 constant vector η , let $\Phi_\eta = \sum_{l=1}^q \frac{1}{2} \eta \Sigma_{-j}^{-1} \Sigma_{\tau_j} \Sigma_{-j}^{-1}$. Then $y_\eta^* = \Phi_\eta y = (y_{\eta,1}^*, \dots, y_{\eta,n}^*)^T$ forms an m -dependent sequence for some constant m .
6. The usual asymptotic behavior of the maximum likelihood estimates of parameter vector $\tau_1, \dots, \tau_{j-1}, \tau_{j+1}, \dots, \tau_q$ holds, including consistency and efficiency.

Proof. Let τ_{-j}^* be the true value of τ_{-j} . First, let's proof the asymptotic normality of $n^{\frac{1}{2}} \eta U(\tau_{-j}^*)$. Let η be a any constant vector of size q . The score test in section 3.1 under H_0 can be written as:

$$\eta U_{\tau_j} = y^T \Phi y - \text{trace} \left(\sum_{l=1}^q \frac{1}{2} \Sigma_{-j}^{-1} \Sigma_{\tau_j} \right),$$

where $\Phi = \sum_{l=1}^m \frac{1}{2} \eta \Sigma_{-j}^{-1} \Sigma_{\tau_j} \Sigma_{-j}^{-1}$. Under conditions 1 and 2, the above equation can be rewritten as:

$$\eta U_{\tau_j} = \sum_{i=1}^n U_{\eta,i} = \sum_{i=1}^n (y_{\eta,i}^* - \Gamma_{\eta,i}),$$

where $y_{\eta,i}^* = \Phi_\eta y = (y_{\eta,1}^*, \dots, y_{\eta,n}^*)^T$ is a weighted sum of the Y_i s and $\Gamma_{\eta,i} = \sum_{l=1}^m \frac{1}{2} (\eta_l) \xi_i$.

Under condition 3, the sequence $\{\xi_i \psi_i^{-1} \xi_{i'} \psi_{i'}^{-1} y_i y_{i'}\}$ for $(i, i' = 1, \dots, n)$ are uniformly bounded. This implies that $U_{\eta,i}$ s are also uniformly bounded for any given η .

Under conditions 4, 5, and an application of Theorem 7.3.1 of Chung (1974), it follows that:

$$n^{-\frac{1}{2}} \eta^T U(\tau_{-j}^*) \rightarrow N(E(U_{\tau_j}), \eta^T I^0(\tau_j^*) \eta),$$

in distribution as $n \rightarrow \infty$. Using Cramer-Wald theorem, we have

$$n^{-\frac{1}{2}} U(\tau_{-j}^*) \rightarrow N(E(U_{\tau_j}), I^0(\tau_j^*))$$

Under condition 6, we have that $n^{-\frac{1}{2}} U(\hat{\tau}_{-j}) \rightarrow N(E(U_{\hat{\tau}_j}), I^0(\hat{\tau}_j))$ and it follows from Slutsky's theorem that:

$$S_{\tau_j}(\hat{v}^T) = U_{\tau_j}(\hat{v}) / (\tilde{\vartheta}_{jj}^{-1}(\hat{v}))^{\frac{1}{2}},$$

where $\tilde{\vartheta}_{jj} = \vartheta_{\tau_j \tau_j} - \vartheta_{v \tau_j}^T \vartheta_{vv}^{-1} \vartheta_{v \tau_j}$.

The above score test follows an asymptotically normal distribution with mean $E(U_{\tau_j})$ and variance 1. And $E(U_{\tau_j}) = \sum_{ii} R_{ii}^* \mu_{2i}^T + 2 \sum_{i \neq j} R_{ij}^{*T} \mu_{1i}^T \mu_{1j}^T$.

