For reprint orders, please contact:
reprints@futuremedicine.com

REVIEW

Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies

Marylyn D Ritchie[†] *& Alison A Motsinger*

[†]*Author for correspondence*
Vanderbilt University Medical Center, Department of Molecular Physiology & Biophysics, 519 Light Hall, Center for Human Genetics Research, Nashville, TN 37232–0700, USA
Tel.: +1 615 343 5851;
Fax: +1 615 343 8619;
E-mail: ritchie@chgr.mc.vanderbilt.edu

In the quest for discovering disease susceptibility genes, the reality of gene–gene and gene–environment interactions creates difficult challenges for many current statistical approaches. In an attempt to overcome limitations with current disease gene detection methods, the multifactor dimensionality reduction (MDR) approach was previously developed. In brief, MDR is a method that reduces the dimensionality of multilocus information to identify polymorphisms associated with an increased risk of disease. This approach takes multilocus genotypes and develops a model for defining disease risk by pooling high-risk genotype combinations into one group and low-risk combinations into another. Cross-validation and permutation testing are used to identify optimal models. While this approach was initially developed for studies of complex disease, it is also directly applicable to pharmacogenomic studies where the outcome variable is drug treatment response/nonresponse or toxicity/no toxicity. MDR is a nonparametric and model-free approach that has been shown to have reasonable power to detect epistasis in both theoretical and empirical studies. This computational technology is described in detail in this review, and its application in pharmacogenomic studies is demonstrated.

One of the biggest challenges in human genetics is identifying polymorphisms, or sequence variations, that lead to an increased risk of disease. In the case of rare, Mendelian single-gene disorders, such as sickle-cell anemia or cystic fibrosis, the genotype–phenotype relationship is easily evident, as the mutant genotype is explicitly responsible for disease. In the case of common, complex diseases such as hypertension, diabetes, or multiple sclerosis, this relationship is extremely difficult to characterize, since disease is likely to be the result of many genetic and environmental factors. In fact, epistasis, or gene–gene interaction, is increasingly assumed to play a crucial role in the genetic architecture of common diseases [1–3]. This challenge is equally present in studies of pharmacogenomics [4].

The dimensionality involved in the evaluation of combinations of multiple variables quickly diminishes the value of traditional, parametric statistical methods. Referred to as the curse of dimensionality [5], as the number of genetic or environmental factors increases and the number of possible interactions increases exponentially, many contingency table cells will be left with very few, if any, data points. This can result in increased Type I errors and parameter estimates with very large standard errors [6–7]. Traditional approaches using logistic regression modeling are limited in their ability to deal with many factors, and simultaneously fail to characterize epistasis models in the absence of main effects due to the hierarchical model-building process [8]. This leads to an increase in Type II errors and decreased power [9]. This is a particular problem with relatively small sample sizes. Since sample collection is time consuming and expensive, the decreased power can make effective studies cost prohibitive with traditional analytical methods.

To deal with these issues, much research is required for improved statistical methodologies. Many researchers are exploring variations and modifications of logistic regression, such as logic regression [10], penalized logistic regression [11], and automated detection of informative combined effects (DICE) [12]. Additional explorations are being conducted in data mining and machine-learning research, including data reduction and pattern recognition approaches. Data reduction involves a collapsing or mapping of the data to a lower dimensional space. Examples of data reduction approaches include the combinatorial partitioning method (CPM) [13], restricted partition method (RPM) [14], and multifactor dimensionality reduction (MDR) [15–17]. In contrast, pattern recognition involves extracting patterns from the data to discriminate between groups using the full dimensionality of the data. Examples of pattern recognition methods include cluster analysis [18], cellular automata (CA) [19], support vector machines (SVM) [20], self-organizing maps (SOM) [21], and neural

Keywords: epistasis, gene–environment interactions, gene–gene interactions, multifactor dimensionality reduction, pharmacogenomics

future medicine

networks (NN) [22]. These methodologies have very different theoretical bases, while they share a common goal of more efficient data exploration and analysis.

Multifactor dimensionality reduction

A novel computational approach for the detection of complex gene–gene and gene–environment interactions has previously been developed. MDR is a data reduction method for detecting multilocus genotype combinations that predict disease risk for common, complex diseases. MDR pools genotypes into high-risk and low-risk or response and nonresponse groups, in order to reduce multidimensional data into only one dimension. It is a nonparametric method, therefore no hypothesis concerning the value of any statistical parameter is made. It is also model free, thus no genetic inheritance model is assumed [16].

