Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions

Jiin Choi, Taesung Park*

From 24th International Conference on Genome Informatics (GIW 2013)
Singapore, Singapore. 16-18 December 2013

Abstract

Background: Recently, one of the greatest challenges in genome-wide association studies is to detect gene-gene and/or gene-environment interactions for common complex human diseases. Ritchie et al. (2001) proposed multifactor dimensionality reduction (MDR) method for interaction analysis. MDR is a combinatorial approach to reduce multi-locus genotypes into high-risk and low-risk groups. Although MDR has been widely used for case-control studies with binary phenotypes, several extensions have been proposed. One of these methods, a generalized MDR (GMDR) proposed by Lou et al. (2007), allows adjusting for covariates and applying to both dichotomous and continuous phenotypes. GMDR uses the residual score of a generalized linear model of phenotypes to assign either high-risk or low-risk group, while MDR uses the ratio of cases to controls.

Methods: In this study, we propose multivariate GMDR, an extension of GMDR for multivariate phenotypes. Jointly analysing correlated multivariate phenotypes may have more power to detect susceptible genes and gene-gene interactions. We construct generalized estimating equations (GEE) with multivariate phenotypes to extend generalized linear models. Using the score vectors from GEE we discriminate high-risk from low-risk groups. We applied the multivariate GMDR method to the blood pressure data of the 7,546 subjects from the Korean Association Resource study: systolic blood pressure (SBP) and diastolic blood pressure (DBP). We compare the results of multivariate GMDR for SBP and DBP to the results from separate univariate GMDR for SBP and DBP, respectively. We also applied the multivariate GMDR method to the repeatedly measured hypertension status from 5,466 subjects and compared its result with those of univariate GMDR at each time point.

Results: Results from the univariate GMDR and multivariate GMDR in two-locus model with both blood pressures and hypertension phenotypes indicate best combinations of SNPs whose interaction has significant association with risk for high blood pressures or hypertension. Although the test balanced accuracy (BA) of multivariate analysis was not always greater than that of univariate analysis, the multivariate BAs were more stable with smaller standard deviations.

Conclusions: In this study, we have developed multivariate GMDR method using GEE approach. It is useful to use multivariate GMDR with correlated multiple phenotypes of interests.
Background

Genome-wide association studies (GWAS) have been successfully conducted to detect disease susceptibility genes for common complex human diseases by focusing on associations between single-nucleotide polymorphisms (SNPs) and phenotypes [1]. While traditional methods for GWAS consider only one SNP at a time, some common complex human diseases such as diabetes, hypertension, and various types of cancers are known to be influenced by multiple genetic variants [2]. In addition, one of the greatest challenges in GWAS is to discover gene-gene and/or gene-environment interactions.

Classic logistic regression can be used to analyze the gene-gene interaction [3]. However, logistic regression suffers from an overfitting problem in high-order interactions [4]. Multifactor dimensionality reduction (MDR) method is a nonparametric, model-free, and combinatorial approach for interaction analysis by identification of a multi-locus model for association in case-control studies [5-9]. MDR method reduces multi-locus genotypes into two disease risk groups: high-risk and low-risk groups. If the ratio of cases and controls in a combination of genotypes is larger than a pre-assigned threshold $T$ (e.g., $T = 1$), the cell of combination is labelled as "high risk", otherwise, "low risk". MDR method shows greater power for testing high-order interactions compared with logistic regression analysis [10]. Several statistical methods have been extended from MDR approach [11-16]. One of the extended methods of MDR is a generalized MDR (GMDR) proposed by Lou et al. [16]. GMDR method allows adjusting for covariates and applying to both dichotomous and continuous phenotypes; it uses the score-based statistic obtained from generalized linear model (GLM) which is implemented in GMDR. Let $y_i = (y_{i1}, \ldots, y_{it})^T$ be the $t \times 1$ vector of the phenotypes for subject $i$ ($i = 1, \ldots, n$), with expectation $E(Y_{it}) = \mu_{it}$. For the multivariate phenotype vector, $y_i$, we assume an underlying generalized linear model which can be written as

