Analysis of multiple related phenotypes in genome-wide association studies

Sohee Oh^{*,‡}, Iksoo Huh^{*,§}, Seung Yeoun Lee^{*,¶} and Taesung Park^{*,||}

**Department of Statistics, Seoul National University, 1 Gwanak-ro
Gwanak-gu, Seoul 08826, South Korea*

*†Department of Mathematics and Statistics, Sejong University, 209 Neungdong-ro
Gwangjin-gu, Seoul 05006, South Korea*

‡oh.sohee@gmail.com

§huhixoo@gmail.com

¶leesy@sejong.ac.kr

||tspark@snu.ac.kr

Accepted 2 September 2016

Published 21 October 2016

Most genome-wide association studies (GWAS) have been conducted by focusing on one phenotype of interest for identifying genetic variants associated with common complex phenotypes. However, despite many successful results from GWAS, only a small number of genetic variants tend to be identified and replicated given a very stringent genome-wide significance criterion, and explain only a small fraction of phenotype heritability. In order to improve power by using more information from data, we propose an alternative multivariate approach, which considers multiple related phenotypes simultaneously. We demonstrate through computer simulation that the multivariate approach can improve power for detecting disease-predisposing genetic variants and pleiotropic variants that have simultaneous effects on multiple related phenotypes. We apply the multivariate approach to a GWA dataset of 8,842 Korean individuals genotyped for 327,872 SNPs, and detect novel genetic variants associated with metabolic syndrome related phenotypes. Considering several related phenotype simultaneously, the multivariate approach provides not only more powerful results than the conventional univariate approach but also clue to identify pleiotropic genes that are important to the pathogenesis of many related complex phenotypes.

Keywords: Genome-wide association study; multivariate approach; pleiotropic factor.

1. Background

As a result of enormous advances in genotyping technologies, genome-wide association studies (GWAS) have become a standard approach for testing association between common human phenotypes and genetic variants using single nucleotide

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC-BY) License. Further distribution of this work is permitted, provided the original work is properly cited.

polymorphism (SNP) across the whole genome.^{1–7} While many susceptibility loci for complex phenotypes have been successfully identified through GWAS, there are still some challenges.

In GWAS, a very stringent genome wide significance threshold is required to account for the large number of tests carried out ($P < 5 \times 10^{-8}$). As a result, only a small number of genetic variants reach genome-wide significance and those explain a small fraction of the heritability of many common complex phenotypes.^{8,9} Furthermore, some variants identified in one GWA study are difficult to replicate in other independent GWA studies due to low power.^{10–12}

To overcome these challenges, there have been many efforts to improve power and detect more genetic variants associated with complex phenotypes. One widely used method is meta-analysis.^{13–17} Combining results from several studies for the same phenotype, meta-analysis has provided more power than a single study analysis owing to the increased sample size. However, there are still many genetic variants that even meta-analysis has limited power to detect due to small effects on the phenotype.

These kinds of challenges provide us a motivation to think about some fundamental questions for GWAS data analysis to improve power. Here, we consider a simple way of increasing statistical power from GWAS. Our idea stems from realization that in many GWAS there are several related phenotypes available but most analyses of current GWAS focus on one phenotype of interest such as a single binary phenotype derived from such related phenotypes. For instance, according to the American Diabetes Association (ADA), type 2 diabetes (T2D) is diagnosed based on fasting plasma glucose level (≥ 126 mg/dL), two hours after oral glucose level (≥ 200 mg/dL), and HbA1c ($\geq 6.5\%$). Instead of using these quantitative multiple phenotypes jointly, most current GWAS analyze many related phenotypes separately.^{18–20} This univariate analysis focusing on only one phenotype has been widely used in GWAS. This analysis is simple to conduct and provides a straightforward description on which variants are related to the phenotype of interest. However, performing univariate analyses on multiple phenotypes separately requires a more stringent significant level to control type I error rates due to the increased number of tests. Additionally, they may lose power in detecting association for one single phenotype due to ignoring the information available from combining phenotypes.^{21,22}

In this paper, we focus on the multivariate approach that considers multiple related phenotypes simultaneously. For example, in association studies to identify T2D associated variants, multivariate analysis takes three phenotypes of fasting plasma glucose level, two hours after oral glucose level, and HbA1c simultaneously in one model. As another example, consider an association study about obesity. There are several obesity related phenotypes available such as body mass index (BMI), weight, waist hip ratio, and etc. Multivariate analysis includes all these phenotypes in one model and identifies the SNPs associated with these phenotypes simultaneously. Taking all these phenotypes in one model has the effect of increasing the power. This type of multivariate analysis is different from the joint identification of

multiple SNPs associated with one phenotype and also from the gene-gene interaction analysis. In general, a multivariate model focuses on identifying single SNPs associated with multiple phenotypes.

Recently, several multivariate approaches have been proposed such as MultiPhen²³ and the linear mixed model.²⁴ MultiPhen evaluates the association between SNP and multiple phenotypes using the proportional odds model. However, it relies on the specific proportional odds assumption and encounters the multicollinearity problem due to correlated predictors such as related phenotypes. While the linear mixed model usually requires a heavy computation due to iterative algorithms to solve equation, Zhou and Stephens proposed the efficient mixed model algorithm for multivariate approach.²⁴ All these approaches demonstrated that the multivariate methods can be much more powerful than univariate approaches.

In this paper, we consider a simple multivariate approach that allows multiple related phenotypes simultaneously while not requiring a specific model assumption and heavy computation time. Our approach is based on the multivariate linear regression model, which is an easily applicable and widely used method.

As a complex phenotype, metabolic syndrome is the result of genetic variation, environmental factors, and interactions between genes and environmental factors.^{25–27} With increasing adoption of western lifestyles, the prevalence of metabolic syndrome is rapidly increasing in Asian countries including South Korea.^{28,29} We select six metabolic syndrome related phenotypes from the Korean GWA study^{5,20} and conduct the GWA test using both univariate and multivariate approaches. Among 327,872 SNPs, 52 SNPs in twelve chromosomal regions passed the genome-wide significance level with much stronger associations in the multivariate approach while 41 SNPs in eleven chromosomal regions passed the genome-wide significance level in the univariate approach. These genetic variants are located in or nearby genes, which are known to be associated with lipids or metabolic syndrome related phenotypes and for which much is known about biological function. Additionally, we present the properties of our multivariate GWA approach and compare its performance with the univariate approach through simulation studies.

