Pair-Wise Multifactor Dimensionality Reduction Method to Detect Gene-Gene Interactions in A Case-Control Study

H. He^a W.S. Oetting^b M.J. Brott^b S. Basu^a

^aDivision of Biostatistics, and ^bDepartment of Experimental and Clinical Pharmacology, College of Pharmacy and Institute of Human Genetics, University of Minnesota, Minneapolis, USA

Key Words

Case-control study • Gene-gene interaction • Multifactor dimensionality reduction • Nonparametric approach

Abstract

Objective: The identification of gene-gene interactions has been limited by small sample size and large number of potential interactions between genes. To address this issue, Ritchie et al. [2001] have proposed multifactor dimensionality reduction (MDR) method to detect polymorphisms associated with the disease risk. The MDR reduces the dimension of the genetic factors by classifying them into high-risk and low-risk groups. The strong point in favor of MDR is that it can detect interactions in absence of significant main effects. However, it often suffers from the sparseness of the cells in high-dimensional contingency tables, since it cannot classify an empty cell as high risk or low risk. **Method:** We propose a pair-wise multifactor dimensionality reduction (PWMDR) approach to address the issue of MDR in classifying sparse or empty cells. Instead of looking at the higher dimensional contingency table, we score each pair-wise interaction of the genetic factors involved and combine the scores of all such pairwise interactions. **Results:** Simulation studies showed that the PWMDR generally had greater power than MDR to detect third order interactions for poly-

morphisms with low allele frequencies. The PWMDR also outperformed the MDR in detecting gene-gene interaction on a kidney transplant dataset. **Conclusion:** The PWMDR outperformed the MDR to detect polymorphisms with low frequencies.

Copyright © 2009 S. Karger AG, Basel

Introduction

Genetic mapping of a trait involves implementation of a number of statistical strategies to identify relative position(s) of gene(s) influencing the trait in the genome. Many complex traits of medical relevance such as diabetes, asthma, and Alzheimer's disease are controlled by multiple genes. Interaction between genes, low penetrance, and environmental factors make the gene discovery difficult for these complex traits. A common study design for genetic mapping of a trait is a case-control study design where genotype data on a large number of single nucleotide polymorphisms (SNPs) are collected for a number of cases and controls to study the association between these SNPs and the trait. There is growing evidence that these SNPs interact with each other in determining the susceptibility to complex traits or diseases. The investigation of such gene-gene interactions in a

case-control study presents new statistical challenges as the number of potential interactions between the SNPs can be large.

The traditional parametric statistical approach to modeling the relationship between disease status and the SNPs is logistic regression which has some obvious limitations. As each additional main effect is included in the logistic model, the number of possible interaction terms grows exponentially. Due to the sparseness of the data in high dimensions, parameter estimates often tend to have large standard errors, making it difficult to detect interaction.

Several methods have been proposed to detect gene-gene interaction in a case-control study design. These methods can be categorized broadly into parametric and nonparametric approaches. The parametric approaches try to address the potential problem of fitting gene-gene interaction models through traditional logistic regression and propose alternative ways to identify interactions. Some examples of parametric approaches are Penalized Logistic Regression [Park and Hastie, 2007], LASSO [Tibshirani, 1996], and logic regression [Kooperberg et al., 2001]. When modeling interaction with a large number of SNPs, the parametric approaches still tend to have too many parameters to estimate compared to the sample size [Concato et al., 1993; Peduzzi et al., 1996]. Moreover, these approaches can hardly capture the complexity of the dependence between the trait and the genes [Page et al., 2003; Cordell, 2002]. Due to these issues, nonparametric methods have been preferred for detection of interactions in many complex human diseases.

In nonparametric methods, several data-mining approaches have been developed and employed in the detection of gene-gene interactions. Some of these approaches are Combinatorial Partitioning Method (CPM) [Nelson et al., 2001], Neural Network, Multifactor Dimensionality Reduction Method (MDR) [Ritchie et al., 2001, 2003]. These methods detect the relevant interactions between SNPs by either reducing the dimension of the vast genetic data or recognizing the useful hidden patterns. These approaches do not assume any specific model (additive/multiplicative) to explain the nature of dependence between the trait and the SNPs. This makes them on parametric methods more flexible compared to the parametric methods. They tend to have lower power if there are multiple SNPs associated with the disease and their effects are indeed additive, because these nonparametric methods can only see them all as an epistatic multifactor interaction.

Among these different nonparametric approaches, there is a growing popularity of the MDR approach and it has been recently extensively used for gene-gene interaction detection in many real studies. The strong point in favor of MDR is that it can detect multiple SNPs associated with a disease. It searches through any level of interaction without considering the significance of the main effects. It is therefore able to detect high-order interactions even when the underlying main effects are not statistically significant. Many studies have shown that MDR can identify putative high-order gene-gene interactions in the absence of any significant independent main effects in sporadic breast cancer [Ritchie et al., 2001] and essential hypertension [Moore and William, 2002]. Ritchie et al. [2003] also evaluated the power of MDR in the presence of genotyping error, missing data, phenocopies and genetic heterogeneity. The impact of different approaches to handle missing data on MDR has been extensively studied in Namkung et al. [2009]. Recently Motsinger-Reif et al. [2008] has compared a number of different models and MDR came out to be performing fairly well across all comparison.

Although the MDR method provides many useful features, MDR suffers from several technical disadvantages. The MDR assigns each genotype combination (cell) as high-risk or low-risk and thus converts the high-dimensional data set into a single dimension. First, cells in high-dimensional tables will often be empty; these cells cannot be labeled as high-risk or low-risk. Second, MDR is prone to errors when the number of cases is similar to the number of controls in a cell, or when the number of both cases and controls is too small in a cell of the contingency table. The classification of these cells is vulnerable to both false positive and false negative errors.

