Genetic Epidemiology 24: 150–157 (2003)

Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity

Marylyn D. Ritchie, Lance W. Hahn, and Jason H. Moore*

Program in Human Genetics and Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, Tennessee

The identification and characterization of genes that influence the risk of common, complex multifactorial diseases, primarily through interactions with other genes and other environmental factors, remains a statistical and computational challenge in genetic epidemiology. This challenge is partly due to the limitations of parametric statistical methods for detecting genetic effects that are dependent solely or partially on interactions with other genes and environmental exposures. We previously introduced multifactor dimensionality reduction (MDR) as a method for reducing the dimensionality of multilocus genotype information to improve the identification of polymorphism combinations associated with disease risk. The MDR approach is nonparametric (i.e., no hypothesis about the value of a statistical parameter is made), is model-free (i.e., assumes no particular inheritance model), and is directly applicable to case-control and discordant sib-pair study designs. Both empirical and theoretical studies suggest that MDR has excellent power for identifying high-order gene-gene interactions. However, the power of MDR for identifying gene-gene interactions in the presence of common sources of noise is not currently known. The goal of this study was to evaluate the power of MDR for identifying gene-gene interactions in the presence of noise due to genotyping error, missing data, phenocopy, and genetic or locus heterogeneity. Using simulated data, we show that MDR has high power to identify gene-gene interactions in the presence of 5% genotyping error, 5% missing data, or a combination of both. However, MDR has reduced power for some models in the presence of 50% phenocopy, and very limited power in the presence of 50% genetic heterogeneity. Extending MDR to address genetic heterogeneity should be a priority for the continued methodological development of this new approach. *Genet. Epidemiol.* 24:150–157, 2003. © 2003 Wiley-Liss, Inc.

Key words: epistasis; multilocus; methods; simulation; complex diseases

Grant sponsor: National Institutes of Health; Grant numbers: HL65234, HL65962, GM31304, AG19085, AG20135, CA78136, LM007450.
*Correspondence to: Jason H. Moore, Ph.D., Program in Human Genetics, Department of Molecular Physiology and Biophysics, 519 Light Hall, Vanderbilt University Medical School, Nashville, TN 37232-0700. E-mail: moore@phg.mc.vanderbilt.edu
Received for publication 29 May 2002; Revision accepted 11 September 2002
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.10218

INTRODUCTION

One goal of genetic epidemiology is to identify polymorphisms associated with common, complex multifactorial diseases. Success in achieving this goal will depend on a research strategy that recognizes and addresses the importance of interactions among multiple genetic and environmental factors in the etiology of diseases such as essential hypertension [Kardia, 2000; Moore and Williams, 2002]. One traditional approach to modeling the relationship between discrete predictors such as genotypes and discrete clinical outcomes is logistic regression [Hosmer and Lemeshow, 2000]. Logistic regression is a parametric statistical approach for relating one or more independent or explanatory variables (e.g., polymorphisms) to a dependent or outcome variable (e.g., disease status) that follows a binomial distribution. However, as reviewed by Moore and Williams [2002], the number of possible interaction terms grows exponentially as each additional main effect is included in the logistic regression model. Thus, logistic regression is limited in its ability to deal with interactions involving many factors. Having too many independent variables in relation to the number of observed outcome events is a well-recognized problem [Concato et al., 1996; Peduzzi et al., 1996] and is an example of the curse of

dimensionality [Bellman, 1961]. In response to this limitation, Ritchie et al. [2001] developed the multifactor dimensionality reduction (MDR) approach that seeks to reduce the dimensionality of multilocus genotype space to facilitate the identification of gene-gene interactions. This approach is nonparametric, is free of a specified genetic model, and is directly applicable to the analysis of case-control and discordant sib-pair study designs [Ritchie et al., 2001]. Further, an MDR software package is available [Hahn et al., 2003].

Empirical studies with both simulated and real data indicate that MDR has good power for identifying high-order gene-gene interactions [Ritchie et al., 2001; Moore and Williams, 2002]. In fact, Ritchie et al. [2001] identified a four-locus interaction that is associated with risk of sporadic breast cancer in the absence of any statistically significant main effects. Further, a theoretical studies study by Hahn and Moore [unpublished findings] has provided a proof that MDR is capable of ideally discriminating between clinical endpoint groups, using multilocus genotypes. This proof shows that no other method will be able to outperform MDR on disease-classification tasks using multilocus genotype data. Thus, MDR is ideally suited for detecting and characterizing gene-gene interactions in epidemiological study designs.