$$\eta_i = g(\mu_i) = X_i \beta + Z_i \gamma,$$

where $g(\cdot)$ denotes a known one-to-one link function that is allowed to change with the characteristics of the different types of phenotype $y_i$. $X_i$ and $Z_i$ represent design matrices of genotype values and known covariate values including the unit component, respectively, and $\beta$ and $\gamma$ are vectors of their corresponding parameters, respectively. We assume that $y_{it}$ has a probability distribution belonging to the exponential family of distributions formed as

$$f(y_{it}; \theta_{it}, \phi) = \exp \left\{ y_{it} \theta_{it} - b(\theta_{it}) \right\} \frac{\exp(c(y_{it}, \phi))}{c(y_{it}, \phi)}.$$

The GEE estimators of $\delta = (\beta^T, \gamma^T)$ for marginal models can be obtained from the solution of a set of following generalized estimating equations:

$$U(\delta) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \delta} \right)^T V_i^{-1} \{ y_i - \mu_i(\delta) \} = 0,$$
where $\partial \mu_i / \partial \delta$ is a matrix of derivatives whose $l$th column is $\partial \mu_i / \partial \delta_l$. $V_i$ is constructed as $V_i = \phi B_i^{1/2} R (\alpha) B_i^{1/2}$, where $B_i = \text{diag} \left( b_i (\theta_{0i}) \right)$ is a diagonal matrix with main diagonal elements of variance function, $b_i (\theta_{0i})$, and $R$ is a correlation matrix. $V_i$ and $R$ are “working” covariance and correlation to distinguish them from the true covariance and correlation among $Y_i$, respectively. When we use canonical link function, $\partial \theta_i / \partial \eta_i$ is the identity matrix. Let $C_i$ be the matrix of predictor values with $X_i$ and $Z_i$ for subject $i$. By the chain rule,

$$\frac{\partial \mu_i}{\partial \delta} = \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \delta} = B_i \Gamma_i C_i.$$ 

Then the score equations for $\delta$ are

$$U (\delta) = \sum_{i=1}^{n} C_i^T B_i V_i^{-1} \{ y_i - \mu_i (\delta) \}.$$

The expression, $B_i V_i^{-1} \{ y_i - \mu_i (\delta) \}$ can be written as a vector of the residual of each phenotype, $y_i$. Thus, the residual score vector for individual $i$ is defined as:

$$S_i = \begin{pmatrix} S_{i1} \\ \vdots \\ S_{in} \end{pmatrix} = B_i V_i^{-1} \{ y_i - \mu_i \},$$

where $\hat{\mu}_i = g^{-1} (Z_i \hat{y})$ and $\hat{y}$ is estimator obtained from estimating equations under the null hypothesis $B_i, B_i$ and $V_i$ are calculated using $\hat{\mu}_i$. Based on this residual score vector, each individual with phenotypes is discriminated between case and control status. From the residual score vector for individual, we propose the aggregation for elements of the score vector, $S_i = \sum_{j=1}^{t} S_{ib}$, and use that as a prediction score for each individual. If the sum of prediction scores over those individuals who have the corresponding genotype combination is greater than a threshold value, assign ‘high-risk’ to the cell corresponding to the genotype combination. Otherwise, assign ‘low-risk’ to the cell.

