Bioinformatics, 32, 2016, i605–i610
doi: 10.1093/bioinformatics/btw424
ECCB 2016

OXFORD

A unified model based multifactor dimensionality reduction framework for detecting gene–gene interactions

Wenbao Yu[1], Seungyeoun Lee[2] and Taesung Park[1,*]

[1]Department of Statistics, Seoul National University, Shilim-Dong, Kwanak-Gu, Seoul 151-742, Korea and [2]Department of Mathematics and Statistics, Sejong University, Seoul 143-747, Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Gene–gene interaction (GGI) is one of the most popular approaches for finding and explaining the missing heritability of common complex traits in genome-wide association studies. The multifactor dimensionality reduction (MDR) method has been widely studied for detecting GGI effects. However, there are several disadvantages of the existing MDR-based approaches, such as the lack of an efficient way of evaluating the significance of multi-locus models and the high computational burden due to intensive permutation. Furthermore, the MDR method does not distinguish marginal effects from pure interaction effects.

Methods: We propose a two-step unified model based MDR approach (UM-MDR), in which, the significance of a multi-locus model, even a high-order model, can be easily obtained through a regression framework with a semi-parametric correction procedure for controlling Type I error rates. In comparison to the conventional permutation approach, the proposed semi-parametric correction procedure avoids heavy computation in order to achieve the significance of a multi-locus model. The proposed UM-MDR approach is flexible in the sense that it is able to incorporate different types of traits and evaluate significances of the existing MDR extensions.

Results: The simulation studies and the analysis of a real example are provided to demonstrate the utility of the proposed method. UM-MDR can achieve at least the same power as MDR for most scenarios, and it outperforms MDR especially when there are some single nucleotide polymorphisms that only have marginal effects, which masks the detection of causal epistasis for the existing MDR approaches.

Conclusions: UM-MDR provides a very good supplement of existing MDR method due to its efficiency in achieving significance for every multi-locus model, its power and its flexibility of handling different types of traits.

Availability and implementation: A R package "umMDR" and other source codes are freely available at http://statgen.snu.ac.kr/software/umMDR/.

Contact: tspark@stats.snu.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Detecting gene–gene interaction (GGI) or epistasis has been recognized as one of the most effective remedies in genome-wide association studies (GWAS) for explaining missing heritability (Eichler *et al.*, 2010; Mackay, 2014). Many efficient approaches have been proposed for GGI analysis and there are generally two types of approaches: model-based approaches and model-free approaches. The model-based approaches assume a statistical model between genotype and phenotype, such as those in Park and Hastie (2008), Wan *et al.* (2010), Wu *et al.* (2009), Yang *et al.* (2009) and Zhang and Liu (2007), while the model-free approaches often have no prior assumptions about the model and data, such as in Dong *et al.* (2008), Li *et al.* (2014), Ritchie *et al.* (2001) and Zhang *et al.* (2010). Van Steen (2012) and Wang *et al.* (2011) provides comprehensive discussions of these approaches.

Among these approaches, the multifactor dimensionality reduction (MDR) approach (Ritchie *et al.*, 2001) is a very popular non-parametric combinatorial approach that reduces the number of dimensions by converting a high-dimensional multi-locus model into a 1D model for case-control studies. MDR reduces multiple genotype combinations into two groups—high risk (H) and low risk (L). There are a number of extensions of MDR: quantitative MDR (QMDR) for quantitative traits (Gui *et al.*, 2013), generalized MDR (GMDR) for both quantitative and binary traits (Lou *et al.*, 2007), Surv-MDR and Cox-MDR for survival data (Gui *et al.*, 2011; Lee *et al.*, 2012), FAM-MDR for family data (Cattaert *et al.*, 2010; Lou *et al.*, 2008), GEE-MDR and Muti-MDR for multivariate phenotypes (Choi and Park, 2013; Yu *et al.*, 2015) and etc. Through these extensions, MDR-based approaches have been demonstrated great power in broad applications of identifying high-order interactions.

