Published in final edited form as:

Ann Hum Genet. 2011 January ; 75(1): 20–28. doi:10.1111/j.1469-1809.2010.00624.x.

A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility

Jiang Gui^b, Angeline S. Andrew^b, Peter Andrews^a, Heather M. Nelson^e, Karl T. Kelsey^d, Margaret R. Karagas^b, and Jason H. Moore^{a,b,c}

^aComputational Genetics Laboratory, Department of Genetics, Dartmouth Medical School, Lebanon, NH

^bDepartment of Community and Family Medicine, Norris-Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH

^cDepartment of Computer Science, University of New Hampshire, Durham, NH, Department of Computer Science, University of Vermont, Burlington, VT, Department of Psychiatry and Human Behavior, Brown University, Providence, RI, Translational Genomics Research Institute, Phoenix, AZ

^dDepartment of Community Health, Brown University, Providence, RI

^eDivision of Epidemiology and Community Health, University of Minnesota School of Public Health, Minneapolis, MN

Abstract

A central goal of human genetics is to identify and characterize susceptibility genes for common complex human diseases. An important challenge in this endeavor is the modeling of gene-gene interaction or epistasis that can result in non-additivity of genetic effects. The multifactor dimensionality reduction (MDR) method was developed as machine learning alternative to parametric logistic regression for detecting interactions in absence of significant marginal effects. The goal of MDR is to reduce the dimensionality inherent in modeling combinations of polymorphisms using a computational approach called constructive induction. Here, we propose a Robust Multifactor Dimensionality Reduction (RMDR) method that performs constructive induction using a Fisher's Exact Test rather than a predetermined threshold. The advantage of this approach is that only those genotype combinations that are determined to be statistically significant are considered in the MDR analysis. We use two simulation studies to demonstrate that this approach will increase the success rate of MDR when there are only a few genotype combinations that are significantly associated with case-control status. We show that there is no loss of success rate when this is not the case. We then apply the RMDR method to the detection of gene-gene interactions in genotype data from a population-based study of bladder cancer in New Hampshire.

To whom correspondence should be addressed: Jason H. Moore, Ph.D., Computational Genetics Laboratory, Norris-Cotton Cancer Center, Dartmouth Medical School, 706 Rubin Bldg, HB7937, One Medical Center Dr., Lebanon, NH 03756, TEL: 603-653-9939, FAX: 603-653-9900, jason.h.moore@dartmouth.edu.

INTRODUCTION

Bladder cancer is the 5th most common cancer in the US and is blamed for about 3% of all cancer deaths. Like many common diseases, the mechanism of bladder cancer is complex, involving many interrelated biochemical pathways under the influence of many genes. The interactions among multiple genetic factors, which are historically referred to as *epistasis*, are likely to play an important role in predicting individual's susceptibility to bladder cancer. Given that interactions are likely to play an important role in the genetic architecture of disease susceptibility, we need to consider interactions of genetic variations in addition to their individual marginal effects. The problem is that interactions are much more difficult to detect and model than marginal effects because of non-additivity in the genotype-to-phenotype mapping relationship and the curse of dimensionality that is inherent in considering the interaction effects of multiple polymorphisms. Addressing this challenge will require a combination of analysis strategies that come from both a bioinformatics and a biostatistics perspective (Cordell, 2009; Moore et al. 2010). We focus here on a bioinformatics approach to modeling gene-gene interactions that uses machine learning to embrace the complexity of the genetic architecture of common human diseases such as bladder cancer.

