Pacific Symposium on Biocomputing 15:327-336(2010)

ENABLING PERSONAL GENOMICS WITH AN EXPLICIT TEST OF EPISTASIS

CASEY S. GREENE, DANIEL S. HIMMELSTEIN
Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756, USA

HEATHER H. NELSON
Division of Epidemiology and Community Health, University of Minnesota School of Public Health, Minneapolis, MN, USA

KARL T. KELSEY
Department of Community Health, Brown University, Providence, RI, USA

SCOTT M. WILLIAMS
Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA

ANGELINE S. ANDREW, MARGARET R. KARAGAS
Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756, USA

JASON H. MOORE
Departments of Genetic and Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756, USA

One goal of personal genomics is to use information about genomic variation to predict who is at risk for various common diseases. Technological advances in genotyping have spawned several personal genetic testing services that market genotyping services directly to the consumer. An important goal of consumer genetic testing is to provide health information along with the genotyping results. This has the potential to integrate detailed personal genetic and genomic information into healthcare decision making. Despite the potential importance of these advances, there are some important limitations. One concern is that much of the literature that is used to formulate personal genetics reports is based on genetic association studies that consider each genetic variant independently of the others. It is our working hypothesis that the true value of personal genomics will only be realized when the complexity of the genotype-to-phenotype mapping relationship is embraced, rather than ignored. We focus here on complexity in genetic architecture due to epistasis or nonlinear gene-gene interaction. We have previously developed a multifactor dimensionality reduction (MDR) algorithm and software package for detecting nonlinear interactions in genetic association studies. In most prior MDR analyses, the permutation testing strategy used to assess statistical significance was unable to differentiate MDR models that captured only interaction effects from those that also detected independent main effects. Statistical interpretation of MDR models required post-hoc analysis using entropy-based measures of interaction information. We introduce here a novel permutation test that allows the effects of nonlinear interactions between multiple genetic variants to be specifically tested in a manner that is not confounded by linear additive effects. We show using simulated nonlinear interactions that the power using the explicit test of epistasis is no different than a standard permutation test. We also show that the test has the appropriate size or type I error rate of approximately 0.05. We then apply MDR with the new explicit test of epistasis to a large genetic study of bladder cancer and show that a previously reported nonlinear interaction between is indeed significant, even after considering the strong additive effect of smoking in the model. Finally, we evaluated the power of the explicit test of epistasis to detect the nonlinear interaction between two XPD gene polymorphisms by simulating data from the MDR model of bladder cancer susceptibility. The results of this study provide for the first time a simple method for explicitly testing epistasis or gene-gene interaction effects in genetic association studies. Although we demonstrated the method with MDR, an important advantage is that it can be combined with any modeling approach. The explicit test of epistasis brings us a step closer to the type of routine gene-gene interaction analysis that is needed if we are to enable personal genomics.

1. Introduction

1.1. *Personal Genomics*

The era of commercial genetic testing and personal genomics was ushered in with help from the discovery and characterization of mutations in *BRCA1* and *BRCA2* that account for between 20% and 40% of all cases of familial breast cancer [1]. Unfortunately, the remaining 60% to 80% of familial breast cancer remains unexplained and the elusive *BRCA3* gene has not yet been identified despite significant efforts using the full spectrum of genetic and genomic tools available [2]. Failure to find the putative *BRCA3* gene is somewhat surprising given the familial nature and high heritability of this type of breast cancer. The current strategy for revealing genetic architecture is to carry out a genome-wide association study (GWAS) with a million or more single nucleotide polymorphisms (SNPs) that capture much of the common single nucleotide variation in the human genome by tagging blocks of variants that are in linkage disequilibrium [3,4]. These SNPs are then individually tested for association with a specific disease state. The GWAS approach is based on the hypothesis that scanning the entire genome for single SNP associations in an unbiased manner that ignores current

understanding about disease etiology will reveal much of the currently unexplained genetic architecture of a particular disease.

Despite the excitement surrounding the GWAS approach, and the time and financial resources already committed, the results have generally been underwhelming. Consider, for example, the application of GWAS to identification of cancer susceptibility genes. A recent review of these studies shows that a number of new susceptibility loci have been identified for several types of cancer, including breast, prostate, colorectal, lung and skin [5]. The identification of new associations is certainly important. However, as Easton and Eeles [5] note, the increase in risk for the susceptibility alleles at each of these loci is generally 1.3-fold or less. For familial breast cancer, Easton et al. [6] reported five significant, replicated associations that were identified by GWAS in a three-stage study design. Four of these variants were in known genes and one was located in a hypothetical gene. Assuming a multiplicative model, these five loci combine to explain only 3.6% of the excess familial risk of breast cancer and, as suggested by Ripperger et al. [2] were not deemed to be suitable for genetic testing due to their small effect sizes [6]. In a recent follow up study with two additional stages of testing and replication two additional susceptibility loci were identified with odds ratios of 1.11 and 0.95, respectively, each accounting for much less than 1% of the familial risk of breast cancer [7]. When combined with the previously known genetic risk factors for familial breast cancer, the estimated fraction of risk explained is approximately 5.9%. This is in stark contrast to *BRCA1* and *BRCA2* mutations that account for between 20% and 40% of familial breast cancer. While the application of GWAS to familial breast cancer has generated new knowledge, it has not resulted in new genetic tests that can be used to predict and prevent familial breast cancer. These results are particularly discouraging for more common diseases such as sporadic breast cancer that are likely to have a much more complex genetic architecture. As Clark et al. [8] predicted, our success with GWAS depends critically on the assumptions we make about disease complexity. It is the goal of this study to develop a new hypothesis testing methodology that can be used to directly confront the challenge of detecting and characterizing epistasis or nonlinear gene-gene interaction that accounts for a portion of the complex etiology of common diseases.