2. Methods

2.1. Study participants

The Korea Association Resource (KARE) project was initiated in 2007 to undertake large-scale GWA analyses. Participants in this project were recruited from two community-based cohorts (i.e., the rural Ansung and urban Ansan cohorts) in the Gyeonggi Province of South Korea.⁵ These two cohorts consist of 5,018 and 5,020 participants, respectively, ranging in age from 40 to 69 years. More than 260

phenotypes have been extensively examined through epidemiological surveys, physical examinations and laboratory tests. Here, we focused on the GWA analysis of metabolic syndrome related six quantitative phenotypes because metabolic syndrome is defined from these phenotypes³⁰: waist circumference (WC), systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein cholesterol (HDLc), triglycerides (TG) and fasting plasma glucose (FPG). Prior to analysis, all phenotypes were examined for departure from normality. In this study, HDLc, TG and FPG were log-transformed. Since HDLc has negative correlation with each of other phenotypes, we use negative HDLc (−HDLc) in the analyses (Table 1). All participants of KARE provided written informed consent. This KARE analysis study was approved by the Institutional Review Board (IRB) of Seoul National University. The de-identified individual-level genotype and phenotype data of KARE samples were provided by the Korea Biobank Network, National Institute of Health, Korea.

2.2. Genotyping and quality control

The genomic DNAs were isolated from peripheral blood drawn from the participants and were genotyped on the Affymetrix Genome-Wide Human SNP array 5.0 containing 500,568 SNPs. For genotype calling, the Bayesian Robust Linear Modeling using Mahalanobis Distance (BRLMM) Genotyping Algorithm was used.⁵ For quality control process, the following procedure was applied: Hardy–Weinberg Equilibrium (HWE) $P \leq 10^{-6}$, missing genotype frequency $> 5\%$ and minor allele frequency (MAF) ≤ 0.01 .⁵ We also removed samples with low call rates, sample contamination, gender inconsistencies, and cryptic relatedness. After filtering out low quality SNPs and samples, we imputed missing genotypes of remaining SNPs using PLINK software³¹ with JPT/CHB reference panel in HapMap. As a result, the rate of total genotyped SNPs in the imputed dataset was 99.91%. Finally a total of 8,842 individuals and 327,872 SNPs were included in the analyses. There was no evidence for possible population stratification in KARE data.⁵

Table 1. Correlation matrix of six metabolic syndrome-related phenotypes in KARE project.

	WC	SBP	DBP	TG	−HDLc	FPG
WC	1	0.293	0.320	0.355	0.287	0.188
SBP	0.293	1	0.812	0.204	0.024	0.143
DBP	0.320	0.812	1	0.218	0.044	0.141
TG	0.355	0.204	0.218	1	0.439	0.181
−HDLc	0.287	0.024	0.044	0.439	1	0.048
FPG	0.188	0.143	0.141	0.181	0.048	1

2.3. Association analyses

For univariate analysis, linear regression analysis was conducted for each phenotype, adjusted for sex, age, and recruitment area. Under an additive genetic model assumption, each SNP genotype was coded as a count of minor alleles with values of 0, 1, and 2. To identify significant genetic variant, a t -statistic was used to examine for the significance of regression coefficient of the SNP effect. For multivariate analysis, a multivariate regression model was fitted with the same adjusting covariates under the additive genetic model assumption. The multivariate linear model is as follows.

$$Y = XB + E$$

$$\begin{bmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{11} & \cdots & \beta_{1k} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pk} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1k} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nk} \end{bmatrix},$$

where n is the number of individuals, k is the number of phenotypes, p is the number of variants, Y is the $n \times k$ phenotype data matrix, X is the $n \times p$ design matrix, B is the $p \times k$ parameter matrix, and E is the $n \times k$ matrix of random errors. The Wilk's Λ statistic is given by

$$\Lambda = \frac{|E|}{|H + E|} = \prod_{i=1}^k \frac{1}{1 + \lambda_i},$$

where H is the hypothesis sums of squares and cross products matrix, E is the error sums of squares and cross products matrix, k is the number of phenotypes, and λ_i is the eigenvalue, was used for testing simultaneous genetic effects on the related phenotypes. Since it is related to likelihood ratio test, its significance level was determined by comparing to the appropriate F -distribution. The test results were obtained using PLINK,³¹ R (version 3.2.3, <http://www.r-project.org>).

2.4. Simulation settings

We conducted simulation studies to evaluate the performance of univariate and multivariate approaches, and investigated the empirical type I error rate at various significance levels and computed the empirical power. For simplicity, we assumed that there are eight phenotypes available. In this simulation study, we considered the following parameters: MAF, marginal correlation coefficient ($\rho = 0, 0.25, 0.5$ and 0.75) between eight phenotypes, genetic effect size (β), and association directions. The genetic effect was evaluated via the linear regression model assuming an additive genetic model.

Prior to analysis, we first generated eight phenotypes from the multivariate normal distribution assuming the same correlation between phenotypes. Their expectations are zero, and variances are one for all eight phenotypes. Then the

genotype was generated under the HWE assumption with a variety of values of MAF (range: 0.01, 0.3). For the simulation of heterogeneous models, the first half of the phenotypes are assumed to have a negative associations with SNPs, and the other half of phenotypes are assumed to have a positive associations. For the homogeneous models, all phenotypes are assumed to be positively associated with the SNPs. The heterogeneous genetic-effect model can be written as follows.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_i + \varepsilon_{ij}, \quad (j = 1, 2, \dots, 8),$$

where y_{ij} is the j th phenotype of individual i , x_i is the genotype score based on the minor allele for the i th individual, β_{1j} is the genetic effect for the j th phenotype, and ε_{ij} is the error term generated from the multivariate normal distribution. This model allows a different genetic effect on each phenotype. For simplicity, we assumed that $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$, $\beta_{15} = \beta_{16} = \beta_{17} = \beta_{18}$, and $\beta_{11} = -\beta_{15}$. In other words, the first four parameters and the other four parameters have opposite (different) genetic association directions. If a common genetic effect is assumed, then the model reduces to the following homogeneous genetic effect model:

$$y_{ij} = \beta_{0j} + \beta_1x_i + \varepsilon_{ij}.$$

This homogeneous genetic effect model may provide more powerful tests than the heterogeneous model because of the reduced degrees of freedom when the homogeneous model is true.