A variety of bioinformatics methods for identifying gene-gene interactions have been developed (reviewed by Motsinger et al. 2008; Cordell, 2009; Moore et al. 2010). One of these methods, multifactor dimensionality reduction or MDR, was developed as a novel machine learning method for detecting and characterizing gene-gene interactions in the absence of significant marginal effects (Ritchie et al., 2001, 2003; Hahn et al., 2003; Hahn and Moore, 2004; Moore, 2004; Moore et al., 2006; Moore, 2007; Moore and Williams, 2009; Moore et al. 2010). The MDR method uses constructive induction to collapse high-dimensional genotype combinations into a one-dimensional variable with two levels, high risk and low risk, using a threshold that is a function of the number of cases and controls (Moore et al. 2006). Cross-validation and permutation testing are used to limit overfitting and false-positives due to multiple testing. Since its initial description by Ritchie et al. (2001), many modifications and extensions to MDR have been proposed. These include entropy-based interpretation methods (Moore et al., 2006), the use of odds ratios (Chung et al., 2007), log-linear methods (Lee et al., 2007), generalized linear models (Lou et al., 2007), methods for imbalanced data (Velez et al., 2007), permutation testing methods (Pattin et al., 2009; Greene et al., 2010), methods for dealing with missing data (Namkung et al., 2009), model-based methods (Calle et al., 2008), parallel implementations (Bush et al., 2006; Sinnott-Arnstrong et al., 2009; Greene et al., 2010) and different evaluation metrics (Mei et al., 2007; Bush et al., 2008; Namkung et al., 2009). These extensions and modifications of MDR have each addressed different limitations of the approach. One of these limitations is the ability to effectively model gene-gene interactions when many of the genotype combinations have case-control ratios that are close to that observed in the overall data set. Extensions proposed by Lee et al. (2007), Lou et al. (2008) and Calle et al. (2008), for example, all address this problem by applying methods such as the generalized linear model that are subject to certain model assumptions and that significantly decrease the computational efficiency of MDR.

The goal of the present study was to develop and evaluate a model-free and computationally efficient extension to MDR, Robust MDR or RMDR, to address the modeling limitation described above. We use two simulation studies to demonstrate the strengths of our proposed method. We also apply RMDR to a large, population-based study of bladder cancer study from New Hampshire and demonstrate that it is able to identify new gene-gene interactions not previously described.

METHODS AND MATERIALS

The MDR algorithm

MDR is a constructive induction or attribute construction approach that seeks to identify combinations of multi-locus genotypes that are associated with either high risk or low risk of disease. The general process of defining a new attribute as a function of two or more other attributes is referred to as constructive induction and was first described by (Michalski et al., 1983). Constructive induction using MDR was accomplished in this study in the following way (Figure 1A):

1. For a given interaction order M , select M SNPs from the dataset.
2. Construct a contingency table using these M SNPs and calculate case-control ratios for each multi-locus genotype.
3. Let T be the ratio of cases to controls in the whole dataset. For each multi-locus genotype, if the ratio of cases to controls exceeds T , it is considered high-risk. Otherwise, it is considered low-risk. Once all genotypes are labeled 'high-risk' and 'low-risk', a new binary attribute is created by pooling the "high risk" genotype combinations into one group and the "low risk" into another group.

MDR uses a simple probabilistic classifier to model the relationship between attributes or variables constructed using MDR and case-control status that is similar to naïve Bayes (Hahn and Moore, 2004). Naïve Bayes classifiers were assessed using balanced accuracy as recommended by Velez et al. (2007). Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity:

$$\frac{1}{2}(\text{TP}/(\text{TP}+\text{FN})+\text{TN}/(\text{TN}+\text{FP}))=(\text{sensitivity}+\text{specificity})/2 \quad (1)$$

where TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives. For each dataset MDR evaluates all possible M -way interactions and identify the best model using balanced accuracy. MDR uses the following ten-fold cross-validation strategy to determine the M that gives the best overall model:

1. Divide the dataset into 10 parts. Using 9/10 of the data as training set and the rest as testing set.
2. Compute training balanced accuracy for each M -way interaction in the training set.
3. Create a MDR attribute for the testing set using the M SNPs that have the largest training balanced accuracy.
4. Repeat the procedure 10 times so that each sample is included in testing set once.
5. Compute the testing balanced accuracy using the new MDR attribute and the case-control status. For the M -way models that are chosen from the training set, record how many times (cross-validation consistency) they are identified as the best model.

The best MDR model was selected with maximum testing accuracy and maximum cross-validation consistency. The latter is used as a tie break. If both statistics are tied, then the more parsimonious model is chosen as the best model.