**Data**

Our primary outcomes are two types of blood pressure, systolic blood pressure (SBP) and diastolic blood pressure (DBP), and hypertension diagnosis of the Korean Association REsource (KARE) Consortium. The measurements of blood pressure were dichotomized at 140 mmHg for SBP and 90 mmHg for DBP, and denoted by SBPB and DBPB, respectively. We defined the hypertensive case as HP = 1 if SBP ≥ 140 mmHg or DBP ≥ 90 mmHg, and HP = 0, otherwise. Several genome-wide association studies (GWAS) have been performed on blood pressure by treating blood pressure as a quantitative trait [24–29]. In this study, we treated blood pressure as a binary trait HP representing whether the hypertension status is yes or no. Among 8,842 KARE subjects, 1,291 subjects were removed in the analysis due to anti-hypertensive therapy and drug treatments that could influence blood pressure. Additionally, 5 subjects were excluded because of missingness in SBP and body mass index (BMI). Of the 7,546 subjects considered in the study, 4,080 (54%) subjects were from urban community Ansan and the others were from rural community Ansung. For the study, the average age is 48.4 years for Ansan and 55.0 years for Ansung. There are three times of bi-yearly measured hypertensive status from 2001 to 2006, denoted by HP1, HP2, and HP3. Among 7,546 subjects, 2,080 subjects did not follow up at time 2 or 3. Subject characteristics are summarized in Table 1. The genomic DNAs were genotyped using Affymetrix Genome-Wide Human SNP Array 5.0. The quality control procedures were adopted such as missing genotype frequency > 0.5% and minor allele frequency (MAF) ≤ 0.01 at least on area. Finally a total of 7,546 individuals and 344,596 SNPs were included in the analysis of dichotomized SBPB and DBPB, while a total of 5,466 individuals and 344,309 SNPs were included in the analysis of repeatedly measured hypertension status.

**Table 1 Subject characteristics of the KARE.**

| Phenotype                        | N(=7,546) | %   |
|----------------------------------|-----------|-----|
| Recruit area                     |           |     |
| Ansung                           | 3,466     | 45.9|
| Ansan                            | 4,080     | 54.1|
| Gender                           |           |     |
| Male                             | 3,743     | 49.6|
| Female                           | 3,803     | 50.4|
| Systolic blood pressure          |           |     |
| ≥ 140                            | 701       | 9.3 |
| < 140                            | 6,845     | 90.7|
| Diastolic blood pressure         |           |     |
| ≥ 90                             | 693       | 9.2 |
| < 90                             | 6,853     | 90.8|
| Age (years)                      |           | Mean SD |
| Overall                          | 51.4      | 8.79 |
| Ansung                           | 55.0      | 8.82 |
| Ansan                            | 48.4      | 7.51 |
| Body mass index (kg/m²)          |           |     |
| Overall                          | 24.4      | 3.08 |
| Hypertensive cases               |           | %   |
| SBP ≥ 140 or DBP ≥ 90            |           |     |
| HP1 (Time 1)                     | 716       | 13.1|
| HP2 (Time 2)                     | 706       | 12.9|
| HP3 (Time 3)                     | 698       | 12.8|

Abbreviations: DBP, diastolic blood pressure; KARE, Korean Association Resource; SBP, systolic blood pressure.
Results

Preliminary analyses

To compare multivariate analysis with univariate analysis, we first separately fit a logistic regression model for each dichotomized blood pressure measurement SBP_B and DBP_B with covariate adjustment for recruitment area, age, sex, and BMI. The correlation between SBP_B and DBP_B is 0.48. The multivariate analysis with two binary phenotypes (SBP_B, DBP_B) was conducted using the GEE approach. For the repeatedly measured hypertension status HP_1, HP_2, and HP_3, we fit logistic models for each HP and fit the GEE model for three HPs simultaneously. The pairwise correlations range from 0.32 to 0.36. In the GEE model, we assumed two types of genetic effect: homogeneous genetic effect and heterogeneous genetic effect for multivariate phenotypes. However, when we compared the effect sizes and p-values of homogeneous model with those of heterogeneous model, there was no strong evidence for supporting the homogeneous genetic effect. So, we present the results of the GEE model with heterogeneous genetic effects for multivariate phenotypes in both of blood pressures and repeatedly measured hypertension status.

To perform gene-gene interaction analysis using GMDR analyses, we first selected SNPs with strong marginal effects in univariate models and among those, we select the ones with strong effects in multivariate models. For SBP_B and DBP_B analysis, we selected the top 50 SNPs for each SBP_B and DBP_B. From these 100 SNPs, we chose 35 SNPs using a p-value criterion (< 1 × 10^{-5}) from the GEE model. In a similar manner, we chose 34 SNPs for HP_1, HP_2, and HP_3 by selecting the top 50 SNPs for each HP, using the same p-value criterion from their GEE model.