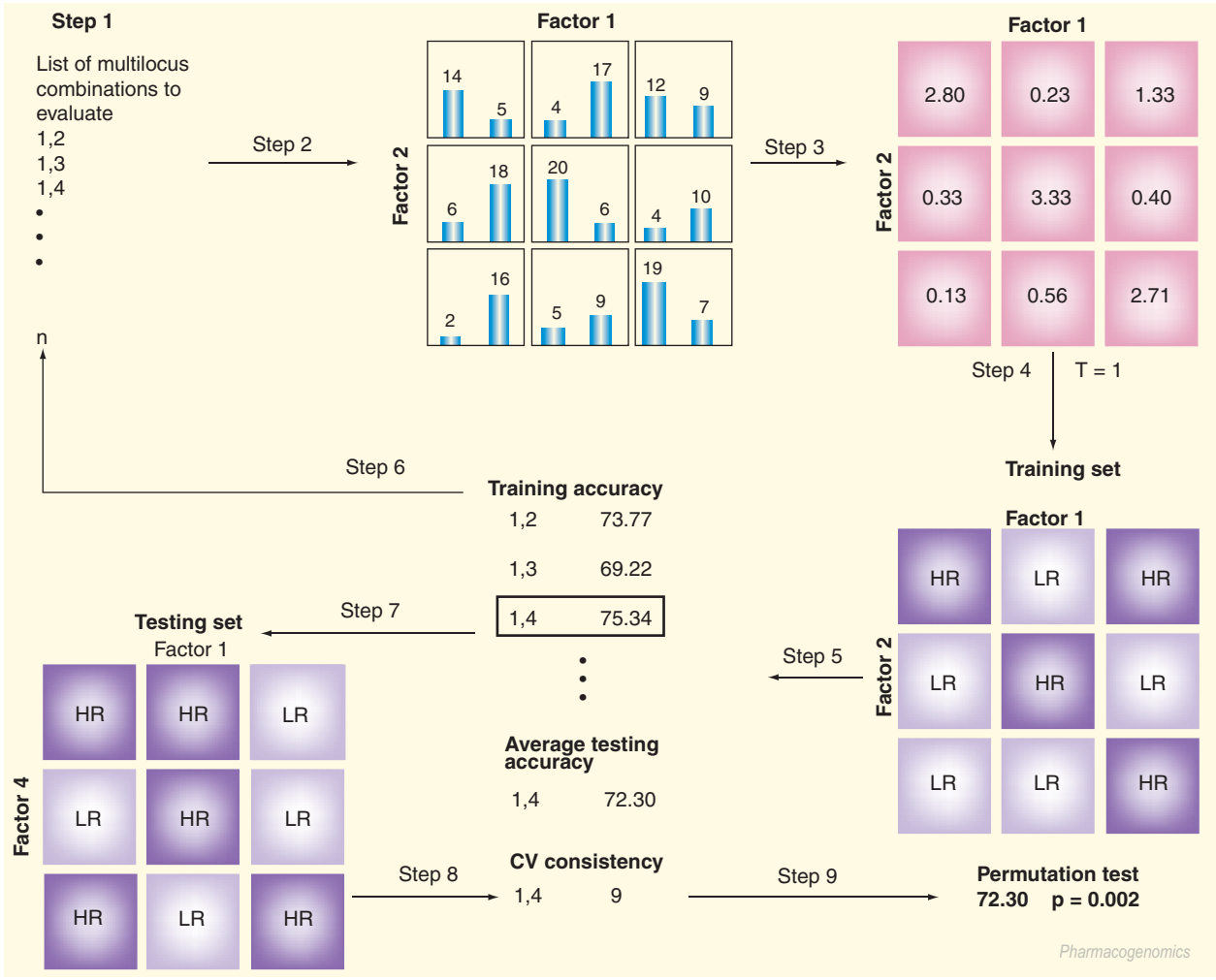
As mentioned earlier, traditional statistical approaches were not successful in detecting gene–gene and gene–environment interactions associated with common, complex diseases and, similarly, pharmacogenomic end points. Many of these challenges were due to the application of the methodologies, more so than the theoretical basis behind them. However, to achieve success in detecting epistasis, our goal was to explore different alternatives and move in a completely different direction. Thus, a nonparametric, model-free approach was highly desirable.

MDR was designed to detect gene–gene or gene–environment interactions in categorical independent variables (such as genotype and environmental data) and a dichotomous dependent variable (such as case/control status or drug treatment response/nonresponse). MDR performs an exhaustive search of all possible single-locus through k-locus interactions (as computationally feasible) to evaluate all possible combinations of loci. As a result of the evaluations, MDR will select a single combination of loci as the optimal model for each level of k, creating a set of models for k = 1 to M (where M is the maximum interaction level analyzed). This set of models can then be compared using various statistics to determine if there is a single best model, or multiple significant models. This process requires a heuristic approach by the end user, as this will differ depending on the goal of the analysis. If one views the application of MDR as a hypothesis-generating exercise, and would prefer to select a few false positives more so than missing a true signal (false negative), then the selection of all statistically significant models would be optimal. However, if one is trying to select a single best model for the purposes of replication or diagnostic test design, then more stringent selection criteria may be used.

MDR algorithm

Figure 1 demonstrates the process for the MDR algorithm. Before the MDR analysis begins, the data set is divided into multiple partitions for cross-validation, if cross-validation is specified. MDR can be performed without cross-validation; however, this is rarely performed due to the potential of overfitting [23]. Cross-validation [21] is an important part of the MDR method, as it aims to find a model that not only fits the given data, but can also predict on future, unseen data. Since attainment of a second data set for testing is time-consuming and often cost-prohibitive, cross-validation produces a testing set from the given data to evaluate the predictive ability of the model produced. Typically, in the case of tenfold cross-validation, the training set is comprised of 9/10 of the data, while the testing set is comprised of the remaining 1/10 of the data. First, an exhaustive list of n combinations of loci to evaluate from the list of all categorical independent variables is generated. These variables can include both genetic and environmental data. There is no set limit on the number of independent variables that can be examined. However, limits due to computation time may arise. In simulations, all single locus and pair-wise interactions in simulated data sets with up to 50,000 single nucleotide polymorphisms (SNPs) genotyped in 1000 individuals have been analyzed [24]. Next, each of the n combinations is arranged in a contingency table in k-dimensional space, with all possible multifactorial combinations represented as cells in the table. The number of cases and controls for each locus combination are counted. In the third step, the ratio of cases:controls within each cell is calculated. Each multilocus genotype combination is then labeled as high risk or low risk based on comparison of the ratio to a threshold (T = 1). Therefore, if the ratio within a multifactor combination is above one it is labeled as high risk for disease, and if it is below one it is labeled as low risk for disease. This step compresses multidimensional data into one dimension with two classes. Similarly, for pharmacogenomic end points, each genotype combination would be labeled response or nonresponse based on the ratio of responders to nonresponders. This threshold, T = 1 has been used previously for both balanced and unbalanced data sets.

Figure 1. Summary of the general steps to implement the MDR method.



Step one: The exhaustive list of n combinations are generated from the pool of all independent variables. Step two: For k = 1 to M, the combinations are represented in k-dimensional space, and the number of cases and controls are counted in each multifactor cell. Step three: The ratio of cases to controls is calculated within each cell. In step four, each multifactor cell in the k-dimensional space is labeled as high risk (HR) if the ratio of affected individuals to unaffected individuals exceeds a threshold (T = 1), and low risk (LR) if the threshold is not exceeded. In step five the training accuracy is calculated. This is then repeated for each multifactor combination. In step seven, the model with the best training accuracy is selected and evaluated in the test set. In step eight, the testing accuracy of the model is estimated. In step nine, a permutation test is conducted to determine the statistical significance of the model(s). Steps 1 through 8 are repeated for each possible cross-validation interval. Bars represent hypothetical distributions of cases (left) and controls (right) with each multifactor combination. Dark-shaded cells represent HR genotype combinations, while light-shaded cells represent LR genotype combinations.