Although empirical and theoretical studies suggest that MDR is a useful method for identifying gene-gene interactions, the power of MDR in the presence of noise that is common to many epidemiological studies is not known. The goal of the present study was to evaluate the power of MDR for identifying gene-gene interactions in the presence of common sources of noise. We evaluated the power of MDR in the presence of noise due to genotyping error, missing data, phenocopy, and genetic or locus heterogeneity.

METHODS

MULTIFACTOR DIMENSIONALITY REDUCTION

Figure 1 illustrates the general procedure to implement the MDR method. In step one, the data are divided into a training set (e.g., 9/10 of the data) and an independent testing set (e.g., 1/10 of

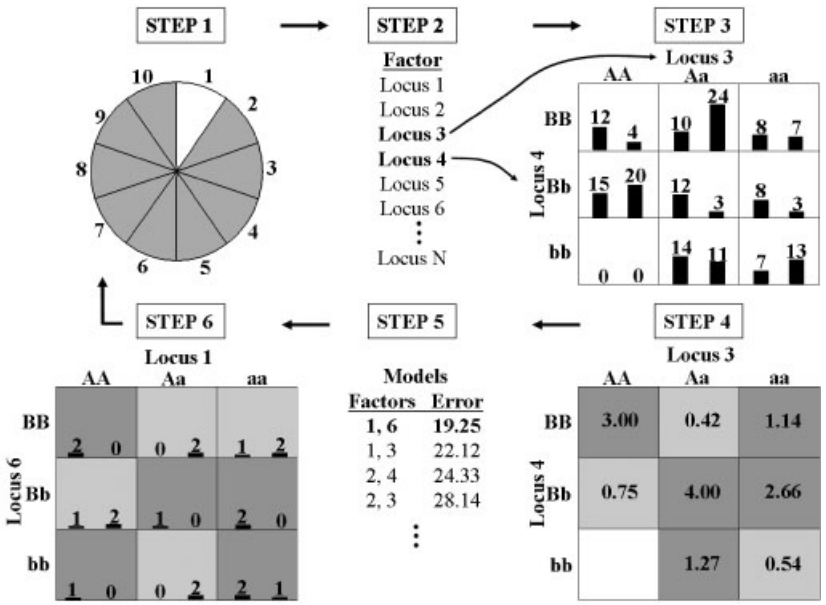

Figure 1. Summary of general steps to implement MDR method [adapted from Ritchie et al., 2001]. In step 1, data are divided into a training set (e.g., 9/10 of the data) and an independent testing set (e.g., 1/10 of the data) as part of cross-validation. In step 2, a set of *n* genetic and/or discrete environmental factors is then selected from the pool of all factors. In step 3, *n* factors and their possible multifactor classes or cells are represented in *n*-dimensional space. In step 4, each multifactor cell in *n*-dimensional space is labeled as high-risk if the ratio of affected individuals to unaffected individuals exceeds some threshold *T* (e.g., *T*=1.0), and low-risk if the threshold is not exceeded. In steps 5 and 6, the model with the best misclassification error is selected, and the prediction error of the model is estimated using the independent test data. Steps 1–6 are repeated for each possible cross-validation interval. Bars represent hypothetical distributions of cases (left) and controls (right) with each multifactor combination. Darker-shaded cells represent high-risk genotype combinations; lighter-shaded cells represent low-risk genotype combinations. No shading or blank cells represent genotype combinations for which no data were observed.

the data) as part of cross-validation. Second, a set of n genetic and/or environmental factors are selected. The n factors and their possible multifactor classes are represented in n-dimensional space, e.g., for two loci with three genotypes each, there are nine possible two-locus-genotype combinations. Then, the ratio for the number of cases to the number of controls is calculated within each multifactor class. Each multifactor cell class in n-dimensional space is then labeled as "high risk" if the cases to controls ratio meets or exceeds some threshold (e.g., ≥ 1), or as "low risk" if that threshold is not exceeded, thus reducing the n-dimensional space to one dimension with two levels ("low risk" and "high risk"). The collection of these multifactor classes composes the MDR model for the particular combination of factors. Among all of the two-factor combinations, a single MDR model that has the fewest misclassified individuals is selected. This two-locus model will have the minimum classification error among the two locus models. In order to evaluate the predictive ability of the model, prediction error is estimated using 10-fold cross-validation. This entire procedure is performed 10 times, using different random number seeds, to reduce the chance of observing spurious results due to chance divisions of the data.