Univariate logistic and multivariate GEE analyses of SBP_B and DBP_B

We report results of GWA studies of dichotomized SBP_B and DBP_B, and their multivariate analyses. For SBP_B and DBP_B, the Manhattan plots are given in Figure 1. As summarized in Table 2, five SNPs for SBP_B (rs1549022, rs2111464, rs12942470, rs2088983, and rs1768145) and three SNPs for DBP_B (rs17045441, rs11866964, and rs7555790) were selected at the 10^{-5} significance level. For multivariate GEE analysis for (SBP_B, DBP_B), six SNPs were selected: rs17045441, rs1378942, rs12942470, rs1549022, rs927833, and rs2111464. Among these six SNPs selected from multivariate GEE analysis, four SNPs were also found by univariate analysis but two SNPs (rs1378942 and rs2111464) were not. A gene CSK in which SNP rs1378942 is located has been reported as a hypertension susceptibility gene in the Korean population [25,26] and also in East Asians [30].

Univariate GMDR and multivariate GMDR analyses of SBP_B and DBP_B

We present GMDR results to discover gene-gene and/or gene-environment interactions. For univariate GMDR analysis, logistic regression models for dichotomized SBP_B and DBP_B were constructed with area, age, sex, and BMI as covariates under the null hypothesis of no genetic effect. For multivariate GMDR analysis, the GEE model with same covariates was constructed. To reduce the computational burden, we focused on 35 SNPs selected from the preliminary analysis. All possible one and two locus models were fit for 35 SNPs. Through 10-fold-cross validation the best combination of loci with maximum train balanced accuracy (BA) which is average of sensitivity and specificity was chosen at each fold. To choose the final model, we considered cross-validation consistency (CVC) among a set of best combinations.

Table 5 summarizes the best model, Train BA, Test BA, and CVC from univariate GMDR and multivariate GMDR. For the purpose of comparison, we computed
Figure 1 Manhattan plots of SBP and DBP in univariate and multivariate analyses. (A) SBP in logistic regression (B) DBP in logistic regression (C) Multivariate model with SBP and DBP (D) Overlay plot of (A)-(C).

Table 2 Selected SNPs of SBP and DBP from univariate and multivariate analyses.