A robust MDR algorithm

When there is an empty cell or the case-control ratio equals T (the ratio of cases to controls in the whole dataset) MDR algorithm randomly assign this genotype to high or low risk group resulting in a balanced accuracy of 0.5 for this cell. When the case-control ratio is close to T in one or more genotypes, MDR's balanced testing accuracy is only slightly over or under 0.5 in those cells. The inclusion of such genotype combinations in the constructive induction process negatively impacts the balanced accuracy of the overall model thus influencing the MDR model fitting process. The goal of RMDR is to provide objective statistical criteria using Fisher's Exact Test for determining whether specific genotype combinations should be included in the overall MDR model with the goal of making the final model more robust. Here, we add an 'unknown risk group' for those genotype combinations with a case-control ratio equal or close to T . We then exclude the 'unknown risk group' when we calculate the balanced accuracy of the MDR model. The detailed algorithm is described as follows (Figure 1B):

1. For a given interaction order M , select M SNPs from the dataset.
2. Construct a contingency table using these M SNPs and calculate case-control ratios for each multi-locus genotype.
3. For each multi-locus genotype, Let a be the number of case and b be the number of control. When $T=1$, Run a Fisher's Exact test on contingency table $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$. If the p-value is larger than α , group it to 'unknown risk group'; otherwise if $a > b$, it is considered high-risk and if $b > a$, it is considered low-risk. If $T \neq 1$, The Fisher's Exact test is applied to $\begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix}$, where a_1, b_1 satisfy, $a_1 + b_1 = a + b$ and $\frac{a_1}{b_1} = \frac{a}{bT}$.
4. Once all genotypes are labeled 'high-risk', 'low-risk' and 'unknown risk group', a new three-level attribute is created by pooling the "high risk" genotype combinations into one group, the "low risk" into one group and the 'unknown risk group' into another group.

In step 3, we add the Fisher's Exact Test as a discriminant rule to determine if the case-control ratio is close to T . There are other alternatives such as Pearson's Chi-Square Test that can achieve the same goal here. We picked Fisher's Exact Test mainly because it handles small sample sizes gracefully and offers an exact p-value based on the hypergeometric distribution. Since we are aiming to tackle high-order interactions, the number of cases and controls in most of the multi-locus genotypes are likely to be small. Fisher's Exact Test is a perfect fit for this situation.

Table 1 shows an example of the discriminant rule set generated by Fisher's Exact Test for an significance level of 0.1 and a balanced number of cases and controls across different sample sizes. For example, when there are a total of 10 samples which have a particular multi-locus genotype combination, we label this combination high risk if there are more than seven cases. If there are less than three cases, we label it low risk. Otherwise, we label it as unknown.

RMDR use the same formula (equation 1) to define balanced accuracy, but is limited to samples in the high or low risk groups. An important concern with this approach is the coverage which is defined as the ratio of number of samples in the high and low risk groups to the total sample size. This is because, in some extreme cases, classification of all but one genotype combination into the 'unknown risk group' could result in a balanced accuracy of one from a very small portion of the dataset. To address this concern, we propose to use:

$$Score = (Balanced\ Accuracy - 0.5) \sqrt{Coverage + 0.5} \quad (2)$$

as a measurement to determine how well a RMDR attribute associates with case-control outcome. The cross-validation steps are similar to MDR's with exception that we use the score in equation 2 to replace balanced accuracy. Note that MDR is a special case of RMDR when we use $\alpha = 1$ as the threshold for the Fisher's Exact Test, because the *Coverage* = 1 and RMDR score is the same as MDR's *Balanced Accuracy*. The RMDR approach has been made freely available as part of the open-source MDR software package is available for download from www.epistasis.org.

Simulation study

The goal of the simulation study was to compare the success rate of RMDR with MDR for identifying true interaction models. We considered two different simulation scenarios. The goal of the first scenario was to generate data from a single gene-gene interaction model for which the majority of genotype combinations had case-control ratios consistent with the overall case-control ratio in the data. A secondary goal was to evaluate the influence of imbalanced numbers of cases and controls on the performance of RMDR. In the second more comprehensive scenario, we simulated datasets from the 70 different gene-gene interaction models described by Velez et al. (2007). In this scenario, our goal was to assess the performance of RMDR on a more general class of gene-gene interaction models. For both scenarios, we only considered two-way combinations of SNPs with both MDR and RMDR since we wanted to evaluate whether the new method was able to identify the true function interacting pair of SNPs without the complication of overfitting.