For each combination of parameter values, both univariate and multivariate regression models were fit. The empirical type I error rates were calculated under the null hypothesis $\beta_{11} = 0$ (or $\beta_{15} = 0$) for univariate analysis, and $(\beta_{11}, \beta_{12}, \dots, \beta_{18}) = (0, 0, \dots, 0)$ for heterogeneous model and $\beta_1 = 0$ for homogeneous model in multivariate analysis by generating 100,000 replicates for 1,000 samples in order to evaluate the likelihood of an elevated chance of false positive effects, especially in extreme tail of the test statistic's. Additional simulation study was performed for the four correlated phenotypes with unstructured correlation structure for 10,000 samples with 10,000,000 replicates. For computing empirical power, we considered various nonzero values of β_{11} and β_{15} for the heterogeneous model and the same β_1 for the homogeneous model by generating 1,000 replications for 1,000 samples.

3. Results

3.1. Simulation settings

The empirical type I error rates were calculated under the null hypothesis $\beta_{11} = 0$ (or $\beta_{15} = 0$) for univariate analysis and $(\beta_{11}, \beta_{12}, \dots, \beta_{18}) = (0, 0, \dots, 0)$ for multivariate analysis by generating 100,000 replicates for 1,000 samples to assess the possibility of false positive in the extreme tail of null distribution. At a variety of value for significance, both univariate and multivariate approaches preserved type I error rates well for various values of MAF. (data not shown). Besides, there were no indications that type I error rate was inflated in the additional simulation studies with four

phenotypes and unstructured correlation study (data not shown). Consequently, we conclude that the multivariate approach well preserved type I error rates under the null hypotheses.

To compute empirical power, we first considered the case when the genetic variant has the same association direction for the multiple phenotypes. For simplicity, we present the results only for the model that has the same genetic effect size for eight phenotypes ($\beta_{11} = \beta_{15} = 0.1$) (Fig. 1). With heterogeneous genetic effect, the power of tests based on both univariate and multivariate models also increases as the MAF increases. Especially when the correlation coefficient between the multiple phenotypes is low (0 or 0.25), the multivariate tests were more powerful than univariate tests. On the other hand, when the multiple responses are highly correlated with each other (0.5 or 0.75), the difference in power between two tests is reduced. As the values of when the genetic effect sizes for the multiple phenotypes are similar,

Fig. 1. Power plots using different number of traits with the same effect sizes and association directions ($\beta_{11} = \beta_{15} = 0.1$) at $\alpha = 0.05$. Powers from the heterogeneous effect model were plotted by solid lines, while those from the homogeneous effect model were plotted by dotted lines. The homogeneous genetic effect models provide more powerful results than the heterogeneous models. For the approaches using only 1, 2, 4 numbers of traits, we used only part of the traits, the detailed approaches are given below. Uni1: Univariate approach using the first trait for both heterogeneous and homogeneous models (positive association with the SNP). Uni2: Univariate approach using the fifth trait. Multi(2): Multivariate approach using two traits: first and fifth traits. Multi(4): Multivariate approach using four traits: first, second, fifth, and sixth traits. Multi(8): Multivariate approach using all eight traits.

the homogeneous genetic effect model can improve power by using a test with 1 degree of freedom (df). As the values of MAF increase, the test based on the homogeneous effect model yields higher power than the tests based on other models (Fig. 1). The homogeneity test for genetic effects can easily be performed. Depending on the test result, the homogeneous genetic effect model can be chosen.

Next, consider the case where the genetic association is in opposite directions for different phenotypes. This can happen, for example, when environmental exposures or other genetic variants induce a positive association that overwhelms the different directions of association seen with the genetic locus in question. In our KARE application, we found that many genetic variants, which were not significant in univariate analyses but showed highly significant results in multivariate analysis, provided different association directions for the phenotypes (Tables 1 and 2). Moreover, they tended to have low MAF or their genetic effect sizes from the univariate analysis were almost zero. They could be false positives.

To investigate this issue, in our simulation model we assumed different directions of genetic effect for the multiple phenotypes and used a heterogeneous genetic effects model. The values of (β_{11}, β_{15}) were chosen close to zero. The empirical power was calculated at the various significance levels. For simplicity, when $(\beta_{11}, \beta_{15}) = (-0.05, 0.05)$, the multivariate approach tends to have slightly higher power than the univariate approach as MAF increases at $\alpha = 0.05$ (Fig. 2). However, at the very stringent significance level ($\alpha = 10^{-4}, 10^{-6}$), the empirical power was closed to zero (data not shown). Figure 3 shows the results when $(\beta_{11}, \beta_{15}) = (-0.10, 0.10)$, where we can see clear patterns. In general, the powers of both approaches increase as MAF increases. However, as the correlation coefficient ρ increases, the power of the test based on the univariate model does not change but the power of the test based on a multivariate model increases. Especially when ρ is large, the power of the multivariate approach becomes very large. We think that this is the main reason why many genetic variants with different association directions showed highly significant results in multivariate analysis for the KARE data, while they were not significant in univariate analyses.

For a simple and intuitive explanation of this phenomenon, we focus on Hotelling's T^2 test which tests for the difference of two sample mean vectors (μ_1 and μ_2). For the given sample sizes n_1 and n_2 , the noncentrality parameter δ , which is related to the power, is defined as follows:

$$\delta = (n_1 n_2 / (n_1 + n_2)) (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2),$$

where Σ is the population variance-covariance matrix. The noncentrality parameter δ tends to become larger when the differences between means have opposite directions and the correlations are high and positive. Thus, the multivariate approach tends to have more power than the univariate approach in this setting.

Further, we compared the performances of our multivariate linear model and MultiPhen. Our multivariate approach has two models: one is multivariate heterogeneous model and the other is multivariate homogeneous model. Since the degrees

Table 2. Significant SNPs from multivariate analysis of KARE data.