| CHR | SNP   | Gene symbol | SBP Beta | P-value | DBP Beta | P-value | Beta1 | Beta2 | P-value |
|-----|-------|-------------|----------|----------|----------|----------|-------|-------|----------|
| 1   | rs7555790 | PEX14       | 0.117    | 4.16E-03 | 0.184    | 4.46E-06 | 0.046 | 0.116 | 2.35E-05 |
| 2   | rs2111464 |            | 0.200    | 1.11E-06 | 0.100    | 1.28E-02 | 0.293 | 0.195 | 8.77E-06 |
| 2   | rs1549022 |            | 0.207    | 6.52E-07 | 0.111    | 5.89E-03 | 0.295 | 0.202 | 5.23E-06 |
| 3   | rs1768145 |            | 0.169    | 8.24E-06 | 0.090    | 2.01E-02 | 0.233 | 0.161 | 7.95E-05 |
| 4   | rs17045441| ANK2        | 0.065    | 1.06E-01 | 0.199    | 7.69E-08 | -0.090 | 0.058 | 3.91E-08 |
| 4   | rs2088983 |            | 0.168    | 6.96E-06 | 0.090    | 1.82E-02 | 0.234 | 0.162 | 4.54E-05 |
| 15  | rs1378942 | CSK         | -0.189   | 2.50E-05 | -0.192   | 1.85E-05 | -0.167 | -0.182 | 3.49E-06 |
| 16  | rs11866964| ZNF423      | -0.089   | 3.66E-02 | -0.206   | 2.78E-06 | 0.036 | -0.087 | 3.26E-05 |
| 17  | rs12942470|            | 0.186    | 4.36E-06 | 0.041    | 3.12E-01 | 0.326 | 0.180 | 4.25E-06 |
| 20  | rs927833 | LOC100270679| -0.127   | 2.31E-02 | 0.074    | 4.53E-02 | -0.343 | -0.130 | 7.43E-06 |
the p-values from the logistic models and GEE model for the SNPs in one-locus model of GMDR methods. The identified SNPs by GMDR methods also have significant p-values from these analyses: 6.51E-07 for SBP_B, 4.21E-05 for DBP_B, and 3.26E-05 for multivariate phenotypes. The best two-locus model of DBP_B included one SNP, rs1378942, in CSK and another SNP, rs11866964, in ZNF423 implying that the interaction between CSK and ZNF423 genes was identified as a significant contributor to dichotomized DBP_B. The test BAs of the one-locus models (two-locus model) for these SNPs were 0.545 and 0.549 (0.566) for rs1378942 and rs11866964, respectively. The best two-locus model from the multivariate GMDR included rs7555790 in PEX14 gene and rs11077135 in A2BP1 gene. The test BA of the one-locus models (two-locus model) for these SNPs were 0.526 and 0.532 (0.546), respectively. It seems that the contribution was from the joint effects of two genes rather than their main effects. The graphical descriptions for test BA are given in Figure 3. The median of test BA for multivariate GMDR is between median of SBP_B and DBP_B in both one and two-locus models. The distribution of test BA for multivariate GMDR is more concentrated than those of SBP_B and DBP_B.
Univariate GMDR and multivariate GMDR analyses of HP1, HP2, and HP3
The results of the univariate and multivariate GMDR are summarized in Table 6 for the repeatedly measured hypertension status HP1, HP2, and HP3. For these hypertension phenotypes, 34 SNPs selected from the preliminary analysis were included to GMDR mechanisms. All possible one and two locus models were fit for 34 SNPs. Not surprisingly, all different SNPs were identified in one-locus model. For the comparison between GMDR methods and classic method of logistic and GEE models, we report the p-values from the logistic models and GEE model for the identified SNPs from GMDR methods in one-locus models: 1.02E-05 for HP1, 1.59E-05 for HP2, 6.33E-06 for HP3, and 8.50E-05 for multivariate phenotypes. The identified SNPs by GMDR methods also have significant p-values from the classic methods. The best two-locus model from multivariate GMDR included rs7791839 in CCDC129 gene and rs7168365 in WDR72 implying that the interaction between CCDC129 and WDR72 genes was identified as a significant contributor to the repeatedly measured hypertension status. Box plots and density plots of test BA for GMDR and multivariate GMDR of HPs are given in Figure 4. Similar to the results of dichotomized SBPB and DBPB, the test BA for multivariate GMDR had a smaller deviation in the both one-and two-locus models.

Comparison of univariate GMDR and multivariate GMDR
We presented the results of univariate and multivariate GMDR by the same phenotypes in the previous two sub-sessions. However, those comparisons are not significantly meaningful to describe the usefulness of multivariate GMDR. Here, we compared the results from multivariate GMDR of SBPB and DBPB with the results

Table 3 Selected SNPs of longitudinal hypertension from univariate and multivariate analyses.

