EXTENSIONS TO AND AN APPLICATION OF THE MULTIFACTOR DIMENSIONALITY REDUCTION PEDIGREE DISEQUILIBRIUM TEST

By

TODD L. EDWARDS

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May, 2008

Nashville, Tennessee

Approved By:

Jonathan Haines

Eden Martin

Dana Crawford

Charles Matthews

Marylyn Ritchie

To my beautiful and brilliant best friend, Digna Velez, who provides me with insight and inspiration

And

To my undergraduate Mentor Dr. Matthew Elrod-Ericson, without whom I would have floundered out of science altogether

And

My Father, Mark Edwards, who instilled in me a love of science fiction and literature which always reminds me I am trying to do something fantastic

ACKNOWLEDGEMENTS

The work presented in the following dissertation is supported in part by National Institutes of Health grant AG20135. Much of these studies would not be possible without the participation of families and patients from the various collection centers from the Collaborative Alzheimer Project.

I would like to thank the members of my dissertation committee (Dr. Jonathan Haines, Dr. Marylyn Ritchie, Dr. Dana Crawford, Dr. Charles Matthews and Dr. Eden Martin) for their time and effort. I would especially like to thank my mentor Dr. Marylyn Ritchie for her encouragement and support during my graduate training. She has provided me with an inspiring example of leadership and scientific success, as well as a host of skills and experiences which exceeded all my hopes and expectations. The resources she has generously made available to me have allowed me to enjoy the synthesis of knowledge, techniques and software, without which, none of my accomplishments would be possible.

I would also like to thank the members of the Ritchie lab for many productive discussions over the years which abetted my progress toward completion of these projects. Of particular importance are the contributions of the lab programmers, Eric Torstenson and Scott Dudek, who have the skills and understanding to cast my ideas into tools and capabilities.

Other notable contributions I acknowledge are those of my fellow graduate students at the Center for Human Genetics Research. Among those, Dr. Kylee Spencer has been particularly helpful, and continues to provide a conspicuous example of intelligence and a work ethic I can only hope to emulate. Additionally, Dr. Chun Li, the finest statistician I have yet known and who's time I have had the honor of wasting improving my understanding and vetting ideas. I also thank my parents, Mark and Linda Edwards, for their patience with my extended adolescence and for teaching me a deep appreciation for education and creativity. Finally, I would like to thank Digna R. Velez, who has been both a professional and intimate personal friend and who has provided me with much more happiness and comfort than I am entitled to.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

A2M	Alpha 2 Macroglobulin
A2MP	Alpha 2 Macroglobulin of Pregnancy
AAE	Age At Examination
AAO	Age At Onset
AAU	2 Affected / 1 Unaffected sibship
AB	Amyloid Beta
ACE	Angiotensinogen Converting Enzyme
AD	Alzheimer’s Disease
ADRDA	Alzheimer’s Disease and Related Disorders Association
AGT	Angiotensinogen
APL	Association in the Presence of Linkage
APOE	Apolipoprotein Epsilon
BA	Balanced Accuracy
CAP	Collaborative Alzheimer’s Project
CART	Classification and Regression Trees
CE	Classification Error
CEPH	Centre d'Etude de Polymorphisme caucasian
CHB	Chinese
COG2	Component of Oligomeric Golgi complex 2
CPM	Combinatorial Partitioning Method
CTNNA3	Alpha 3 Catenin

CV	Cross-Validation
CVC	Cross-Validation Consistency
DICE	Detection of Informative Combined Effects
DNA	Deoxyribo Nucleic Acid
DSP	Discordant Sibling Pair
EM	Estimation Maximization
FBAT	Family-Based Association Test
FITF	Focused Interaction Testing Framework
FPG	Forward Population Genetic
FREGENE	Forward Evolution of Genomic Regions
genoPDT	Genotype Pedigree Disequilibrium Test
HWE	Hardy-Weinberg Equilibrium
I/D	Insertion / Deletion
IBD	Identical By Descent
IU	Indiana University
JPT	Japanese
LD	Linkage Disequilibrium
LOD	Log Odds of Disequilibrium
LR	Logistic Regression
LRP1	Low-density lipoporotein Receptor related Protein 1
LRRTM3	Leucine Rich Repeat Transmembrane 3
MAF	Minor Allele Frequency
MARS	Multiple Adaptive Regression Splines