Locus	SNP	Position	Nearby gene	Minor allele	Multivariate			Univariate					
					MAF	F-statistic	P	Phenotype	Beta	P	Phenotype	Beta	P
2p23.3a	rs780094	27594741	GCKR	C	0.463	13.997	6.45×10^{-16}	WC	0.002	8.77×10^{-01}	−HDLc	0.012	4.33×10^{-01}
								SBP	−0.003	8.34×10^{-01}	TG	−0.102	1.46×10^{-11}
								DBP	0.003	8.16×10^{-01}	FPG	0.044	3.38×10^{-03}
8p21.3c	rs10503669	19891970	LDL	A	0.124	20.157	1.54×10^{-23}	WC	0.017	4.28×10^{-01}	−HDLc	−0.196	6.42×10^{-18}
								SBP	−0.031	1.46×10^{-01}	TG	−0.178	3.21×10^{-15}
								DBP	−0.018	4.10×10^{-01}	FPG	−0.021	3.57×10^{-01}
8q24.13e	rs2001945	126547160		G	0.416	7.965	1.38×10^{-08}	WC	−0.004	7.66×10^{-01}	−HDLc	−0.007	6.25×10^{-01}
								SBP	−0.004	8.00×10^{-01}	TG	0.077	2.77×10^{-07}
								DBP	−0.007	6.19×10^{-01}	FPG	−0.031	3.73×10^{-02}
9q31.1d	rs12686004	106693247	ABCA1	A	0.214	11.720	4.00×10^{-13}	WC	−0.016	3.69×10^{-01}	−HDLc	0.125	1.03×10^{-11}
								SBP	−0.028	9.67×10^{-02}	TG	0.000	9.99×10^{-01}
								DBP	−0.016	3.55×10^{-01}	FPG	0.033	7.17×10^{-02}
11q12.2b	rs174547	61327359	FADS1	C	0.324	9.106	5.91×10^{-10}	WC	−0.008	5.91×10^{-01}	−HDLc	0.055	6.32×10^{-04}
								SBP	−0.020	1.86×10^{-01}	TG	0.068	2.08×10^{-05}
								DBP	−0.005	7.54×10^{-01}	FPG	−0.073	4.42×10^{-06}
12q24.11d	rs12229654	109898844	MYL2	G	0.144	18.168	4.52×10^{-21}	WC	−0.092	1.29×10^{-05}	−HDLc	0.132	1.04×10^{-09}
								SBP	−0.066	1.03×10^{-03}	TG	−0.057	8.43×10^{-03}
								DBP	−0.071	6.78×10^{-04}	FPG	−0.098	5.04×10^{-06}

Table 2. (Continued)

Locus	SNP	Position	Nearby gene	Minor allele	Multivariate			Univariate					
					MAF	F-statistic	P	Phenotype	Beta	P	Phenotype	Beta	P
12q24.12a	rs2238153	110423930	ATXN2	A	0.460	7.754	2.46×10^{-08}	WC	-0.033	2.79×10^{-02}	-HDLc	0.066	1.70×10^{-05}
								SBP	-0.037	9.84×10^{-03}	TG	-0.026	8.45×10^{-02}
								DBP	-0.021	1.67×10^{-01}	FPG	-0.024	1.15×10^{-01}
12q24.13a	rs11066280	111302166	C12orf51	A	0.173	24.880	2.02×10^{-29}	WC	-0.089	4.86×10^{-06}	-HDLc	0.145	5.81×10^{-13}
								SBP	-0.074	6.49×10^{-05}	TG	-0.077	1.03×10^{-04}
								DBP	-0.079	5.17×10^{-05}	FPG	-0.086	1.49×10^{-05}
12q24.13b	rs2072134	111893559	OAS3	A	0.115	13.942	7.55×10^{-16}	WC	-0.070	2.50×10^{-03}	-HDLc	0.143	1.88×10^{-09}
								SBP	-0.077	4.40×10^{-04}	TG	-0.049	3.69×10^{-02}
								DBP	-0.077	8.30×10^{-04}	FPG	-0.066	4.95×10^{-03}
15q22.1b	rs16940212	56481312		T	0.340	17.299	5.41×10^{-20}	WC	0.002	9.18×10^{-01}	-HDLc	-0.132	1.37×10^{-16}
								SBP	-0.012	4.25×10^{-01}	TG	0.020	2.11×10^{-01}
								DBP	-0.005	7.25×10^{-01}	FPG	-0.002	8.94×10^{-01}
16p12.3c	rs7197218	17319136	XYLT1	G	0.014	7.935	1.49×10^{-08}	WC	-0.033	6.07×10^{-01}	-HDLc	-0.191	3.69×10^{-03}
								SBP	-0.012	4.25×10^{-01}	TG	0.106	1.04×10^{-01}
								DBP	0.061	3.34×10^{-01}	FPG	0.352	6.29×10^{-08}
16q13b	rs6499863	rs6499863	CETP	A	0.104	9.116	5.76×10^{-10}	WC	0.014	5.60×10^{-01}	-HDLc	0.152	1.06×10^{-09}
								SBP	-0.003	8.99×10^{-01}	TG	-0.005	8.25×10^{-01}
								DBP	0.030	2.18×10^{-01}	FPG	0.016	5.09×10^{-01}

Fig. 2. Power plots using different number of traits with the same effect sizes (absolute value of 0.05) and opposite directions of association (half are positive and another half are negative associated) at $\alpha = 0.05$. The heterogeneous multivariate models seem to be more powerful than univariate models, as MAF and correlation coefficient increase. However, two univariate models show no differences even though correlation coefficients are increasing. It needs a caution to interpret result in the presence of genetic effects with different directions in a multivariate analysis. Legends of power lines have same meanings that are used in the Fig. 1.

of freedom of the test based on Multiphen are equal to the number of phenotypes, these two models have different degrees of freedom for the association test. Thus, it is more reasonable to compare Multiphen with the multivariate heterogeneous model. As expected, the multivariate heterogeneous model and Multiphen have similar

Fig. 3. Different association direction, effect sizes for multiple phenotypes with absolute value of 0.10 at $\alpha =$ (a) 0.05, (b) 0.01, (c) 10^{-4} , and (d) 10^{-6} . Power of multivariate approach is increasing as MAF and correlation coefficient ρ increase.

Fig. 3. (Continued)

power (Fig. 4). However, the multivariate homogeneous model has higher power than MultiPhen when the true model has homogeneous associations. It implies that our multivariate model can be applied to more flexible situations.

3.2. Identifying metabolic syndrome associated SNPs by univariate and multivariate approaches

Prior to analysis, all six phenotypes were standardized to have a mean of zero and a variance of one. The correlation matrix of six phenotypes is given in Table 1. For the univariate approach, we fit a linear model relating each phenotype to genotype separately with covariate adjustment for sex, age, and recruitment area. For the multivariate approach, we applied a multivariate regression model allowing for heterogeneous genetic effects and including the same covariates under the unstructured covariance structure assumption for the phenotypes. The Manhattan plots from these analyses are given in Figs. 5 and 6.