However, MDR has several shortcomings. First, it is hard to obtain the significance of a multi-locus model without a subsampling procedure such as permutation, which introduces a heavy computational burden. Second, MDR does not distinguish marginal effects from pure interaction effects. To account for such kinds of problems, Calle *et al.* (2008) and Cattaert *et al.* (2011) have proposed a model-based MDR approach (MB-MDR) that first classifies all the genetic combination cells into H or L and an intermediate group based on statistical testing, and then evaluate the significance of each genetic model through another testing. Unfortunately, MB-MDR needs a lot of tests even in the classification stage, and the significance of the final model still needs a computationally intensive permutation scheme.

To circumvent the above drawbacks, we propose a novel two-step unified model MDR approach (UM-MDR). UM-MDR first classifies all the genetic combination cells into H and L groups using a simple classification rule as in MDR-based approaches and then evaluates the significance of each genetic model by regression or a penalized regression framework. There are three main differences between our approach and MB-MDR. (i) We avoid using significant tests in the classification step, and instead we use simple classification rules as in traditional MDR approaches. (ii) We suggest applying ridge regression or logistic ridge regression to adjust the marginal effects. (iii) Finally, a simple and easily computed semi-parametric procedure is proposed for correcting the raw *P*-values from the regression model instead of an intensive permutation process.

2 Method: UM-MDR

In this section, we introduce UM-MDR for detecting GGI. Basically, this framework includes two steps, namely, a classification (or dimension reduction) step and a modeling step. The classification step reduces the multi-level genotype combinations into a 1D variable, as the traditional MDR approaches do. The modeling step evaluates the significance of the GGI model, while adjusting for the covariate effects and/or the marginal effects simultaneously.

First, we classify each genotype combination (cell) into high (H) or low (L) risk group, and we use S, which stands for the H/L status. There are several methods available for classification. For example, for a dichotomous trait, we can classify a cell into H(L) group if the ratio of the number of case to the number of control in the cell is greater (smaller) than a given threshold, e.g. the global ratio of the numbers of cases and controls. For a quantitative trait, we can simply classify a cell into H(L) group if the mean value of the trait in that cell is greater (smaller) than a given threshold, such as the global mean, as QMDR does.

Second, we can model the effect of the H/L status S defined in the first stage by fitting a generalized linear model:

$$g(\mu) = \alpha_0 + \ \beta S + \ \gamma^T \boldsymbol{X}, \tag{1}$$

where μ is the mean vector of phenotype Y, $g(\cdot)$ is the link function, and $\boldsymbol{X}$ is the covariate vector such as some environmentally related variables. The model parameters can be estimated by maximum likelihood (ML) estimation.

The main idea of the UM-MDR is to define an indicator variable, say S, for high- or low-risk groups classified by the given v-order single nucleotide polymorphism (SNP) pair and to test the significance of effect of S on the response variable to check the GGI, instead of selecting the best combination of SNPs by the intensive cross-validation procedure. Since S is an indicator of the high or low risk groups which is reduced from the multi-level genetic combination cells constructed by the given v-order SNP pair, the non-zero effect of S implies that there is an interactive effect of the corresponding pair of SNPs on the response variable. Therefore, we can easily obtain the significance of the corresponding multi-locus model by testing the null hypothesis $\mathrm{H}_0: \ \beta = 0$, in which we check the association between S and the phenotype Y.

To account for the strong marginal effect, we can modify Model (1) to

$$g(\mu) = \alpha_0 + \ \beta S + \ \gamma^T \boldsymbol{X} + \sum_{i=1}^{v} \alpha_i SNP_i. \tag{2}$$

A potential problem of (2) is that Y and some SNPs may be highly correlated, and so the ordinary ML estimator may suffer from the multi-collinearity problem. To overcome such problem, we recommend using the ridge regression model (Cule *et al.*, 2011; Vago and Kemeny, 2006).

The advantages of UM-MDR include the following. (i) The significance of every genetic model is easily obtained through the standard regression analysis (or regularized regression), while adjusting for the covariant effects and/or the marginal effects. (ii) The flexibility is reflected through the following two aspects: it can handle both quantitative traits and qualitative traits, and a lot of existing classification methods can be used to define H/L in the first step.