MDR	Multifactor Dimensionality Reduction
MDR-PDT	Multifactor Dimensionality Reduction Pedigree Disequilibrium Test
MOR	Matched Odds Ratio
NCSTN	Nicastrin
NIMH	National Institutes of Mental Health
NINDS	National Institute of Neurological Disorders and Stroke
NN	Neural Networks
OR	Odds Ratio
PD	Parkinson’s Disease
PDT	Pedigree Disequilibrium Test
PE	Prediction Error
PRP	Patterning and Recursive Partitioning
PZP	Pregnancy Zone Protein
RAM	Random Access Memory
RPM	Restricted Partitioning Method
SD	Standard Deviation
SIMLA	Simulation of Linkage and Association
SNP	Single Nucleotide Polymorphism
S-TDT	Sibling Transmission Disequilibrium Test
T/UT	Transmitted / Untransmitted
TDT	Transmission Disequilibrium Test
WGA	Whole-Genome Association
YRI	Yoruba African

CHAPTER I

OVERVIEW

Interactive effects are currently an area of interest to genetic epidemiologists. Genetic interaction has long been of interest to biologists studying yeast, worms, mice, and other model organisms. However, in humans, these effects must usually be detected using samples from real populations and statistical analyses. Statisticians have been working on approaches to detect and characterize interactions throughout the history of statistics. Methods for detecting these interactions usually feature searches through large spaces of possible multilocus models, followed either by statistical adjustment or permutation testing. Herein, some issues common to many methods that search for interactions are explored, and some solutions are proposed.

Chapter II lists and describes the Specific aims of this dissertation. Each of these tasks was completed and approved by committee before this document was compiled.

Chapter III provides introduction and background information on topics relevant to the projects executed for this dissertation. Some of the literature that preceded and was necessary for this dissertation to be completed is reviewed. Also, statistical methods that are relevant to this work are discussed. The pedigree disequilibrium test, multifactor dimensionality reduction pedigree disequilibrium test, transmission disequilibrium test, multifactor dimensionality reduction algorithm, and genotype pedigree disequilibrium test are described in detail. Other methods relevant to studying interactions and association are also included.

Chapter IV describes some initial work done by members of the Ritchie lab with genomeSIMLA (SIMulation of Linkage and Association), a software package used for simulation studies presented later. Genome wide association studies and other simulation software packages are reviewed in the introduction and background sections. The immediate predecessors, genomeSIM and SIMLA are also discussed. The techniques used to simulate realistic data for human populations are presented in the methods section, as well as the means used to scan parameters and evaluate linkage disequilibrium. We also provide details about the implementation of the software and some of its capabilities. In the results section we show some examples of results from the parameter sweep.

Chapter V describes a series of simulation studies which were performed to evaluate the cross-validation extension to the multifactor dimensionality reduction pedigree disequilibrium test (MDR-PDT). The revised algorithm is a valid test of multilocus association in pedigree data when searching through several orders of genetic models. It is demonstrated that performing regression on the best model after MDR-PDT analysis is a procedure that is strongly biased to the alternate hypothesis. Also, simulation demonstrates that the specificity of the hypothesis test for MDR-PDT is low when strong main effects are present in the data. This property of MDR-PDT is more problematic as the sample size increases, as the signal to noise ratio increases for the model loci as sampling precision increases. To solve these problems, a regression-based test of interaction was proposed, where a saturated conditional logistic regression model was fitted for the best model from the MDR-PDT analysis. For hypothesis testing, permutations are run as usual and likelihood ratio statistics from regression are calculated for the best multilocus model from each permutation. This provides a distribution of statistics which represent the null hypothesis of no interaction after an exhaustive search. This procedure was performed on data simulated with genomeSIMLA in which purely epistatic disease susceptibility models were simulated alongside MDR-PDT with cross-validation. The power of MDR-PDT with cross validation was higher than that of MDR-PDT with cross validation followed by regression; however, this is offset by the advantages in interpretability of results. It is also shown in this chapter that when two main effects without interaction are present in the data, MDR-PDT without the regression procedure rejects the null hypothesis much more than the nominal rate, while the regression procedure maintains very high specificity for these scenarios.

