Lee et al. BMC Medical Genomics 2019, 12(Suppl 5):100
https://doi.org/10.1186/s12920-019-0517-4

BMC Medical Genomics

RESEARCH Open Access

Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis

Sungyoung Lee[1†], Sunmee Kim[2†], Yongkang Kim[3], Bermseok Oh[4], Heungsun Hwang[2] and Taesung Park[3,5*]

From The 8th Annual Translational Bioinformatics Conference
Seoul, South Korea. 31 October - 2 November 2018

Abstract

Backgrounds: Recent large-scale genetic studies often involve clustered phenotypes such as repeated measurements. Compared to a series of univariate analyses of single phenotypes, an analysis of clustered phenotypes can be useful for substantially increasing statistical power to detect more genetic associations. Moreover, for the analysis of rare variants, incorporation of biological information can boost weak effects of the rare variants.

Results: Through simulation studies, we showed that the proposed method outperforms other method currently available for pathway-level analysis of clustered phenotypes. Moreover, a real data analysis using a large-scale whole exome sequencing dataset of 995 samples with metabolic syndrome-related phenotypes successfully identified the glyoxylate and dicarboxylate metabolism pathway that could not be identified by the univariate analyses of single phenotypes and other existing method.

Conclusion: In this paper, we introduced a novel pathway-level association test by combining hierarchical structured components analysis and penalized generalized estimating equations. The proposed method analyzes all pathways in a single unified model while considering their correlations. C/C++ implementation of PHARAOH-GEE is publicly available at http://statgen.snu.ac.kr/software/pharaoh-gee/.

Keywords: Generalized estimating equations, Clustered phenotypes, Pathway analysis, Rare variants

Backgrounds

The history of Genome-Wide Association Studies (GWAS) now has reached two decades, and those GWAS have identified almost 60,000 unique associations of over 3000 traits [1]. However, despite the steeply increasing GWAS discoveries, those discoveries explain only a small portion of expected phenotypic variations [2, 3], a phenomenon known as "missing heritability" [2]. Some of the possible explanation for such phenomenon include gene-gene interaction, pleiotropic effect, and rare variants [3].

For the analysis of rare variants, the low statistical power caused by the sparseness of rare variants is one of the major issues. The use of biological information such as genes or pathways has been proven to escalate the statistical power and improve the biological interpretation, for identifying statistically significant genes and pathways associated with complex traits such as high-density lipoprotein levels, obesity, schizophrenia, and multiple cancers [4–8]. Taking the advantages of the pathway-level analysis, we have developed statistical methods PHARAOH that investigates pathway-level associations [9] and PHARAOH-multi that extends PHARAOH to the analysis of multiple continuous phenotypes [10]. Our PHARAOH method has two exclusive

* Correspondence: heungsun.hwang@mcgill.ca; tspark@stats.snu.ac.kr
Sungyoung Lee, and Sunmee Kim are the authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
[3]Department of Statistics, Seoul National University, Seoul, Korea
[5]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
Full list of author information is available at the end of the article

features. First, it employs the hierarchy of biological process by constructing a hierarchical structural model of the rare variants, genes, pathways, and phenotype(s). Second, it considers all pathways within a single unified model with statistical regularization, hence effectively controlling the correlations between genes and pathways.

Another approach to improving the statistical power is a simultaneous analysis of clustered phenotypes. For example, the analysis of repeatedly measured phenotypes outperforms the analysis of cross-sectionally observed phenotypes, since the information on the temporal differences within a subject improves the power [11]. Many recent GWAS have analyzed the repeatedly measured phenotypes and discovered many novel associations, such as fasting glucose, body mass index, and lung function [12–14]. In the repeated measures analysis, a consideration of the correlations between the repeated measurements is crucial. Neglecting the nature of clustered phenotypes may result in loss of statistical power [15].

The Generalized Estimating Equations (GEE) approach is one of the most commonly used methods for the analysis of clustered and correlated phenotypes [15]. The major advantages of GEE include that it can handle a wide class of phenotypes such as binary, count, and continuous traits from an exponential family distribution and that its estimator is consistent regardless of the specification of the working correlation structure. In these respects, the GEE approach has been contributed to the discovery of genetic components from various studies including association studies of lung cancer [16], ophthalmological measurements [16, 17], and gene-drug interaction analysis [18]. For the analysis of expression datasets, various extensions of GEE have been proposed such as the repeated microarray experiment and penalized GEE for microRNA dataset [17, 18]. For gene-level tests, several GEE methods have been developed, including Longitudinal Genetic Random Field (LGRF) and GEE-KM [19, 20].

However, unlike the gene-level analyses, to the best of our knowledge, only one method based on GEE has employed the pathway-level analysis of the correlated phenotypes [21] with the R package GEEaSPU. Note that GEEaSPU employs the adaptive Sum of Powered score (aSPU) and adapts the GEE framework to enable pathway-level analysis of genetic variants [21]. However, the GEEaSPU method cannot handle the correlations between the pathways, which can result in the biased results.

In order to address this problem, we propose a novel pathway-level association test for clustered and correlated phenotypes such as repeated measurements, Pathway-based approach using HierArchical component of collapsed RAre variants Of High-throughput sequencing data using Generalized Estimating Equations (PHARAOH-GEE). While the existing GEE based pathway-level method GEEaSPU implements the individual "pathway-wise" test assuming all tests are independent, the proposed PHARAOH-GEE method implements a "global test" that considers the correlation among the pathways into account by putting all pathways simultaneously into a single model. Moreover, PHARAOH-GEE can handle various types of phenotypes (e.g., binary), and it also retains the advantages of PHARAOH, such as the hierarchical model that mimics the natural biological processes. By providing PHARAOH-GEE program using a powerful and fast C/C++ based framework WISARD [22], it supports various genetic data formats and provides affordable performance.

Results

We used a workstation system consists of two Intel Xeon E5–2640 CPUs and 256GiB of RAM. Due to the limitation of the compared method, the R version 3.4.0 and R package 'GEEaSPU' were used with default settings.