References

1. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2012. *Diabetes Care*. 2013 Mar.2013 Epub.
2. Algul H, Tando Y, Beil M, Weber C, Von Weyhern C, Schneider G, Adler G, Schmid R. Different modes of NF-kappaB/Rel activation in pancreatic lobules. *J Physiol Gastrointest Liver Physiol*. 2002; 283:G270–G281.
3. Baldi C, Cho S, Ellis R. Mutations in two independent pathways are sufficient to create hermaphroditic nematodes. *Science*. 2009; 326:1002–1005. [PubMed: 19965511]
4. Beinborn M, Worrall C, McBride E, Kopin A. A human glucagon-like peptide-1 receptor polymorphism results in reduced agonist responsiveness. *Regul Pept*. 2005; 130:1–6. [PubMed: 15975668]
5. Buse J, Hirst K. The HEALTHY study: introduction. *Int J Obes*. 2003; 33(Suppl 4):S1–S2.
6. Canty T, Boyle E Jr, Farr A, Morgan E, Verrier E, Pohlman T. Oxidative stress induces NF-kappaB nuclear translocation without degradation of IkappaBalpha. *Circulation*. 1999; 100:II361–II364. [PubMed: 10567330]
7. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services. Vol. 2011. Atlanta, GA: 2011. National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2011.
8. Chakrabarti S, Varghese S, Vitseva O, Tanriverdi K, Freedman J. D40 ligand influences platelet release of reactive oxygen intermediates. *Arterioscler Thromb Vasc Biol*. 2005; 25:2428–2434. [PubMed: 16141403]
9. Chen C, Chai H, Wang X, Jiang J, Jamaluddin M, Liao D, Zhang Y, Wang H, Bharadwaj U, Zhang S, Li M, Lin P, Yao Q. Soluble CD40 ligand induces endothelial dysfunction in human and porcine coronary artery endothelial cells. *Blood*. 2008; 112:3205–3216. [PubMed: 18658029]
10. Chung, K. A course in probability theory. 2nd ed.. New York: Academic Press; 1974.
11. Croom K, McCormack P. Liraglutide: a review of its use in type 2 diabetes mellitus. *Drugs*. 2009; 69:1985–2004. [PubMed: 19747013]
12. Dettling M. BagBoosting for tumor classification with gene expression data. *Bioinformatics*. 2004; 20:3583–3593. [PubMed: 15466910]
13. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven P, Zieve F, Marks J, Davis S, Hayward R, Warren S, Goldman S, McCarren M, Vitek M, Henderson W, Huang G. VADT Investigators. Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med*. 2009; 360:129–139. [PubMed: 19092145]
14. Gerstein H, Miller M, Byington R, Goff D Jr, Bigger J, Buse J, Cushman W, Genuth S, Ismail-Beigi F, Grimm R Jr, Probstfield J, Simons-Morton D, Friedewald W. Effects of intensive glucose lowering in type 2 diabetes. Action to Control Cardiovascular Risk in Diabetes Study Group. *N Engl J Med*. 2008; 358:2545–2559. [PubMed: 18539917]
15. Goeman J, van de Geer S, de Kort F, van Houwelingen H. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004; 20:93–99. [PubMed: 14693814]
16. Goeman J, Oosting J, Cleton-Jansen A, Anninga J, van Houwelingen H. Testing association of a pathway with survival using gene expression data. *Bioinformatics*. 2005; 21:1950–1957. [PubMed: 15657105]
17. Henderson C, Kempthorne O, Searle S, von Krosigk C. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*. 1959; 15:192–218.
18. Henderson C. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 1975; 31:432–447.
19. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004; 32:D277–D280. [PubMed: 14681412]
20. Ke Z, Calingasan N, DeGiorgio L, Volpe B, Gibson G. CD40-CD40L interactions promote neuronal death in a model of neurodegeneration due to mild impairment of oxidative metabolism. *Neurochem Int*. 2005; 47:204–215. [PubMed: 15885854]