In this paper, we have tried to address these drawbacks in MDR for detecting high dimensional interactions and proposed the pair-wise multifactor dimensionality reduction (PWMDR) approach. We have focused into balanced case-control studies to compare the performance of MDR and PWMDR. We have studied the power of MDR and PWMDR for detecting third order gene-gene interaction through extensive simulations. We have also studied their performance on a real dataset. The ROC analysis on the real dataset suggested that the PWMDR outperformed the MDR in detecting gene-gene interaction for polymorphisms with low minor allele frequencies.

Methods

Multifactor Dimensionality Reduction

The MDR method, proposed by Ritchie et al. [2001, 2003], is widely used for detecting gene-gene interactions associated with common complex genetic diseases. As an alternative to traditional logistic regression, MDR is nonparametric and genetic model free. Ritchie et al. [2003] demonstrated that the MDR method identified a four-locus interaction on the risk of sporadic breast cancer and was able to detect a high-order interaction in simulated data in the absence of any statistically significant main effects. Although the MDR method cannot distinguish between main effects and interactions, a major strength of the MDR method is its ability to detect higher order interactions in the absence of main effects.

Consider a case-control dataset with N SNPs on n individuals with equal number of cases as controls. Suppose M ($M \leq N$) is the highest order interaction we want to address. With MDR, multilocus genotypes are pooled into high risk and low risk groups, effectively reducing the dimensionality of the predictors from M dimensions to one dimension [Ritchie et al., 2001, 2003]. MDR carries out an exhaustive search of all possible 1-way, 2-way, 3-way, up to M -way combinations of predictors (SNPs). The prediction error of each model is estimated using k -fold (usually $k = 10$) cross-validation. First, we randomly divide the data into k equal parts. The model is developed using each $(k - 1)$ parts of the data, the training data, and then it is used to make predictions about the disease status on there maining part of the data. The process is repeated k times and the prediction errors are averaged to reduce the bias in the estimation of prediction error.

For any given m ($m \leq M$), the general procedure to implement the MDR method to detect the optimum m -way interaction, is illustrated as the following:

- 1 Run k -fold (often $k = 10$) cross-validation to find the best set of m -way interactions. For each cross-validation fold, repeat the following steps: (a) Use every possible part as the test data and the other remaining 9 parts as the training data. (b) A set of m SNPs is then selected from the pool of all SNPs. (c) m SNPs and their possible multifactor cells are represented in m -dimensional space (The m -way contingency table is formed on the training data.). (d) Each multifactor cell is labeled as high-risk if case/control ratio exceeds or equal to some threshold T (e.g., $T = 1.0$), and low-risk otherwise. (e) Compute the training error for the 9/10 data, by classifying high-risk as a case, low-risk a control. (f) Search for all possible choices of m SNPs. Totally we need to construct $\binom{N}{m}$ contingency tables and correspondingly get $\binom{N}{m}$ training errors. The model with the lowest training error (misclassification error) is selected, and the prediction error (test error) of the model is estimated using the independent test data.
- 2 Calculate the averaged k prediction errors as the prediction error for model size m . Obtain the cross-validation consistency, a measure of the number of times a particular set of SNPs is identified across cross-validations.

We have selected the final model size with the lowest prediction error and the final model with the chosen size is selected based on the largest cross-validation consistency and/or lowest prediction error. That is, the model that minimizes the prediction error and/or maximizes the cross-validation consistency is selected as the final MDR model.

There is a growing popularity of the MDR approach and it has been recently extensively used for gene-gene interaction detection in many real studies. The strong point in favor of MDR is that it can detect multiple SNPs associated with a disease. It searches through any level of interaction without considering the significance of the main effects. It is therefore able to detect high-order interactions even when the underlying main effects are statistically not significant. On the other hand, MDR suffers from several technical disadvantages. MDR assigns each genotype combination (cell) as high-risk or low-risk and thus converts the high-dimensional data set to a single dimension. First, cells in high-dimensional tables will often be empty; these cells cannot be labeled as high-risk or low-risk. Second, the binary assignment (high-risk/low-risk) is highly unstable and prone to high errors when the proportions of cases and controls are similar or both the number of cases and controls are too small in a cell of the contingency table.

Pair-Wise Multifactor Dimensionality Reduction

MDR faces several technical problems such as the sparseness of data in high dimensions. To address the problems faced by MDR, we have proposed a PWMDR approach. In this approach, we model the interaction between m SNPs by $\binom{m}{2}$ pair-wise comparisons among m SNPs. For example, when we look for 3-way interaction, instead of looking at a 3-way contingency table, we look at three 2-way tables. In this way, we could avoid the sparseness in higher dimensions. For every person, we go through every 2-way table (totally $\binom{m}{2}$ tables), if the ratio of the number of case over the number of control meets or exceeds some threshold T (usually we set $T = 1$), we increase that person's risk by 1; otherwise we decrease that person's risk by 1.