CV: Cross-validation; HR: High risk; LR: Low risk; MDR: Multifactor dimensionality reduction.

Traditionally for unbalanced data, oversampling and undersampling has been used as opposed to altering the threshold value. Further research is being conducted to fully understand the implications of adjusting the threshold, as well as over and undersampling for unbalanced data.

The disease risk distribution for each of the multifactorial combinations represents the MDR model for a particular combination of multilocus genotypes. The classification error, or one minus the training accuracy, for each model is calculated based on the number of individuals within the model that are actually cases in genotype combinations classified as low risk and the number of individuals that are actually controls in the genotype combinations classified as high risk. The best k locus model is selected and the model is evaluated against the testing group and testing accuracy is calculated. Prediction error, or one minus the testing accuracy, is based on the

number of misclassified individuals in the testing set, based on the model developed in the training set. Error and accuracy are used interchangeably in the MDR literature, as error is calculated by: 1 minus accuracy. This is repeated for each cross-validation interval (i.e., training set and testing set) and the average training accuracy and testing accuracy are calculated across all intervals. Among all of the k-locus models created, the single model with the highest cross-validation consistency is chosen as the best k-locus model. This process is completed for k = 1 to M loci combinations that are computationally feasible. An optimal k-locus model is chosen for each level of k considered; thus if k = 1–3 is tested, a one-locus model, two-locus model, and three-locus model will each comprise the final set. Traditionally, once this set of models is completed, a final model or set of models are chosen. The final model is selected based on maximization of both testing accuracy and cross-validation consistency. Testing accuracy is how well the model predicts risk/disease status in independent testing sets generated through cross-validation. The average testing accuracy for the model is calculated by subtracting the average prediction error based on the ten testing sets from one. The testing accuracy is averaged across the cross-validation intervals regardless of the cross-validation consistency. This will give a measure of testing accuracy for the k-level model. Once it is determined that certain k-level interactions are of interest, a more accurate estimate of the testing accuracy for that particular combination of loci can be calculated by forcing only those variables into an MDR model. Cross-validation consistency is the number of times a model is identified as the best model across the cross-validation sets. Therefore, for tenfold cross-validation, the consistency can range from one to ten. The higher the cross-validation consistency is, the stronger the support for the model. When testing accuracy and cross-validation consistency indicate different models, the rule of parsimony can be used to choose between them. Here, one might be interested in selecting the simpler model (i.e., the model with fewer factors). However, as discussed below, it is often not advantageous to select only one best model.

Once the best/final model is chosen, permutation testing is used to test the significance of the hypothesis generated. Permutation testing involves creating multiple permuted data sets by randomizing the disease status labels. Typically, at least 1000 randomized data sets are generated. The theory behind permutation testing is to create a distribution of a statistic, here testing accuracy, that could be expected simply by chance [25]. For MDR, this distribution must be created for each individual data set, mimicking the configuration parameters and data set characteristics of the original MDR analysis. Thus, the entire MDR procedure is repeated for each randomized data set. The single best model from all levels of interaction is extracted for each random data set as described above, which generates a distribution of 1000 testing accuracies that could be expected by chance alone. This would be considered the omnibus permutation test. Some users of MDR have used an alternative k-locus permutation test, where a separate distribution is created for each level of k [26–27]. The significance of the final model is determined by comparing the testing accuracy of the final model to the distribution. A p-value is extracted for the model by its location in this empirical distribution. This omnibus permutation test may be a conservative method. However, it is more likely to control for Type I error while not limiting power.

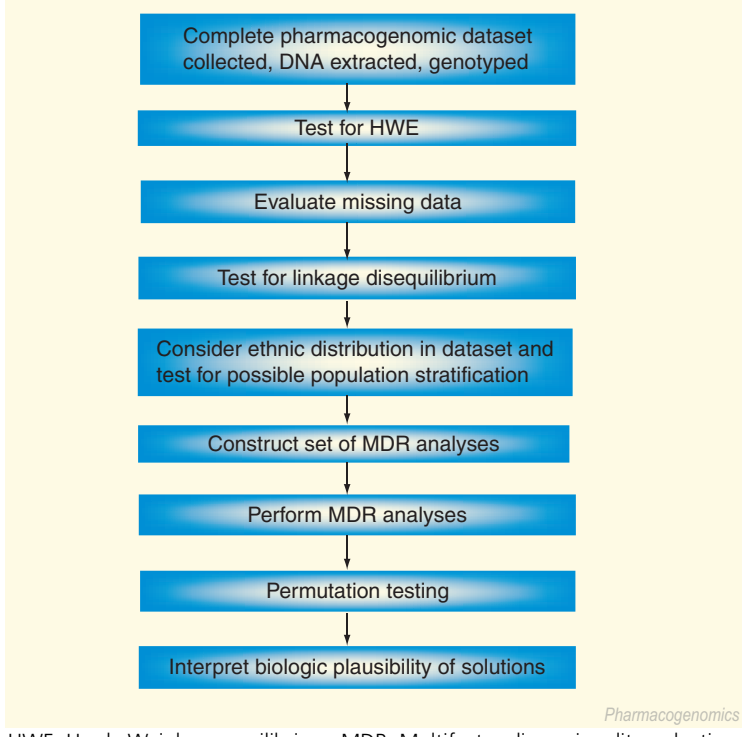
More recently, less emphasis has been placed on choosing a single final model, and instead a set of interesting models are generated. Significance levels are assigned to each model in the final set using the procedure described above and then all significant models are reported. This new approach attempts to use all information within the final set of models. If the end goal of the MDR method is hypothesis generation, this approach may be preferred to reduce the risk of false negatives.

MDR data analysis flow

Application of MDR to a data analysis plan is merely one component of a complex process. **Figure 2** shows an example of a possible workflow for a pharmacogenomic analysis. This workflow assumes that the study subjects have already been collected, DNA samples extracted and genotyping has been performed. Each of these components of pharmacogenomic studies are huge challenges in themselves and a discussion of the details of each step of the study design is out of the scope of this report. Here, the process once the data set has been collected and assembled is described.