Fig. 4. Power plots of MultiPhen and multivariate model at $\alpha = 0.10$ in the homogeneous effect setting. The multivariate heterogeneous effect model has similar power with MultiPhen, while the multivariate homogeneous model has higher power than MultiPhen.

Among 327,872 directly genotyped SNPs, a total of 41 SNPs on eleven chromosomal regions reached the genome-wide significance ($P < 5 \times 10^{-8}$) for at least one of the six metabolic syndrome related phenotypes. 34 SNPs in 7 chromosomal regions and 19 SNPs in 3 chromosomal regions showed highly significant associations with

Fig. 5. Manhattan plots of metabolic syndrome-related phenotypes from multivariate analysis. In a genome-wide significance level $\alpha = 5.0 \times 10^{-8}$, multivariate model provided more genetic variants (52) in 12 chromosomal regions that are associated with metabolic syndrome related phenotypes.

Fig. 6. Manhattan plots of metabolic syndrome-related phenotypes from univariate analyses (A)–(F) show results from univariate analyses for each phenotype. In a genome-wide significance level $\alpha = 5.0 \times 10^{-8}$, 11 chromosomal regions reached this significance level in a univariate framework.

HDLc ($P_{\min} = 6.42 \times 10^{-18}$) and TG ($P_{\min} = 2.12 \times 10^{-15}$), respectively. For SBP, DBP and FPG, one locus for each phenotype reached genome-wide significance. On the other hand, none of the variants reached the genome-wide significance level for WC. All eleven chromosomal regions have been identified from the previous GWAS to be associated with blood pressure, lipids, HDLc, TG and T2D.^{2,20,25–39}

In the multivariate analysis, we identified 52 variants in twelve chromosomal regions that passed the genome-wide significance (Fig. 5). The number of detected SNPs in the multivariate analysis is larger than the number of those in six univariate analyses. Among them, nine chromosomal regions were at the same chromosomal regions identified from univariate analyses (Figs. 5 and 6, and Table 2). The remaining three chromosomal regions did not reach the univariate genome-wide significance, but satisfied the genome-wide significance in the multivariate approach. Table 2 summarizes the SNPs with the smallest multivariate analysis P -value along with univariate P -values in each chromosomal region. For example, the multivariate analysis P for rs2238153 was 2.46×10^{-8} while its smallest univariate P was 1.70×10^{-5} (Table 2). Furthermore, we unexpectedly observed that the directions of genetic effects for the phenotypes are not the same. For instance, the minor allele of the most significant SNP (rs11066280) showed negative associations both with TG and FPG. On the other hand, this variant showed a positive association with $-HDLc$

while all these phenotypes showed strong positive correlations among them (Table 2).

Note that the univariate approach provided 41 genetic variants in eleven chromosomal regions at the genome-wide significance level, while the multivariate analysis yielded 52 genetic variants in twelve chromosomal regions with much higher association signals (Fig. 5 and Table 2). Among them, nine chromosomal regions have been identified in both univariate and multivariate approaches, which have been reported to be associated with lipids and metabolite related phenotypes and their biological functions are known.

We performed the functional enrichment analysis for a further biological interpretation for the SNPs given in Table 2. Among 12 SNPs, 10 SNPs have nearby genes. We then performed functional enrichment analysis using Database for Annotation, Visualization and Integrated Discovery (DAVID) via the functional annotation clustering tool (<http://david.abcc.ncifcrf.gov/>). As a result, one cluster had an enrichment score of 1.48 higher than the significant level of 1.3 and shown to be related to lipid metabolism and metabolic process (Table 3). This enrichment analysis results support the biological interpretation of our findings.

The variant rs780094 ($P = 6.45 \times 10^{-16}$) is located in 2p23.3, intronic region of *GCKR*, glucokinase regulator. *GCKR* encodes glucokinase regulatory protein (GKRP), a predominantly nuclear protein that inhibits hepatic glucokinase and plays a critical role in glucose homeostasis. It has been considered a candidate gene as one type of maturity-onset diabetes of the young (MODY2, MIM = 606391),²⁰ and reported to be associated with fasting glucose, insulin, low-density lipoprotein cholesterol (LDLc) and metabolic syndrome in previous GWA studies^{32–35} and the KARE study.³⁶ Two variants, rs10503669 on 8q21.3 and rs2001945 on 8q24.13, showed a strong evidence of association with metabolic syndrome related phenotypes ($P = 1.54 \times 10^{-23}$ and $P = 1.38 \times 10^{-08}$). Both variants have been reported to be associated with TG and HDLc in a previous KARE study²⁰ and the European population GWA study.³² An identified association signal in the 9q31.1 region included the variant rs12686004 ($P = 4.00 \times 10^{-13}$), which is located in intronic region of *ABCA1*, ATP-binding cassette, the gene functioning as a cholesterol efflux pump in the cellular lipid removal pathway.²⁰ This result was previously reported in the KARE study.²⁰

One variant rs174547 on chromosome 11q12.2 reached the genome-wide significance level ($P = 5.91 \times 10^{-10}$) while it has not been identified in previous KARE studies. This variant is located in the intronic region of *FADS1*, fatty acid

Table 3. Functional enrichment analysis for the genes in Table 2.

Enrichment score: 1.48	Gene	Count	P-value	Benjamini
Lipid metabolism	<i>ABCA1, CETP, FADS1</i>	3	8.8E-3	4.0E-1
Small molecule metabolic process	<i>ABCA1, CETP, FADS1, GCKR</i>	4	2.8E-2	9.5E-1
Transport	<i>ABCA1, CETP, FADS1</i>	3	1.4E-1	8.4E-1

desaturase1. *FADS1* encodes a key enzyme in metabolism of unsaturated fatty acids, and insulin secretion differs in response to meals with varying fatty acid composition. It has been reported to be associated with lean mass, lipids, HDLc, and TG from GWAS of European and Chinese populations.^{37–39}

rs12229654 near *MYL2*, myosin light chain 2, on chromosome 12 showed highly significant association with metabolite phenotypes ($P = 4.52 \times 10^{-21}$). *MYL2* encodes the regulatory light chain associated with the cardiac myosin β (or slow) heavy chain. This variant has been identified in a previous KARE study, which is related to HDLc and gamma glutamyl transferase (GGT),²⁰ and a predisposing risk factor for T2D.⁴⁰ Other three signals on chromosome 12 were also detected (rs2238153 in *ATXN2* on 12q24.12, $P = 2.46 \times 10^{-08}$; rs11066280 near *C12orf51* on 12q24.13, $P = 2.02 \times 10^{-29}$; rs2072134 near *OAS3*, $P = 7.55 \times 10^{-16}$). While rs2238153 did not reach the genome-wide significance in univariate analyses, it was newly identified in this study. Chromosome 12q24.13 locus was associated with metabolite levels such as HDLc, alanine aminotransferase (ALT) and aspartate aminotransferase (AST).^{20,41,42}