The cumulative risk score over the $\binom{m}{2}$ tables can then be used to predict disease status. If there is no association between the SNPs and the disease, we expect that the cumulative score will have the same distribution in cases and controls. For example, for a 3 SNP interaction model, the possible values of the cumulative score are $-3, -1, 1$ and 3 . If there is no association, the score will be symmetrically distributed around 0 in cases as well as in controls. In presence of interaction, we will see a deviation from this symmetric distribution. Then, for cases, there is will more positive values than expected under no association. On the other hand, for the controls, the cumulative score will take more negative values than expected under no association. In other words, if we use a cut-off of zero, we expect to see equal number of cases and controls on the positive and negative side of zero under no association. In the presence of an association, there will be more cases with cumulative score greater than 0 and there will be more controls with cumulative score less than zero. Hence, our decision rule is to classify a person as case if the cumulative risk score meets or exceeds 0 and control otherwise. Thus, the PWMDR approach is exactly the same as the MDR approach for 2-way interaction.

Just like MDR, we carry out an exhaustive search of all possible 2-way, 3-way, up to m -way combinations of SNPs to identify SNPs that are significant in determining the disease status. A final model is selected by minimizing the prediction error across k -fold cross-validations. Just as in MDR, with k -fold (usually $k = 10$) cross-validation we first randomly divide the data into k equal parts. The model is fitted on each possible $(k - 1)$ parts of the data and then is used to make predictions about the disease status of the remaining part of the data.

For any given m ($2 \leq m \leq M$), the general procedure to implement the PWMDR method is detailed below:

- 1 Run k -fold (often $k = 10$) cross-validation to find the best set of interactions of m predictors. For each cross-validation fold, repeat the following steps: (a) Use every 1 possible part as the test data and the other remaining 9 parts as the training data. (b) A set of m SNPs is then selected from the pool of all SNPs. (c) m SNPs are represented by $\binom{m}{2}$ 2-way table. (d) Label the cells of each 2-way table either high-risk or low-risk depending on the cases/control ratio in the training part. (e) For each 2-way table, if a person falls into a high-risk cell, increase that person's risk by 1; otherwise decrease by 1. (f) Classify a person as a case if the cumulative risk score meets or exceeds 0 and control otherwise. (g) Search for all possible choices of m SNPs. Select the model with the lowest training error. The prediction error of the model is estimated using the independent test data.
- 2 Obtain the cross-validation consistency. The model with the largest consistency and/or lowest prediction error is selected. Calculate the average of the prediction errors across cross validations.

For each m , we get a PWMDR model. The result is a set of models, one for each model size considered. From this set, the model size that minimizes the prediction error is selected as the final model size the model with the chosen model size is selected based on the minimum prediction error and/or the maximum cross-validation consistency.

The proposed approach is solely based on the idea that most of the SNPs with high-order interaction would have lower order effects. Hence the high-order interaction among a set of SNPs could be captured by studying the pair-wise interaction between each pair of SNPs in the set. Under the no association between the SNPs and the binary trait, the joint distribution of the SNPs will be the same in cases and controls. Let us denote the joint distribution of the p SNPs among cases, $\Pr[X_1 = x_1, X_2 = x_2, \dots, X_p = x_p | Y = 1]$, and controls, $\Pr[X_1 = x_1, X_2 = x_2, \dots, X_p = x_p | Y = 0]$ by $f_1(x_1, x_2, x_3, \dots, x_p)$ and $f_0(x_1, x_2, x_3, \dots, x_p)$, respectively. If there is no association between the SNPs and the disease, then

$$f_1(x_1, x_2, x_3, \dots, x_p) \equiv f_0(x_1, x_2, x_3, \dots, x_p) \quad \forall x_1, x_2, \dots, x_p \quad (1)$$

For demonstration purpose, let us assume that there is no linkage disequilibrium between the SNPs. That is the unconditional distribution of the SNPs is a product of their marginal distributions.

$$\Pr[X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] = \prod_{i=1}^p \Pr[X_i = x_i] \quad (2)$$

Consider a 3 SNP model where the genotype combinations (x_1, x_2, x_3) is associated with the disease. Let

$$\log \left[\frac{\Pr[Y = 1]}{\Pr[Y = 0]} \right] = \beta_0 + \beta, \quad (3)$$

where β_0 is the intercept and β is the log odds ratio corresponding to genotype combination (x_1, x_2, x_3) . Then the expected case-control ratio for the genotype combination (x_1, x_2, x_3) in a balanced case-control study is given by

$$\begin{aligned} & \frac{\Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3 | Y = 1]}{\Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3 | Y = 0]} \\ &= \frac{\frac{1}{\Pr[Y = 1]} \left[\left(\prod_{i=1}^3 \Pr[X_i = x_i] \right) \Pr[Y = 1 | (x_1, x_2, x_3)] \right]}{\frac{1}{\Pr[Y = 0]} \left[\left(\prod_{i=1}^3 \Pr[X_i = x_i] \right) \Pr[Y = 0 | (x_1, x_2, x_3)] \right]} \\ &= \frac{(1-p)}{p} \exp^{\beta_0 + \beta} \\ &> 1 \end{aligned} \quad (4)$$

for large value of β . Here p is the weighted average of the disease probability given all genotype combinations. Hence, if we look at the 3-way contingency table, the expected case-control ratio for the genotype combination (x_1, x_2, x_3) will be much higher than 1 for a large value of β and hence the cell will be classified as a high-risk cell. For the remaining cells, the expected case-control ratio will be

$$\left(\frac{1-p}{p} \right) \exp^{\beta_0}$$

and will generally be < 1 and hence will be classified as low-risk cells. Instead of looking at the 3-way contingency table, we propose to capture the association by looking at three 2-way contingency tables. For a cell (x'_1, x'_2) other than (x_1, x_2) , the expected case-controls ratio would be