First, the integrity of the data set should be evaluated. There are a number of mechanisms by which quality control can be performed on a data set. In a genomic study examining unrelated individuals, the two main concerns that should be evaluated are the possibility of genotyping

Figure 2. Flow chart of workflow in an MDR analysis.



HWE: Hardy-Weinberg equilibrium; MDR: Multifactor dimensionality reduction.

error and the inherent patterns of missing data. The most common test performed to check for genotyping error in case–control data is a test for Hardy-Weinberg equilibrium (HWE) [28–30]. Theoretically, disease-free control groups should follow HWE. Similarly, a combined group of cases and controls with the same disease, such as in a pharmacogenomic study where all individuals have the disease and case/control status is determined by response to treatment, should also follow HWE [28]. Often, if the test for HWE shows that a marker is out of HWE in the entire data set, this may be evidence for genotyping error. However, this test is challenging when deviations are observed from expected HWE frequencies in either the case or control group. Studies have shown that this may be a direct result of a biologically meaningful result, and not the presence of genotyping error [28–32]. Thus, caution should be taken when evaluating these results. The other quality control measure that is important to evaluate is the missing data distribution. Data missing at random is inevitable in almost all studies consisting of large, complex data sets [21]. Likewise, it is rarely problematic when the quantity of missing data is sparse, and the distribution of the missing data is random. However, if the data are missing not-at-random, this can lead to enormous problems in the analysis of statistical results. In addition, the manner in which the missing data are dealt with, including data imputation, sample deletion and variable deletion, will heavily depend on the amount and distribution of the missing data [21–33].

Once the data are clean and ready for analysis, it is important to understand the genetic patterns in the data set. For example, an analysis of the linkage disequilibrium (LD) patterns in the data set should be performed if multiple markers from the same genes have been genotyped [34–37]. This can be done using one of many freely available software packages, such as Haploview [38]. It is important to understand the degree of correlation between the markers before any statistical analysis, including an MDR analysis, is performed, so that the interpretation of results can be maximized. Further research is currently being conducted to understand the impact that LD has on an MDR analysis. An additional consideration is the ethnicity distribution of the individuals in the data set. If the data are comprised of multiple ethnic groups, it is important to consider this in the analyses. If the frequency of alleles and the frequency of the clinical end point are similar, then the data may be able to be merged for MDR analysis. However, if there are differences in allele frequency or in the frequency of response/nonresponse status, then the groups should be analyzed separately. This will prevent any spurious findings due to population stratification. Testing for population stratification can be performed in a variety of ways, including genomic control [39] and using software such as structure and STRAT [40], as well as Bayesian methods [41].

Determination of the analyses to be performed is the next step before an MDR analysis takes place. It has been shown that creating subsets of the complete data set based on the biochemical pathways that the different genes belong to is a successful strategy for an MDR analysis [42]. The creation of this list of separate analyses should be compiled before any analyses are performed to ensure that the user has planned possible scenarios that have the opportunity to produce biologically meaningful results. Once this list has been assembled, the MDR analyses can be performed.

The user should run the MDR analysis on each data set with or without cross-validation depending on their preference. The interpretation of results and selection of best models will be different depending on whether cross-validation

was used. For example, if cross-validation is used, then the best models are those that have high cross-validation consistency (the number of times the same model is selected across the cross-validation intervals) and maximum testing accuracy. All models with statistically significant testing accuracy can be reported as interesting results, but the cross-validation consistency may provide the final determination of a single best model. If cross-validation is not used, training accuracy or classification error is the typical metric of model fitness.

Once the analyses are completed, all data sets that are of interest can be evaluated through permutation testing. Here, the models that have statistically significant results, not expected by chance alone, can be determined [25]. The stringency of the cutoff value selection is again dependent on the goal of the analysis. In most data-mining exercises, a p-value of 0.05, or even 0.10, is used as a first stage cutoff. All models that meet this level of significance can be evaluated in a second independent data set, and those that replicate at a more stringent p-value can be considered of greatest interest.

Finally, the interpretation of results concludes the MDR analysis. Unfortunately, this is often the most challenging aspect and, in many cases, still cannot be fully understood. The distribution of responders and nonresponders can be visualized as part of the output of the MDR software (**Figure 3**). This provides an idea of the way in which the high/low risk genotype combinations are distributed. However, making biologic interpretations of this model will require follow-up studies in model organisms, cell culture or other *in vitro* experiments. However, it is important to note which genes are selected in the best models when multiple data sets, or many levels of interaction, have been performed. For example, if markers in the same gene continuously result as the best model, this could indicate that there is a high degree of LD in that gene and the solutions are indicating a single biologic signal. Similarly if cross-validation was used, and the best two-locus model had a low cross-validation consistency, but an excellent testing accuracy, it would be wise to explore the raw results to see what other models were selected in the different cross-validation intervals. It is conceivable that there is LD between the markers and this yields a result where some of the intervals show one model and other intervals show the other model. The final result is a low consistency for both models, where in reality, it is one signal.

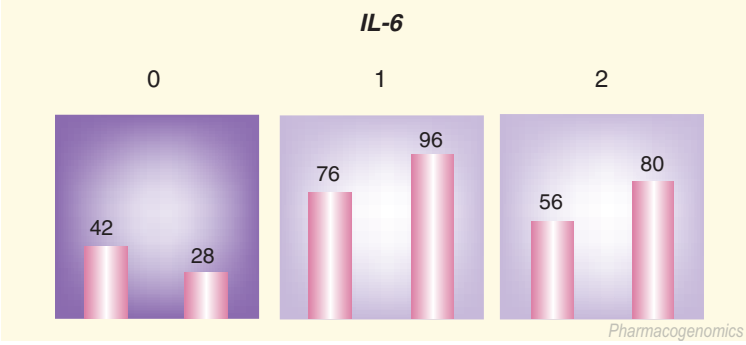
Another important consideration is whether or not the interaction models detected have been observed in other genetic association studies, or more importantly in any biochemical or model organism experiments. If a biomolecular interaction is known to occur, and the best MDR model consists of markers in genes coding for the two proteins, then this will help to interpret/explain the statistical results. However, it is important to note that evidence for statistical epistasis detected by MDR does not necessarily mean that biologic epistasis exists for these genes [43].