| CHR | SNP   | Gene symbol | HP1 Beta | P-value   | HP2 Beta | P-value   | HP3 Beta | P-value   | Multivariate Beta | P-value |
|-----|-------|-------------|----------|-----------|----------|-----------|----------|-----------|-------------------|---------|
| 1   | rs4908736 |             | 0.111    | 6.02E-03  | 0.178    | 8.83E-06  | 0.079    | 5.17E-02  | 0.110             | 0.178   |
| 4   | rs17675997 |             | 0.176    | 6.16E-06  | 0.051    | 2.11E-01  | 0.175    | 2.27E-01  | 0.051             | 0.048   |
| 4   | rs2411259  | Loci152578  | 0.176    | 5.33E-06  | 0.051    | 2.06E-01  | 0.174    | 1.28E-01  | 0.051             | 0.061   |
| 5   | rs12054837  | ARSB        | -0.029   | 4.83E-01  | -0.042   | 3.18E-01  | 0.162    | 2.12E-05  | -0.031            | -0.043  |
| 5   | rs294082   |             | 0.067    | 1.02E-01  | 0.087    | 3.28E-02  | 0.181    | 5.80E-06  | 0.068             | 0.087   |
| 5   | rs17677051 |             | -0.086   | 3.79E-02  | -0.188   | 7.84E-06  | -0.089   | 3.16E-02  | -0.081            | -0.188  |
| 5   | rs4867707  |             | -0.088   | 3.22E-02  | -0.189   | 7.00E-06  | -0.091   | 2.77E-02  | -0.083            | -0.188  |
| 5   | rs4084097  |             | 0.163    | 9.61E-06  | -0.004   | 9.27E-01  | 0.092    | 1.68E-02  | 0.019             | 0.005   |
| 7   | rs7751214  | EPHA7       | -0.191   | 9.16E-06  | -0.009   | 8.26E-01  | 0.099    | 1.85E-02  | 0.008             | 0.100   |
| 8   | rs4495407  |             | 0.038    | 3.60E-01  | 0.012    | 7.74E-01  | 0.185    | 8.40E-06  | 0.036             | 0.189   |
| 8   | rs10956596 |             | -0.044   | 2.82E-01  | -0.047   | 2.58E-01  | -0.185   | 8.82E-06  | -0.043            | -0.188  |
| 8   | rs6470947  |             | 0.053    | 1.94E-01  | 0.023    | 5.69E-01  | 0.187    | 6.69E-06  | 0.053             | 0.230   |
| 8   | rs4615555  |             | 0.051    | 2.17E-01  | 0.030    | 4.69E-01  | 0.191    | 3.81E-06  | 0.049             | 0.209   |
| 8   | rs4279577  |             | 0.052    | 2.06E-01  | 0.031    | 4.65E-01  | 0.192    | 3.44E-06  | 0.051             | 0.196   |
| 8   | rs746333   |             | 0.050    | 2.31E-01  | 0.031    | 4.57E-01  | 0.189    | 6.33E-06  | 0.048             | 0.193   |
| 11  | rs550214   |             | 0.081    | 4.39E-02  | 0.075    | 3.23E-02  | 0.175    | 6.09E-06  | 0.102             | 0.077   |
| 15  | rs11636344 | FBN1        | 0.075    | 5.81E-02  | 0.167    | 6.51E-06  | 0.035    | 3.88E-01  | 0.073             | 0.166   |
| 16  | rs17722281 | WWOX        | -0.142   | 7.68E-04  | -0.160   | 1.52E-04  | 0.034    | 4.16E-01  | -0.140            | 0.034   |

Table 4 Transition of hypertensive case over time.

| HP1 Time 1 (716) | HP2 Time 2 (706) | HP3 Time 3 (288) | HP3 Time 3 (410) |
|------------------|------------------|------------------|------------------|
| Hypertension     | Normal           | Hypertension     | Normal           |
| 166              | 154              | 147              | 239              |
| Normal           | Hypertension     | Normal           |                  |
| 122              | 274              | 263              | 410              |

Note that numbers within parentheses are the number of hypertensive case at each time point.