For studies with more than two factors, the steps of the MDR method are repeated for each possible model size (i.e., each number of loci and/or environmental factors), if computationally feasible. The result is a set of models, one for each model size considered. From this set, the model with the combination of loci and/or discrete environmental factors that maximizes the cross-validation consistency and minimizes the prediction error is selected. Cross-validation consistency is a measure of the number of times a particular set of loci and/or factors is identified across the cross-validation subsets [Ritchie et al., 2001; Moore et al., 2002b], and is measured in the following way. For each 10-fold cross-validation, the number of times the same set of loci/factors was identified across the 10 data subsets is recorded. The minimum value is 1 if the combination of factors occurs in only one subset, and the maximum value is 10 if the same combination of loci/factors is identified across all 10 subsets. Thus, there are 10 possible values of the measure of consistency ranging from 1–10, where 10 is considered strong evidence in favor of a multifactor association. Prediction error is a measure of how well the MDR model predicts risk status in independent test sets. The prediction error is calculated as the average of prediction errors across each of the 10 cross-validation subsets. When cross-validation consistency is maximal for one model and prediction error is minimal for another model, the model with the fewest loci/factors is selected. Hypothesis testing of this final best model can then be performed by evaluating the magnitude of cross-validation consistency and prediction error. We determined statistical significance by comparing the average cross-validation consistency and average prediction error from the observed data to the distribution of average consistencies or errors under the null hypothesis of no associations, derived empirically from 1,000 permutations. Here, consistency and error is evaluated for the best model identified by MDR in each permuted dataset. The null hypothesis was rejected when the upper-tail Monte Carlo P-value derived from the permutation test was 0.05.

The MDR method is currently limited to loci and environmental factors with 2 or 3 levels, but no more. Thus, it is ideally suited for single-nucleotide polymorphisms (SNPs) with two alleles and discrete environmental factors or other covariates. No assumptions about the independence or biological relevance of SNPs or any other factor are made, since MDR selects the combination of factors providing the most information. The MDR software will currently support disease models with up to 15 factors at a time from a list of up to 500 total factors and a maximum sample size of 4,000 subjects. The MDR method is described in further detail by Ritchie et al. [2001] and reviewed by Moore and Williams [2002]. An MDR software package is available from the authors by request, and is described in detail by Hahn et al. [2003]. More information can be found at http://phg.mc.vanderbilt.edu/Software/MDR

DATA SIMULATION

To evaluate the power of MDR for detecting gene-gene interactions, we simulated case-control data using six different two-locus epistasis models in which the functional loci are single-nucleotide polymorphisms (SNPs). The first model was initially described by Li and Reich [2000], and later by Moore et al. [2002a]. This model is based on the nonlinear XOR function [Anderson, 1995] that generates an interaction effect in which high risk of disease is dependent on inheriting a heterozygous genotype (*Aa*) from one locus or a heterozygous genotype (*Bb*) from a second locus,

but not both. The high-risk genotype combinations are *AaBB*, *Aabb*, *AABb*, and *aaBb*, all with penetrances of 0.1 (Fig. 2A). The second model was initially described by Frankel and Schork [1996] and later by Moore et al. [2002a]. In this model, high risk of disease is dependent on inheriting exactly two high-risk alleles (*A* and/or *B*) from two different loci. For this model, the high-risk genotype combinations are *AAbb*, *AaBb*, and *aaBB*, with penetrances of 0.1, 0.05, and 0.1, respectively (Fig. 2B). The remaining four models were generated using the epistasis model discovery method of Moore et al. [2002a], using allele frequencies of p=0.25 and q=0.75 for models 3 (Fig. 2C) and 4 (Fig. 2D), and allele frequencies of p=0.1 and q=0.9 for models 5 (Fig. 2E) and 6 (Fig. 2F). All of these models were selected because they exhibited interaction effects in the absence of any main effects when genotypes were generated according to Hardy-Weinberg proportions. Interactions without main effects are desirable, because they provide a high degree of complexity to challenge the ability of a method to identify gene-gene interactions. If main effects were present, it could be difficult to evaluate whether particular loci were detected because of the main effects, or because of the interactions, or both.