Summary of previous applications

MDR has been used to identify higher order interactions in the absence of any significant main effects in simulated data [16–17]. Studies with simulated data (of multiple models of different allele frequencies and heritability) have also shown that MDR has a high power to identify interactions in the presence of many types of noise commonly found in real data sets (including missing data and genotyping error), while errors such as heterogeneity (genetic or locus), and phenocopy diminish the power of MDR [17]. Additionally, a mathematical proof has shown that no other method will discriminate between clinical end points using multilocus genotype data more efficiently than MDR [44]. In addition, MDR has demonstrated gene–gene interactions in a variety of different real data sets, including sporadic breast cancer [16], essential hypertension [8], Type II diabetes [45], atrial fibrillation [46], amyloid polyneuropathy [47], coronary artery calcification [48], autism [26] and schizophrenia [49].

The first application of MDR to pharmacogenomics data was in a study of atorvastatin-induced muscle damage [50]. Atorvastatin-induced muscle damage can be monitored clinically through lab tests of circulating blood creatine kinase levels. Atorvastatin-induced myopathy is typically rare when atorvastatin is prescribed as monotherapy, but increases as additional drugs metabolized by members of the cytochrome P450 (*CYP*)*3A* gene family are prescribed. Thus, the goal was to determine if polymorphisms in *CYP3A* could predict the risk of developing atorvastatin-induced muscle damage. This retrospective study consisted of 68 cases of muscle damage (increased serum creatine kinase [CK] levels) and 69 controls that experience no muscle damage (normal serum CK levels). Here MDR identified a statistically significant interactive effect between gender and lipid status associated with differential elevation in serum CK levels ($p < 0.05$). This result suggested a potential

Figure 3. Distribution of cases and controls in the application of MDR to postoperative atrial fibrillation.



Single-locus MDR model demonstrating the effect of *IL-6*, which is able to correctly predict disease status with 58.8% accuracy. Light purple cells are low-risk, while dark purple cells are high risk. The number of cases is shown in the histogram on the left in each cell, while controls are shown by the histogram on the right. The genotype labels indicate number of *IL-6* promoter -174 G alleles. IL: Interleukin; MDR: Multifactor dimensionality reduction.

genotype–lipid status interaction; however, this could not be confirmed in this study [50].

One of the previously mentioned successful applications of MDR was reported in a study of atrial fibrillation (AF) in a Chinese population. A case–control study was conducted in 250 patients with documented AF and 250 matched controls. Eight polymorphisms in genes from the renin–angiotensin system were selected due to previous associations with cardiovascular phenotypes [46]. Several main effects were detected, including three polymorphisms in the angiotensinogen gene, as well as a three-locus interaction between two polymorphisms in angiotensinogen and the angiotensin converting enzyme gene insertion/deletion. This three-locus model correctly predicted disease status 62.74% of the time ($p < 0.001$) [46].

Another recent application of MDR involved a treatment response phenotype. Postoperative atrial fibrillation (PoAF) is the most common arrhythmia following heart surgery, and continues to be a major cause of morbidity [51–53]. Due to the complexity of this condition, many genes and/or environmental factors may play a role in susceptibility. Previous findings have shown several clinical and genetic risk factors for the development of PoAF [54]. The goal of this study was to determine whether interactions among candidate genes and a variety of clinical factors are associated with PoAF [55].

Using the steps outlined in **Figure 2**, the authors have conducted a study of PoAF. The MDR method was applied to detect interactions in a sample of 940 adult subjects undergoing elective procedures of the heart or great vessels, requiring general anesthesia and sternotomy or thoracotomy, where 255 developed PoAF. A random sample of controls matched to the 255 AF cases was taken for a total sample size of 510 individuals. Polymorphisms in three (interleukin [*IL*]-*6*, angiotensin I converting enzyme [*ACE*], and apolipoprotein E [*APOE*]) candidate genes, all of which were previously implicated in PoAF risk, and 36 clinical factors were chosen for analysis.

The data set was collected, DNA extracted and genotyping conducted as part of the Pharmacogenetics Research Network (PGRN): Pharmacogenetics of Arrhythmia Therapy center at Vanderbilt University (U01: HL65962). HWE was tested in the data set and no statistically significant deviations from HWE were detected. The amount of missing data did not warrant further data manipulations or imputations. The missing genotypes were coded as a separate categorical level. Since the polymorphisms were in different genes, there was no need to test for linkage disequilibrium in this data set. In addition, the data set consisted of over 94% Caucasians, thus no stratified analyses by race were performed.

Three separate MDR analyses were conducted: a genetic analysis, clinical risk factor analysis, and a gene–risk factor analysis. The gene–risk factor analysis results were identical to the clinical risk factor only analysis. Thus, only the results of the two analyses were reported: genetic analysis and clinical risk factor analysis. The results of the analyses are shown in **Table 1 & 2**. In the genetic analysis, a single locus effect of *IL-6* that is able to correctly predict disease status with 58.8% ($p < 0.001$) accuracy was detected. To verify that this was the best genetic model, the cross-validation consistency and testing accuracy statistics were used. The goal is to find the model with the maximum cross-validation consistency and the maximum testing accuracy. In addition, if multiple models have statistically significant testing accuracies, all significant models might be reported as interesting for follow up. In this study, the single locus model of *IL-6* had both the maximum testing accuracy (58.8%) and the maximum cross-validation consistency (10), and was the only statistically significant model. **Figure 3** shows the distribution of cases and controls for the single locus MDR model generated. Based on logistic regression, *IL-6* has an odds ratio of 1.144 (95% confidence interval: 1.035–1.264). That *IL-6* was shown to have an association with disease risk replicates the findings of Gaudino and

Table 1. Results of MDR analysis on genetic factors.