The UM-MDR procedure is summarized as follows:

i. For each given v-order SNP combination, classify the genotype combination cell into H(L), and let S stands for the H/L status.
ii. Fitting either model (1) or model (2) and then testing $\mathrm{H}_0: \ \beta = 0$ using the following model based framework.

2.1 Ridge regression and logistic ridge regression

We use the ridge regression (for a quantitative trait) or logistic ridge regression (for a binary trait to adjust for the marginal effects (Cule *et al.*, 2011; Vago and Kemeny, 2006). Specifically, consider a standard linear regression model as follows:

$$Y = Z\beta + \epsilon.$$

Here Y and Z stand for the response vector and design matrix of the covariate variables, respectively, and we assume that $\epsilon \sim N(0, \ \sigma^2 I)$. The ordinary least squares estimator for β is $(Z^{'}Z)^{-1} Z^{'}Y$ and the ridge regression estimator is $\widehat{\beta}^{\lambda} = (Z^{'}Z + \lambda I)^{-1} Z^{'}Y$, where λ a positive tuning parameter, and I is the identity matrix, and note that

$$\mathrm{var}\left(\widehat{\beta}^{\lambda}\right) = \sigma^2(Z'Z + \lambda I)^{-1}Z'Z(Z'Z + \lambda I)^{-1}.$$

For a binary response variable, the logistic regression model is defined as

$$\log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = X\beta.$$

Then the corresponding ridge regression estimator can be found by maximizing the log-likelihood function with a L_2 norm penalty of β, equivalently,

$$\widehat{\beta}^{\lambda} = \mathrm{argmin}_{\beta}\ \{-\log(L(\beta)) + \lambda||\beta||_2\},$$

where L(β) is the likelihood function. The Newton-Raphson algorithm can be used to find $\widehat{\beta}^{\lambda}$ and the variance is estimated by

$$\mathrm{var}\left(\widehat{\beta}^{\lambda}\right) = (Z'WZ + 2\lambda I)^{-1}Z'WZ(Z'WZ + 2\lambda I)^{-1},$$

where $W = \mathrm{diag}[\widehat{p}_i(1-\widehat{p}_i)]$ and $\widehat{p}_i = \frac{e^{X_i\widehat{\beta}^{\lambda}}}{1+e^{X_i\widehat{\beta}^{\lambda}}}$.

2.2 Significance test of a multi-locus model and a semi-parametric *P*-value correction procedure

In step 2, we use the Wald type statistic $W = \widehat{\beta}^2/\mathrm{var}\left(\widehat{\beta}\right)$ for testing the significance of a multi-locus model, but the null distribution may not necessarily follow the chi-square distribution (Calle *et al.*, 2008). Instead of using a computationally intensive permutation process, we suggest the following semi-parametric procedure to correct the raw *P*-value. First, assume that the null distribution follows a non-central chi-square distribution with degree of freedom one and estimate the non-central parameter through only a few numbers of permutations (e.g. 5 or 10). Then re-calculate the *P*-value based on the non-central chi-square distribution. This approach avoids the heavy computational cost compared to the conventional permutation method for estimating *P*-values. The rationale of this procedure can be justified as follows: consider a very simple regression model defined as