The variant rs16940212 mapped to chromosome 15q22.1 satisfied the genome-wide significance ($P = 5.41 \times 10^{-20}$), which showed a strong association with HDLc in previous KARE study.²⁰ On chromosome 16, two signals were identified (rs7197218 in intronic region of *XYLT1*, xylosyltransferase I, on 16p12.3 and rs6499863 near *CETP*, cholesteryl ester transfer protein, on 16q13b). *CETP* is a plasma protein that facilitates the transport of cholesteryl esters and triglycerides between the lipoproteins. In previous GWA studies, this gene has been reported to be associated with HDLc and TG.^{20,33,43–45}

Note that three SNPs (rs174547, rs2238153, rs7197218) are only detected by the multivariate approach using 5×10^{-8} criterion. Although rs2001945 is not significant by univariate analysis, another SNP (rs4351435) in the same chromosomal region was detected to be associated with TG. Among results of the three SNPs, the multivariate analysis P for rs2238153 was 2.46×10^{-08} while its smallest univariate P was 1.70×10^{-05} (Table 2). Although the minimum of the univariate P -values did not satisfied genome-wide significance criterion, three of the six univariate P -values are under 5×10^{-2} . Moreover, we also unexpectedly observed that the directions of genetic effects on the phenotype are not the same. The minor allele of the rs2238153 showed negative associations both with WC and SBP, while it showed a positive association with $-HDLc$. In addition, these phenotypes showed considerably positive correlations among them (Table 2). Other two SNPs also have similar results with the rs2238153. The multivariate analysis for this type of SNPs tended to yield high power, as shown in the simulation study (Fig. 2).

On the other hand, two chromosomal regions are detected only by the univariate approach. The first one is rs7754840 ($P = 4.73 \times 10^{-8}$) located in 6p22.3, and the second one is rs17249754 ($P = 9.86 \times 10^{-10}$) located in 12q21.3. These two SNPs are known to be associated with FPG and SBP, respectively.^{5,46} The reasons why the two associations are not detected in multivariate analysis are as follow. In general,

Fig. 7. Quantile-quantile plot from a multivariate approach in Korean GWAS data. Black points represented the total number of 327,872 SNPs and blue points represented the SNPs after excluding significant SNPs at $\alpha = 5 \times 10^{-8}$. Because the multivariate approach provides more power, QQ plot tends to deviate from the reference line. The genomic control inflation factor was estimated to be 1.03.

Fig. 8. Quantile-quantile plots for univariate approach. Black points represented the total number of 327,872 SNPs and blue points represented the SNPs after excluding significant SNPs at $\alpha = 5 \times 10^{-8}$.

Fig. 9. Quantile–quantile plots from univariate analyses with 100 PCs, and w/o 100PCs. (a) QQ plot of TG phenotype (b) QQ plot of $-HDL$ phenotype.

Multivariate analysis has lower power than univariate analysis when many phenotypes are not associated with SNPs. In case of rs7754840, among the six phenotypes, only FPG phenotype is statistically significant and the others are nonsignificant with large P -values. For rs17249754, although SBP and DBP are statistically significant, they are very highly correlated (Pearson's $r = 0.812$, see Table 1) and have the same direction of effects. Therefore, their contribution to multivariate statistic is relatively small, which resulted in non-rejection of multivariate hypothesis. This trend was observed in the simulation of homogeneous effect setting, as shown in Fig. 1.

4. Conclusion

In this study, we investigate a simple and more powerful approach to GWA analysis with multiple related phenotypes simultaneously. In our multivariate application of KARE data, 52 variants in twelve chromosomal regions were identified with stronger association signals than the univariate analysis (Table 2). Among them, nine chromosomal regions were reported to be associated with lipids, diabetes-related and metabolic related phenotypes in the previous GW studies. The remaining three chromosomal regions were newly identified by our multivariate analysis, and are reported to have biological functions related to metabolic syndrome. Besides, most chromosomal regions have been known related with various metabolic phenotypes from GWAS catalog (<http://www.ebi.ac.uk/gwas>). Among 52 variants identified by multivariate analysis, 36 variants were also detected by univariate analyses. The remaining 16 variants were only identified by the multivariate approach by the increase of power due to combining all available information from the univariate approaches.

In our study, the genomic control inflation factor (λ) was estimated to be 1.03. Although its value is close to 1 and did not seem to be inflated, the observed quantiles tended to deviate from the theoretical quantiles (Figs. 7 and 8). Such deviations from the reference line could be occurred by the result of polygenic nature of metabolic syndrome. In addition, several reasons such as incorrect null distribution, population

stratification, cryptic relatedness, or a noticeable proportion of positive variants can be also possible reasons.¹ The prior quality control procedure showed that there was no strong evidence of population stratification in KARE data.⁵ All samples with cryptic relatedness were removed before analysis.⁵ Related to deviation of quantiles in GWAS, in the presence of polygenic inheritance, substantial genomic inflation is expected in the quantile-quantile (QQ) plot.⁴⁷ Since metabolic syndrome is a polygenic phenotype, quantiles may tend to deviate more from the reference line in the QQ plot. Further, the simulation study showed that our univariate and multivariate analyses produced the correct null distributions, and additional analyses with principal components (PCs) showed very similar result compared to our original analyses (Fig. 9). Therefore, the deviation of the multivariate analysis might be due to a noticeable proportion of positive variants.

Our simulation study revealed the following characteristics of the multivariate approach. First, when the directions of genetic effects for the phenotypes are the same and of similar magnitude, the homogeneous genetic effect model is more efficient and provides more power than the heterogeneous genetic effect model. Second, when the directions of genetic effects for the phenotypes are different, the multivariate approach provides much higher power than the univariate approach. In our KARE data analysis, the application of multivariate methods identified highly significant genetic variants, which were not identified by the univariate approach. We think it is occasionally possible to identify genetic associations in different directions from positively correlated phenotypes. Because phenotypes are marginally correlated with each other with same direction, it is possible that the conditional association between genetic variants and phenotypes shows different association direction due to its different biological function, hormone and environmental factors on each phenotype. For this reason, it is important to understand underlying biological background between gene and phenotype. In addition, compared with a meta analysis that combines *P*-values from univariate analyses,⁴⁸ the multivariate method detected more significant SNPs (data not shown). This may be due to the fact that the meta analysis only uses the summary data. The multivariate analysis uses all raw data available and can produce more accurate results because it can estimate covariance part for each of tests after adjusting information of covariates and SNP markers. Therefore, if we can get all information of raw dataset, it is recommended to use multivariate analysis rather than meta-analysis.