1.2. *Genetic Architecture of Common Diseases*

When designing and executing a genetic association study of disease susceptibility it is very important to consider the assumptions that are being made about the genetic architecture of the disease [8]. The questions that we ask, the hypotheses that we formulate, the analytical tools selected for data analysis and the inferences we make from the results are all limited by the assumptions we make about genetic architecture. Weiss [9] has defined genetic architecture as 1) the set of genes and DNA sequence involved in the disease, 2) their variation in the population and 3) their specific effects on the phenotype. It was initially thought that much of the genetic risk of familial breast cancer could be explained by three genes (*BRCA1*, *BRAC2* and the hypothetical *BRCA3*). However, it is now clear that the remaining 60% to 80% of risk is likely to be explained by many genes each with multiple variations that have very small effects. It also likely that each variant contributes to risk of sporadic breast cancer through nonlinear interactions with other variants in the genome and with multiple environmental factors such as diet and smoking. We focus here on epistasis or gene-gene interaction that is expected to be a ubiquitous component of the genetic architecture of common diseases.

William Bateson coined the word epistasis in the early 1900s to explain deviations from Mendelian inheritance [10]. The term literally means “standing upon”, and Bateson used it to describe characters that were layered on top of other characters thereby masking their expression. Since Bateson there have been many different and evolving definitions of epistasis or gene-gene interaction [e.g. 11-17]. For example, Fisher [18] defined epistasis in a statistical manner as an explanation for deviation from additivity in a linear model. This non-additivity of genetic effects measured mathematically is different than the more biological definition of epistasis from Bateson. We have previously made the distinction between Bateson's biological epistasis and Fishers statistical epistasis [16]. This distinction is important to keep in mind when thinking about the genetic architecture of common human diseases because biological epistasis happens at the cellular level in an individual while statistical epistasis is a pattern of genotype to phenotype relationships that results from genetic variation in a human population. This distinction becomes important when attempting to draw a biological conclusion from a statistical model that describes a genetic association. Moore and Williams [16] and Phillips [13] have discussed the idea that more modern definitions of epistasis may be needed in light of our new knowledge about gene networks and biological systems. However, the classic definitions provided by Bateson [10] and Fisher [18] still provide a good starting point for thinking about gene-gene interactions.

1.3. *A Multifactor Dimensionality Reduction Approach to Detecting Epistasis*

As discussed above, one of the early definitions of epistasis was deviation from additivity in a linear model [18]. The linear model plays a very important role in modern genetic epidemiology because it has a solid theoretical foundation, is easy to implement using a wide-range of different software packages, and it is easy to interpret.

Despite these good reasons to use linear models [14,15], they do have limitations for explaining genetic models of disease because they have limited ability to detect nonlinear patterns of interaction [19]. Here, a nonlinear interaction is defined as a synergistic or nonadditive effect of multiple genetic variants that is greater than the independent effects of the variants considered alone. It is well documented that linear models have greater power to detect main effects than interactions [20-22]. The limitations of the linear model and other parametric statistical approaches have motivated the development of computational approaches such as those from machine learning and data mining that make fewer assumptions about the functional form of the model and the effects being modeled [23-25]. Several recent reviews highlight the need for new methods [26] and discuss and compare different strategies for detecting statistical epistasis [15,27].

As reviewed recently by Cordell [15], multifactor dimensionality reduction or MDR has emerged as one important new and novel method for detecting and characterizing patterns of statistical epistasis in genetic association studies that complements the linear modeling paradigm. Multifactor dimensionality reduction (MDR) was developed as a nonparametric (i.e. no parameters are estimated) and genetic model-free (i.e. no genetic model is assumed) data mining and machine learning strategy for identifying combinations of discrete genetics and environmental factors that are predictive of a discrete clinical endpoint [28-34]. Unlike most other methods, MDR was designed to detect interactions in the absence of detectable main effects and thus complements other statistical approaches such as logistic regression and other machine learning methods such as random forests and neural networks. At the heart of the MDR approach is a feature or attribute construction algorithm that creates a new variable or attribute by pooling genotypes from multiple SNPs (see Figure 1). The general process of defining a new attribute as a function of two or more other attributes is referred to as constructive induction, or attribute construction, and was first described by Michalski [35]. Constructive, induction using the MDR kernel, is accomplished in the following way. Given a threshold *T*, a multilocus genotype combination is considered high-risk if the ratio of cases (subjects with disease) to controls (healthy subjects) exceeds *T*, otherwise it is considered low-risk. Genotype combinations considered to be high-risk are labeled G_1 while those considered low-risk are labeled G_0. This process constructs a new one-dimensional attribute with values of G_0 and G_1. It is this new single variable that is assessed, using any classification method. The MDR method is based on the idea that changing the representation space of the data will make it easier for methods such as logistic regression, classification trees, or a naive Bayes classifier to detect attribute dependencies. As such, MDR complements any classification methods such as those reviewed by Hastie et al. [24]. Cross-validation is used to prevent overfitting while permutation testing is used to assess statistical significance and to control for false-positives due to multiple testing. This method has been confirmed in numerous simulation studies and a user-friendly open-source MDR software package written in Java is freely available from www.epistasis.org.