$$Y_i = \alpha + \beta S_i + \varepsilon_i,$$

for which the least square estimator is given as $\widehat{\beta} = N(\bar{Y} - \widehat{Y}_H)/N_L$, where $\bar{Y}$ and $\widehat{Y}_H$ are the global mean and mean of H group, and N and N_L are the total sample size and sample size of the L group, respectively. Therefore, testing $\beta = 0$ is equivalent to testing $E(Y) = E(Y_H)$. Note that for quantitative trait, like QMDR, we classify cell into H if its mean is larger than the global mean, which means that we have implicitly set $\widehat{Y}_H > \bar{Y}$ in the first step. Consequently, the null distribution of z-statistic $\widehat{\beta}/\sqrt{\mathrm{var}(\widehat{\beta})}$ is presumed to have a nonzero mean due to the classification step. Therefore, it is natural to assume that the z-statistic follows approximately a normal distribution with a nonzero mean and standard deviation 1, i.e. the statistic W follows a non-central chi-square distribution with one degree of freedom and non-centrality parameter k. Because the mean of the non-central chi-square distribution is $k+1$, we can estimate the non-centrality parameter as $\widehat{k} = \max(0,\ \widehat{\mu} - 1)$, where $\widehat{\mu}$ is estimator for the mean of W under the null distribution. To estimate $\widehat{\mu}$, we can permute the trait a few times, say 5 or 10, repeat step 1 and step 2, and take the sample mean for W statistic as $\widehat{\mu}$. We can estimate the non-centrality parameter for each multi-locus model or pool all the statistics and then estimate the common non-centrality parameter for all multi-loci models.

Note that the proposed semi-parametric correction procedure does not necessarily need a very large number of permutations as the conventional method of obtaining the null distribution by permutation. Through the following simulation studies, we demonstrate that such a correction procedure controls the Type I error rate well.

3 Simulation studies

3.1 Type I error

In this section, we first check whether the Type I error rate is well controlled for in the proposed two-step approach. To do this, we randomly generate two SNPs and a binary (or quantitative) trait under the null hypothesis of no association. We permute the phenotype variable five times to estimate the non-centrality parameter of the null distribution in all the following simulation studies. The sample size is 400 and 100 data sets are simulated under various minor allele frequencies (MAFs). The nominal size is set to be 0.05 and the proportion of the corrected *P*-values that are smaller than the nominal size is defined as the Type I error rate. As shown in Table 1, the Type I error rates are well controlled after using the proposed semi-parametric correction procedure. We also present the QQ plot of the raw and corrected *P*-value in Figure 1.

We then study whether the proposed method can identify the causal interaction in different contexts, such as binary or quantitative trait with/without marginal effects. We also examine whether our approach can avoid detecting those multi-locus models just because one locus has a significant marginal effect. For comparison, MDR or QMDR analysis are also performed correspondingly. We consider the following four different scenarios: Case 1) binary trait without marginal effect, Case 2) quantitative trait without marginal effect, Case 3) quantitative trait with a marginal effect SNP and Case 4) a three-order GGI analysis.

Table 1. Type I error rates

	Binary trait		Quantitative trait	
MAF	Raw	Corrected	Raw	Corrected
0.05	0.14	0.03	0.14	0.04
0.10	0.26	0.04	0.22	0.04
0.20	0.38	0.03	0.39	0.02
0.30	0.51	0.03	0.51	0.02
0.40	0.65	0.02	0.63	0.03

Raw and Corrected stand for using the raw *P*-value and the proposed corrected *P*-value, respectively.

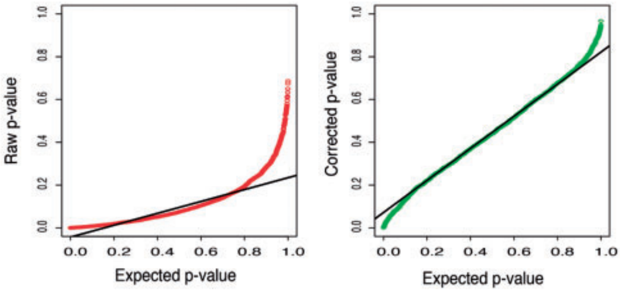

Fig. 1. QQ plots for the raw *P*-value (left panel) and the corrected *P*-value (right panel)

3.2 Case 1: binary trait without marginal effect

In this scenario, we generate data by following the model introduced in Velez *et al.* (2007). 70 different penetrance functions and a probabilistic relationship between the trait and SNPs, are used, where the trait is dependent on two SNPs, say SNP_1 and SNP_2, in the absence of any marginal effects. There are two different MAFs 0.2 and 0.4 and seven different heritability 0.01, 0.02, 0.05, 0.1, 0.2, 0.3 and 0.4. Five models for each of these 14 combinations are generated. We generate m (i.e. 10 or 20) SNPs, but we only present the results for 10 SNPs, since very similar patterns are found for the different numbers of SNPs. All the simulation results for 20 SNPs are shown in Supplementary Figures S1–S4.