Simulation study

For our simulation study, we generated 300 replicates from the simulated data pool. Each replicate consisted of 10 pathways in which the first pathway was causal and the other nine were non-causal (i.e., no effect). For each replicate, the proposed PHARAOH-GEE method was applied to the 10 pathways simultaneously, whereas GEEaSPU was applied to each pathway individually. Here we assumed that the first pathway is causal and the others are non-causal. For the causal pathway, we considered three different parameter settings: four gene-level effects (w = 0.1, 0.2, 0.5 and 1.0), three pathway-level effects (β =0.15, 0.2 and 0.25), two correlations of phenotypes (ρ =0.25 and 0.5). For all test results, we applied the BH step-up procedure to control the False Discovery Rate (FDR) at 5% level [23]. Details on simulation procedure can be found on Methods section.

First, we evaluated the type 1 errors of PHARAOH-GEE and GEEaSPU. For the given parameter settings for the causal pathway, we evaluated the type 1 errors using 9 non-causal pathways with significance level $\alpha = 0.01$. As shown in Fig. 1, all methods controlled the type 1 error rates appropriately, regardless of the parameter values.

Second, we evaluated statistical power of the methods where power was computed as a proportion of the causal pathway being statistically significant at the FDR < 0.05 over 300 replicates. In addition to three parameter settings for the causal pathway, we consider two cases when the numbers of significant genes within the causal pathway are only one ($H_1 = 1$) and two ($H_1 = 2$) out of ten simulated genes, respectively. As shown in Fig. 2, PHARAOH-GEE outperforms GEEaSPU in all simulation scenarios.

In the power analysis, there were two additional interesting findings. First, when the proportion of significant genes in the causal pathway became smaller, the proposed method

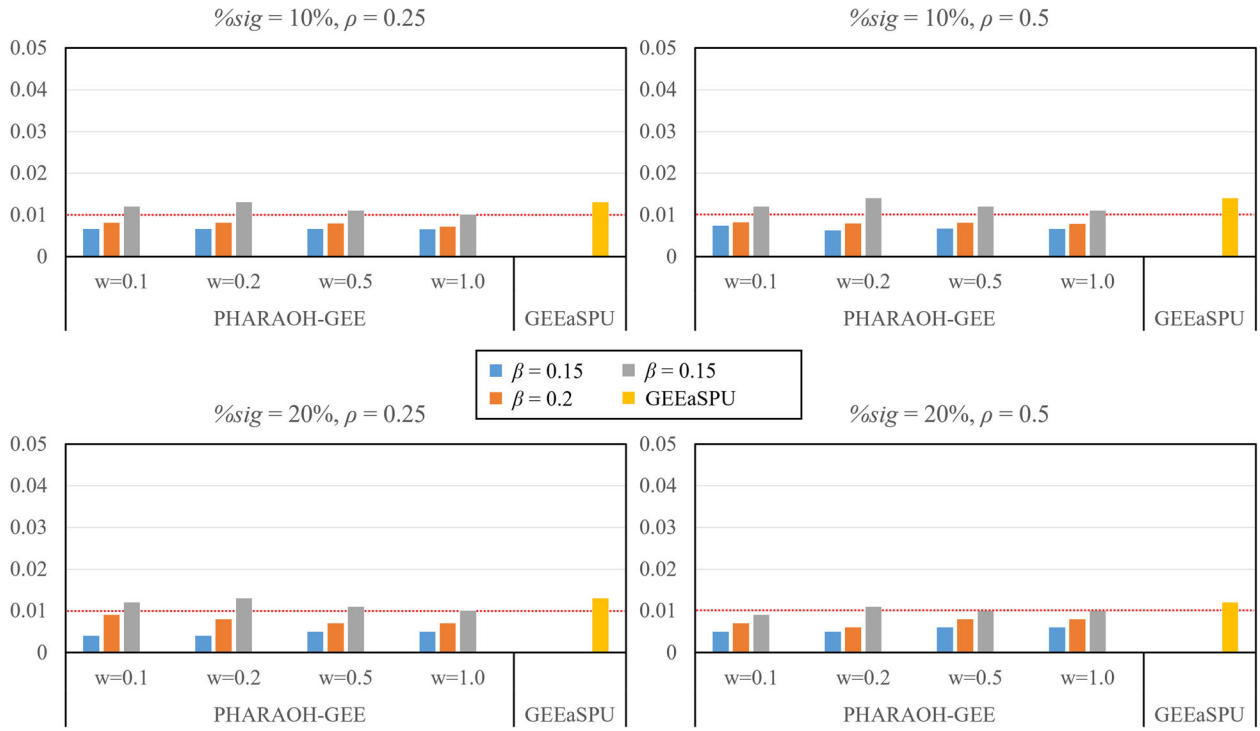

Fig. 1 Results of type 1 error simulation. Rows represent the proportions of significant genes within the causal pathway (10 and 20%), and columns represent different phenotypic correlation (0.25 and 0.5). For each plot, type 1 errors of PHARAOH-GEE are shown with varying gene-level effects (0.1, 0.2, 0.5 and 1.0) and pathway-level effects (0.15, 0.2 and 0.25), and type 1 errors of GEEaSPU are shown with orange bars

tended to outperform GEEaSPU. Second, PHARAOH-GEE showed less reduction of statistical power than GEEaSPU when the phenotypic correlation ρ increased. In real practical situation where only a fraction of genes is likely related to phenotypes and that the correlations among clustered phenotypes are high, these findings suggest that PHARAOH-GEE would be more powerful for detecting true biological signals than GEEaSPU.