Although MDR is a powerful method for detecting nonlinear interactions in the absence of independent main effects it, like other machine learning methods, does not explicitly disentangle these two types of genetic effects. In other words, a statistically significant MDR model could capture interactions, main effects or both interactions and main effects. It may not be immediately apparent to the user which types of effects are represented in a high-order MDR model. This has been previously addressed through post-hoc analysis methods that use entropy-based measures of interaction information to identify evidence of nonlinear interactions [33]. These information theoretic approaches work well but do not reveal directly which genetic effects made a meaningful contribution to the statistical significance. We propose here a new explicit test of epistasis that can be used in conjunction with MDR or any other method to directly test for nonlinear gene-gene interaction while holding the independent main effects constant.

1.4. *Redefining the Null Hypothesis in Genetic Association Studies*

The present study is motivated by the need to greatly improve our knowledge of biological and statistical epistasis and its role in human health and disease. We know very little about the role of epistasis in human biology and public health because the focus for so long as has been on the effects of single genes and single genetic variants in biological and clinical endpoints. Given the ubiquity of complexity in genetic architecture, with epistasis as a central component, we propose a rephrasing of our research questions. Instead of asking which single SNP is associated with disease, we propose asking which combination of SNPs is associated with disease. Rephrasing the question in this manner necessitates a redefinition of the null hypotheses that needs to be tested using statistical and computational methods. Given the reality of complexity, and this specific research question, we propose the following logical set of hypotheses as a starting point for retooling our analytical approach to this problem. First, we propose testing the null hypothesis that the associations in the data are only linear and additive using methods such as MDR and the explicit test of epistasis that were designed specifically for this purpose. Once there is significant evidence for rejecting the null hypothesis of linearity, it is then a logical next step to test the universal null hypothesis of no association using linear statistical methods such as logistic regression that are powered to model the independent and additive main effects. Rejection of the universal null in addition to the linear null provides a set of results generated in a systematic manner that

addresses complexity that can then be interpreted biologically using experimental methods or that can be interpreted statistically using approaches such as parsimony. Is the evidence generated by testing the linear null more compelling than the evidence generated by testing the universal null? Answering this question will help further our understanding of genetic architecture. We propose here a new 'explicit test of epistasis' that allows us to directly test the linear null hypothesis using MDR or any other method.

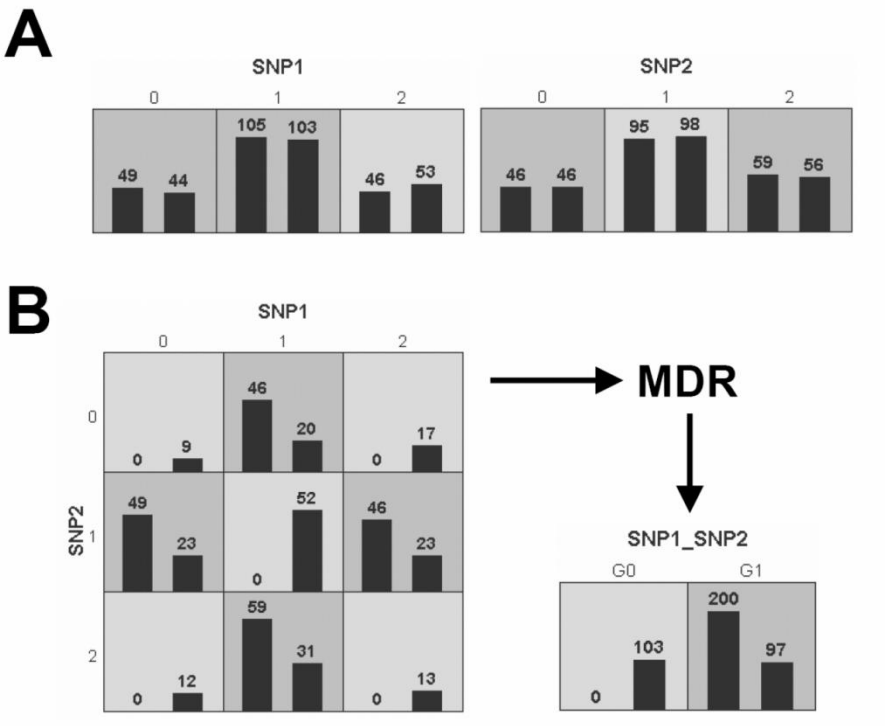

Figure 1. MDR attribute construction. A) Illustrates distribution of cases (left bars) and controls (right bars) for each of the three genotypes of SNP1 and SNP2. The dark-shaded cells have been labeled 'high-risk' using a threshold of $T = 1$. The light-shaded cells have been labeled 'low-risk'. B) Illustrates the distribution of cases and controls when the two functional SNPs are considered jointly. A new single attribute is constructed by pooling the "high-risk" genotype combinations into one group (G1) and the low-risk" into another group (G0).