$$\begin{aligned} & \frac{\Pr[X_1 = x'_1, X_2 = x'_2 | Y = 1]}{\Pr[X_1 = x'_1, X_2 = x'_2 | Y = 0]} \\ &= \frac{\sum_{x_3^*} \Pr[X_1 = x'_1, X_2 = x'_2, X_3 = x_3^* | Y = 1]}{\sum_{x_3^*} \Pr[X_1 = x'_1, X_2 = x'_2, X_3 = x_3^* | Y = 0]} \\ &= \left(\frac{1-p}{p} \right) \left[\frac{\sum_{x_3^*} \Pr[X_1 = x'_1, X_2 = x'_2, X_3 = x_3^*] \Pr[Y = 1 | (x'_1, x'_2, x_3^*)]}{\sum_{x_3^*} \Pr[X_1 = x'_1, X_2 = x'_2, X_3 = x_3^*] \Pr[Y = 0 | (x'_1, x'_2, x_3^*)]} \right] \\ &= \left(\frac{1-p}{p} \right) \left[\frac{\sum_{x_3^*} \Pr[X_3 = x_3^*] \Pr[Y = 1 | (x'_1, x'_2, x_3^*)]}{\sum_{x_3^*} \Pr[X_3 = x_3^*] \Pr[Y = 0 | (x'_1, x'_2, x_3^*)]} \right] \\ &= \frac{(1-p)}{p} \exp^{\beta_0} \\ &< 1 \end{aligned} \quad (5)$$

and hence will be classified as a low-risk cell. The remaining cell (x_1, x_2) will then have the expected case-control ratio > 1 in our balanced case-control study. For the above 3-way interaction model, the expected case-control ratios of the genotype combina-

tions (x_1, x_2) in a balanced case-control study can be derived from the equation below:

$$\begin{aligned} & \frac{\Pr[X_1 = x_1, X_2 = x_2 | Y = 1]}{\Pr[X_1 = x_1, X_2 = x_2 | Y = 0]} \\ &= \frac{\sum_{x_3^*} \Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3^* | Y = 1]}{\sum_{x_3^*} \Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3^* | Y = 0]} \\ &= \left(\frac{1-p}{p} \right) \left[\frac{\sum_{x_3^*} \Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3^*] \Pr[Y = 1 | (x_1, x_2, x_3^*)]}{\sum_{x_3^*} \Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3^*] \Pr[Y = 0 | (x_1, x_2, x_3^*)]} \right] \\ &= \left(\frac{1-p}{p} \right) \left[\frac{\sum_{x_3^*} \Pr[X_3 = x_3^*] \Pr[Y = 1 | (x_1, x_2, x_3^*)]}{\sum_{x_3^*} \Pr[X_3 = x_3^*] \Pr[Y = 0 | (x_1, x_2, x_3^*)]} \right] \\ &= \left(\frac{1-p}{p} \right) \left[\frac{\Pr[X_3 = x_3] \left(\frac{\exp^{\beta_0 + \beta}}{1 + \exp^{\beta_0 + \beta}} \right) + (1 - \Pr[X_3 = x_3]) \frac{\exp^{\beta_0}}{1 + \exp^{\beta_0}}}{\Pr[X_3 = x_3] \left(\frac{1}{1 + \exp^{\beta_0 + \beta}} \right) + (1 - \Pr[X_3 = x_3]) \frac{1}{1 + \exp^{\beta_0}}} \right] \end{aligned}$$

Hence if the genotype $(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ is positively associated with the disease, the expected case-control ratio for the genotype combinations (x_1, x_2) , (x_2, x_3) and (x_1, x_3) will all be >1 and hence our PWMDR approach will capture the association between the trait and the genotype combination $(X_1 = x_1, X_2 = x_2, X_3 = x_3)$. However, this PWMDR approach will lose power if there are multiple cells in the 3-way contingency table having opposite direction of association.

Note that, both MDR and PWMDR approaches have the same probability of selecting a subset of SNPs under the null hypothesis of no association between a group of SNPs and the disease. For example, consider a model with 4 SNPs such as SNP1, SNP2, SNP3 and SNP4, where none of the SNPs are associated with the disease. Assume that the SNPs have same minor allele frequency. If we perform a search with 3 SNPs, the chance of choosing the combination (SNP1, SNP2, SNP3) as the best set of SNPs would be $1/\binom{4}{3} = 0.25$ for both these approaches.

In this paper, we have studied the power of MDR and PWMDR for detecting gene-gene interaction for various 3-way interaction models. We have also studied their performance on a real dataset. The PWMDR faces same challenges of choice of optimal threshold for defining high-risk and low-risk genotype combinations for imbalanced case-control study as in MDR [Velez et al., 2007]. Hence we have focused on balanced case-control studies for comparison between these two methods. As expected, none of these two methods were consistently better for all data scenarios, but, generally PWMDR outperformed MDR, especially for allele frequency of 0.20 or 0.30 of the associated SNPs. Moreover, the ROC analysis on the real dataset suggested that the PWMDR outperformed the MDR in detecting gene-gene interaction for SNPs with low minor allele frequencies.

Simulations and Results

We simulated datasets under the null hypothesis of no association between the disease status and the SNPs. We simulated data on 4 SNPs, each with minor allele frequency of 0.5. We simulated 5,000 datasets under the null hypothesis. Each dataset contained 400 samples (200 cases and 200 controls) and 4 SNPs such as SNP1, SNP2, SNP3 and SNP4, none of which were associated with the disease. We applied MDR and PWMDR to select the best 3-SNP model among these 4 SNPs for each dataset. The distribution of the average cross-validation prediction error for the MDR approach had mean 0.50 and standard deviation 0.038. The distribution of the average cross-validation prediction error for the PWMDR approach had mean 0.50 and standard deviation 0.035. We also computed the number of times each of the 4 combinations of a 3-SNP model were selected as the best model. The difference in the frequency distribution of the 3-SNP models between the MDR and the PWMDR approach was not statistically significant (p value of 0.23).