Scenario I—In the first scenario, we simulated datasets based on a nonlinear gene-gene interaction model for which there are by design no marginal effects. The model we selected for illustration purposes is similar to model M186 described by Li and Reich (2000). We simulated 500 datasets from each of the penetrance functions described in table 2. Here k the prevalence of the disease and h is a function of the broad-sense heritability (H) and the case-

control ratio (T) where $h = \sqrt{4TH/(1+T)^2}$. All loci have allele frequencies of 0.5 and genotype frequencies that are consistent with expected Hardy-Weinberg proportions. We simulated data with 200 cases and either 200, 400 or 600 controls. The imbalanced datasets were specifically generated to evaluate the impact of this feature on RMDR. Here, we used case-control ratios of $T = (1, 0.5, .33)$. We picked small to moderate heritabilities of $H = (0.01, 0.025, 0.05, 0.1)$. Each dataset consisted of two functional interacting SNPs and 18 independent non-functional SNPs. Both MDR and RMDR were applied to each dataset as described above. The success rate was estimated as the number of times MDR or RMDR correctly identified the two functional SNPs out of each set of 500 balanced and imbalanced datasets.

Scenario II—In the second scenario we wanted to evaluate RMDR on a more comprehensive and more general set of gene-gene interaction models. Here, we used a comprehensive set of 70 different two-locus penetrance functions that were previously developed by Velez et al. (2007) to evaluate MDR-based methods. The models were distributed evenly across seven broad-sense heritabilities (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, and 0.4) and two different minor allele frequencies (0.2 and 0.4). A total of five models for each of the 14 heritability-allele frequency combinations were generated for a total of 70 models. Genotype frequencies for all 70 epistasis models were consistent with Hardy-Weinberg proportions. A total of 500 datasets were generated for each heritability-allele

frequency combination with 100 total per model. In this study we used a case-control ratio of $T=1$ since the results from scenario I and previous results from Velez et al. (2007) suggest that imbalanced data don't influence MDR or RMDR results when balanced accuracy is used. Each pair of functional polymorphisms was embedded within a set of 20 independent SNPs. A total of 7,000 datasets were generated and analyzed. We estimated success rate as the number of times MDR or RMDR correctly identified the two functional SNPs out of each set of 100 dataset then averaged these results over the five models with the heritability-allele frequency combination.

Real data analysis

We demonstrated the RMDR method with real data by applying it to a genetic epidemiology study that examined the relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility that was previously analyzed using MDR and a 1000-fold permutation test (Andrew et al. 2006). The study analyzed 355 bladder cancer cases and 559 controls ascertained from the state of New Hampshire and focused specifically on genes that play an important role in the repair of DNA sequences that have been damaged by chemical compounds (e.g. carcinogens). Seven SNPs were measured including two from the *X-ray repair cross-complementing group 1* gene (*XRCC1*), one from the *XRCC3* gene, two from the *xeroderma pigmentosum group D* (*XPB*) gene, one from the *nucleotide excision repair* gene (*XPC*), and one from the *AP endonuclease 1* gene (*APE1*). Each of these genes plays an important role in DNA repair. Smoking is a known risk factor for bladder cancer and was included in the analysis along with gender and age for a total of 10 attributes. Age was discretized to $>$ or $\leq$ 50 years. Pack years of smoking (Pkys1) was discretized to never smoked, smoked < 35 pack-years and smoked ≥ 35 pack-years.

A parametric linear statistical analysis of each attribute individually revealed that none of the measured SNPs were significant predictors of bladder cancer individually ($P > 0.05$). Andrew et al. (2006) used MDR to exhaustively evaluate all possible two-, three-, and four-way interactions among the genetic and environmental attributes. For each combination of attributes a single constructed attribute was evaluated using a naïve Bayes classifier. Training and testing accuracy were estimated using 10-fold cross-validation. A best model was selected that maximized the testing accuracy. The best model had a testing accuracy of approximately 0.63 and included two SNPs from the *XPB* gene and smoking. They statistically evaluated this model with a 1000-fold permutation test and determined these results to be highly significant ($p < 0.001$). Post-hoc analysis of the MDR model using entropy-based measures of interaction information revealed that the two *XPB* polymorphisms had evidence of nonlinear interaction or synergy in the near complete absence of marginal effects. Interestingly, the joint effect of the two *XPB* SNPs was larger than the independent from the effect of smoking. The goal of the present study was to determine whether RMDR finds the same best model, and, if so, to compare how the MDR and RMDR models differ.