2. Methods

2.1. *An Explicit Test of Epistasis*

The goal of our proposed explicit test of epistasis is provide a hypothesis testing framework that will allow us to directly test the null hypothesis that the only genetic effects in the data are linear and additive. As described in detail by Pattin et al. [36], the current hypothesis testing framework for MDR is based on a permutation test that randomizes the class (i.e. case and control) labels so that the only genetic associations in the permuted data are there by chance (see Figure 1A and 1B). Permutation testing is used because it doesn't assume we know the null distribution of the test statistic (e.g. testing accuracy) and it controls for false-positives due to multiple testing. However, the current permutation testing framework provides a global p-value for an MDR model that might have main effects, gene-gene interactions, or a combination of both. Significance tells us nothing about the nature of the MDR model and only reflects the fact that the model predicts class better than chance.

We propose here an explicit test of interaction that has all the same advantages of the permutation testing framework but that is able to provide a p-value that reflects only the nonlinear interaction or epistasis component of the model. To accomplish this, we first sort the data rows (i.e. the subjects) by class into cases and controls (see Figure 1C). We then randomize each column (i.e. the SNPs) within each class. This removes any relationship between genotypes within class but preserves the overall genotype frequency difference between the classes. This new type of permutation randomizes any interaction effects while keeping the independent main effects as defined by class differences in genotype frequency. This allows us to generate permuted datasets under the null hypothesis that the only genetic associations in the data are linear or additive in nature and that any nonlinear interaction effects are only there by chance. This yields an explicit test of epistasis when combined with a method such as MDR that is capable of modeling nonlinear interactions.

We have included the explicit test of interaction in the MDR permutation testing (MDRpt) module that is open-source and freely available from www.epistasis.org.

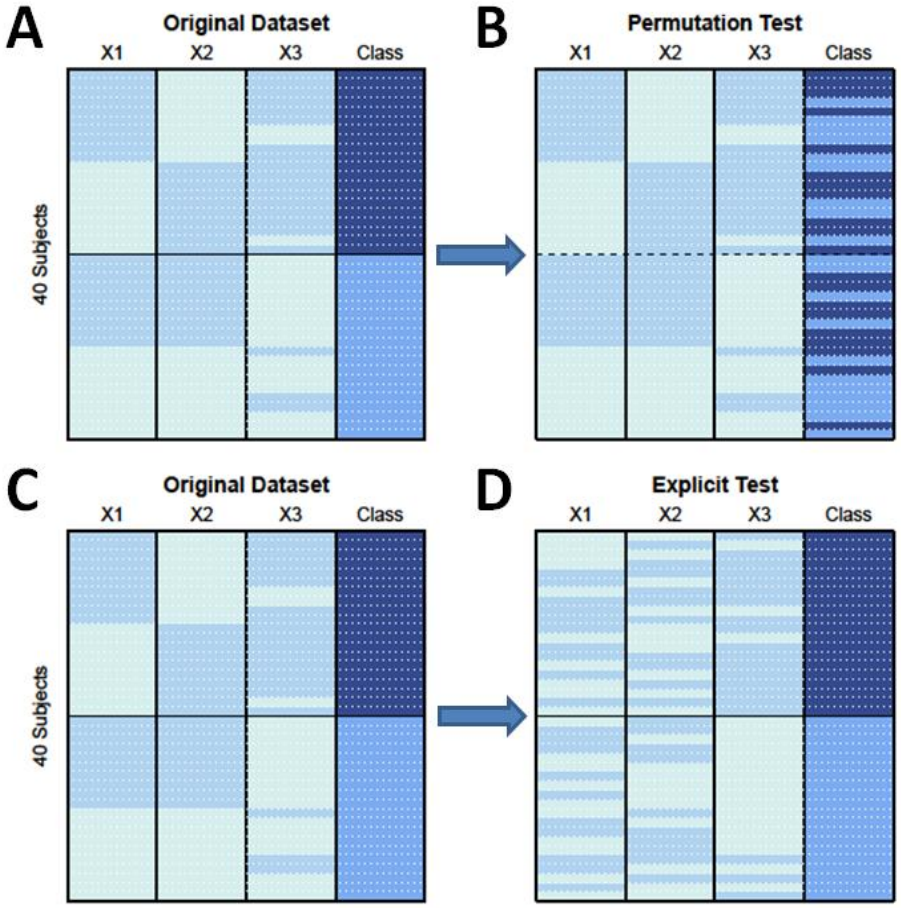

Figure 2. Overview of the explicit test of epistasis. Shown on the left (panels A and C) is a hypothetical dataset with three attributes (e.g. SNPs) coded X1, X2 and X3 and class (i.e. case-control status). Each row of the dataset is one of 40 subjects with hypothetical binary genotypes colored in light shades of blue and case-control status coded darker shades of blue. In this simple example, X1 and X2 effect disease risk through a nonlinear interaction while X3 has an independent main effects that is reflected by a frequency difference in genotypes between cases and controls. Panel B shows the process of randomizing class labels in a standard permutation test. Panel D shows the same data randomized for the explicit test of interaction. Here, the columns are randomized within each class. Note that the genotype frequencies within each class remain fixed. This preserves the independent main effects while randomizing any nonlinear interactions.