Our multivariate approach can be extended into the more general situation. First, the multivariate model with a single genetic variant can be easily extended to the model with multiple genetic variants.⁴⁹ The model with a single genetic variant tends to be hard to explain and predict the accumulated and/or joint effects of multiple genetic variants on the phenotypes. Second, in GWA analysis, linkage disequilibrium (LD) among nearby SNPs can provide valuable information in understanding genetic architecture. Recently, many univariate methods have been proposed to handle LD information in the multiple regression framework such as the regularization method.^{17,50,51} The multivariate approach can also be easily extended to association

studies with the multiple genetic variants and to examine genetic effects for all related phenotypes. This extension of the multivariate approach might provide more power and better prediction for complex diseases. The multivariate approach can also be applied to cohort studies in which several repeated measurements are obtained. In cohort studies, the same phenotypes are repeatedly measured through the follow-up periods. Instead of focusing only on one phenotype observed at the specific time point, the multivariate approach can focus on the multiple phenotypes repeatedly measured at the different time points, and allow us to perform profile analysis. Finally, the multivariate approach can easily be applied to detect rare variants that are associated with more than one phenotype. One of the competing hypotheses for common disease is Common Disease–Rare Variant (CDRV) hypothesis.^{10,52} That is, multiple rare variants have moderate and large genetic effects on the common disease. Considering several related phenotype simultaneously, the multivariate approach is expected to be quite powerful in detecting rare variants.

The conducting GWA analysis using multivariate approach has several advantages against the univariate approach. First, the multivariate approach can consider several phenotypes simultaneously in one model and hence it can reduce multiple comparison problems by reducing the number of tests. Second, considering correlation between related-phenotypes in one model, the multivariate approach would be more suitable to detect pleiotropic genes. Additionally, by taking into account the correlation among phenotypes, the multivariate method can identify genetic variants not easily detected by the univariate approach. Lastly, the multivariate approach helps us to understand genetic architecture for complex diseases and to identify pleiotropic genes that are important to the pathogenesis of many related complex diseases.

Acknowledgments

This work was supported by a grant funded by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of science, ICT, and Future Planning, also through the NRF, Korea and by a NRF grant (2015R1A5A6001906).

References

1. The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common disease and 3,000 shared controls, *Nature* **447**:661–678, 2007.
2. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, R. Saxena *et al.*, Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels, *Science* **316**:1331–1336, 2007.
3. Sladek R, Rocheleau G, Rung J, Dina C *et al.*, A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature* **445**:881–885, 2007.
4. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V *et al.*, Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity, *Nat Genet* **41**:18–24, 2009.

5. Cho YS, Go MJ, Kim YJ, Heo JY *et al.*, A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits, *Nat Genet* **41**:527–534, 2009.
6. Ripke S, Snader AR, Kendler KS, Levinson DF *et al.*, Genome-wide association study identifies five new schizophrenia loci, *Nat Genet* **43**:969–976, 2011.
7. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K *et al.*, Genome-wide association studies identify four ER negative-specific breast cancer risk loci, *Nat Genet* **45**:392–398, 2013.
8. Hirschhorn JN, Daly MN, Genome-wide association studies for common diseases and complex traits, *Nat Rev Genet* **6**:95–108, 2005.
9. Chatterjee N, Wheeler B, Sampson J, Hartge P *et al.*, Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies, *Nat Genet* **45**:400–405, 2013.
10. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J *et al.*, Genome-wide association studies for complex traits: Consensus, uncertainty and challenges, *Nat Genet Rev* **9**:356–369, 2008.
11. Kraft P, Replication in genome-wide association studies, *Stat Sci* **24**:561–573, 2009.
12. Fall T and Ingelsson E, Genome-wide association studies of obesity and metabolic syndrome, *Mol Cell Endocrinol* **382**:740–757, 2014.
13. Houlston RS, Webb E, Broderick P, Pittman AM *et al.*, Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer, *Nat Genet* **40**:1426–1435, 2008.
14. Perry JR, Stolk L, Franceschini N, Lunetta KL *et al.*, Meta-analysis of genome-wide association data identifies two loci influencing age at menarche, *Nat Genet* **41**:648–650, 2009.
15. Heid IM, Jackson AU, Randall JC, Winkler TW *et al.*, Meta-analysis identifies 13 new loci associated with waist-hip ratio reveals sexual dimorphism in the genetic basis of fat distribution, *Nat Genet* **42**:949–960, 2010.
16. Cho YS, Chen CH, Hu C, Long J *et al.*, Meta-analysis of genome-wide association studies identified eight new loci for type 2 diabetes in East Asians, *Nat Genet* **44**:67–72, 2012.
17. Evangelou E, Ioannidis JP, Meta-analysis methods for genome-wide association studies and beyond, *Nat Rev Genet* **14**:379–389, 2013.
18. Yuan X, Waterworth D, Perry JR, Lim N *et al.*, Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes, *Am J Hum Genet* **83**:520–528, 2008.
19. Sabatti C, Service SK, Hartikainen AL, Pouta A *et al.*, Genome-wide association analysis of metabolic traits in a birth cohort from a founder population, *Nat Genet* **41**:35–46, 2009.
20. Kim YJ, Go MJ, Hu C, Hong CB *et al.*, Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* **43**:990–995, 2011.
21. Schmitz S, Cherny SS, Fulker DW, Increase in power through multivariate analyses, *Behav Genet* **28**:357–363, 1998.
22. Liu YZ, Pei YF, Liu JF, Yang F *et al.*, Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males, *PLoS One* **4**:e6827, 2009.
23. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC *et al.*, MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS, *PLoS One* **7**:e34861, 2012.
24. Zhou X, Stephens M, Efficient multivariate linear mixed model algorithms for genome-wide association studies, *Nat Methods* **11**:407–409, 2014.