The power of UM-MDR can be defined as the proportion of the P-values (after Bonferroni correction) for the causal model that are less than a given nominal size, for instance, 0.05. We denote such power as *PBonf*. On the other hand, the power of the traditional MDR approach is defined as a successful detection rate in identifying the true causal model as the best model, so it is not feasible to compare the power of the traditional MDR with that of our approach directly. For a fair comparison, we suggest defining another power of UM-MDR denoted as *PRank* as the rate of the causal model with the smallest P-value among all possible multi-locus models. We present both powers *PBonf* and *PRank* for UM-MDR in Figure 2. It is clear that *PRank* and the power of the traditional MDR approach showed very similar patterns, with MDR approach providing slightly higher powers for some models. However, the *PBonf* showed quite low power for many cases among the 70 models (Fig. 2).

3.3 Case 2: quantitative trait without marginal effects

In this scenario, we consider quantitative traits instead of binary traits. The phenotype is generated in the same way as in (Gui *et al.*, 2013), that is, $Y|(\mathrm{SNP}_1 = i,\ \mathrm{SNP}_2 = j) \sim N(\mu f_{ij},\ 1)$, where μ is set to be 1 and $f_{ij} = P(\text{high risk}|\mathrm{SNP}_1 = i,\ \mathrm{SNP}_2 = j)$ is the penetrance function, which is the same penetrance function used in Case 1. From Figure 3, we can see similar pattern as in Case 1; UM-MDR, especially when *PRank* is used, can achieve almost the same power as QMDR.

We also calculate the ratio of the computation time for QMDR to that for UM-MDR, which has mean 1.06 across the 70 models, with 95% CI (1.01, 1.11), suggesting that the two approaches take similar computation times. However, since the QMDR approach only provides the best model without evaluating its significance, an extra conventional permutation procedure (Gui *et al.*, 2013) is needed to obtain the significance of the selected model, which reruns the QMDR procedure for a large number times, for instance 1000 times to achieve a P-value as small as 0.001, indicating that QMDR takes 1000 times more computation cost than that of UM-MDR.