2.2. *Multifactor Dimensionality Reduction Analysis*

As described above, the goal of MDR is to change the representation space of the data using constructive induction to make nonlinear interactions easier to detect. This is accomplished by combining two or more variables or attributes into a single attribute that can be modeled using a discrete data classifier. Here, we used a simple probabilistic classifier that is similar to naïve Bayes [31] to model the relationship between variables constructed using MDR and case-control status. Naïve Bayes classifiers were assessed using balanced accuracy as recommended by [37]. For each dataset we evaluated all possible pairwise combinations of SNPs using MDR. The model with the maximum training accuracy as assessed with ten-fold cross validation was selected as the best model. The testing accuracy (i.e. predictive ability) of the single best MDR model was then assessed using the cross-validation hold-out data. We used the open-source MDR software package that is freely available from www.epistasis.org. A tutorial on MDR can be found in the November and December 2006 postings at compgen.blogspot.com.

2.3. *Evaluation of Power and Type I Error Using Simulated Data*

The goal of the simulation study was to generate artificial datasets that could be used to evaluate the power of the MDR within the explicit test of epistasis framework to detect nonlinear gene-gene interactions. We developed a total of 35 different penetrance functions that define a probabilistic relationship between genotype and phenotype where susceptibility to disease is dependent on genotypes from two loci in the absence of any marginal effects. 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) with minor allele frequencies of 0.4. A total of five models for each of the seven heritabilities were generated for a total of 35 models. More information about the mathematics of penetrance functions and heritability can be found in Culverhouse et al. [38]. A heritability of 0.01 is a very small genetic effect size while 0.4 is a very large genetic effect size. The details of the 35 penetrance functions used here have been previously described in detail by Velez et al. [37]. Genotype frequencies for all 35 epistasis models were consistent with Hardy-Weinberg proportions. One hundred data sets were generated for each model with three sample sizes (400, 800, and 1600 total individuals) with case-control proportions of 1:1. Each pair of functional polymorphisms was embedded within a set of 20 independent single-nucleotide polymorphisms (SNPs). A total of 7,000 artificial datasets were generated and analyzed. For evaluating the type I error of the explicit test of epistasis, null data sets with no functional SNPs were generated by permuting the case-control labels of the data sets described above. All simulated data are available upon request.

The power of MDR using the explicit test of epistasis test was estimated as the percentage of times MDR correctly identified the two functional SNPs in the best model out of each set of 100 datasets for which the result was statistically significant at the 0.05 level (i.e. the testing accuracy was equal to or higher than the top 5% highest testing accuracies in the permuted data). Type I error was estimated as the proportion of times that the permutation test indicated a statistically significant MDR model in data consistent with the null hypothesis of no association.

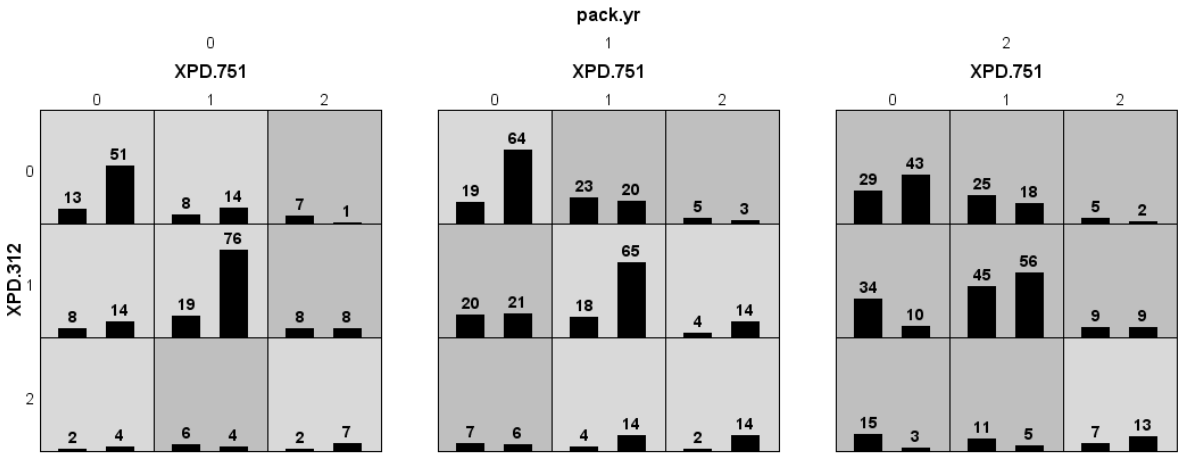

Figure 3. Distribution of cases (left bars) and controls (right bars) for each XPD genotype (coded 0, 1, 2) and for pack years of smoking (pack.yr) in the bladder cancer example. Dark shaded cells indicate high-risk for disease while light-shaded cells indicate low-risk. The p-value from a standard permutation test for this model was <0.001. Note that it is difficult to tell which attribute has a main effect and which are interacting and how these different effects contribute to the statistical significance.

2.4. *Application to Bladder Cancer*

We demonstrated use of the explicit test of epistasis 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 [39]. The study analyzed 355 bladder cancer cases and 559 controls ascertained from the state of New Hampshire. This study 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 (XPD)* 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 ≤ 50 years.

A parametric linear statistical analysis of each attribute individually revealed a significant independent main effect of smoking as expected ($P < 0.05$). However, none of the measured SNPs were significant predictors of bladder cancer individually ($P > 0.05$). Andrew et al. [39] 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 *XPD* gene and smoking. The distribution of cases and controls with each genotype/smoking combination is illustrated above in Figure 3. 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 *XPD* polymorphisms had evidence of nonlinear interaction or synergy in the near complete absence of main effects. Interestingly, the joint effect of the two *XPD* SNPs was larger than the independent from the effect of smoking. As such, these data provide an ideal test case for the proposed explicit test of interaction. Is the nonlinear interaction between the two *XPD* SNPs statistically significant after holding the effects of smoking constant in the new permutation test or was the significance only due to the large effect of smoking? To answer this question we applied MDR with the explicit test of interaction to the bladder cancer data and determined the statistical significance of the model comprised of the two SNPs from the *XPD* gene and smoking.