Our goal was to simulate data under these six epistasis models in the absence or presence of noise due to genotyping error, missing data, phenocopy, and genetic (locus) heterogeneity. These sources of noise were selected because they

Model 1

	BB	Bb	bb
AA	0	.10	0
Aa	.10	0	.10
aa	0	.10	0

(A) $p = 0.5, q = 0.5$

Model 2

	BB	Bb	bb
AA	0	0	.10
Aa	0	.05	0
aa	.10	0	0

(B) $p = 0.5, q = 0.5$

Model 3

	BB	Bb	bb
AA	.08	.07	.05
Aa	.10	0	.10
aa	.03	.10	.04

(C) $p = 0.25, q = 0.75$

Model 4

	BB	Bb	bb
AA	0	.01	.09
Aa	.04	.01	.08
aa	.07	.09	.03

(D) $p = 0.25, q = 0.75$

Model 5

	BB	Bb	bb
AA	.07	.05	.02
Aa	.05	.09	.01
aa	.02	.01	.03

(E) $p = 0.1, q = 0.9$

Model 6

	BB	Bb	bb
AA	.09	.001	.02
Aa	.08	.07	.005
aa	.003	.007	.02

(F) $p = 0.1, q = 0.9$

Figure 2. Multilocus penetrance functions and allele frequencies (p, q) used to simulate case-control data exhibiting gene-gene interactions in absence of main effects. Marginal penetrances for each genotype are not shown, but are all equal for each specific model.

are commonly encountered in genetic epidemiology studies. Thus, for each epistasis model, we simulated 100 datasets with no noise, and 100 datasets for each noise type (5% genotyping error, 5% missing data, 50% phenocopy, or 50% genetic heterogeneity). In addition, we simulated 100 datasets with each combination of 2, 3, or all 4 sources of noise to evaluate their combined effect on power. Thus, a total of 96 different sets of 100 datasets was generated. Genotyping error was simulated using a directed-error model [Akey et al., 2001]. This model simulates systematic genotyping errors that result in overrepresentation of one allele. For each locus, a bias towards the *a* or the *A* allele was proscribed. Five percent of the genotypes were selected and, unless it was already homozygous in the biased direction, the genotype was changed so that it had one more of the overrepresented alleles. Phenocopies were simulated such that 50% of the affected individuals had genotype combinations that were consistent with low risk according to the epistasis model. These individuals were assumed to be affected due to random environmental factors. Fifty percent genetic heterogeneity was simulated such that there were actually two different two-locus combinations that increased the risk of disease. Half of the affected individuals had one high-risk genotype combination, and the other half had the other high-risk genotype combination. Each dataset consisted of 200 cases and 200 controls, each with 10 SNPs, 2 of which were functional. Each SNP had two alleles with the common allele having a frequency of 0.5, 0.75, or 0.9, as described above for the six different models. Genotypes were generated according to Hardy-Weinberg proportions. All datasets are available from the authors by request.

DATA ANALYSIS

The datasets were first analyzed using a chi-square test of independence with two degrees of freedom to characterize the independent main effects of the 10 simulated SNPs. We estimated the power of the chi-square test for each of the functional SNPs as the proportion of significant chi-square results in each group of 100 datasets, using a significance level of 0.05. The type I error rate for the nonfunctional SNPs was also estimated. Next, we analyzed each of the datasets with MDR by conducting an exhaustive search of all possible 2–5-locus interactions. We applied the MDR algorithm as described above, using

a cases-to-controls threshold ratio of 1:1 as described by Ritchie et al. [2001]. Each dataset was analyzed using 10-fold cross-validation. In addition, the 10-fold cross-validation was performed 10 times, and the prediction error and cross-validation consistency were averaged across the 10 trials. This reduced the possibility of biased estimates due to chance divisions of the data. We estimated the power of MDR under each epistasis model as the number of times MDR identified the functional SNPs out of each set of 100 datasets.

RESULTS

The results of the chi-square test of independence indicate that the epistasis models exhibit little or no independent main effects. As summarized in Table I, the power to detect each functional SNP was very close to 5%. Further, the observed type I error rate for the nonfunctional SNPs was also very close to the expected type I error rate of 0.05. For the functional SNPs, these results were expected due to the deliberate absence of main effects in the six epistasis models.