Table 5 Comparison of results for SBP and DBP by GMDR and multivariate GMDR.

| No. of Loci | Method   | Best model   | Train BA | Test BA | CVC |
|------------|----------|--------------|----------|---------|-----|
| 1          | GMDR_SBP | rs1549022    | 0.544    | 0.544   | 6   |
|            | GMDR_DBP | rs11077135   | 0.548    | 0.547   | 7   |
|            | Multivariate GMDR | rs11866964 | 0.539 | 0.536 | 8   |
| 2          | GMDR_SBP | rs2111464, rs12942470 | 0.566 | 0.566 | 7   |
|            | GMDR_DBP | rs1378942, rs11866964 | 0.566 | 0.566 | 3   |
|            | Multivariate GMDR | rs7555790, rs11077135 | 0.551 | 0.546 | 2   |
from the GMDR of HP, including the same individuals and candidate SNPs (Table 7). Because hypertension was defined by SBP or DBP, we can directly compare the performance of multivariate GMDR and univariate GMDR through those analyses. Multivariate GMDR and GMDR yielded the same best two-locus model. However, multivariate GMDR shows slightly better test BA than GMDR. Box plots of test BA for multivariate GMDR and GMDR from those two analyses are given in Figure 5. The test BA of multivariate model has smallest deviation also.

**Conclusions**

In this paper, we have developed multivariate analysis for discovering gene-gene interaction, namely multivariate GMDR. Our multivariate GMDR analysis was developed by utilizing a GEE approach to multivariate phenotypes. Many studies emphasized the importance and the increase of power for multivariate analysis in GWAS [33-35]. Although MDR method has been developed in variety of manners [5-9], there have been no extensions to the multivariate analysis. We proposed multivariate GMDR analysis by utilizing the GEE model to calculate the prediction score to be a tool for reducing the multifactor dimensionality. The GEE approach is an extension of generalized linear models to the longitudinal data and handles both discrete and continuous phenotypes. Thus, our multivariate GMDR can be applicable to both discrete and continuous phenotypes.

**Table 6 Comparison of results for longitudinal hypertension by GMDR and multivariate GMDR.**

| No. of Loci | Method       | Best model          | Train BA | Test BA | CVC |
|-------------|--------------|---------------------|----------|---------|-----|
| 1           | GMDR_HP1     | rs11097953          | 0.542    | 0.543   | 9   |
|             | GMDR_HP2     | rs11115097          | 0.545    | 0.546   | 5   |
|             | GMDR_HP3     | rs7465333           | 0.540    | 0.542   | 5   |
|             | Multivariate GMDR | rs7168365       | 0.529    | 0.528   | 9   |
| 2           | GMDR_HP1     | rs11097953, rs7751214 | 0.555    | 0.540   | 6   |
|             | GMDR_HP2     | rs11115097, rs1772281 | 0.566    | 0.566   | 8   |
|             | GMDR_HP3     | rs7791839, rs6470947 | 0.563    | 0.563   | 9   |
|             | Multivariate GMDR | rs7791839, rs7168365 | 0.544    | 0.544   | 7   |
Though real GWAS data analysis, we investigated the properties of multivariate GMDR. Firstly, the result of multivariate GMDR does not always coincide with that of GEE approach. That is, the best SNP set selected by multivariate GMDR does not always have the smallest p-value from GEE model. In our analysis, note that the SNP set selected by multivariate GMDR still tends to have quite a small p-value. Secondly, the test BAs of the multivariate GMDR is not always larger than those of univariate GMDR. As shown in Figures 3 to 5, the distribution of test BAs from the multivariate GMDR is different from those of univariate GMDR. The test BAs of multivariate GMDR are more densely distributed with a smaller standard deviation than those of univariate GMDR. Thus, a direct comparison of test BAs between multivariate GMDR and univariate GMDR may lead a misleading conclusion.

The proposed multivariate GMDR can be extended in many different ways. The modified version BAs which takes account for the distributional difference is expected to improve the performance of multivariate GMDR. The testing procedure using the modified BAs under the null distribution would enable us to demonstrate the increase of power of multivariate GMDR. A prediction score is defined as the sum of elements of

| No. of Loci | Method                          | Best model          | Train BA | Test BA | CVC |
|------------|---------------------------------|---------------------|----------|---------|-----|
| 1          | Multivariate GMDR with BPs      | rs11866964          | 0.539    | 0.536   | 9   |
|            | GMDR with HP1                   | r54811719           | 0.542    | 0.541   | 4   |
| 2          | Multivariate GMDR with BPs      | rs1338574, r4811719 | 0.560    | 0.557   | 7   |
|            | GMDR with HP1                   | rs1338574, r4811719 | 0.560    | 0.554   | 7   |
the score vector from GEE model. We are currently working on several different weighting schemes for accounting various relationships between phenotypes. The weighted prediction score is also expected to improve the performance of multivariate GMDR. In the future studies, all these extensions will be evaluated through extensive simulation studies.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
JC and TP designed the study and JC carried out statistical analysis. TP coordinated the study. JC and TP wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MISP)(2012R1A1A2026438, 2008-0062618).