We have compared the power of MDR and PWMDR for detecting 3-way interactions in a balanced case-control study through extensive simulations. We generated 4 epistatic models with different magnitudes of interaction effect. For each model, we simulated 100 datasets. Each dataset contained 400 samples (200 cases and 200 controls) and 8 SNPs, only 3 of which were associated with the disease. If we denote the associated SNPs by SNP1, SNP2 and SNP3 and the associated alleles of SNP1, SNP2 and SNP3 by A, B and C, then the four interaction models were

- (1) Model 1: $\text{Logit}(p) = -5 + 3 \text{I}(\text{SNP1} = \text{Aa}, \text{SNP2} = \text{Bb}, \text{SNP3} = \text{Cc}) + 3 \text{I}(\text{SNP1} = \text{AA}, \text{SNP2} = \text{BB}, \text{SNP3} = \text{CC})$
- (2) Model 2: $\text{Logit}(p) = -5 + 3 \text{I}(\text{SNP1} = \text{AA}) + 3 \text{I}(\text{SNP2} = \text{BB}) + 3 \text{I}(\text{SNP3} = \text{CC})$
- (3) Model 3: $\text{Logit}(p) = -5 + 3 \text{I}(\text{SNP2} = \text{Bb}) + 3 \text{I}(\text{SNP1} = \text{AA}, \text{SNP2} = \text{BB}, \text{SNP3} = \text{CC})$
- (4) Model 4: $\text{Logit}(p) = -5 + 3 \text{I}(\text{SNP1} = \text{AA}) + 3 \text{I}(\text{SNP2} = \text{BB}) + 3 \text{I}(\text{SNP3} = \text{CC}) - 6 \text{I}(\text{SNP1} = \text{AA}, \text{SNP2} = \text{BB})$.

The conditional probabilities of being diseased for these 3 SNPs are listed in table 1. The levels of the remaining 5 SNPs were in Hardy-Weinberg equilibrium for cases and controls, separately. Figure 1 displays the log-odds of being diseased for all possible combinations of levels of the 3 associated SNPs.

The heritability with the minor allele frequency of 0.30 for all SNPs were 7, 29, 6 and 23% for Model 1, 2, 3 and 4, respectively. Model 1 was a purely epistatic model

Table 1. Penetrance tables: 3-way interactions

	3rd factor = cc			3rd factor = Cc			3rd factor = CC		
	bb	Bb	BB	bb	Bb	BB	bb	Bb	BB
Model 1									
aa	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Aa	0.01	0.01	0.01	0.01	0.12	0.01	0.01	0.01	0.01
AA	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.12
Model 2									
aa	0.01	0.01	0.12	0.01	0.01	0.12	0.12	0.12	0.73
Aa	0.01	0.01	0.12	0.01	0.01	0.12	0.12	0.12	0.73
AA	0.12	0.12	0.73	0.12	0.12	0.73	0.73	0.73	0.98
Model 3									
aa	0.01	0.12	0.01	0.01	0.12	0.01	0.01	0.12	0.01
Aa	0.01	0.12	0.01	0.01	0.12	0.01	0.01	0.12	0.01
AA	0.01	0.12	0.01	0.01	0.12	0.01	0.01	0.12	0.12
Model 4									
aa	0.01	0.01	0.12	0.01	0.01	0.12	0.12	0.12	0.73
Aa	0.01	0.01	0.12	0.01	0.01	0.12	0.12	0.12	0.73
AA	0.12	0.12	0.73	0.12	0.12	0.73	0.01	0.01	0.12

Penetrance tables of four different models with 3-way interactions. SNP1 has three genotypes AA, Aa, aa and SNP2 has three different genotypes BB, Bb, bb and SNP3 has three different genotypes CC, Cc, cc. Each cell of a penetrance table represents the probability of being affected given the cell genotype.

with no main effect or second order interactions. Model 2 was an additive model with a main effect for each of the 3 SNPs. Model 3 had a main effect term for the 2nd SNP and a third order interaction term. For Model 4, each of the 3 SNPs had a main effect and a pair-wise interaction term between SNP1 and SNP2 in opposite direction to the main effect.

We conducted the analyses for the paper by coding PWMDR in R [R Development Core Team, 2005]. The code was run on a Linux workstation with two Intel Xeon quad core processors. It took 15 s to identify the best 4-way interaction model among the 8 SNPs in our simulation study for a single dataset.

For both MDR and PWMDR, we searched through 2-way to 4-way interaction using a 10-fold exhaustive cross-validation and a final model was selected based on the minimum averaged prediction error and/or maximum cross validation consistency. We estimated the power of each method as the number of times (out of 100) the correct SNPs were identified. We also varied the allele frequencies of the associated SNPs to check the impact of allele frequencies on the power of MDR and PWMDR.

Figure 2 summarizes the power comparisons of MDR and PWMDR when $P(A) = P(B) = P(C) = 0.1$, $P(A) = P(B) = P(C) = 0.2$, $P(A) = P(B) = P(C) = 0.3$ under each of these 4 models in table 1, where the associated alleles of SNP1, SNP2 and SNP3 are denoted by A, B and C, respectively.