Analysis of whole exome sequencing (WES) dataset using clustered phenotypes

To demonstrate the usefulness of PHARAOH-GEE, we analyzed a large-scale sequencing dataset with six phenotypes related to the metabolic syndrome: systolic blood pressure (SBP), diastolic blood pressure (DBP), triglycerides (TG), fasting glucose (FASTGLU), waist circumference (WAIST), and high-density lipoprotein (HDL). Before the analysis, we binarized these phenotypes according to the metabolic syndrome criteria of International Diabetes Federation (IDF) consensus worldwide definition of the metabolic syndrome (https://www.idf.org). Metabolic syndrome is diagnosed as the presence of three or more of the following criteria: (1) WAIST ≥90 cm in males and ≥ 80 cm in females; (2) elevated TG ≥ 150 mg/dL or taking medication; (3) HDL-cholesterol < 40 mg/dL in males and < 50 mg/dL in females or taking lipid-lowering agents; (4) systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 85 mmHg or taking antihypertensive medications; and (5) elevated FASTGLU ≥100 mg/dL or oral hypoglycemic agents use. From these six metabolic syndrome related phenotypes, we derived five clustered binary traits. Especially, we combined two blood pressure phenotypes (SBP & DBP) into a single phenotype, named BP, by setting 1 if either SBP or DBP satisfied the diagnosis criteria of metabolic syndrome and 0 otherwise. All other phenotypes were binarized if the diagnosis criteria of metabolic syndrome was satisfied and 0 otherwise.

We applied PHARAOH for the univariate analysis of each binary phenotype and applied PHARAOH-GEE and GEEaSPU for the multivariate analysis of the five binary phenotypes. We conducted the multiple testing adjustment to both univariate and multivariate analyses by using the BH step-up procedure [23]. The unstructured covariance structure of the phenotypes was assumed for both PHARAOH-GEE and GEEaSPU. Figure 3 presents quantile-quantile (Q-Q) plots showing that PHARAOH and PHARAOH-GEE led to no substantial deflation or inflation of p-values.

Table 1 exhibits the pathways with the five smallest q-values identified by PHARAOH-GEE, as well as their q-values under PHAROH and GEEaSPU. PHARAOH-GEE was able to identify one KEGG pathway, the glyoxylate and

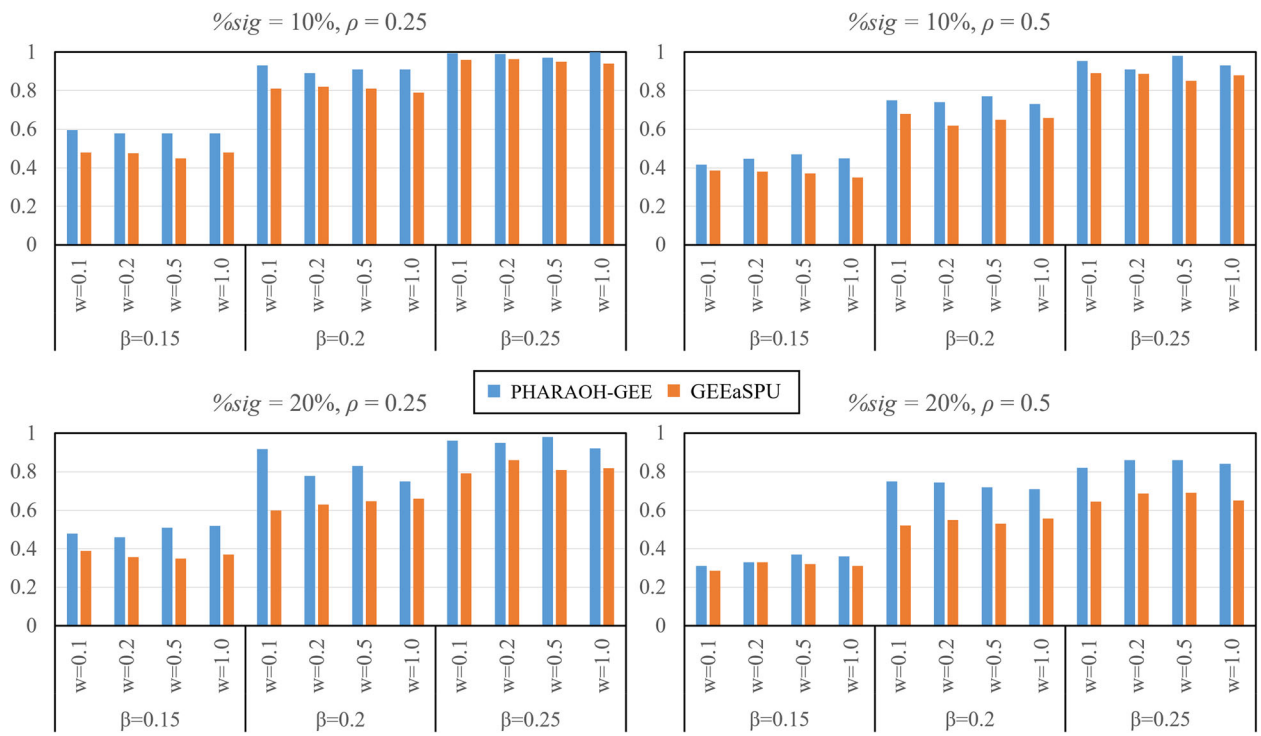

Fig. 2 Result of power analysis. Columns and rows represent different phenotypic correlations (0.25 and 0.5) and proportions of significant genes within the causal pathway (10 and 20%). For each plot, estimated statistical powers from 300 simulation datasets are shown with combinations of gene-level effects (0.1, 0.2, 0.5 and 1) and pathway-level effects (0.15, 0.2 and 0.25)

dicarboxylate metabolism, at the *q*-value threshold of 0.1. None of these pathways turned out to be statistically significant in the univariate analyses of PHARAOH, always resulting in larger *q*-values than those from PHARAOH-GEE. Although the same glyoxylate and dicarboxylate pathway had the lowest *p*-value by GEEaSPU, it failed to pass the *q*-value threshold of 0.1, after the multiple testing adjustment. Thus, our real data analyses showed the relatively superior performance of PHARAOH-GEE.

Among the five pathways identified by PHARAOH-GEE, a recent study suggests a strong relationship between the metabolic syndrome and two pathways (glyoxylate and dicarboxylate, and fatty acid metabolisms), through their role in abdominal obesity [24]. In addition, the glycosphingolipid biosynthesis and MAPK signaling pathways are reported to be related to the metabolic syndrome via insulin resistance that plays a critical role in manifestation of the metabolic syndrome [25, 26].