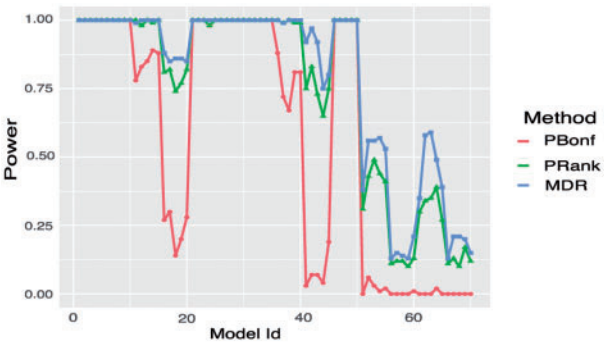

Fig. 2. Power of *PBonf, PRank and MDR* for Case 1 over 70 models

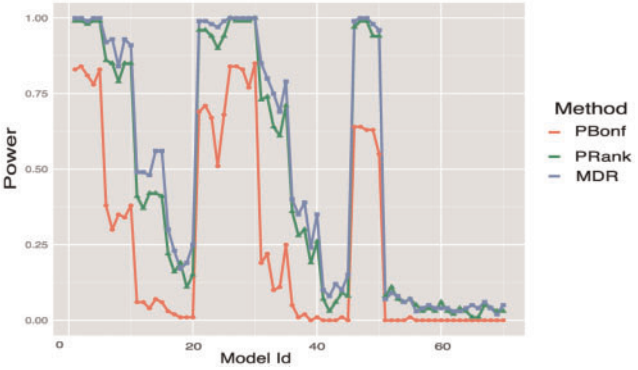

Fig. 3. Power of *PBonf, PRank and QMDR* for Case 2 over 70 models

3.4 Case 3: quantitative trait with a marginal SNP effect

In this case, we consider a scenario in which a 'noisy' SNP, say, SNP_3, exists and it has only a marginal effect. This is a very realistic scenario, since there are possibly many SNPs, which have only marginal effects and the true epitasis is less frequently observed. This simulation is designed to check whether either our approach or the MDR approach can detect the causal multi-locus model ($\mathrm{SNP}_1,\ \mathrm{SNP}_2$). Here, the trait is generated as $Y|(\mathrm{SNP}_1 = i,\ \mathrm{SNP}_2 = j) \sim N(\mu f_{ij},\ 1) + N(\alpha \mathrm{SNP}_3,\ 1)$, and we set $\mu = \alpha = 1$. Figure 4 shows that UM-MDR, especially with *PRank*, is very powerful in identifying the causal interaction model, while MDR, as expected, fails completely to identify the causal model for all 70 models.

3.5 Case 4: a three-order GGI analysis

We consider the three-order interaction model introduced by Ritchie *et al.* (2001) for a binary trait. For certain genotype combinations, there is an increased disease risk. We use a similar strategy as in Case 3 to make the penetrance function for quantitative traits and add a marginal effect of another SNP. That is, we generate $Y|(\mathrm{SNP}_1 = i,\ \mathrm{SNP}_2 = j, \mathrm{SNP}_3 = k) \sim N(\mu f_{ijk},\ 1) + N(\alpha \mathrm{SNP}_4,\ 1)$, where f_{ijk} is the penetrance function, which is specified in the Supplementary document, and we set $\mu = 1,\ \alpha = 0.5$. The results are summarized in Figure 5. Although all of these three powers are very low when MAF is small, UM-MDR with *PRank* always gives highest power in identifying the causal three-order model across all MAF values, especially when the MAF is large.

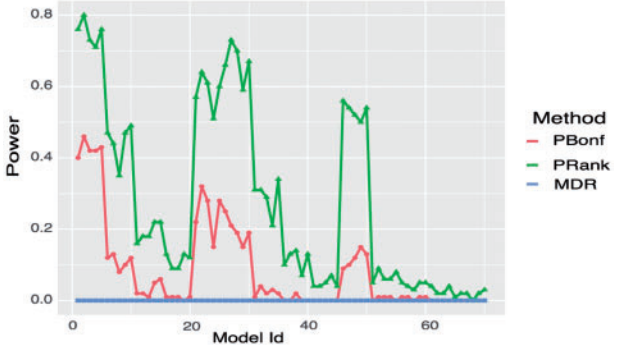

Fig. 4. Power of *PBonf, PRank and QMDR* for Case 3 over 70 models

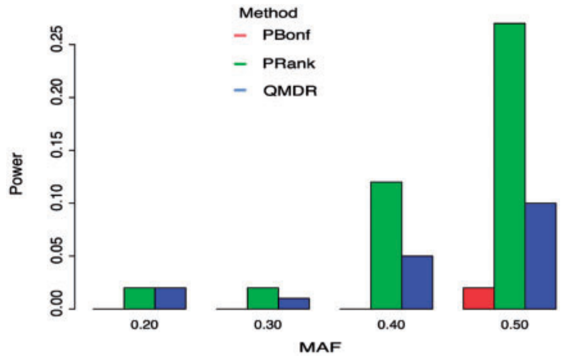

Fig. 5. Power of *PBonf, PRank and QMDR* for Case 4: a three-order model

In summary, UM-MDR approach with *PRank* can achieve at least the same power as MDR (or QMDR) for most scenarios, and it outperforms MDR (or QMDR) approaches when there are some SNPs only having marginal effects (Case 3) and/or when there is higher-order causal model (Case 4). We have shown that while the computation times for UM-MDR and QMDR are similar, UM-MDR provides significance for each multi-locus model automatically while QMDR does not. To present the significance, QMDR needs an extra computationally intensive permutation scheme, which results in QMDR taking a computation time exceeding that of UM-MDR by a thousand times.