RESULTS

Simulation results

The success rate of RMDR and MDR to detect the true SNP pairs for the M186 model from scenario I is presented in table 3. For RMDR, we used a more liberal significance threshold of $\alpha = 0.1$ and a standard significance threshold of $\alpha = 0.05$. We can see that when the broad-sense heritability or genetic effect size is low (e.g. 0.01 or 0.025) RMDR has nearly double the success rate as MDR in most cases across different case-control ratios (T). The more conservative significance criteria of 0.05 yielded slightly better success rate for RMDR. Both RMDR and MDR performed similarly at the moderate heritability of 0.1. As

described in the methods, the increased success rate of RMDR observed at the lower heritabilities can be attributed to the chosen simulation model described for simulation scenario I. Here, the number of genotype combinations where the ratio of cases to controls was equal to that of the overall dataset exceeded the number of genotype combinations that accounted for the epistasis effect. Under this specific scenario we expected RMDR to outperform MDR and this is what we observed in the simulation for the lower heritabilities.

The success rate of RMDR and MDR to detect the true SNP pairs across the 70 two-locus epistasis models from scenario II is presented in table 4. There was no appreciable difference in success rate between MDR and RMDR across different heritabilities and allele frequencies. These results document that RMDR has good success rate to detect gene-gene interactions across a wide-range of models that were not selected to have many genotype combinations where the ratio of cases to controls was equal to that of the overall dataset. This confirms that RMDR is generally applicable, as MDR is, to detecting gene-gene interactions in the absence of marginal effects.

Real data analysis results

We applied RMDR with a threshold $\alpha = 0.05$ to the bladder cancer study described above and summarize the top one-way, two-way, three-way, and four-way models in table 5. Here Pkyr1 is a discrete variable with three levels representing increasing pack years of smoking. As reported by Andrew et al. (2006) with MDR, RMDR found that the best model consisted of Pkyr1 and two *XPD* polymorphisms. We verified the statistical significance of this model using a 1000-fold permutation test and found that this best model had a p-value < 0.001 . In other words, an RMDR model with a score as good or better than our best model was never found in 1000 permutations of the data. The level of significance is consistent with that observed by Andrew et al. (2006).

Figure 2A summarizes the original MDR model published by Andrew et al. (2006) while Figure 2B summarizes the new RMDR model. Note the greatly reduced number of genotype-smoking combinations that are labeled high risk or low risk by RMDR compared to MDR. In fact, the number of cells assigned a risk has decreased from 27 to 13. It is much easier to see the marginal effect of smoking since all the cells with Pkyr1=2 (smoked ≥ 35 pack-years) are now labeled high risk. It is also easier to see the genotype-dependent effects that comprise the synergistic gene-gene interaction that was described in detail by Andrew et al. (2006).

DISCUSSION

The goal of this study was to develop a simple method for implementing the MDR method for modeling gene-gene interactions in genetic association studies that takes into account the ratio of cases to controls for each genotype combination. We introduced here a robust MDR or RMDR method that uses Fisher's Exact Test to assess the statistical significance of each multilocus cell for its association with case-control status before including it in an MDR model. Here, we used two different simulation scenarios to demonstrate RMDR's ability to identify complex associations between genotype and phenotype. In the first simulation, RMDR had a clear advantage over MDR when only four out nine genotype combinations from a two-way interaction model are associated with case-control status. In the second set of simulations, we demonstrated that RMDR and MDR perform similarly across a wide range of different gene-gene interactions models confirming that RMDR retains a good success rate even when most of the genotype combinations contribute to the association. Finally, we applied RMDR to a genetic association study of bladder cancer susceptibility that had been previously analyzed using MDR (Andrew et al. 2006). The RMDR analysis confirmed the previous finding of MDR at a similar significance level but resulted in a much

simpler model that reduced the number of genotype combinations from 27 to only 13 thus making this model easier to interpret. We see the improved simplicity of MDR models a clear advantage of the RMDR approach using Fisher's Exact Test.