Number of loci	Polymorphism in model	Cross validation consistency	Testing accuracy
1	*IL-6*	10	58.80*
2	*IL-6, APOE4*	5	53.20
3	*IL-6, ACE, APOE3*	9	53.99

**P < 0.001. ACE: Angiotensin-converting enzyme; APOE: Apolipoprotein E; IL: Interleukin; MDR: Multifactor dimensionality reduction.*

colleagues [56], providing support for a postulated role for activation of inflammatory pathways in this [57] and perhaps other forms of AF [58]. This underscores a possible role for anti-inflammatory approaches for the prevention of this common complication.

In the clinical risk factor analysis, an interaction between history of AF and length of hospital stay that predicted disease status with 68.54% ($p < 0.001$) accuracy was detected. Again, we used cross-validation consistency and testing accuracy. Here, all three models have statistically significant testing accuracies. Thus, we have selected the two-factor model as the best model, since it has the highest testing accuracy (68.54%). However, all three models are significant and should be considered of interest for follow up. **Figure 4** shows the distribution of cases and controls for the interactive model. Based on a logistic regression analysis of these two main effect terms and their interaction term, all three terms are statistically significant (history of AF $p < 0.001$, length of stay $p < 0.001$, interaction $p < 0.04$). PoAF is known to prolong length of hospital stay, and preoperative history of AF is a risk factor for postoperative AF [59–60]. The interaction model detected using MDR is consistent with these findings. The occurrence of multiple significant models also demonstrates the extreme complexity of the phenotype and could imply the importance of complicating issues such as heterogeneity and phenocopy.

These findings demonstrate the utility of novel computational approaches for the detection of disease susceptibility genes. These results also showcase the value of being able to detect both main effects and interactions. While each of these results appears to be of interest, they only explain part of PoAF susceptibility. It will be important to collect a larger set of candidate genes and environmental factors to better characterize the development of PoAF. Applying this approach, we were able to elucidate potential associations with postoperative atrial fibrillation [55].

Outlook

The original distributed MDR software was available as Linux®, UNIX®, and MAC OS® command line software. Presently, MDR software is being distributed in a Java software package with a user-friendly graphical interface or a command-line option. The most current open-source version is available at [101]. MDR has also been added to Weka-CG, which is available from the same website. The MDR software will continue to be distributed through the website [101] and includes modules for data manipulation for preprocessing and permutation testing.

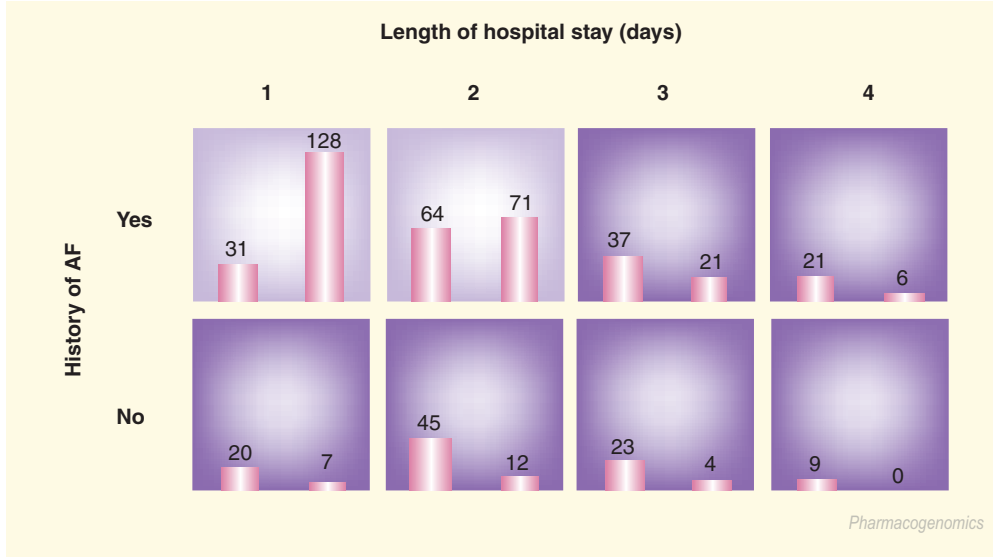
In addition to advances being made in the software, more research questions continue to be followed-up. For example, questions regarding adequate sample size and statistical power for a given study arise as each new study begins. For MDR, there is currently no theoretical power calculation that can be performed. Thus

Table 2. Results of MDR analysis on clinical factors.

Number of loci	Variable in model	Cross validation consistency	Testing accuracy
1	Length of stay	10	66.94*
2	History of AF, length of stay	8	68.54*
3	AF at time of surgery, age, length of stay	3	61.36*

**p < 0.001. AF: Atrial fibrillation; MDR: Multifactor dimensionality reduction.*

Figure 4. MDR model demonstrating an interaction between history of AF and length of hospital stay that predicted disease status with 68.54% ($p < 0.001$) accuracy.



Light purple cells are low-risk, while dark purple cells are high risk. The number of cases is shown in the histogram on the left in each cell, while controls are shown by the histogram on the right. Length of stay was defined as the date difference between surgery date and discharge date. It is coded as follows: 1 = 0–4 days, 2 = 5–7 days, 3 = 8–13 days, 4 = 14 or more days. These cutoffs correspond to the 40th, 80th and 95th percentile cutoffs. AF: Atrial fibrillation; MDR: Multifactor dimensionality reduction.

far, only empirical power studies have been conducted. In previously published studies, it has been shown that a sample size of 200 cases and 200 controls is sufficient for detecting two-locus interaction models [17]. Continued simulation studies are being conducted to evaluate adequate sample sizes for detecting higher order interactions. As part of this set of simulation studies, the power, Type I error, and bias/variance statistics on the prediction error estimates for a variety of sample sizes and effect sizes are continuously evaluated. Classification error, or training accuracy, is the current fitness function utilized for selection of the best MDR model to be evaluated during testing. However, a number of other possible functions could be used, including sensitivity, specificity, χ^2, and others. The authors are currently evaluating which, if any, of these additional functions improves the power of MDR. Cross-validation is an optional technique to use in conjunction with MDR. Bootstrapping is an additional technique that may provide the ability to place confidence intervals on the prediction error estimates for the best models, as well as perform an MDR analysis effectively without cross-validation. This technique is currently being evaluated. Additionally, we have merged MDR with a more traditional technique in statistical genetics, the pedigree-disequilibrium test (PDT) [61], to create MDR-PDT. MDR-PDT will allow for the evaluation of complex epistasis models in studies of family data including discordant sibships and families with an affected child and known parental genotype data. The original version of MDR is capable of evaluating family data as a matched case–control study, but MDR-PDT has the added benefit of calculating the geno-PDT statistic and using all information provided in family data [62].