To assess the power of the explicit test of epistasis to detect the joint effect of the two *XPD* SNPs in the bladder cancer we simulated 100 datasets using three different MDR models from the bladder cancer data analysis described above. First, we simulated 100 datasets using the MDR model containing the two *XPD* SNPs. Second, we simulated 100 datasets using the MDR model containing just smoking. Third, we simulated 100 datasets using the MDR model containing the two *XPD* SNPs with smoking. The total number of simulated attributes was the same as the original data. We applied MDR along with the explicit test of interaction to each simulated dataset and recorded the power to detect an interaction. We expect the results of this study to provide realistic power estimates for real data with a detectable interaction and a strong independent main effect.

3. Results

3.1. *The Power and Type I Error of the Explicit Test of Epistasis*

Table 1 summarizes the power and the type I error (in parentheses) of the explicit test of epistasis to detect nonlinear interactions in the simulated data using MDR models. The power exceeded 0.80 for all sample sizes for data with moderate to large genetic effect sizes (heritability > 0.025). Power also exceeded 0.80 at sample sizes of 1600 and 800 for the small genetic effect sizes of 0.01 and 0.025, respectively. It is important to note that these power estimates are extremely close (± 0 to 0.01) to those estimated using a standard permutation test by Pattin et al. [36]. These results demonstrate that the new explicit test of interaction does not lose power to detect nonlinear interactions as compared to a standard permutation test.

Also shown in Table 1 in parentheses are the estimates of the false-positive rate or type I error. Note that in each case the type I error rate was approximately 0.05 suggesting that the explicit test of interaction is an appropriately sized test. As with power, this is not different than has been previously reported for standard permutation tests with MDR [36]. This is important given that MDR is a machine learning algorithm that looks at the data in a combinatorial manner.

Table 1. Summary of the power and type I error (parentheses) of the explicit test of interactions when combined with MDR.

	Heritability						
Sample Size	0.01	0.025	0.05	0.10	0.20	0.30	0.40
400	0.22 (0.06)	0.65 (0.06)	0.87 (0.04)	0.95 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)
800	0.51 (0.04)	0.88 (0.07)	0.98 (0.04)	1.00 (0.06)	1.00 (0.06)	1.00 (0.04)	1.00 (0.07)
1600	0.87 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.04)	1.00 (0.05)	1.00 (0.05)	1.00 (0.04)

3.2. *Application to Bladder Cancer*

As described above, the bladder cancer study of Andrew et al. [39] makes an ideal test case for the new explicit test of interaction because a statistically significant MDR model was detected that consisted of two interacting SNPs and smoking that appeared to have an independent main effect. This model was determined to be significant at the 0.001 level using a standard permutation test and, at the time, it wasn't clear the degree to which the significance was due to the main effect of smoking, the nonlinear gene-gene interaction, or both. We applied MDR with the explicit test of interaction and found the same best model with the p-value of 0.005. This is a highly significant result that confirms the important role of a nonlinear interaction between the two *XPD* polymorphisms. This synergistic interaction was still highly significant even after controlling for the contribution made by smoking, a known risk factor for bladder cancer.

Figure 4 below illustrates the distribution of testing accuracies for best MDR models from the standard permutation test and the explicit test of interaction. First, note that the center for the permutation distribution is approximately 0.50. This is the result that is expected if a fair coin were used to predict who is a case and who is a control. Now note that the distribution for the explicit test of epistasis is shifted to the right. This shift is due to the factors with independent main effects in the data such as smoking that are fixed during the randomization process used by the explicit test of epistasis.