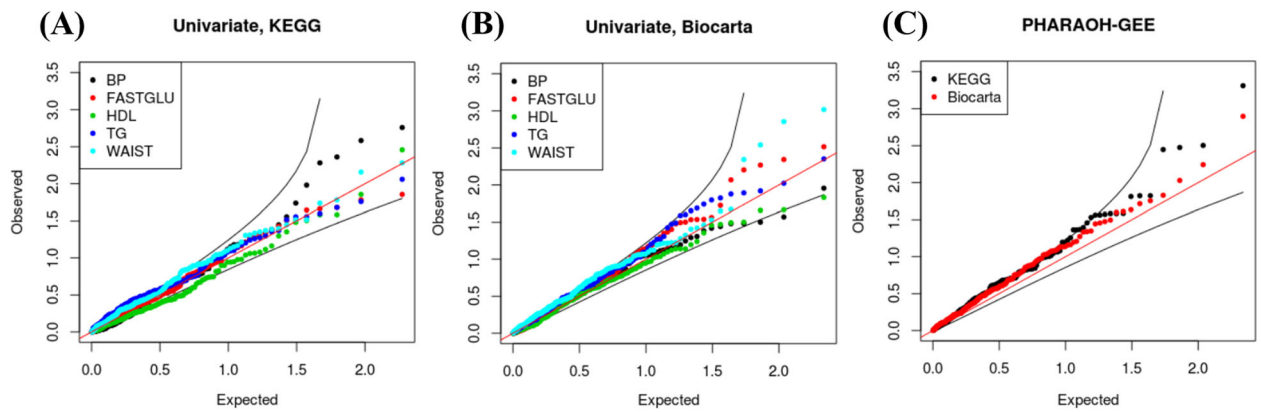

Fig. 3 Q-Q plots of the real data analyses. **a** Q-Q plot of the univariate analyses using KEGG pathway database and **b** Biocarta pathway database. **c** Q-Q plot of the analysis of the five phenotypes using PHARAOH-GEE

Table 1 Top five pathways from PHARAOH-GEE. The *q*-values after the multiple testing adjustment are presented in each cell, with their corresponding *p*-values within the brackets. The results of univariate PHARAOH are also provided on the right side of the table

Pathway	PHARAOH-GEE	GEEaSPU	Univariate PHARAOH				
			HDL	TG	FASTGLU	WAIST	BP
Glyoxylate and dicarboxylate metabolism	0.0929 (0.00063)	0.16 (0.00099)	0.987 (0.902)	0.721 (0.021)	0.772 (0.023)	0.91 (0.842)	0.916 (0.202)
Glycosphingolipid biosynthesis ganglio series	0.159 (0.0038)	0.979 (0.804)	0.987 (0.79)	0.805 (0.658)	0.855 (0.137)	0.805 (0.359)	0.695 (0.067)
MAPK signaling pathway	0.159 (0.00404)	0.468 (0.126)	0.987 (0.327)	0.721 (0.234)	0.855 (0.072)	0.953 (0.901)	0.997 (0.45)
Valine-leucine and isoleucine biosynthesis	0.159 (0.0043)	0.979 (0.797)	0.987 (0.242)	0.871 (0.779)	0.999 (0.801)	0.91 (0.813)	0.695 (0.067)
Fatty acid metabolism	0.436 (0.0173)	0.977 (0.459)	0.987 (0.834)	0.721 (0.143)	0.999 (0.893)	0.903 (0.647)	0.997 (0.909)

Conclusion

An analysis of the clustered phenotypes provides more information than the cross-sectional studies. Recent large cohort studies keep producing repeatedly measured phenotypes. We introduced a novel statistical method for the pathway analysis of the large-scale genetic dataset with clustered phenotypes. While our previous PHARAOH-multi method can handle only continuous phenotypes, the proposed PHARAOH-GEE can handle various phenotypes such as clustered binary and count phenotypes under the various correlation structures. Through the comparison study using the simulated datasets, we demonstrated that the proposed PHARAOH-GEE method outperforms an existing pathway method. Furthermore, our application to the large-scale WES dataset successfully identified one pathway that has not been discovered in the analyses of individual phenotype with the multiple testing adjustments.

Discussion

Compared to GEEaSPU the only currently available method for pathway-level test of clustered phenotypes, the proposed method has many advantages. First, PHARAOH-GEE effectively controls the complex correlations among the pathways by constructing a unified hierarchical, doubly-penalized statistical model. Second, it successfully reflects the nature of biological process from GSCA framework and takes clustered phenotypes into account from GEE framework. In conclusion, we hope that PHARAOH-GEE can serve as a main tool for the pathway-level analysis of clustered phenotypes in genetic studies.

Currently, we have a number of considerations for our future research. Although we considered many possible combinations of parameters in the simulation setting, a further extensive simulation study is required for more comprehensive comparison with existing pathway-based methods. In addition, we will perform a replication study using other independent datasets with the metabolic syndrome phenotypes. Finally, we will employ other penalization methods such as lasso and elastic-net.

Methods

PHARAOH-GEE method

Technically, the proposed method is an extension of the doubly-regularized Generalized Structured Component Analysis into the GEE framework [27] that imposes ridge penalties [28] on both gene-pathway and pathway-phenotype relationships. From the previous studies, we successfully demonstrated that those two ridge penalties effectively control the correlations between genes and pathways [9, 10]. PHARAOH-GEE aims to identify associations between Q clustered phenotypes and K pathways, each of which is linked to T_k genes ($k = 1, \cdots, K$). An example of the PHARAOH-GEE model is depicted in Fig. 4.