4 Real example

We analyze a real dataset from the Korean Association Resource project to demonstrate the proposed approach. Three phenotypes, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol and triglyceride were measured. A total of 8581 samples are available after removing the subjects with at least one missing phenotype value. The genomic DNAs are genotyped using Affymetrix Genome-Wide Human SNP Array 5.0. For our GGI analysis, we only use 19 candidate SNPs identified in an earlier study (Willer *et al.*, 2008) from the single SNP GWAS analysis. For the purpose of demonstration, we use HDL in the GGI analysis.

In the second step of UM-MDR, we also consider the covariate adjustments for sex, age, and recruitment area. Figure 6 displays the *P*-value profile for all second-order genetic models and shows that the significance of a multi-locus model depends largely on adjusting for the marginal effects. After a Bonferroni correction, a lot of significant models were identified without adjusting for the marginal effects whereas no significant model was found with a marginal effect adjustment. This implies that the significance of most models identified without considering marginal effects might be false-positive.

Table 2 displays the top five second-order interaction models after marginal effect adjustment. The top five models were significant at the 5% significance level, although these models were not significant after multiple testing adjustments. These models may have a higher chance of detecting the true epitasis, since we have already adjusted for the marginal effects. On the other hand, the QMDR method identifies a pair of SNPs (rs12596776, rs17321515) as the best second-order model with Cross Validation Consistency (CVC) = 6. However, the *P*-value of this model is estimated to be 0.34 by UM-UMDR. This model is selected as the best model by QMDR perhaps due to the strong marginal effect of the SNP (rs12596776). For comparison, we also list the top detected models when marginal effects adjustment is not considered (Table 3). As shown in Tables 2 and 3, the top five models are quite different and the corresponding *P*-values are substantially different. This implies that it is difficult to detect the significant epistasis when there are SNPs with strong marginal effects.

Table 2. Top 5 second-order interaction models for real example study, with marginal adjustment

Top	SNP1	SNP2	*P*-value
1	rs2156552 (FH0D3)	rs17145738 (TBL2)	0.0018
2	rs2144300 (GALNT2)	rs2338104 (KCD10)	0.0128
3	rs2144300 (GALNT2)	rs1748195 (DOCK7)	0.0267
4	rs2338104 (KCD10)	rs17145738 (TBL2)	0.0303
5	rs2144300 (GALNT2)	rs10402271 (ZNF107)	0.0322

The corresponding gene name related to each SNP is presented in the parentheses.

Table 3. Top 5 second-order interaction models for real example study, without marginal adjustment

Top	SNP1	SNP2	*P*-value
1	rs780049 (CNTNAP5)	rs10402271 (ZNF107)	5.62×10^{-9}
2	rs780049 (CNTNAP5)	rs12596776 (SLC12A3)	7.19×10^{-9}
3	rs780049 (CNTNAP5)	rs4149268 (ABCA1)	8.44×10^{-9}
4	rs780049 (CNTNAP5)	rs1566439 (HPR)	1.18×10^{-8}
5	rs17321515 Unknown)	rs12596776 (SLC12A3)	1.82×10^{-8}

The corresponding gene name related to each SNP is presented in the parentheses.

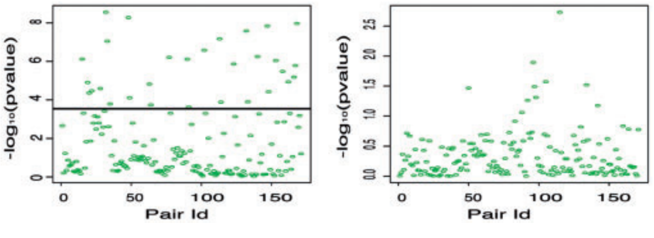

Fig. 6. Plot of *P*-value over all possible SNP pairs in real example: negative log(*P*-value) plots in the real example study for all second-order multi-locus models without marginal effect adjustment (left panel) and with marginal effect adjustment (right panel). The sold blank horizontal line corresponds to the significant level 0.05 after Bonferroni correction