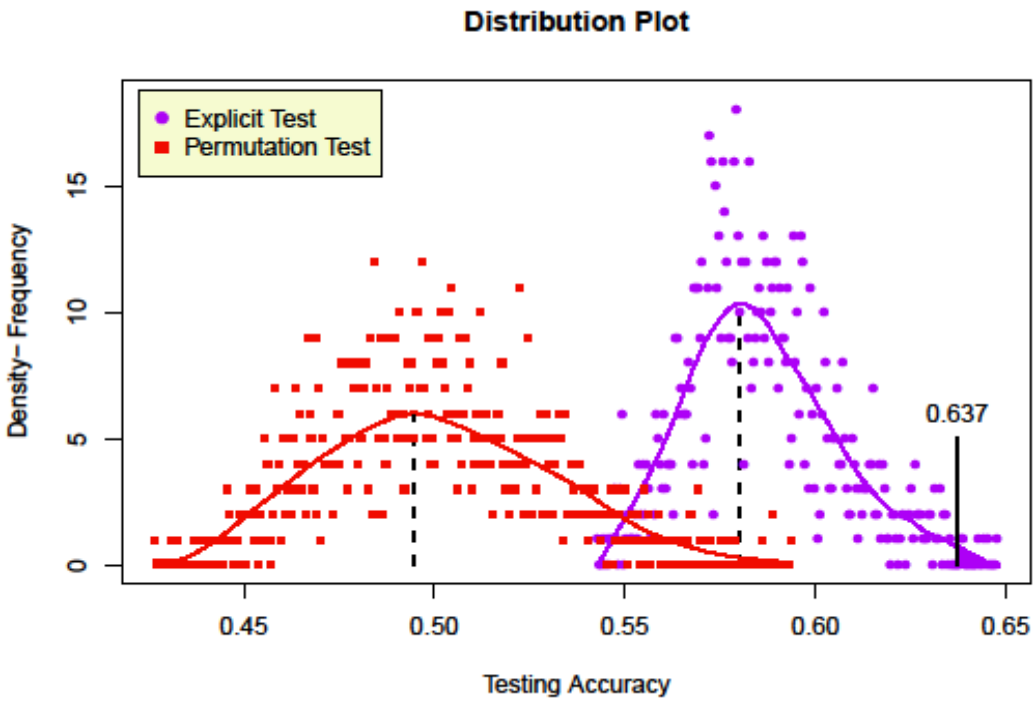

Figure 4. Distribution of testing accuracies from best MDR models obtained from 1000 datasets randomized using a standard permutation test (red squares) and the explicit test of epistasis (purple circles). Note that the permutation distribution is centered (dashed line) at approximately 0.50, as expected. However, the center null distribution derived from the explicit test of epistasis is shifted to the right. This new center is consistent with the fixed main effects in the data. The testing accuracy for the best MDR model from the bladder cancer data is shown on the right (solid line). The area to the right of 0.637 is shaded purple and is equivalent to the p-value of 0.005.

What was the power to detect this effect? As described above, we evaluated power by simulating data from MDR models of the two *XPD* SNPs, just smoking and the two *XPD* SNPs with smoking. We found that the power to detect just the interaction was 1.00 while the power to detect the interaction with the effect of smoking in the model was 0.94. This reduction in power is not surprising given the increase in MDR model size from two factors (two dimensions) to three factors (three dimensions). As expected, the power to detect an interaction for the model with just smoking was 0.06. This is approximately equal to the type I error rate of 0.05 since there was no interaction to find. These findings confirm the results from the earlier simulation study. The power results are not relevant to the actual bladder cancer data analysis since a highly significant model was detected with a p-value of 0.005. However, they do help to reveal the operating characteristics of the explicit test of interaction.

4. Discussion

Epistasis or gene-gene interaction is expected to be a ubiquitous component of the genetic architecture of common human diseases [40]. As such, it has very important implications for the success of personal genomics which is currently based almost entirely on results from genetic association studies that only consider one SNP or one gene at a time. Moore and Williams [41] have suggested that personal genomics will not reach its full potential to impact human health until the full complexity of the genotype-to-phenotype relationship is addressed in all genetic studies. What can be done to improve the usefulness of personal genomics? Moore and Williams [41] offer the following five recommendations:

1. We need to greatly improve our understanding of biological and statistical epistasis and their roles in human health and disease.
2. We need powerful analytical tools that are designed to address the complexity of genetic architecture due to epistasis and other phenomena.
3. We need better experimental methods for confirming statistical models of epistasis in animal models or in human cell culture.
4. We need to remember the principles of classical genetics as we immerse ourselves in the excitement of cutting-edge genotyping technology and emerging methods for rapidly sequencing an entire genome.

5. We need to continue to integrate systems biology into human genetics in a meaningful manner.

The goal of the present study was to develop a hypothesis testing framework and methodology that can be used with methods such as MDR that were designed specifically for detecting and characterizing nonlinear or nonadditive gene-gene interactions in genetic association studies. As such, this study is consistent with the first two recommendations listed above. We have introduced an explicit test of epistasis that can be used to test the null hypothesis that the only genotype-to-phenotype relationships in the data are linear and additive. This is important because until now methods such as MDR could only perform a universal test of the null hypothesis of no association [36]. Inferences about nonadditive interactions were made from post-hoc analyses using methods based on information theory [33]. We demonstrated using simulated data that this approach retains the power of a standard permutation test to detect epistasis across a range of effects sizes and sample sizes. Further, we demonstrated that this new approach has a reasonable type I error rate of approximately 0.05. Finally, we applied this new approach to a large genetic study of bladder cancer and were able to confirm a previously reported nonadditive gene-gene interaction in the presence of the large independent effect of smoking [39].

In addition to introducing a new method for epistasis analysis, we have also introduced a new hypothesis testing framework that redefines the null hypothesis of no genetic association into component parts that are more consistent with the assumption that the genetic architecture of common diseases is complex (see Section 1.4). It is important to note that idea of testing the null hypothesis of linearity using nonlinear statistical methods is not new. For example, Theiler et al. [42] introduced the method of surrogate data in the context of time series analysis as way to test for nonlinear patterns with the confounding of linear patterns. With the method of surrogate data, a discrete Fourier transform of a time series is taken, the phases are randomized and a new time series generated using an inverse discrete Fourier transform. The resulting phase-randomized time series has the same linear patterns as the original time series with all other patterns randomized. This procedure makes it possible to test the null hypothesis of linearity using any statistic that is capable of measure nonlinear patterns. As reviewed by Moore [43], the method of surrogate data is a type of permutation and thus has many similarities to the explicit test of interaction introduced here.