Chapter VI describes a series of analyses performed on an Alzheimer's disease dataset consisting of both case-control and family-based samples. The data were collected for 10 candidate genes on 47 markers common to both datasets. Univariate tests of allele and genotype frequency differences were performed at each locus and further confirmed the association of *APOE* in both samples. All of the results that were nominally significant had effect size estimates calculated in both case-control and family samples separately. Haplotype association tests revealed associated haplotypes in the *ACE* gene for both samples. This result is strongly supported by the literature. Additionally, MDR and MDR-PDT analyses were performed in case-control and family-based samples respectively to search for multilocus models. MDR-PDT found significant 2 and 3 locus models, but these were not confirmed by MDR analysis of the case-control data.

Chapter VII describes the conclusions of this work and the future directions for further investigations. Extensions to case-control data and studies comparing the MDR regression procedure to FITF are proposed.

CHAPTER II

INTRODUCTION AND BACKGROUND

Epistasis

Gene-gene and gene-environment interactions are long-recognized phenomena that are becoming a focus of the method-development arm of human genetics. Descriptions of epistasis have been developing throughout the history of genetics. Sewell Wright described the relationship between genotype and phenotype as dependent on dynamic interactive networks of genes and environmental factors (Wright 1932).

As individual susceptibility loci displaying main effects are discovered, many disease etiologies remain unknown; supporting the hypothesis that statistical epitasis plays a role in many disease models. Statistical epistasis was defined by Fisher as deviation from additivity in a mathematical model (Fisher 1918). In genetics a statistically epistatic relationship between multilocus genotypes and phenotype is not discernable through consideration of a linear combination of variables. This is a population-level phenomenon, as opposed to biological epistasis, which is the result of the physical interactions of biomolecules in individuals (Moore and Williams 2005). Variation in biologically epistatic processes among individuals within populations causes epistatic statistical signals (Moore and Williams 2005).

The simultaneous selection of factors makes detecting and characterizing statistical epistasis difficult. Methods designed to perform searches involving many comparisons in high dimensions, and studies powered to capture subtle effects are required to unravel statistically epistatic disease models. Otherwise, given data containing a statistically epistatic disease model where two or more risk variables do not have significant main effects, a highly powered study searching for single-locus effects would fail and potentially present type I errors as findings (Hirschhorn et al. 2002). Even if the loci do have detectable main effects, failure to find an interaction among them would leave trait variance unexplained, leading to doomed future studies searching elsewhere for more susceptibility loci. The difficulty experienced by genetic epidemiologists when attempting to replicate association studies is perhaps evidence of this phenomenon (Ioannidis 2007). For interactions in particular, clear and specific statistical hypotheses must be tested to determine where efforts to replicate should be expended.

Traditional parametric statistics such as logistic regression (LR) (Hosmer and Lemeshow 2000) have limited utility when searching for interactive effects in a large search space, whether they are genes (Templeton 2000) or environmental exposures (Schlicting and Pigliucci 1998). Like other parametric methods, LR does not adjust well for high dimensionality in analysis of interactions. When modeling high-order interactions with insufficient data, this can lead to coefficient and standard error inflation, leading to inaccurate estimates of effect sizes and very wide beta parameter confidence intervals due to high sampling variance in cells with low counts (Hosmer and Lemeshow 2000).