Let y_{iq} be the value of the q^{th} phenotype measured on the i^{th} individual ($i = 1, \ldots, N$; $q = 1, \ldots, Q$) and $\tilde{y}_i = [y_{i1}, \cdots, y_{iQ}]'$ be a $Q \times 1$ vector of the clustered phenotypes of the i^{th} individual. Similar to the previous description of the PHARAOH model [9], we assume that y_{iq} follows an exponential family distribution with a mean μ_{iq}. Let Σ_i be the $Q \times Q$ covariance matrix of $\tilde{y}_i$. Then,

$$\operatorname{cov}(\tilde{y}_i) = \Sigma_{i\,(Q\times Q)} = A_i^{1/2} R_i(\alpha) A_i^{1/2}, \tag{1}$$

where $R_i(\alpha)$ is a so-called "working correlation matrix", α is a parameter vector that fully characterizes $R_i(\alpha)$, and $A_i^{1/2} = \operatorname{diag}[\operatorname{var}(\mu_{ij})]$, i.e., a $Q \times Q$ diagonal matrix with the marginal variance of responses. Liang and Zeger [29] suggested various choices for $R_i(\alpha)$, e.g., the independence covariance structure, $R_i(\alpha) = \mathbf{I}_Q$, where $\mathbf{I}_Q$ is the identity matrix of order Q.

Let $\tilde{x}_i' = [1, \cdots, 1; x_{i11}, \cdots, x_{i1T_1}; \cdots; x_{iK1}, \cdots, x_{iKT_K}]$ be a $(T+1) \times 1$ vector consisting of all gene-level collapsed variables for the i^{th} individual across K pathways, where $T = \Sigma_{k=1}^{K} T_k$. The gene-level collapsed variables are generated as the weighted sums of rare variants. Let $\mathbf{X}$ be an $N \times (T+1)$ matrix of the gene-level collapsed variables for N observations, as expressed in (2).

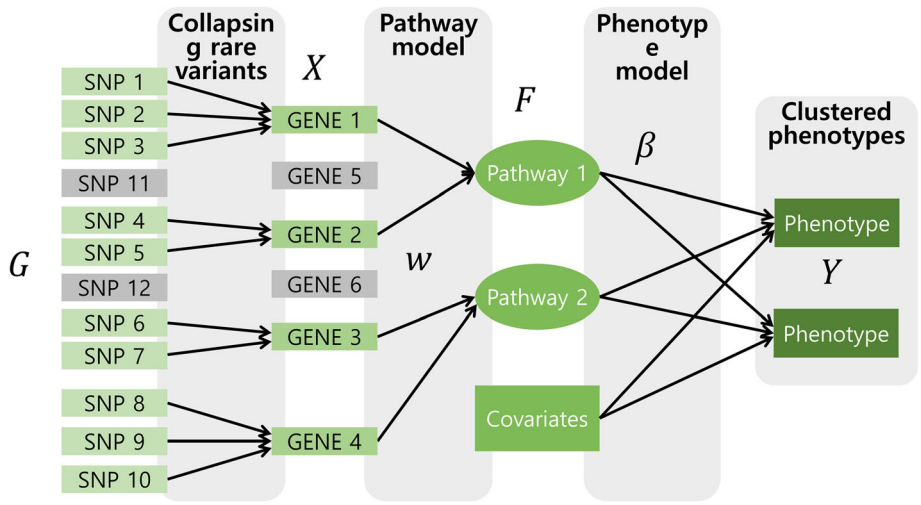

Fig. 4 An example of the PHARAOH-GEE model

$$\mathbf{X}_{N\times(T+1)} = \begin{bmatrix} 1 & x_{111} & x_{112} & \cdots & x_{1KT_K} \\ 1 & x_{211} & x_{212} & \cdots & x_{2KT_K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N11} & x_{N12} & \cdots & x_{NKT_K} \end{bmatrix} = \begin{bmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \vdots \\ \tilde{x}_N' \end{bmatrix}. \quad (2)$$

As in the previous methods [9], we standardize $\mathbf{X}$ to satisfy the conventional scaling constraint $\mathrm{diag}(\mathbf{X}'\mathbf{X}) = N\mathbf{I}$. Each element of $\mathbf{X}$, x_{ikt}, denotes a gene-level summary of the i^{th} sample for the t^{th} gene ($t = 1, \cdots, T_k$) in the k^{th} pathway and is generated by the weighted sum of rare variants that is same as the previous work [9, 10]. Let W denote a $(T+1)\times(K+1)$ matrix consisting of component weights w_{tk}, which are assigned to x_{ikt}. This matrix can be generally expressed as

$$\boldsymbol{W}_{(T+1)\times(\mathrm{K}+1)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & w_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & w_{1T_1} & 0 & \cdots & 0 \\ 0 & 0 & w_{21} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & w_{2T_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & w_{K1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_{KT_K} \end{bmatrix}. \quad (3)$$

Let η_{iq} and $g(\cdot)$ denote the i^{th} linear predictors of the q^{th} phenotype and a link function, respectively. We define the proposed PHARAOH-GEE model as

$$g(\mu_{iq}) = \eta_{iq} = \beta_{0q} + \sum_{k=1}^{K}\left(\sum_{t=1}^{T_k} x_{ikt}w_{tk}\right)\beta_{kq} = \beta_{0q} + \sum_{k=1}^{K} f_{ik}\beta_{kq} = \tilde{\boldsymbol{f}}_i\tilde{\boldsymbol{\beta}}_q, \quad (4)$$

where $f_{ik} = \sum_{t=1}^{T_k} x_{ikt}w_{tk}$ is the component score of the i^{th} individual for the k^{th} pathway $\tilde{\boldsymbol{f}}_i = [1, f_{i1}, \cdots, f_{iK}]$, and $\tilde{\boldsymbol{\beta}}_q = [\beta_{0q}\,\beta_{1q}\cdots\beta_{Kq}]$ is a vector of coefficients linking K pathways to the q^{th} phenotype. We can statistically examine the joint effects of the k^{th} pathway on Q phenotypes by testing the null hypothesis H_0: $\beta_{k1} = \ldots = \beta_{kQ} = 0$. Moreover, it is possible to evaluate the effect of one gene on a single phenotype mediated by its corresponding pathway.