21. Kim I, Pang H, Zhao H. Semiparametric methods for evaluating pathway effects on clinical outcomes using gene expression data. *Stat Medicine*. 2012; 10:1633–1651.
22. Kingwell B, Formosa M, Muhlmann M, Bradley S, McConell G. Nitric oxide synthase inhibition reduces glucose uptake during exercise in individuals with Type 2 diabetes more than in control subjects. *Diabetes*. 2002; 51:2572–2580. [PubMed: 12145173]
23. Lin X. Variance component testing in generalised linear models with random effects. *Biometrika*. 1997; 84:309–326.
24. Lin J, Wu H, Tarr P, Zhang C, Wu Z, Boss O, Michael L, Puigserver P, Isotani E, Olson E, Lowell B, Bassel-Duby R, Spiegelman B. Transcriptional co-activator PGC-1 alpha drives the formation of slow-twitch muscle fibres. *Nature*. 2002; 418:797–801. [PubMed: 12181572]
25. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
26. Malhotra R, Liu Z, Vincenz C, Brosius F 3rd. Hypoxia induces apoptosis via two independent pathways in Jurkat cells: differential regulation by glucose. *Am J Physiol Cell Physiol*. 2001; 281:C1596–C1603. [PubMed: 11600423]
27. Mandrup-Poulsen T. Apoptotic signal transduction pathways in diabetes. *Biochem Pharmacol*. 2003; 66:1433–1440. [PubMed: 14555218]
28. Mansmann U, Meister R. Testing differential gene expression in functional groups. Goemans global test versus an ANCOVA approach. *Methods of Information in Medicine*. 2003; 44:449–453. [PubMed: 16113772]
29. Mootha V, Lindgren C, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrle M, Laurila E, Houstis N, Daly M, Patterson N, Mesirov J, Golub T, Tamayo P, Spiegelman B, Lander E, Hirschhorn J, Altshuler D, Groop L. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 2003; 34:267–273. [PubMed: 12808457]
30. Pande V, Sharma R, Inoue J, Otsuka M, Ramos M. A molecular modeling study of inhibitors of nuclear factor kappa-B (p50)–DNA binding. *J Comput Aided Mol Des*. 2003; 17:825–836. [PubMed: 15124931]
31. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H. Pathway analysis using random forests classification and regression. *Bioinformatics*. 2006; 22:2028–2036. [PubMed: 16809386]
32. Pang H, Zhao H. Building pathway clusters from Random Forests classification using class votes. *BMC Bioinformatics*. 2008; 9:87. [PubMed: 18254968]
33. Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*. 2010; 26:250–258. [PubMed: 19933158]
34. Raz I, Hanefeld M, Xu L, Caria C, Williams-Herman D, Khatami H. Sitagliptin Study 023 Group. Efficacy and safety of the dipeptidyl peptidase-4 inhibitor sitagliptin as monotherapy in patients with type 2 diabetes mellitus. *Diabetologia*. 2006; 49:2564–2571. [PubMed: 17001471]
35. Robinson G. That BLUP is a good thing: the estimation of random effects. *Statistical Science*. 1991; 6:15–32.
36. Ryan G, Jobe L, Martin R. Pramlintide in the treatment of type 1 and type 2 diabetes mellitus. *Clin Ther*. 2010; 27:1500–1512. [PubMed: 16330288]
37. Shackelford D, Shaw R. The LKB1-AMPK pathway: metabolism and growth control in tumor suppression. *Nat. Rev. Cancer*. 2009; 9:563–575. [PubMed: 19629071]
38. Shaik Z, Fifer E, Nowak G. Akt activation improves oxidative phosphorylation in renal proximal tubular cells following nephrotoxicant injury. *Am J Physiol Renal Physiol*. 2010; 294:F423–F432. [PubMed: 18077599]
39. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005; 122:957–968. [PubMed: 16169070]

40. Wang X, Shaw S, Amiri F, Eaton D, Marrero M. Inhibition of the JAK/STAT signaling pathway prevents the high glucose-induced increase in TGF- β and fibronectin synthesis in mesangial cells. *Diabetes*. 2002; 51:3505–3509. [PubMed: 12453907]
41. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*. 2007; 8:265–284. [PubMed: 16772399]
42. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*. 2004; 27:1047–1053. [PubMed: 15111519]
43. Yang J, Yeh H, Lin K, Wang P. Insulin stimulates Akt translocation to mitochondria: implications on dysregulation of mitochondrial oxidative phosphorylation in diabetic myocardium. *J Mol Cell Cardiol*. 2009; 46:919–926. [PubMed: 19249309]
44. Zeitler P, Epstein L, Grey M, Hirst K, Kaufman F, Tamborlane W, Wilfley D. Treatment options for type 2 diabetes in adolescents and youth: a study of the comparative efficacy of metformin alone or in combination with rosiglitazone or lifestyle intervention in adolescents with type 2 diabetes. *Pediatr Diabetes*. 2007; 8:74–87. [PubMed: 17448130]
45. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*. 2003; 4:57–74. [PubMed: 12925330]
46. Zhang L, Lon S, Subramani S. Two independent pathways traffic the intraperoxisomal peroxin PpPex8p into peroxisomes. *Mol Biol Cell*. 2006; 17:690–699. [PubMed: 16319171]