Despite these advantages, it is important to point out some disadvantages of this approach. For example, running Fisher's Exact Test for every genotype combination does result in a significant computational burden. For example, MDR takes less than one second to search over all one-way, two-way and three-ways models in a datasets with 20 SNPs and a sample size of 400. RMDR takes about three seconds to do the same job. We expect that RMDR will take much more time than MDR when focusing on higher order combinations. On the other hand, the extra computation burden can be significantly reduced if we pre-calculate the discriminant rule (table 1) for all possible sample sizes in any genotypes. For instance, if we have a dataset with sample size of 1000, we could run Fisher's Exact Test 1000 times to get the discriminant rule. We could store the results in a lookup table and then apply it to any genotype combination. Pre-computing the discriminant rules will bring the computation burden of RMDR to a level very close to that of MDR. This will be important for genome-wide association studies, for example.

Future studies with RMDR will investigate measures of entropy as an approach for eliminating multilocus genotype combinations with case-control ratios close to that of the whole dataset. We have previously explored the use of interaction information as a measure of gene-gene interactions (Moore et al. 2006). Bush et al. (2008) have previously used entropy-based contingency table measures to assess overall association of MDR models. However, these measures have yet to be applied to the problem that RMDR addresses. An advantage that entropy might have over Fisher's Exact Test is computational efficiency. A disadvantage is that it typically does not provide a measure of statistical significance. These benefits and potential drawback will be explored and compared within the context of RMDR. It is clear that additional extensions and modifications of the MDR approach will be needed to fully address the complexity of genetic association data and the mapping relationship between genotype and phenotype.

Acknowledgments

This work was funded by grant #IRG-82-003-22 from the American Cancer Society and NIH grants LM009012, LM010098, AI59694, CA57494 and ES007373.

References

- Andrew AS, Nelson HN, Kelsey KT, Moore JH, Meng A, Casella DP, Tosterson TD, Schned AR, Karagas MR. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility. *Carcinogenesis* 2006;27:1030–1037. [PubMed: 16311243]
- Bush WS, Dudek SM, Ritchie MD. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics* 2006;22:2173–2174. [PubMed: 16809395]
- Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics* 2008;9:238. [PubMed: 18485205]
- Calle ML, Urrea V, Malats N, van Steen K. MB-MDR: Model-Based Multifactor Dimensionality Reduction for Detecting Interactions in High-dimensional Genomic Data. Technical report. 2008
- Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics* 2007;23:71–76. [PubMed: 17092990]
- Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;10:392–404. [PubMed: 19434077]

- Greene CS, Himmelstein DS, Kelsey KT, Williams SM, Andrew AS, Karagas MR, Moore JH. Enabling personal genomics with an explicit test of epistasis. *Pacific Symposium on Biocomputing* 2010;327–336. [PubMed: 19908385]
- Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003;19:376–82. [PubMed: 12584123]
- Hahn LW, Moore JH. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol* 2004;4:183–194. [PubMed: 15107022]
- He H, Oetting WS, Brott MJ, Basu S. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC Medical Genetics* 2009;10:127. [PubMed: 19961594]
- Lee SY, Chung Y, Elston RC, Kim Y, Park T. Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions. *Bioinformatics* 2007;23:2589–2595. [PubMed: 17872915]
- Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene by gene and gene by environment interactions with application to nicotine dependence. *Am J Hum Genet* 2007;80:1125–1137. [PubMed: 17503330]
- Mei H, Cuccaro ML, Martin ER. Multifactor dimensionality reduction-phenomics: a novel method to capture genetic heterogeneity with use of phenotypic variables. *Am J Hum Genet* 2007;81:1251–1261. [PubMed: 17999363]
- Michalski RS. A theory and methodology of inductive learning. *Artif Intel* 1983;20:111–161.
- Moore JH. Computational analysis of gene-gene interactions in common human diseases using multifactor dimensionality reduction. *Expert Rev Mol Diagn* 2004;4:795–803. [PubMed: 15525222]
- Moore JH. A global view of epistasis. *Nat Genet* 2005;37:13–14. [PubMed: 15624016]
- Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden W, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 2006;241:252–261. [PubMed: 16457852]
- Moore JH. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zhu, X.; Davidson, I., editors. *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*. Hershey: IGI Press; 2007. p. 17-30.
- Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;85:309–20. [PubMed: 19733727]
- Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010;26:445–55. [PubMed: 20053841]
- Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD. A Comparison of Analytical Methods for Genetic Association Studies. *Genet Epidemiology* 2008;32:767–778. [PubMed: 18561203]
- Motsinger AA, Ritchie MD, Reif DM. Novel methods for detecting epistasis in pharmacogenomics studies. *Pharmacogenomics* 2007;8:1229–1241. [PubMed: 17924838]
- Namkung J, Kim K, Yi S, Chung W, Kwon M, Park T. New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. *Bioinformatics* 2009;25:338–345. [PubMed: 19164302]
- Pattin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KR, Andrew AS, Karagas MR, Moore JH. A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genetic Epidemiology* 2009;33:87–94. [PubMed: 18671250]
- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147. [PubMed: 11404819]
- 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. [PubMed: 12548676]
- Sinnott-Armstrong NA, Greene CS, Cancare F, Moore JH. Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Res Notes* 2009;2:149. [PubMed: 19630950]