Parameter estimation

For simplicity, we describe the propose method, assuming that the phenotype $\tilde{y}_i$ is continuous. It is technically straightforward to extend the method to other phenotypes from exponential distributions. In parameter estimation, we add two L_2 penalty terms to control for potential adverse influences of high correlations between genes and/or pathways. Specifically, to estimate the parameters $\boldsymbol{W}$ and $\boldsymbol{B}$, we seek to minimize the following penalized estimating equations.

$$\phi_{\alpha,B,W} = \sum_{i=1}^{N} U_i + \lambda_P tr(\boldsymbol{B}'\boldsymbol{B}) + \lambda_G tr(\boldsymbol{W}'\boldsymbol{W}), \quad (5)$$

where U is the estimating equation for the parameters, $\boldsymbol{B}$ is a matrix consisting of all regression coefficients $\tilde{\beta}_q$, $tr(\cdot)$ denotes the trace of matrix, and λ_G and λ_P denote ridge parameters on the L_2 penalty terms for the weights

and regression coefficients, respectively. A more detail on the estimating equation and solving process can be found on elsewhere [9].

To minimize $\phi_{\alpha,\ B,\ W}$, we use an iterative algorithm that repeats the following steps until no substantial changes in parameter estimates occur.

Step 1: We update $\boldsymbol{B}$ for fixed $\boldsymbol{W}$ and $R_i(\alpha)$. Let $\boldsymbol{b} = \mathrm{vec}(\boldsymbol{B})$ denote a vector formed by stacking all columns of B one below another. This is equivalent to minimizing the following estimating equations

$$\begin{aligned}\phi_1 &= \sum_{i=1}^{N} U(\boldsymbol{b}) + \lambda_P \boldsymbol{b}'\boldsymbol{b} \\ &= \sum_{i=1}^{N} \left(\boldsymbol{f}_i' \otimes \mathbf{I}\right)\Sigma_i^{-1}\left(y_i - \left(\boldsymbol{f}_i' \otimes \mathbf{I}\right)\boldsymbol{b}\right) + \lambda_P \boldsymbol{b}'\boldsymbol{b} \\ &= \sum_{i=1}^{N} \boldsymbol{Q}_i\Sigma_i^{-1}(y_i - \boldsymbol{Q}_i\boldsymbol{b}) + \lambda_P \boldsymbol{b}'\boldsymbol{b}, \end{aligned} \tag{6}$$

where $\boldsymbol{Q}_i = \boldsymbol{f}_i' \otimes \mathbf{I}$ and $\otimes$ denotes Kronecker product. Then, $\boldsymbol{b}$ can be estimated by $\hat{\boldsymbol{b}} = \left(\sum_{i=1}^{N} \boldsymbol{Q}_i'\Sigma_i^{-1}\boldsymbol{Q}_i + \lambda_P \mathrm{I}\right)^{-1}\left(\sum_{i=1}^{N} \boldsymbol{Q}_i'\Sigma_i^{-1} y_i\right)$, and $\hat{\boldsymbol{B}}$ is reconstructed from $\hat{\boldsymbol{b}}$.

Step 2: We update $\boldsymbol{W}$ for fixed $\boldsymbol{B}$ and $R_i(\alpha)$. Let $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{W})$. Similar to step 1, it is equivalent to minimizing

$$\begin{aligned}\phi_2 &= \sum_{i=1}^{N} U(\boldsymbol{w}) + \lambda_G \boldsymbol{w}'\boldsymbol{w} \\ &= \sum_{i=1}^{N} \left(\tilde{x}_i' \otimes B'\right)'\Sigma_i^{-1}\left(y_i - \left(\tilde{x}_i' \otimes B'\right)w\right) + \lambda_G w'w \\ &= \sum_{i=1}^{N} M_i'\Sigma_i^{-1}(y_i - M_i w_*) + \lambda_G w_*'w_*, \end{aligned} \tag{7}$$

where $M_i = \tilde{x}_i' \otimes \boldsymbol{B}'$, $\boldsymbol{w}^*$ is the vector formed by eliminating all zero elements of $\boldsymbol{w}$, and M_i is the matrix formed by removing the columns of $\tilde{x}_i' \otimes B'$ corresponding to the zero elements of w. Then, w_* can be estimated by $\hat{\boldsymbol{w}}_* = \left(\sum_{i=1}^{N} M_i'\Sigma_i^{-1}M_i + \lambda_G \mathrm{I}\right)^{-1}\left(\sum_{i=1}^{N} M_i'\Sigma_i^{-1} z_i\right)$. Then, the estimated $\boldsymbol{W}$ is reconstructed from $\hat{\boldsymbol{w}}_*$.

Step 3: We update $R_i(\alpha)$ from the updated $\boldsymbol{B}$ and $\boldsymbol{W}$ using Pearson residuals with the variance function of the distribution v,

$$r_{ij} = \left(y_{ij} - \hat{\mu}_{ij}\right)/v^{1/2}\left(\hat{\mu}_{ij}\right). \tag{8}$$

where $\hat{\mu}_{ij} = \beta_{0q} + \sum_{k=1}^{K} f_{ik}\hat{\beta}_{kq}$. Finally, the dispersion parameter ϕ is estimated consistently by

$$\hat{\phi} = \left(NQ - \left(K + \sum_{k=1}^{K} T_k\right)\right)^{-1} \sum_{i=1}^{N}\sum_{j=1}^{Q} \hat{r}_{ij}^2. \tag{9}$$

We apply k-fold cross-validation (CV) to estimate the values of λ_G and λ_P, which compares the quasi-deviance values [30] of a two-dimensional grid of candidate values of λ_G and λ_P.

Significance testing and multiple correction

Resampling methods can be used to test the statistical significance of the estimated effects of all pathways on a given set of clustered phenotypes. In the proposed method, we utilize a permutation test to obtain p-values. By permuting the phenotypes, the method first generates the empirical null distributions of both pathways and gene coefficients. By computing the quantile of the estimated pathway and gene coefficients from the non-permuted dataset with the corresponding null distribution, we can obtain an empirical p-value for any specific pathway and gene.