The power of MDR to detect the correct 2 or 4 (in the case of genetic heterogeneity) functional SNPs from each of the epistasis models, with and without common sources of noise, is presented in Table II. For all of the models, the power to identify functional SNPs ranged from 80–100% in the absence of noise. The presence of 5% genotyping error, 5% missing data, or the combination of both had very little effect on power for any of the six models. The presence of 50% phenocopy had the largest impact on the power for models 3–6, with an overall decrease in power from 99% to 45% for model 3, 99% to 32% for model 4, 82% to 30% for model 5, and 84% to 32% for model 6. There did not appear to be much of a synergistic effect of phenocopy and noise due to genotyping error and/or missing data on power.

Of all the sources of noise, 50% genetic or locus heterogeneity had the greatest impact on power. For every model except model 2, the power dropped from greater than 80% to less than 5%. For model 2, the power dropped from 100% to 41%. When combined with phenocopy, the power for model 2 dropped to 1%. These results suggest that locus heterogeneity has a very large impact on power regardless of the particular epistasis model. In fact, when combined with phenocopy, power never exceeded 5%.

DISCUSSION

Multifactor dimensionality reduction is a promising new approach for overcoming some of the limitations of logistic regression [Moore and Williams, 2002] for the detection and

TABLE I. Power and Type I error rates for Chi-square test of independence in single-locus tests of association[a]

Source of noise	Range of power (%) to detect functional SNPs across six models				Range of type I error (%) for nonfunctional SNPs across six models					
	1	2	3[b]	4[b]	5	6	7	8	9	10
None	1–9	2–10	3–10	1–7	5–13	3–7	3–6	1–6	3–7	1–7
GE	4–10	2–7	2–8	1–8	4–10	0–8	3–11	2–6	3–6	2–7
GH	4–8	3–7	2–11	1–6	3–10	2–6	1–7	1–6	2–8	2–9
PC	2–8	5–9	4–8	0–8	3–7	2–6	2–7	3–10	2–6	1–6
MS[1]	2–9	2–7	2–11	2–8	4–10	4–8	3–10	2–8	2–9	5–7
GE+GH	4–7	2–5	2–11	1–6	3–10	2–8	1–8	1–6	2–8	2–9
GE+PC	4–9	2–9	4–8	2–5	5–11	2–8	3–9	1–7	3–8	4–7
GE+MS	1–9	1–8	3–9	1–9	3–8	2–6	3–10	1–6	3–11	4–7
GH+PC	1–10	2–7	4–8	1–8	6–10	2–7	2–6	3–8	4–8	3–8
GH+MS	4–9	3–7	1–10	2–4	3–10	1–7	0–5	2–9	1–8	1–9
PC+MS	1–9	3–7	5–9	2–6	5–7	2–7	4–7	2–8	2–6	3–6
GE+GH+PC	2–7	2–9	2–9	3–6	5–11	2–7	3–8	2–6	2–6	1–9
GE+GH+MS	4–5	2–7	1–11	2–5	3–9	3–7	0–8	1–6	2–7	2–7
GH+PC+MS	1–6	2–5	4–13	1–5	4–7	2–9	4–7	3–5	2–8	1–7
GE+PC+MS	2–4	2–5	2–7	3–5	6–9	2–12	3–10	3–7	2–9	1–6
GE+GH+PC+MS	2–8	3–9	3–11	3–6	5–9	1–8	3–9	1–10	3–8	3–7

[a]GE, 5% genotyping error; GH, 50% genetic heterogeneity; PC, 50% phenocopy; MS, 5% missing data.
[b]Note that these two loci are only functional in GH datasets. Therefore, for non-GH datasets, the power value should be interpreted as type I error.