25. Bell CG, Walley AJ, Froguel P, The genetics of human obesity, *Nat Rev Genet* **6**:221–234, 2005.
26. Eckel RH, Grundy SM, Zimmet PZ, The metabolic syndrome, *Lancet* **365**:1415–1428, 2005.
27. Pollex RL and Hegele RA, Genetic determinants of the metabolic syndrome, *Nat Clin Pract Cardiovasc Med* **3**:482–489, 2006.
28. Nestel P, Lyu R, Low LP, Sheu WH, Nitiyanant W, Saito I, Tan CE, Metabolic syndrome: Recent prevalence in East and Southeast Asian populations, *Asia Pac J Clin Nutr* **16**:362–367, 2007.
29. Lim S, Shin H, Song JH, Kwak SH *et al.*, Increasing prevalence of metabolic syndrome in Korea, *Diabetes Care* **34**:1323–1328, 2011.
30. Grundy SM, Cleeman JJ, Daniels SR, Donato KA *et al.*, Diagnosis and management of the Metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institutes Scientific Statement, *Circulation* **112**:2735–2752, 2005.
31. Purcell S, Neale B, Todd-Brown K, Thomas L *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet* **81**:559–575, 2007.
32. Willer CJ, Sanna S, Jackson AU, Scuteri A *et al.*, Newly identified loci that influence lipid concentrations and risk of coronary artery disease, *Nat Genet* **40**:161–169, 2008.
33. Kathiresan S, Melander O, Guiducci C, Surti A *et al.*, Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans, *Nat Genet* **40**:189–197, 2008.
34. Kristiansson K, Perola M, Tikkanen E, Kettunen J *et al.*, Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits, *Circ Cardiovasc Genet* **5**:242–249, 2012.
35. Coram MA, Duan Q, Hoffmann TJ, Thornton T *et al.*, Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations, *Am J Hum Genet* **92**:904–916, 2013.
36. Lee HJ, Jang HB, Kim HJ, Ahn Y *et al.*, The dietary monounsaturated to saturated fatty acid ratio modulates the genetic effects of GCKR on serum lipid levels in children, *Clin Chim Acta* **450**:155–161, 2015.
37. Lemaitre RN, Tanaka T, Tang W, Manichaikul A *et al.*, Genetic loci associated with plasma phospholipid n-3 fatty acids: A meta-analysis of genome-wide association studies from the CHARGE consortium, *PLoS Genet* **7**:e1002193, 2011.
38. Suhre K, Shin SY, Petersen AK, Mohny RP *et al.*, Human metabolic individuality in biomedical and pharmaceutical research, *Nature* **31**:54–60, 2011.
39. Han Y, Pei Y, Liu Y, Zhang L *et al.*, Bivariate genome-wide association study suggests fatty acid desaturase genes and cadherin DCHS2 for variation of both compressive strength index and appendicular lean mass in males, *Bone* **51**:1000–1007, 2012.
40. Go MJ, Hwang JY, Kim YJ, Oh JH *et al.*, New susceptibility loci in MYL2, C12orf51 and OAS1 associated with 1-h plasma glucose as predisposing risk factors for type 2 diabetes in the Korean population, *J Hum Genet* **58**:362–365, 2013.
41. Kato N, Takeuchi F, Tabara Y, Kelly TN *et al.*, Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in East Asians, *Nat Genet* **43**:531–538, 2011.
42. Lu X, Wang L, Chen S, He L *et al.*, Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease, *Nat Genet* **44**:890–894, 2012.
43. Ridker PM, Paré G, Parker AN, Zee RY *et al.*, Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among

- 18245 initially healthy women from the Women's genome health study, *Circ Cardiovasc Genet* **2**:26–33, 2009.
44. Sull JW, Lee JE, Lee M, Jee SH, Cholesterol ester transfer protein gene is associated with high-density lipoprotein cholesterol levels in Korean population, *Genes Genom* **34**:231–235, 2012.
45. Lee Y, Park S, Moon S, Lee J *et al.*, On the analysis of a repeated measure design in genome-wide association analysis, *Int J Environ Res Public Health* **11**:12283–12303, 2014.
46. Go MJ, Hwang JY, Park TJ, Kim YJ *et al.*, Genome-wide association study identifies two novel loci with sex-specific effects for type 2 diabetes mellitus and glycemic traits in a Korean population, *Diabetes Metab J* **38**:375–387, 2014.
47. Yang J, Weedon MN, Purcell S, Lettre G *et al.*, Genomic inflation factors under polygenic inheritance, *Eur J Hum Genet* **19**:807–812, 2011.
48. Poole W, Gibbs DL, Shmulevich I, Bernard B *et al.*, Combining dependent P-values with an empirical adaptation of Brown's method, *bioRxiv* doi: <http://dx.doi.org/10.1101/029637>.
49. Johnson AR, Wichern DW, *Applied Multivariate Statistical Analysis*, 6th edn., Prentice Hall, 2007.
50. Cho S, Kim K, Kim YJ, Lee JK *et al.*, Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis, *Ann Hum Genet* **74**:416–428, 2010.
51. Austin E, Pan W, Shen X, Penalized regression and risk prediction in genome-wide association studies, *Stat Anal Data Min* **6**:315–328, 2013.
52. Schork NJ, Murray SS, Frazer KA, Topol EJ, Common vs. rare allele hypotheses for common diseases, *Curr Opin Genet Dev* **19**:212–219, 2009.

Sohee Oh received her Ph.D. degree in Statistics from Seoul National University, South Korea, in 2011. She is currently a Research Assistant Professor in the Department of Biostatistics, Seoul Metropolitan Government –Seoul National University Boramae Medical Center, Seoul, South Korea.

Iksoo Huh received his Ph.D. degree in Statistics from Seoul National University, South Korea, in 2015. Now, he is a postdoctoral fellow at the School of Biological Sciences, Georgia Institute of Technology, USA.

Seung-Yeoun Lee received her Ph.D. degree in Biostatistics from University of Michigan, USA, in 1990. From August 1991 to August 1992, she was a visiting scholar at the National Cancer Institute, USA. Since 1997, she is a professor in the Department of Applied Statistics, Sejong University, Seoul, Korea. Now she serves as the Chairman of International Biometric Society, Korean region.

Taesung Park received his Ph.D. degree in Biostatistics from University of Michigan, USA, in 1990. He was a visiting fellow at National Institute of Child Health and Development National Institutes of Health, USA from August 1991 to August 1992. In 1992, he joined the Department of Statistics, Hankuk University of Foreign Studies, South Korea. Since 1999, he is a professor in the Department of Statistics, Seoul National University, Seoul, South Korea.