5 Discussion

We proposed a UM-MDRto overcome the shortcomings of the existing MDR approaches for GGI analysis. The main contribution of UM-MDR is to provide a *P*-value for the testing of GGI with a simple and effective semi-parametric procedure instead of a computationally intensive permutation. The UM-MDR can handle flexibly different types of traits by adjusting for the marginal effects along with covariate effects through a regularized regression approach, the ridge regression or logistic ridge regression, and by correcting the raw *P*-value through a simple and effective semi-parametric procedure, which avoids using the conventional computationally intensive permutation method to achieve the *P*-value of every multi-locus model. Through simulation studies, UM-MDR can provide at least the same power as MDR for most scenarios and outperforms MDR for some cases, especially when there are some SNPs with only marginal effects, which can mask the detection of causal epistasis.

The classification step of our approach has many options, in that all classification strategies used by MDR and its extensions can be applied. Consequently, the proposed UM-MDR can provide the significance of many existing MDR extensions for different types of traits, even for multivariate phenotypes, as in Choi and Park (2013) and Yu *et al.* (2015).

In the second step of UM-MDR, we use an alternative way of modeling by reversing the roles of response (Y) and H/L status (S), that is, we treat S as the response variable and Y as the explanatory variable. Such modeling enjoys more flexibility because it uses a logistic regression framework for all different types of traits, which can even handle multivariate traits. Furthermore, this kind of multivariate analysis can also present significance for each trait on every multi-locus model and provide more information when detecting epistasis.

Acknowledgements

Thanks to Soohyun Kim for very helpful English on the article.

Funding

This work was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation (NRF) Korea and by a NRF grant (2013R1A1A3010025).

Conflict of Interest: none declared.

References

Calle,M.L. *et al.* (2008) MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data. Technical Report 24, Department of Systems Biology, Universitat de Vic, Vic,: Spain; available at http://repositori.uvic.cat/xmlui/handle/10854/408.

Cattaert,T. *et al.* (2011) Model-based multifactor dimensionality reduction for detecting epistasis in case–control data in the presence of noise. *Ann. Hum. Genet.*, **75**, 78–89.

Cattaert,T. *et al.* (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. *PLoS One*, **5**, e10304.

Choi,J. and Park,T. (2013) Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions. *BMC Syst. Biol.*, **7**(Suppl 6), S15.

Cule,E. *et al.* (2011) Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, **12**, 372.

Dong,C. *et al.* (2008) Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.*, **16**, 229–235.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Gui,J. *et al.* (2011) A novel survival multifactor dimensionality reduction method for detecting gene–gene interactions with application to bladder cancer prognosis. *Hum. Genet.*, **129**, 101–110.

Gui,J. *et al.* (2013) A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS One*, **8**, e66545.

Lee,S. *et al.* (2012) Gene–gene interaction analysis for the survival phenotype based on the Cox model. *Bioinformatics*, **28**, i582–i588.

Li,J. *et al.* (2014) A model-free approach for detecting interactions in genetic association studies. *Brief. Bioinformatics*, **15**, 1057–1068.

Lou,X.Y. *et al.* (2008) A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am. J. Hum. Genet.*, **83**, 457–467.

Lou,X.Y. *et al.* (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.*, **80**, 1125–1137.

Mackay,T.F. (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **15**, 22–33.

Park,M.Y. and Hastie,T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30–50.

Ritchie,M.D. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Vago,E. and Kemeny,S. (2006) Logistic ridge regression for clinical data analysis (a case study). *Appl. Ecol. Environ. Res.*, **4**, 171–179.

Van Steen,K. (2012) Travelling the world of gene–gene interactions. *Brief. Bioinformatics*, **13**, 1–19.

Velez,D.R. *et al.* (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.*, **31**, 306–315.

Wan,X. *et al.* (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **87**, 325–340.

Wang,Y. *et al.* (2011) An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics*, **27**, 2936–2943.

Willer,C.J. *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.*, **40**, 161–169.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yang,C. *et al.* (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, **25**, 504–511.

Yu,W. *et al.* (2015) Multivariate Quantitative Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions. *Hum. Hered.*, **79**, 168–181.

Zhang,X. *et al.* (2010) TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, **26**, i217–i227.

Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.