In our study, we want to test the joint effects of pathways on clustered phenotypes. In our previous study, we introduced two approaches to test β_{k1}, ..., β_{kQ} simultaneously and suggested the Wald-type statistics [10]. Similarly, we construct a single statistic that combines all Q coefficients. Here, we define a Wald-type statistic T as.

$$T = \tilde{\boldsymbol{\beta}}_k' \ \mathrm{cov}^{-1}\left(\tilde{\boldsymbol{\beta}}_k\right)\tilde{\boldsymbol{\beta}}_k. \tag{10}$$

Under penalized GEE, the estimated covariance $\mathrm{cov}(\widehat{\tilde{\beta}_k})$ can be obtained in two ways. One way is to calculate it directly, as introduced by Wang et al. [31] as follows.

$$\mathrm{cov}\left(\widehat{\tilde{\boldsymbol{\beta}}}\right) = \left(H_{\widehat{\tilde{\beta}}} + nE_{\widehat{\tilde{\beta}}}\right)^{-1} M_{\widehat{\tilde{\beta}}} \left(H_{\widehat{\tilde{\beta}}} + nE_{\widehat{\tilde{\beta}}}\right)^{-1}, \tag{11}$$

where $H_{\widehat{\tilde{\beta}}} = \sum_{i=1}^{N} \tilde{x}_i' A_i^{1/2} R_i^{-1}(\alpha) A_i^{1/2} \tilde{x}_i$, $E_{\widehat{\tilde{\beta}}} = tr(B'B)$, and $M_{\widehat{\tilde{\beta}}} = \sum_{i=1}^{N} \tilde{x}_i' A_i^{1/2} R_i^{-1}(\alpha) e_{\widehat{\tilde{\beta}}} e_{\widehat{\tilde{\beta}}}' R_i^{-1}(\alpha) A_i^{1/2} \tilde{x}_i$ with $e_{\widehat{\tilde{\beta}}} = A_i^{1/2}(\tilde{y}_i - \tilde{\mu}_i)$. The other indirect way is to calculate it as the sample covariance of $\tilde{\boldsymbol{\beta}}_k$ from permutations. We use this indirect way to reduce computational burden.

For the calculated p-values, we implemented two types of multiple testing procedure as we discussed earlier [10]. In short, we applied two approaches: Westfall & Young permutation procedure [32] that effectively considers the correlation of p-values, and the Benjamini-Hochberg (BH) step-up procedure [23] that computes q-values by False Discovery Rate (FDR) adjustment.

Simulation study

We conducted a simple simulation study to investigate the performance of PHARAOH-GEE and to compare the proposed method with the existing methods. We first simulated a large pool of rare genetic variants using SimRare [33]. All simulation settings were unchanged except for the 1Kbp of gene length. From the pool, one thousands of replicates were generated, each of those consists of 1000 individuals and 10 pathways. Finally,

the phenotypes were simulated from the below model that assumes only the first pathway is causal:

$$g\left(\mu_{iq}\right) = \eta_{iq} = \beta_{1q}\tilde{f}_{i1} = \beta_{1q}\sum_{t=1}^{H_1} w_{1t}x_{i1t} = \beta_{1q}\sum_{t=1}^{H_1}\left(w_{1t}\sum_{j=1}^{M_{1t}}\gamma_{1tj}g_{i1tj}\right), \quad (12)$$

where H_1 and M_{1t} denote the number of causal genes in the causal pathway and the number of causal rare variants in the t^{th} causal gene, respectively. Note that M_{1t} was the number of rare variants in the simulated gene varies and was used as an input variable in our simulation study. We set γ_{1tj} to $|\log_{10}MAF_{tj}|$, which represents the effect of the j^{th} genetic variant of the t^{th} gene. For the simplicity, we generated the phenotypes from the simulated linear predictor η_{iq}, by using it as a binarization threshold from the randomly generated variables from the multivariate normal distribution MVN$(0, \Sigma)$. For each replicate, all rare variants were collapsed into genes.

Exome sequencing dataset with clustered phenotypes

In order to illustrate PHARAOH-GEE for investigating associations between multiple pathways and the clustered phenotypes, we analyzed a large-scale WES dataset from a Korean population cohort. Our WES dataset consists of next-generation sequencing data of 1087 individuals' genomes, using the Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA), as a part of the T2D-GENES consortium [34]. All individuals of the dataset were originated from a large Korean cohort named the Korean Association REsource (KARE) study [35]. For our analysis, we selected six phenotypes related to the metabolic disease: SBP, DBP, TG, FASTGLU, WAIST and HDL. In our analysis, we considered 995 individuals with complete phenotypes of interest. We then applied two pathway databases Biocarta and KEGG from Molecular Signatures Database [36], which is a curated collection of multiple pathway databases.

Abbreviations

BH: Benjamini-Hochberg; CV: Cross-validation; DBP: Diastolic blood pressure; FASTGLU: Fasting glucose; FDR: False Discovery Rate; GEE: Generalized estimating equations; GWAS: Genome-wide association studies; HDL: High-density lipoprotein; IDF: International Diabetes Federation; KARE: Korean Association REsource; SBP: Systolic blood pressure; TG: Triglycerides; WAIST: Waist circumference; WES: Whole exome sequencing

Acknowledgements

Not applicable.

Funding

Publication costs are funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant (HI16C2037). Also, this work was supported by the Bio & Medical Technology Development Program of the National Research Foundation of Korea (NRF) grant (2013M3A9C4078158) and by grants of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C2037, HI15C2165, HI16C2048).

Availability of data and materials

We provide PHARAOH-GEE method as a program from the website (http://statgen.snu.ac.kr/software/pharaoh-gee). The KARE exome sequencing dataset is a part of T2D-GENES consortium, and is available upon approval of T2D-GENES project committee.

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 5, 2019: Selected articles from the 8th Translational Bioinformatics Conference: Medical Genomics*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-5.