The advantages of the explicit test of epistasis include its simplicity and its flexibility. First, the explicit test of interaction is simply a modified permutation test that randomizes the attribute columns within each class. Thus, it can be easily implemented in a Perl or Python script or in a data analysis package such as R. We have also provided the method in the open-source MDR permutation testing module. Second, the approach is very flexible in that it can be generally applied to any method that is designed for detecting nonlinear gene-gene interactions. Thus, it could be combined with other machine learning methods such as decision trees, neural networks or support vector machines. The only disadvantage of the approach is that permutation testing can add a significant amount of computational time. This will be important for application of these methods to GWAS. Approaches such as the extreme value distribution (EVD) that can reduce the number of permutations that need to be performed are likely to help address this problem [36].

We recommend several future studies with the explicit test of epistasis. First, it will be interesting to use the explicit test of epistasis to compare the power of different methods for detecting gene-gene interactions in the presence of independent main effects. This will be important because some methods may be confounded by any linear additive patterns in the data. Second, it will be important to demonstrate that the EVD approach described by Pattin et al. [36] could be combined with the explicit test of epistasis without violating the distributional assumptions of the EVD. This will be important in the context of GWAS where computational efficiency is extremely important. Finally, it will be very important to implement the explicit test of interactions with other real datasets where both interactions and independent main effects are present. Reanalysis of published epistasis results to confirm nonlinear interactions will be helpful for determining statistical significance. We anticipate the explicit test of epistasis will play an important role in the detection, characterization and interpretation of nonlinear gene-gene interactions in genetic association studies. As such, it will play an important role in improving the impact of personal genomics and other healthcare endeavors that depend critically on published genetic association results that reflect the underlying genetic architecture of the disease in question.

Acknowledgments

This work was supported by NIH grants LM009012, LM010098, HD047447, AI59694 and ES007373. We would like to thank the anonymous reviewers for their very helpful comments.

References

1.Narod SA, Foulkes WD (2004) Nat Rev Cancer 4:665-76.
2.Ripperger T, Gadzicki D, Meindl A, Schlegelberger B (2009) Eur J Hum Genet 17:722-31.

3.Hirschhorn JN, Daly MJ (2005) Nat Rev Genet 6:95-108.
4.Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Nat Rev Genet 6:109-18.
5.Easton DF, Eeles RA (2008) Hum Mol Genet 17:R109-15.
6.Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. (2007) Nature 447:1087-93.
7.Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. (2009) Nat Genet 41:585-90.
8.Clark AG, Boerwinkle E, Hixson J, Sing CF (2004) Genome Res 15:1463-7.
9.Weiss KM (1993) Genetic variation and human disease. Cambridge University Press, New York.
10.Bateson W (1909) Mendel's Principles of Heredity. Cambridge University Press, Cambridge.
11.Hollander WF (1955) J Hered 46:222-225.
12.Phillips PC (1998) Genetics 149:1167-71.
13.Phillips PC (2008) Nat Rev Genet 9:855-67.
14.Cordell HJ (2002) Hum Mol Genet 11:2463-8.
15.Cordell HJ (2009) Nat Rev Genet, in press.
16.Moore JH, Williams SW (2005) BioEssays 27:637-46.
17.Tyler AL, Asselbergs FW, Williams SM, Moore JH (2009) Bioessays 31:220-7.
18.Fisher RA (1918) Trans R Soc Edinb 52:399-433.
19.Moore JH, Williams SW (2002) Ann Med 34:88-95.
20.Lewontin RC (1974) Am J Hum Genet 26:400-411.
21.Lewontin RC (2006) Int J Epidemiol 35:536-7.
22.Wahlsten D (1990) Behav Brain Sci 13:109-161.
23.Mitchell T (1997) Machine Learning. McGraw-Hill, New York.
24.Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer-Verlag, New York.
25.McKinney BA, Reif DM, Ritchie MD, Moore JH (2006) Appl Bioinformatics 5:77-88.
26.Thornton-Wells TA, Moore JH, Haines JL. (2004) Trends Genet 20:640-7.
27.Motsinger AA, Ritchie MD, Reif DM (2007) Pharmacogenomics 8:1229-41.
28.Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Am J Hum Genet 69:138-147.
29.Ritchie MD, Hahn LW, Moore JH (2003) Genet Epidemiol 24:150-157.
30.Hahn LW, Ritchie MD, Moore JH (2003) Bioinformatics 19:376-82.
31.Hahn LW, Moore JH (2004) In Silico Biol 4:183-94.
32.Moore JH (2004) Expert Rev Mol Diagn 4:795-803.
33.Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White, B. C. (2006) J Theor Biol 241, 252-261.
34.Moore JH (2007) In: Zhu, X., Davidson, I. eds. Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data, IGI Global, pp. 17-30.
35.Michalski RS (1983) Artif Intell 20:111-161.
36.Pattin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KT, Andrew AS, Karagas MR, Moore JH (2009) Genet Epidemiol. 33(1):87-94.
37.Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH (2007) Genet Epidemiol 31:306-315.
38.Culverhouse R, Suarez BK, Lin J, Reich T (2002) Am J Hum Genet 70:461-71.
39.Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR (2006) Carcinogenesis 27:1030-7.
40.Moore JH (2003) Hum Hered 56:73-82.
41.Moore JH, Williams SM (2009) Am J Hum Genet, in press.
42.Theiler J, Eubank S, Longtin A, Galdrikian B, Farmer JD (1992) Physica D 58:77-94.
43.Moore JH (1999) Phys Med Biol 44:L11-2.