For Model 1, none of the methods had any power to detect interaction at allele frequency of 0.10 of the associated SNPs, but for allele frequencies of 0.20 and 0.30, the power of PWMDR was substantially higher than the MDR. For Model 3, power to detect association was low for both these approaches, but still PWMDR consistently outperformed MDR for different allele frequencies of the associated SNPs. Both Model 1 and Model 3 had a third order interaction term, where the genotype combination (AA, BB, CC) was associated with the disease. The MDR suffered from the sparseness of the cells in this high dimensional contingency table and lost substantial power compared to PWMDR.

In Model 2, the SNPs had only main effect, and the effects of the three SNPs were the same and they were additive on log scale. The performance of MDR and PWMDR were fairly similar across different allele frequencies of the associated SNPs. In Model 4, we added a 2nd order interaction term, additional to the main effects in Model 2 in the reverse direction. The PWMDR approach can suffer when the pair-wise interactions among SNPs are acting in reverse directions (fig. 1). The PWMDR had lower power than the MDR, but we did not see any substantial loss in power in PWMDR across different allele frequencies. We also checked if the power differences between these two methods were statistically significant. Using a normal approximation to the two binomial distributions, the p values corresponding to the null hypothesis that the expected probability of rejection is the same for MDR and PWMDR were 0.0007, 0.506, 0.028 and 0.347 for Model 1, 2, 3 and 4 respectively for an allele frequency of 0.30.

We have also reported in table 2 the average prediction error and the corresponding standard deviation over 100 datasets simulated under each of these 4 models in table 1. The PWMDR had in general lower prediction error than the MDR approach. The MDR is prone to errors when the number of both cases and controls is too small in a high dimensional contingency table. The classification of these cells is vulnerable to both false positive and false negative errors. Table 2 indicated that the PWMDR generally had improved error rates as compared to the MDR. We also computed the expected prediction error under the true models, which were used to simulate the data-

Fig. 1. The patterns of log-odds for class 1 (affected) for different levels of the three SNPs (SNP1, SNP2 and SNP3) under four different models with 3-way interactions. SNP1 has three levels AA, Aa, aa; SNP2 has three different levels BB, Bb, bb and SNP3 has three levels CC, Cc, cc.

Fig. 2. Figure shows the power of MDR and PWMDR for detection of interaction under four different 3-way interaction models. Each plot represents the power of MDR and PWMDR under different allele frequencies of the associated SNPs.

Table 2. Comparison of prediction errors between MDR and PWMDR for an allele frequency of 0.30 for all associated SNPs and for all 4 different interaction models

	Model 1	Model 2	Model 3	Model 4
MDR	0.254 (0.023)	0.163 (0.021)	0.229 (0.023)	0.174 (0.02)
PWMR	0.244 (0.020)	0.156 (0.019)	0.268 (0.01)	0.168 (0.017)

Table 3. Comparison of prediction performance between MDR and PWMDR on real dataset

	Best model	Misclassification error	Sensitivity	Specificity
MDR	ABCB1*ABCC1*SETX	0.29	0.80	0.61
PWMR	ABCB1*ABCC1*SETX *F5	0.26	0.80	0.68

sets. For allele frequency of 0.30, the expected prediction errors and the corresponding standard deviations were 0.240 (0.017), 0.157 (0.018), 0.230 (0.019) and 0.167 (0.021) for Model 1, 2, 3 and 4, respectively. The PWMDR produced least biased prediction errors under Model 1, 2 and 4. The MDR approach has better accuracy than PWMDR for Model 3.

Real Data Analysis

We have compared the performance of MDR and PWMDR to detect gene-gene interaction associated with acute rejection (AR) in kidney transplant patients. Whole blood was obtained with informed consent and DNA isolated from 271 kidney allograft recipients, 136 of whom had AR within 6 months of transplant, and 135 of whom did not have any detectable AR after at least 8 years post-transplant. All received Ab induction and CNI, with either MMF or sirolimus. DNA variants were genotyped using a Affymetrix custom genotyping chip containing 3,590 SNPs, many of which are thought to be functional variants within biologically relevant genes to acute rejection including genes in pathways associated with immunity, cell signaling, ADME, cell growth and proliferation [Van Ness et al. 2008]. Genotyping was performed using the Affymetrix Gene Chip Scanner 3000 Targeted Genotyping System (GCS 3000 TG System), which utilizes molecular inversion probes to simultaneously identify the 3404 pre-selected SNPs. Methods for genotyping have been previously described and were performed in strict adherence to the manufacturer's protocol [Hardenbol et al. 2003]. For this comparison study, we randomly selected 120 caucasian patients with acute rejection within 6 months of transplant, and 120 caucasian patients without any detectable AR after at least 8 years post-transplant.

Of the 3,404 SNPs typed, 80 SNPs had no data and were hence excluded from the analysis. Of the remaining 3,324 SNPs, the call rate was 98.6%. We also excluded the SNPs with minor allele frequency less than 5% and the SNPs which had more than 10% missing values. Our goal here was to detect any evidence of interaction among the SNPs associated with AR in kidney allografts. For all the SNPs, we did Fisher's exact test and selected only the top 77 SNPs with p values ≤ 0.01 for the interaction detection purpose. One could consider the whole list of SNPs for interaction detection, but with this small sample size, we decided to limit our search to the top SNPs. Among these 77 SNPs, we imputed the missing data for each SNP from the observed genotype distribution in order to imple-

ment MDR and PWMDR for interaction detection purpose. In order to emphasize the issue of sparseness of MDR in high dimensions and demonstrate how PWMDR could address it, we selected only the SNPs with allele frequencies lower than 0.30. We finally had 48 SNPs for the analysis. We applied MDR and PWMDR methods to detect evidence of interaction among the 48 SNPs and AR.