Figure 5 Graphical descriptions for Test BA from univariate GMDR for hypertension at time 1 (HP1) and multivariate GMDR for SBP and DBP. (A) Box plots of Test BA in one-locus model (B) Box plots of Test BA in two-locus model (C) Density plots of Test BA in one-locus model (D) Density plots of Test BA in two-locus model.

Declarations
The publication cost for this article was supported by the Seoul National University.

This article has been published as part of BMC Systems Biology Volume 7 Supplement 6, 2013: Selected articles from the 24th International Conference on Genome Informatics (GIW2013). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/7/S6.

Published: 13 December 2013

References
1. Foukkes AJ: Applied Statistical Genetics with R: For Population-based association Studies Springer; 2009.
2. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, et al: A genome-wide search for human type 1 diabetes susceptibility genes. Nature 1994, 371:130-136.
3. Kraft P, Yen YC, Stram DO, Morris J, Gauderman WG: Exploiting gene environment interaction to detect genetic associations. Hum Hered 2007, 63:111-119.
4. Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics 2005, 37:413-417.
28. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2003; 73:183-194.

26. Hong KW, Jin HS, Lim JE, Lee JY, Han BG, Hwang SY, Lee SH, et al. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003; 25:150-157.

24. Hong KW, Jin HS, Cho YS, Lee JY, Lee JE, Cho NH, et al. Power of multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003; 19:376-382.

22. Liu J, Pei Y, Papasian CJ, Deng HW, et al. Multifactor dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics* 2007; 23:2589-2595.

20. Liang K, Zeger SL, et al. Longitudinal binary responses. In *Generalized Linear Models* by Davidian M, Harville DA, 1996; pp. 469-501.

18. Laird NM, Ware JH, et al. Multivariate analysis of mixed data using generalized estimating equations. *Biometrics* 1982, 38:631-641.

16. Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD, et al. Power of multifactor dimensionality reduction method for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003; 25:150-157.

14. Calle ML, Uretas V, Velikova G, Malats N, Van Steen K, et al. Finding the missing heritability of complex diseases. *Nature* 2009; 461:747-753.

12. Chung Y, Lee SY, Elston RC, Park T, et al. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics* 2007; 23:71-76.

10. Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der A DL, Feskens EJ, et al. Evidence for the multifactor-dimensionality reduction method to detect gene-gene interactions in hypertension. *Am J Hum Genet* 2007, 80:1125-1137.

8. Ritchie MD, Hahn LW, Moore JH, et al. Multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003; 25:150-157.

5. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2003; 73:183-194.

3. Hong KW, Jin HS, Lim JE, Lee JY, Cho YS, Lee JE, et al. Genetic variations in ATP2B1, CSK, ARSG and CSMD1 loci are related to blood pressure and/or hypertension in two Korean cohorts. *J Hum Genet* 2010, 55:336-341.

1. Hong KW, Jin HS, Lim JE, Kim S, Go MJ, Oh B, et al. Recapitulation of two genomewide association studies on blood pressure and essential hypertension in the Korean population. *J Hum Genet* 2010; 55:336-341.

2. Wang Y, O’Connell JR, McArthu HF, Wade JE, Dorff SE, Shah SJ, et al. From the cover: whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc Natl Acad Sci USA* 2009; 106:226-231.

1. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009; 41:666-676.

Cite this article as: Choi and Park: Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions. *BMC Systems Biology* 2013; 7(Suppl 6):S15.