TABLE II. Power of MDR to detect the correct two functional interacting loci[a]

Source of noise	Power (%)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
None	100	100	99	99	82	84
GE	100	100	100	97	80	92
GH	3	41	2	3	4	4
PC	90	99	45	32	30	32
MS	100	100	99	97	82	87
GE+GH	4	41	2	3	4	6
GE+PC	94	99	41	48	28	33
GE+MS	100	100	98	98	74	84
GH+PC	0	1	0	0	0	0
GH+MS	5	38	0	2	4	6
PC+MS	96	99	42	43	14	16
GE+GH+PC	1	1	0	0	0	0
GE+GH+MS	6	34	2	1	3	7
GH+PC+MS	0	0	0	0	0	0
GE+PC+MS	94	100	48	42	18	16
GE+GH+PC+MS	0	1	0	1	0	0

[a]GE, 5% genotyping error; GH, 50% genetic heterogeneity; PC, 50% phenocopy; MS, 5% missing data.

characterization of gene-gene and gene-environment interactions. Previous empirical studies demonstrated that MDR has good power for identifying high-order interactions in simulated data [Ritchie et al., 2001], as well as real data from case-control studies of sporadic breast cancer [Ritchie et al., 2001] and essential hypertension [Moore and Williams, 2002]. Additionally, a theoretical study provided proof that MDR is ideally suited for discriminating between binary clinical endpoints using multilocus genotypes [Hahn and Moore, unpublished findings], and a software package is available [Hahn et al., 2003]. The main conclusion that can be drawn from the present study is that MDR has excellent power to identify gene-gene interactions even in the presence of 5% genotyping error, 5% missing data, or a combination of both for a wide range of different two-locus epistasis models with varying allele frequencies. The presence of 50% phenocopy had a significant effect on power for models 3–6, but not for models 1 or 2. The greatest reduction in power was observed in the presence of 50% genetic or locus heterogeneity. In fact, in the presence of 50% locus heterogeneity, the power was less than 5% for all but one model. These results suggest that more empirical and theoretical work is needed to improve the power of MDR for identifying functional loci when genetic heterogeneity is present. Since it is anticipated that genetic heterogeneity is likely to be common for complex multifactorial diseases such as type 2 diabetes mellitus [Busch and Hegele, 2001], we suggest that extending MDR to deal with this complicating factor should be a priority.

Why was the power to detect functional loci from model 2 much greater than for the other models in the presence of genetic heterogeneity? This was most likely due to the relative simplicity of model 2 compared to the other models. For example, model 2 has only three high-risk genotype combinations compared to at least four for all the other models. Additionally, the high-risk cells are symmetric about the diagonal and, as a result, each single genotype is associated with high-risk only in the presence of a single genotype from the other locus. This is perhaps one of the simplest two-locus epistasis models that exhibit no independent main effects [Moore et al., 2002a]. In contrast, model 1 has four high-risk cells, and each single genotype is associated with high risk in the presence of two genotypes from the other locus. Model 1 is an example of the XOR function, a classic nonlinear Boolean operator that is not linearly separable [Anderson, 1995]. It is interesting to note that the addition of 50% phenocopy to the heterogeneity dataset for model 2 dropped the power to less than 5%, a value similar to all the other models. Thus, even though model 2 is a simpler model, the combined effect of both genetic heterogeneity and phenocopy significantly reduced power, while phenocopy alone did not impact power.

It should also be noted that the power decreased slightly as the frequency of the rare allele decreased. The largest drop in power occurred

for models 5 and 6, that had a rare allele frequency of 0.1. Under Hardy-Weinberg equilibrium, there are only four two-locus genotype combinations that are common enough to be represented in a modest size case-control study as was simulated here. This results in a decrease in size of the interaction effect, even though the model satisfies the mathematical criteria for an interesting epistasis model exhibiting no independent main effects, as defined by Moore et al. [2002a]. These studies preliminarily suggest that increased sample size may be necessary to detect gene-gene interactions when the interacting loci have a relatively rare allele.

What can be done to improve the power of MDR in the presence of phenocopy? Since phenocopies are the result of an environmental disease etiology, it is unlikely that any methodological changes to MDR will improve power. Rather, power is likely to improve if the appropriate environmental risk factors can be included in the analysis. For example, the genetic effect may be stronger in a particular environmental context in which case a stratified MDR analysis could be carried out. Alternatively, the environmental factor could be included in the MDR analysis. The challenge, of course, is to identify the appropriate environmental risk factors beforehand.