We considered up to 4-way interactions. Ten-fold cross validation was used to obtain the best model for each given number ($n = 1, 2, 3, 4$) of SNPs. For one SNP model (main effect model), 4 out of 10 times MDR chose rs1128503 (ABCB1). For two-SNPs model (2-way interaction model), the best one was rs1128503 (ABCB1) * rs875740 (ABCC1); and for three-SNPs model (3-way interaction model), the best model was rs1128503 (ABCB1) * rs215095 (ABCC1) * rs2296869 (SETX) with a cross-validation consistency of 3 out of 10. For 4-way interaction model, the best model was rs1128503 (ABCB1) * rs215095 (ABCC1) * rs2296869 (SETX) * rs1805332 (RAD23B) with a cv consistency of 2 out of 10. The average prediction errors for 1-way, 2-way, 3-way and 4-way interaction models were 0.525, 0.512, 0.437 and 0.454, respectively. The average prediction error was minimum for the 3-way interaction model. Hence the 3-way interaction model was selected as the best model. According to table 3, the specificity and sensitivity of the final MDR model (also the best overall model) were 0.61 and 0.80, respectively. Total misclassification error for the chosen 3-way interaction model was 0.29.

The PWMDR approach was used for 2-way, 3-way and 4-way interaction model selection. For the 2-way interaction model, as expected, the best model was rs1128503 (ABCB1) * rs875740 (ABCC1), which was also identified by MDR. For the 3-way interaction model, PWMDR selected a different model rs875740 (ABCC1) * rs1128503 (ABCB1) * rs4072037 (MUC1) with a cross-validation consistency of 3 out of 10. For the 4-way interaction model, PWMDR selected rs1128503 (ABCB1) * rs215095 (ABCC1) * rs2296869 (SETX) * rs6030 (F5). This 4-way interaction model was selected 4 out of 10 times. The average prediction errors for the 3-way and 4-way interaction model were 0.550 and 0.390, respectively. The average prediction error was minimum for the 4-way interaction model. Hence the 4-way interaction model was selected as the best model. These SNPs rs1128503 (ABCB1), rs215095 (ABCC1), rs6030 (F5) and rs2296869 (SETX) were ranked 9, 13, 22, and 46, respectively, in the single SNP association analysis with 48 SNPs using Fisher's exact test.

According to table 3, the specificity and sensitivity for the selected model by PWMDR was 0.68 and 0.80, respectively. The total misclassification error for the chosen model was 0.26. According to the total misclassification rate, specificity and sensitivity, the PWMDR approach outperformed the MDR approach to detect evidence of interaction among the SNPs and AR.

We also performed a data permutation testing procedure to assess the statistical significance of the prediction errors for the best model selected by MDR and PWMDR. We generated 1,000 randomized datasets by randomizing the disease status labels, while maintaining the multifactor matrix. We repeated the entire MDR procedure and PWMDR procedure for each dataset. The best overall model was extracted for each random dataset, which generated a distribution of 1,000 prediction errors that could be expected by chance alone for both MDR and PWMDR approach. The significance of the final model is determined by comparing the prediction error of the final model to the distribution. The permutation p value for the MDR approach was 0.10. The permutation p value for the PWMDR approach was 0.023.

Both ABCB1 (ATP-binding cassette, sub-family B member 1) and ABCC1 (ATP-binding cassette, sub-family C, member 1) gene products are involved in small molecule transport and have broad substrate specificity [Sharon, 2008]. These proteins are also involved in the transport and distribution of immunosuppressants, used in kidney transplantation, and reports have shown that variation within these proteins is associated with kidney transplant outcomes, especially when calcineurin inhibitors are used [Naesens et al., 2009]. The SNP rs1128503 (ABCB1) results in a synonymous amino acid substitution for glycine at codon 412. The two SNPs in ABCC1 produces nucleotide substitutions in intron1 (rs215095) and in intron 5 (rs875740). The gene SETX (senataxin) is associated with autosomal recessive spinocerebellar ataxia-1 and is thought to be involved in RNA processing and may be protective against oxidative DNA damage [Sureweera et al., 2007]. The polymorphism (rs2296869) results in a synonymous amino acid substitution for asparagine at codon 1937. The SNP rs1805335 is within the fifth intron (IVS5-15A>G) of the human homolog of *Saccharomyces cerevisiae* Rad23, a protein involved in nucleotide excision repair (NER). RAD23 has recently been implicated in the regulation of targeting protein to the proteasome for degradation [Madura, 2004]. The SNP rs4072037 results in a synonymous amino acid substitution of threonine at codon 22 and 31 (depending on the transcription product) of the mucin 1, cell surface associ-

ated protein (MUC1). The protein is expressed in the kidney and this polymorphism results in alternative splicing of the gene [Moehle et al., 2008; Imbert et al., 2008]. The protein has been shown to suppress the Toll-like receptor (TLR) 5 signaling pathway, and may play an anti-inflammatory role [Ueno et al., 2008]. Factor 5 (F5) is involved in blood coagulation. The SNPrs6030 produces a valine to methionine amino acid substitution at codon 1764. This has been shown to affect factor 5 resulting in a loss of activity [Shinozawa et al., 2007]. Variations within F5 have been associated with adverse outcomes in kidney allograft recipients [Celik et al., 2006]. Those SNPs that appear to be in proteins with obvious biological relevance to kidney allograft outcomes are those within ABCB1, ABCC1, F5 and MUC1.