Finally, whole-genome association studies are the future of pharmacogenomics. Many studies will likely adopt this unbiased approach for association analysis. The development of the International HapMap Project has provided a wealth of information regarding the common variation of the human genome [63]. Perhaps one of the most exciting implications of this work is the development of high-throughput genotyping technologies designed to facilitate whole-genome association analyses. By surveying a large number of SNPs densely scattered throughout the genome, case–control studies of common complex diseases and pharmacogenomics may utilize these technologies to capture much of the

Highlights

- Epistasis, or gene–gene interaction, is an important aspect of common, complex disease and pharmacogenomic studies.
- Detecting epistasis is challenging for traditional statistical methods, thus new methods are being developed.
- Multifactor dimensionality reduction (MDR) is a novel computational approach for detecting gene–gene and gene–environment interactions.
- MDR has high power for detecting interactions and has been demonstrated in both simulated and real data sets.
- MDR can be used for pharmacogenomic end points, such as drug treatment response/nonresponse.
- MDR has successfully detected single factor and two-factor effects in postoperative atrial fibrillation.

variation present in affected individuals. It is the goal of these large-scale studies to pinpoint genetic variations that contribute to the susceptibility of complex diseases. The hundreds of thousands of markers typed in these studies constitute massive amounts of data, and the analytical challenge of data analysis on this scale is a major hurdle for their success [64].

Epistasis is often found when properly investigated; however, whole-genome studies prove especially difficult for the analysis of interactions, as the combinatorial nature of the problem exponentially increases the number of statistical tests required. MDR might have the potential for application in population-based case–control whole-genome association studies. Further investigations into the computational feasibility of such studies are warranted. One advance that is currently being evaluated is a parallel implementation of the MDR algorithm. Parallel MDR is a new implementation of the MDR algorithm that can scale to handle extremely large data sets, dramatically decreases single-processor runtimes, and can also use a parallel software framework to allow operation in a clustered computing environment to further reduce runtime. Research is currently underway to estimate the number of factors that will be practical to analyze with this software. It is hoped that these advances will provide further enhancement of this powerful computational methodology.

Acknowledgments

This work was supported by National Institutes of Health grants GM62758, AG20135, and in part by HL65962, the Pharmacogenomics of Arrhythmia Therapy U01 site of the Pharmacogenetics Research Network.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Sing CF, Stengard JH, Kardia SL: Dynamic relationships between the genome and exposures to environments as causes of common human diseases. *World Rev. Nutr. Diet* 93, 77–91 (2004).
2. Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* 56, 73–82 (2003).
• **Describes the rationale and importance of studying epistasis.**
3. Thornton-Wells TA, Moore JH, Haines JL: Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.* 20, 640–647 (2004).
4. Wilke RA, Reif DM, Moore JH: Combinatorial pharmacogenetics. *Nature Rev. Drug Discov*. (2005) (In press).
5. Bellman R: Adaptive Control Processes. Princeton University Press, Princeton, USA (1961).
6. Peduzzi P, Concato J, Feinstein AR, Holford TR: Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J. Clin. Epidemiol.* 48, 1503–1510 (1995).
7. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49, 1373–1379 (1996).
8. Moore JH,Williams SM: New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.* 34, 88–95 (2002).
9. Moore JH: Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert. Rev. Mol. Diagn.* 4, 795–803 (2004).
10. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L: Sequence analysis using logic regression. *Genet. Epidemiol.* 21(Suppl. 1), S626–S631 (2001).
11. Zhu J, Hastie T: Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5, 427–443 (2004).
12. Tahri-Daizadeh N, Tregouet DA, Nicaud V, Manuel N, Cambien F, Tiret L: Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Res.* 13, 1952–1960 (2003).
13. Nelson MR, Kardia SL, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11, 458–470 (2001).
14. Culverhouse R, Klein T, Shannon W: Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* 27, 141–152 (2004).
15. Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19, 376–382 (2003).
• **Describes the original distribution of the multifactor dimensionality reduction (MDR) software.**
16. Ritchie MD, Hahn LW, Roodi N *et al.*: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147 (2001).
• **Original MDR paper discussing the algorithm, some simulations, and the first real data application.**
17. 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.* 24, 150–157 (2003).

• **Describes a more thorough simulation study of MDR in data with diferent sources of error.**