How should MDR be extended to account for genetic heterogeneity? When applied to real data, it might be possible to identify clusters of individuals that have a similar genetic background prior to analysis by MDR. The use of cluster analysis in case-control studies to identify genetic heterogeneity was suggested by Schork et al. [2001], and is based on the idea that genetic heterogeneity appears as differences in genetic background that may reflect different populations of origin. Schork et al. [2001] proposed that if significant evidence for clusters was identified, it would then be possible to incorporate information about the clusters into an association study. In a preliminary study of renal failure subjects, Schork et al. [2001] clustered cases and controls into three distinct groups, using 44 microsatellite markers. An association analysis indicated that allelic variation in their candidate gene was not statistically independent of case and control clusters. A subsequent logistic regression analysis indicated that the allelic variation was still significantly associated with renal failure, even after considering cluster status. Thus, it may be possible to first perform a cluster analysis on cases and controls using the available genetic markers, and then perform a subsequent MDR analysis on the resulting clusters or by using cluster status as a covariate.

Methods and approaches other than cluster analysis have been proposed for deriving more homogeneous subsets of data for genetic analysis [Province et al., 2001]. For example, Shannon et al. [2001] explored the use of recursive partitioning tree-based methods for identifying homogeneous subgroups of sib-pairs to increase the power of linkage analysis. This approach is based on the classification and regression tree models of Breiman et al. [1984]. With classification trees, observations are subdivided into groups, using information about covariates such that observations within a group are more similar to each other than they are to observations from other groups. Ideally, covariates would subdivide the data such that observations within each group were homogeneous with respect to the outcome variable class. Using simulated data, Shannon et al. [2001] showed that this sort of recursive partitioning could greatly improve the overall power to detect linkage. Although specifically designed for sib-pair linkage analysis, it may be possible to extend this approach to case-control data and integrate it with MDR.

Both cluster analysis and recursive partitioning are promising methods for dealing with genetic heterogeneity by creating more homogeneous subgroups that can be incorporated into an association analysis. Future studies will need to investigate which methods are most appropriate for use with case-control association studies. Further, the most appropriate strategy for incorporating the results of a cluster- or recursive partitioning-type analysis with the MDR approach will need to be investigated. Regardless, this is a promising direction for beginning to address the genetic heterogeneity issue. Other ideas, such as incorporating an OR function into the MDR method, should also be explored. It is clear that the development and evaluation of new methods for addressing gene-gene interactions and other complicating issues in genetic association studies such as genetic heterogeneity should be a priority in genetic epidemiology.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

Akey JM, Zhang K, Xiong M, Doris P, Jin L. 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 68:1447–56.

Anderson JA. 1995. An introduction to neural networks. Cambridge, MA: MIT Press.

Bellman R. 1961. Adaptive control processes. Princeton: Princeton University Press.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and regression trees. New York: Chapman and Hall.

Busch CP, Hegele RA. 2001. Genetic determinants of type 2 diabetes mellitus. Clin Genet 60:243–54.

Concato J, Feinstein AR, Holford TR. 1996. The risk of determining risk with multivariable models. Ann Intern Med 118:201–10.

Frankel WN, Schork NJ. 1996. Who's afraid of epistasis? Nat Genet 14:371–3.

Hahn LW, Ritchie MD, Moore JH. 2003. Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions. Bioinformatics (in press).

Hosmer DW, Lemeshow S. 2000. Applied logistic regression. New York: John Wiley & Sons, Inc.

Kardia SLR. 2000. Context-dependent genetic effects in hypertension. Curr Hypertens Rep 2:32–8.

Li W, Reich J. 2000. A complete enumeration and classification of two-locus disease models. Hum Hered 50:334–49.

Moore JH, Williams SM. 2002. New strategies for identifying gene-gene interactions in hypertension. Ann Med 34:88–95.

Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC. 2002a. Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics. In: Langdon WB, Cantu-Paz E, Mathias K, Roy R, Davis D, Poli R, Balakrishnan K, Honavar V, Rudolph G, Wegener J, Bull L, Potter MA, Schultz AC, Miller JF, Burke E, Jonoska N, editors. Proceedings of the Genetic and Evolutionary Computation Conference. San Francisco: Morgan Kaufmann Publishers. p 1150–5.

Moore JH, Parker JS, Olsen NJ, Aune T. 2002b. Symbolic discriminant analysis of microarray data in autoimmune disease. Genet Epidemiol 23:57–69.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. 1996. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49:1373–9.

Province MA, Shannon WD, Rao DC. 2001. Classification methods for confronting heterogeneity. Adv Genet 42:273–86.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–47.

Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D. 2001. The future of genetic case-control studies. Adv Genet 42:191–212.

Shannon WD, Province MA, Rao DC. 2001. Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. Genet Epidemiol 20: 293–306.