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. A balanced accuracy metric for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol* 2007;31:306–315. [PubMed: 17323372]

Figure 1. MDR (A) and RMDR (B) attribute construction. The white cells are labeled as unknown, light grey cells are labeled as low risk and dark grey cells are labeled as high risk. For MDR, we use a threshold $T=1$ to determine the high and low risk group. For RMDR, we use $\alpha = 0.1$ as the threshold for the Fisher's Exact Test.

Figure 2. The best MDR (A) and RMDR (B) interaction models of bladder cancer susceptibility. The white cells are labeled as unknown, light grey cells are labeled as low risk and dark grey cells are labeled as high risk.

Table 1

Discriminant rule of Fisher's Exact Test for different sample number. ($\alpha=0.1$ and $T=1$).

Number of samples in a genotype	10	20	30	40	50	70	100
High risk group case number	8-10	14-20	19-30	25-40	30-50	41-70	57-100
Unknown group case number	3-7	7-13	12-18	16-24	21-29	30-40	44-56
Low risk group case number	0-2	0-6	0-11	0-15	0-20	0-29	0-43

Table 2

Penetrance function for a two-locus epistatic model. The values in parentheses show genotype frequencies that are consistent with expected Hardy-Weinberg proportions.

Genotype	AA (0.25)	Aa (0.5)	aa (0.25)
BB (0.25)	k-h	k	k+h
Bb (0.5)	k	k	k
Bb (0.25)	k+h	k	k-h

Table 3

Success rate of RMDR and MDR to detect the correct two-locus model in scenario I

<i>Heritability (H)</i>	<i>T=1, n=400</i>			<i>T=0.5, n=600</i>			<i>T=0.33, n=800</i>		
	MDR	RMDR-0.05	RMDR-0.1	MDR	RMDR-0.05	RMDR-0.1	MDR	RMDR-0.05	RMDR-0.1
0.01	0.034	0.076	0.064	0.062	0.084	0.09	0.092	0.186	0.184
0.025	0.1	0.242	0.224	0.236	0.468	0.448	0.342	0.646	0.648
0.05	0.35	0.642	0.64	0.678	0.946	0.95	0.824	0.986	0.99
0.1	0.984	1	1	0.998	1	1	1	1	1

Table 4
Success rate of RMDR and MDR to detect the correct two-locus model in scenario II

Heritability (H)	MAF =0.4				MAF =0.2			
	MDR	RMDR-0.05	RMDR-0.1	RMDR-0.1	MDR	RMDR-0.05	RMDR-0.1	RMDR-0.1
0.01	0.474	0.484	0.486	0.486	0.076	0.07	0.07	0.074
0.025	0.832	0.862	0.866	0.866	0.5	0.498	0.504	0.504
0.05	0.958	0.98	0.984	0.984	0.734	0.714	0.71	0.71
0.1	0.994	0.992	0.99	0.99	0.992	0.984	0.988	0.988
0.2	1	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1

Table 5

Top 1-4 ways model identified by RMDR-0.05 for bladder cancer study.

Model	Training score	Testing score	CV consistency
Pkyr1	0.6139	0.6139	10/10
xpd.751 xpd.312	0.6401	0.6291	10/10
xpd.751 xpd.312 Pkyr1	0.6764	0.6429	10/10
xpd.751 xpd.312 xpcinsrtb Pkyr1	0.6840	0.6178	4/10