18. Kaufman L,Rousseeuw PJ: Finding groups in data: an introduction to cluster analysis.Wiley-Interscience Publication, New York, NY, USA (1990).
19. Wolfram S: *Cellular Automata and Complexity.* Addison-Wesley Company, Reading, MA, USA (1994).
20. Cristianini N, Shawe-Taylor J: *An introduction to support vector machines*. Cambridge University Press, Cambridge, UK (2000).
21. Hastie T, Tibshirani R, Friedman JH: *The elements of statistical learning*. Springer Series in Statistics. Springer Verlag, Basel, Switzerland (2001).
22. Ripley BD: *Pattern Recognition via Neural Networks.* Cambridge University Press, Cambridge, UK (1996).
23. Coffey CS, Hebert PR, Ritchie MD *et al.*: An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene–gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics* 5, 49 (2004).
24. Dudek S, Motsinger AA, Velez D, Williams SM, Ritchie MD: Data simulation software for whole-genome association and other studies in human genetics. *Pac. Symp. Biocomput.* (2006) (In press).
25. Good P: Permutation Tests: *A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, New York, USA (2000).
26. Ma DQ, Whitehead PL, Menold MM *et al.*: Identification of significant association and gene–gene interaction of GABA receptor subunit genes in autism. *Am. J. Hum. Genet.* 77, 377–388 (2005).
27. Mei H, Ma D, Ashley-Koch A, Martin ER: Extension of multifactor dimensionality reduction for identifying multi-locus effects in the GAW14 simulated data. *BMC Genetics* (2005) (In press).
28. Salanti G, Amountza G, Ntzani EE, Ioannidis JP: Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *Eur. J. Hum. Genet.* 13, 840–848 (2005).
29. Sham PC: Statistics in Human Genetics. Arnold Publishers, London, UK (2001).
30. Khoury MJ: Fundamentals of genetic epidemiology. Oxford University Press, New York, NY, USA (1993).
31. Wittke-Thompson JK, Pluzhnikov A, Cox NJ: Rational inferences about departures from Hardy-Weinberg equilibrium. *Am. J. Hum.Genet.* 76, 967–986 (2005).
32. Weir BS, Hill WG, Cardon LR: Allelic association patterns for a dense SNP map. *Genet. Epidemiol.* 27, 442–450 (2004).
33. Little R,Rubin D: *Statistical Analysis with Missing Data*.Wiley, New York, USA (1987).
34. Lonjou C, Zhang W, Collins A *et al.*: Linkage disequilibrium in human populations. *Proc. Natl Acad. Sci. USA* 100, 6069–6074 (2003).
35. Hill W: Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33, 229–239 (1974).
36. Lewontin R: On measures of gametic disequilibrium. *Genetics* 120, 849–852 (1988).
37. Lewontin R,Kojima K: The evolutionary dynamics of complex polymorphisms. *Evolution* 14, 450–472 (1960).
38. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265 (2005).
39. Devlin B,Roeder K: Genomic control for association studies. *Biometrics* 55, 997–1004 (1999).
40. Pritchard JK,Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 65, 220–228 (1999).
41. Hoggart CJ, Parra EJ, Shriver MD *et al.*: Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* 72, 1492–1504 (2003).
42. Williams SM, Ritchie MD, Phillips JA *et al.*: Multilocus analysis of hypertension: a hierarchical approach. *Hum. Hered.* 57, 28–38 (2004).
43. Moore JH,Williams SM: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27, 637–646 (2005).
44. Hahn LW, Moore JH: Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol.* 4, 183–194 (2004).
45. Cho YM, Ritchie MD, Moore JH *et al.*: Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47, 549–554 (2004).
46. Tsai CT, Lai LP, Lin JL *et al.*: Renin–angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 109, 1640–1646 (2004).
47. Soares ML, Coelho T, Sousa A *et al.*: Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. *Hum. Mol. Genet.* 14, 543–553 (2005).
48. Bastone L, Reilly M, Rader DJ, Foulkes AS: MDR and PRP: a comparison of methods for high-order genotype–phenotype associations. *Hum. Hered.* 58, 82–92 (2004).
49. Qin S, Zhao X, Pan Y *et al.*: An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (*GRIN1*) and NR2B subunit gene (*GRIN2B*) in schizophrenia with universal DNA microarray. *Eur. J. Hum. Genet.* 13, 807–814 (2005).
50. Wilke RA, Moore JH, Burmester JK: Relative impact of *CYP3A* genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. *Pharmacogenet. Genomics.* 15, 415–421 (2005).

•• **First pharmacogenomic application of MDR.**

51. Aranki SF, Shaw DP, Adams DH *et al.*: Predictors of atrial fibrillation after coronary artery surgery. Current trends and impact on hospital resources. *Circulation* 94, 390–397 (1996).
52. Hravnak M, Hoffman LA, Saul MI, Zullo TG, Whitman GR, Griffith BP: Predictors and impact of atrial fibrillation after isolated coronary artery bypass grafting. *Crit. Care Med.* 30, 330–337 (2002).
53. Ommen SR, Odell JA, Stanton MS: Atrial arrhythmias after cardiothoracic surgery. *N. Engl. J. Med.* 336, 1429–1434 (1997).
54. Darbar D, Herron KJ, Ballew JD *et al.*: Familial atrial fibrillation is a genetically heterogeneous disorder. *J. Am. Coll. Cardiol.* 41, 2185–2192 (2003).
55. Motsinger AA, Donahue BS, Brown NJ, Roden DM, Ritchie MD: Risk factor Interactions and genetic effects associated with post-operative atrial fibrillation. *Pac. Symp. Biocomput.* (2005) (In press).

• **Second pharmacogenomics application, which is also reviewed more thoroughly in this second review paper.**

56. Gaudino M, Andreotti F, Zamparelli R *et al.*: The -174G/C interleukin-6 polymorphism influences postoperative interleukin-6 levels and postoperative atrial fibrillation. Is atrial fibrillation an

inflammatory complication? *Circulation* 108(Suppl. 1), II195–II199 (2003).

57. Aviles RJ, Martin DO, Apperson-Hansen C *et al.*: Inflammation as a risk factor for atrial fibrillation. *Circulation* 108, 3006–3010 (2003).
58. Chung MK, Martin DO, Sprecher D *et al.*: C-reactive protein elevation in patients with atrial arrhythmias: inflammatory mechanisms and persistence of atrial fibrillation. *Circulation* 104, 2886–2891 (2001).
59. DiDomenico RJ, Massad MG: Pharmacologic strategies for prevention of atrial fibrillation after open heart surgery. *Ann. Thorac. Surg.* 79, 728–740 (2005).
60. Mathew JP, Fontes ML, Tudor IC *et al.*: A multicenter risk index for atrial fibrillation after cardiac surgery. *JAMA* 291, 1720–1729 (2004).
61. Martin ER, Monks SA, Warren LL, Kaplan NL: A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.* 67, 146–154 (2000).
62. Martin ER, Ritchie MD, Hahn LW, Kang S, Moore JH: A novel method to identify gene-gene effects in nuclear families: The MDR-PDT. *Genet. Epidemiol.* (2005) (In press).
63. The International HapMap Project. *Nature* 426, 789–796 (2003).
64. Moore JH, Ritchie MD: STUDENTJAMA. The challenges of whole-genome approaches to common diseases. *JAMA* 291, 1642–1643 (2004).

Website

101 www.epistasis.org
The Computational Genetics Laboratory (CGL) at Dartmouth Medical School homepage (Accessed November 2005).