Table 1

Simulations results for two pathways part 1 and part 2

	n	g_1	g_2	τ_1	$E(\hat{\tau}_1)$	$var(\hat{\tau}_1)$	$mse(\hat{\tau}_1)$	τ_2	$E(\hat{\tau}_2)$	$var(\hat{\tau}_2)$	$mse(\hat{\tau}_2)$	σ^2	$E(\hat{\sigma}^2)$	$var(\hat{\sigma}^2)$	$mse(\hat{\sigma}^2)$
1	100	50	50	$\sqrt{2}$	1.41	0.09	0.09	$\sqrt{2}$	1.41	0.10	0.10	1	0.93	0.34	0.34
2	100	50	50	$\sqrt{9}$	3.02	0.37	0.37	$\sqrt{9}$	3.09	0.43	0.43	1	0.90	0.42	0.43
3	100	50	50	$\sqrt{2}$	1.43	0.10	0.10	$\sqrt{9}$	3.01	0.37	0.37	1	0.94	0.40	0.41
4	100	50	50	$\sqrt{9}$	2.97	0.40	0.40	$\sqrt{30}$	5.41	1.33	1.33	4	4.13	0.88	0.90
5	70	35	35	$\sqrt{5}$	2.23	0.36	0.36	$\sqrt{5}$	2.22	0.37	0.37	1	0.95	0.44	0.44
6	200	50	50	$\sqrt{5}$	2.24	0.19	0.19	$\sqrt{5}$	2.24	0.21	0.21	1	1.00	0.02	0.02
7	100	30	30	$\sqrt{5}$	2.26	0.38	0.38	$\sqrt{5}$	2.21	0.32	0.32	1	0.98	0.05	0.05

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Simulations - Type I error for testing $\tau_2 = 0$

n	g_1	g_2	τ_1	τ_2	σ^2	Type I error
70	35	35	$\sqrt{2} \approx 1.414$	0	1	0.045
100	30	30	$\sqrt{5} \approx 2.236$	0	1	0.013
100	50	50	$\sqrt{2} \approx 1.414$	0	1	0.022
100	50	50	$\sqrt{5} \approx 2.236$	0	1	0.025
100	50	50	$\sqrt{9}$	0	1	0.023
100	50	50	$\sqrt{30} \approx 5.477$	0	1	0.025
200	50	50	$\sqrt{2} \approx 1.414$	0	1	0.034

Table 3

Simulations - Power for testing $\tau_2 > 0$

n	g_1	g_2	τ_1	τ_2	σ^2	Power
70	35	35	$\sqrt{2} \approx 1.414$	$\sqrt{2} \approx 1.414$	1	1
100	30	30	$\sqrt{5} \approx 2.236$	$\sqrt{5} \approx 2.236$	1	1
100	50	50	$\sqrt{2} \approx 1.414$	$\sqrt{2} \approx 1.414$	1	1
100	50	50	$\sqrt{5} \approx 2.236$	$\sqrt{5} \approx 2.236$	1	1
100	50	50	$\sqrt{5} \approx 2.236$	$\sqrt{9}$	1	1
100	50	50	$\sqrt{9}$	$\sqrt{30} \approx 5.477$	1	1
200	50	50	$\sqrt{2} \approx 1.414$	0.1	1	1
70	35	35	$\sqrt{2} \approx 1.414$	0.1	1	0.95
100	30	30	$\sqrt{5} \approx 2.236$	0.1	1	0.998
100	50	50	$\sqrt{2} \approx 1.414$	0.1	1	0.996
200	50	50	$\sqrt{2} \approx 1.414$	0.1	1	1
70	35	35	$\sqrt{2} \approx 1.414$	0.05	1	0.764
100	30	30	$\sqrt{5} \approx 2.236$	0.05	1	0.97
100	50	50	$\sqrt{2} \approx 1.414$	0.05	1	0.94
200	50	50	$\sqrt{2} \approx 1.414$	0.05	1	1

Table 4

One pathway ranked by τ values.

Rank	Num of genes	$\hat{\tau}$	$\hat{\sigma}^2$	Pathway Names
1	133	76.5618	3.0465	Oxidative phosphorylation
2	39	44.5528	4.3937	Actions of Nitric Oxide in the Heart
3	49	44.0789	4.3535	ATP synthesis
4	36	27.2932	4.4335	c25 U133 probes
5	71	26.3367	4.0365	JAK-Stat Signaling Pathway
6	40	25.576	4.5308	Parkinson's disease
7	121	25.1714	4.4349	OXPHOS HG-U133A probes
8	46	14.3665	4.8516	Ubiquitin mediated proteolysis
9	41	12.8057	4.7578	gamma-Hexachlorocyclohexane degradation
10	44	12.004	4.8305	Signaling of Hepatocyte Growth Factor Receptor