Authors' contributions

SL and SK performed all analyses and developed the software implementation. SL, SK and TP conducted the entire study, developed the methodology, and wrote the manuscript. YK and BO helped with the performing of analyses. HH helped developing the methodology. All of the authors have read and approved of the final manuscript.

Ethics approval and consent to participate

We used the exome sequencing data of 1,037 samples from KARE. KARE study is a part of Korean Genome Epidemiology Study (KoGES), and the dataset was used under the partnership of T2D-GENES. All participants of KARE study provided written informed consent. The study using KARE samples was approved by two independent institutional review boards at Seoul National University.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Center for Precision Medicine, Seoul National University Hospital, Seoul, Korea. [2]Department of Psychology, McGill University, Montreal, Canada. [3]Department of Statistics, Seoul National University, Seoul, Korea. [4]Department of Biochemistry and Molecular Biology, School of Medicine, Kyung Hee University, Seoul, Korea. [5]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea.

Published: 11 July 2019

References

1. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res. 2017;45(D1):D896–901.
2. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008;456(7218):18–21.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
4. Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, et al. Medical sequencing at the extremes of human body mass. Am J Hum Genet. 2007;80(4):779–91.
5. Brunham LR, Singaraja RR, Hayden MR. Variations on a gene: rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis. Annu Rev Nutr. 2006;26:105–29.
6. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004;305(5685):869–72.

7. Slatter TL, Jones GT, Williams MJ, van Rij AM, McCormick SP. Novel rare mutations and promoter haplotypes in ABCA1 contribute to low-HDL-C levels. Clin Genet. 2008;73(2):179–84.
8. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008;320(5875):539–43.
9. Lee S, Choi S, Kim YJ, Kim BJ, T2D-GENES Consortium, Hwang H, Park T. Pathway-based approach using hierarchical components of collapsed rare variants. Bioinformatics. 2016;32(17):i586–94.
10. Lee S, Kim Y, Choi S, Hwang H, Park T. Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes. BMC Bioinformatics. 2018;19(Suppl 4:79.
11. Landerman LR, Mustillo SA, Land KC. Modeling repeated measures of dichotomous data: testing whether the within-person trajectory of change varies across levels of between-person factors. Soc Sci Res. 2011;40(5):1456–64.
12. Rasmussen-Torvik LJ, Alonso A, Li M, Kao W, Kottgen A, Yan Y, Couper D, Boerwinkle E, Bielinski SJ, Pankow JS. Impact of repeated measures and sample selection on genome-wide association studies of fasting glucose. Genet Epidemiol. 2010;34(7):665–73.
13. Mei H, Chen W, Jiang F, He J, Srinivasan S, Smith EN, Schork N, Murray S, Berenson GS. Longitudinal replication studies of GWAS risk SNPs influencing body mass index over the course of childhood and adulthood. PLoS One. 2012;7(2):e31470.
14. Tang W, Kowgier M, Loth DW, Soler Artigas M, Joubert BR, Hodge E, Gharib SA, Smith AV, Ruczinski I, Gudnason V, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. PLoS One. 2014;9(7):e100776.
15. Mukherjee B, Ko YA, Vanderweele T, Roy A, Park SK, Chen J. Principal interactions analysis for repeated measures data: application to gene-gene and gene-environment interactions. Stat Med. 2012;31(22):2531–51.
16. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. Am J Hum Genet. 2013;92(5):744–59.
17. Fan Q, Teo YY, Saw SM. Application of advanced statistics in ophthalmology. Invest Ophthalmol Vis Sci. 2011;52(9):6059–65.
18. Sitlani CM, Rice KM, Lumley T, McKnight B, Cupples LA, Avery CL, Noordam R, Stricker BH, Whitsel EA, Psaty BM. Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. Stat Med. 2015;34(1):118–30.
19. He Z, Zhang M, Lee S, Smith JA, Guo X, Palmas W, Kardia SL, Diez Roux AV, Mukherjee B. Set-based tests for genetic association in longitudinal studies. Biometrics. 2015;71(3):606–15.
20. Wang X, Zhang Z, Morris N, Cai T, Lee S, Wang C, Yu TW, Walsh CA, Lin X. Rare variant association test in family-based sequencing studies. Brief Bioinform. 2017;18(6):954–61.
21. Kim J, Zhang Y, Pan W, Alzheimer's Disease Neuroimaging I. Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data. Genetics. 2016;203(2):715–31.
22. Lee S, Choi S, Qiao D, Cho M, Silverman EK, Park T, Won S. WISARD: workbench for integrated superfast association studies for related datasets. BMC Med Genet. 2018;11(Suppl 2):39.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
24. Chen G, Ye G, Zhang X, Liu X, Tu Y, Ye Z, Liu J, Guo Q, Wang Z, Wang L, et al. Metabolomics reveals protection of resveratrol in diet-induced metabolic risk factors in abdominal muscle. Cell Physiol Biochem. 2018;45(3):1136–48.
25. Gehart H, Kumpf S, Ittner A, Ricci R. MAPK signalling in cellular metabolism: stress or wellness? EMBO Rep. 2010;11(11):834–40.
26. Aerts JM, Boot RG, van Eijk M, Groener J, Bijl N, Lombardo E, Bietrix FM, Dekker N, Groen AK, Ottenhoff R, et al. Glycosphingolipids and insulin resistance. Adv Exp Med Biol. 2011;721:99–119.
27. Hwang H, Takane Y. Generalized structured component analysis. Psychometrika. 2004;69(1):81–99.
28. Hoerl AE, Kennard RW. Ridge regression - biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55.
29. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.
30. Li B. A deviance function for the quasi-likelihood method. Biometrika. 1993;80(4):741–53.
31. Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics. 2012;68(2):353–60.
32. Westfall PH, Young SS. Resampling-based multiple testing : examples and methods for P-value adjustment. New York: Wiley; 1993.
33. Li B, Wang G, Leal SM. SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. Bioinformatics. 2012;28(20):2703–4.
34. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–7.
35. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, Yoon D, Lee MH, Kim DJ, Park M, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet. 2009;41(5):527–34.
36. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.