Discussion

Statistical methods for the detection of gene-gene interactions in a case-control study can be categorized broadly into parametric and nonparametric approaches. The MDR and PWMDR are both non-parametric data-mining approaches which do not make any assumption about the nature of dependence between the trait and the SNPs. Empty cell and sparseness problems become a more serious issue for MDR when the number of loci in the model is large. The PWMDR approach provides an alternative to the MDR approach for detection of higher order interaction, since PWMDR is less likely to suffer from sparseness in higher dimension interaction models. However, the PWMDR approach is solely based on the assumption that the set of SNPs with a high-order (≥ 3) interaction would have lower order effects, and hence we would be able to capture the high-order interaction by looking into each pair of SNP in the set. If the high-dimensional interaction effects nullify each other in a way that there is no pair-wise interaction effects for the SNPs, the PWMDR approach would have no power to detect such an interaction.

We performed many simulation studies with three loci in which several combinations of main effects and interaction effects were considered. The PWMDR generally outperformed the MDR approach, especially for low allele frequencies. The prediction error rate was generally lower for PWMDR as compared to MDR. This indicates that MDR suffered from the sparseness in three dimensional contingency tables and could not classify the nearly empty cells accurately.

When applying the MDR and PWMDR methods, the presence of missing observations reduces the number of observations available in the analysis. The most appropriate approach at present is to use only subjects with complete observations. However, as the number of genotypes increases, the number of subjects with complete observations decreases rapidly. Currently, there are several approaches to handle missing data and can be used to implement MDR or PWMDR to detect gene-gene interaction in presence of missing data [Namkung et al.,

2009]. One solution to handle this situation is to impute missing observations which we did for our real data analysis.

Acknowledgements

The research was supported by the Minnesota Medical Foundation grant. The authors would like to thank the two anonymous referees for valuable comments.

References

- Celik A, Tekis D, Saglam F, tunali S, Kabakci N, Ozaksoy D, Manisali M, Ozcan M, Meral M, Gülay H, Camsari T: Association of corticosteroids and factor V, prothrombin, and MTHFR gene mutations with avascular osteonecrosis in renal allograft recipients. *Transplant Proc* 2006;38:512–516.
- Concato J, Feinstein AR, Holford TR: The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118:210–210.
- Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;11:2463–2468.
- Hardenbol P, Baner J, Jain M, et al: Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 2003;21:673–678.
- Imbert Y, Foulks G, Brennan M, Jumblatt M, John G, Shah H, Newton, C, Pouranfar, F, Young WJ: Muc1 and estrogen receptor alpha gene polymorphisms in dry eye patients. *Exp Eye Res* 2008 (Epub ahead of print).
- Kooperberg C, Ruczinski I, LeBlanc M, Hsu L: Sequence analysis using logic regression. *Genet Epidemiol* 2001;21:S626–S631.
- Madura K: Rad23 and rpn10: perennial wall-flowers join the melee. *Trends Biochem Sci* 2004;29:637–640.
- Moehle C, Ackermann N, Langmann T, Aslanidis C, Kel A, Kel-Margoulis O, Schmitz-Madry A, Zahn A, Stremmel W, Schmitz G: Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *J Molec Med* 2008;84:1055–1066.
- Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD: A comparison of analytical methods for genetic association studies. *Genet Epidemiol* 2008;32:767–778.
- Naesens M, Kuypers D, Sarwal M: Calcineurin inhibitor nephrotoxicity. *Clin J Am Soc Nephrol* 2009;4:481–508.
- Namkung J, Elston R, Yang J, Park T: Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. *Genet Epidemiol* 2009 (Epub ahead of print).
- Nelson MR, Kardia SLR, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;11:458–470.
- Page GP, George V, Go RC, Page PZ, Allison DB: Are we there yet? Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* 2003;73:711–719.
- Park M, Hastie T: Penalized logistic regression for detecting gene interactions. *Biostatistics* 2007;9:30–50.
- Peduzzi P, Concato JE, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–1379.
- R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-9000051-07-0. URL <http://www.R-project.org>
- Ritchie MD, Hahn LW, Moore JH: 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;24:150–157.
- Ritchie MD, Hahn LW, Roodi N, Bailey LE, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals highorder interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147.
- Sharom F: ABC multidrug transporters: structure, function and role in chemoresistance. *Pharmacogenomics* 2008;9:105–127.
- Shinozawa K, Amano K, Suzuki T, Tanaka A, Iijima K, Takahashi H, Inaba H, Fukutake K: Molecular characterization of 3 factor v mutations, r21731, v1813m, and a 5-bp deletion, that cause factor v deficiency. *Int J Hematol* 2007;86:407–413.
- Suraweera A, Becherel O, Chen P, Rundle N, Woods R, Nakamura J, Gatei M, Criscuolo C, Filla A, Chessa L, Fusser M, Epe B, Gueven N, Lavin M: Senataxin, defective in ataxia oculomotor apraxia type 2, is involved in the defense against oxidative dna damage. *J Cell Biol* 2007;177:969–979.
- Tibshirani R: Regression shrinkage and selection via the lasso. *J R Stat Soc* 1996;58:267–288.
- Ueno K, Koga T, Kato K, Golenbock D, Gendler S, Kai H, Kim K: MUC1 mucin is a negative regulator of toll-like receptor signaling. *Am J Respir Cell Mol Biol* 2008;38:263–268.
- Van Ness B, Ramos C, Haznadar M, et al: Genomic variation in myeloma: design, content, and initial application of the bank on a cure snp panel to detect associations with progression-free survival. *BMC Medicine* 2008;6:26.
- Velez D, White B, Motsinger A, Bush W, Ritchie M, Williams S, Moore J: A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol* 2007;31:306–315.