As the number of predictor variables increases, the number of comparisons necessary to explore the entire statistically epistatic search space expands rapidly. Variations of LR have been considered for these situations. Applying stepwise LR to this problem by only including factors that exhibit a significant marginal or main effects in the final model seems promising (Marchini, Donnelly, and Cardon 2005); however, factors participating in purely interactive effects with no main effects will not be detected.

Several derivations of the logistic regression procedure have been developed to detect statistical epistasis. Some of these are logic regression (Kooperberg et al. 2001), Monte Carlo logic regression (Kooperberg and Ruczinski 2005), automated detection of informative combined effects (DICE) (Tahri-Daizadeh et al. 2003), patterning and recursive partitioning (PRP) (Bastone et al. 2004), classification and regression trees (CART) (Breiman et al. 1984) and multiple adaptive regression splines (MARS) (Cook, Zee, and Ridker 2004).

Statistical epistasis has also been pursued via data mining and machine learning approaches such as pattern recognition and data reduction. Pattern recognition describes the use of an algorithm to discover data patterns that discern groups from fully dimensionalized data. An example is neural networks (NN) (Ripley 1996). Data reduction is the collapsing or mapping of data into lower-dimensional space. Examples of data reduction are the combinatorial partitioning method (CPM) (Nelson et al. 2001), restricted partitioning method (RPM) (Culverhouse, Klein, and Shannon 2004) for continuous outcomes, multifactor dimensionality reduction (MDR) (Hahn, Ritchie, and Moore 2003; Hahn and Moore 2004; Moore 2004; Moore et al. 2006; Moore 2007; Ritchie et al. 2001; Ritchie, Hahn, and Moore 2003), and generalized MDR (Lou et al. 2007).

MDR was developed to address the shortcomings of other methods used to assess effects only observable within high orders of dimensionality, such as that which occurs in statistical epistasis. MDR collapses data into simpler patterns relevant to interaction and makes comparisons between interactions to evaluate their effects. The method seeks to reduce dimensionality of a search space by classifying multilocus genotypes into high or low risk classes. These classes are based on ratios of affected to unaffected subjects in population data or discordant sib pairs (DSPs), with a cases/control ≥ 1 ratio usually considered high-risk. The result is compared to an empirical null distribution from a permutation test. A significant result indicates that cases are more likely than controls to have a given multilocus genotype, and can be a useful hypothesis-generating tool when searching for interactive effects.

MDR was shown to find genetic models optimized with regard to classification error by functioning similar to a naïve Bayes classifier using the ratio of status counts within multilocus genotypes to detect interactions (Hahn and Moore 2004). A naïve Bayes classifier treats each measured variable as a statistically independent element and finds the highest posterior probability that an outcome belongs to an observation. This posterior is found as the probability of some outcome times the product of the conditional probabilities of each measured variable given that outcome. For genotypes predicting affection status, the standard application of the naïve Bayes classifier would be the maximum posterior among the two hypotheses (Hahn and Moore 2004):

$$\mathbf{p(aaBB\ is\ low\ risk) = p(low\ risk)\ p(aa \mid low\ risk)\ p(BB \mid low\ risk)} \tag{2-1}$$

$$\mathbf{p(aaBB\ is\ high\ risk) = p(high\ risk)\ p(aa \mid high\ risk)\ p(BB \mid low\ risk)} \tag{2-2}$$

This application of the naïve Bayes classifier only observes main effects of measured variables, making it insensitive to interactive effects. To overcome this shortcoming, MDR collapses the set of independent variables into one aggregate variable and calculates the probabilities of the aggregate instead of the probability for each variable:

$$\mathbf{p(aaBB\ is\ low\ risk) = p(low\ risk)\ p(aaBB \mid low\ risk)} \tag{2-3}$$

$$\mathbf{p(aaBB\ is\ high\ risk) = p(high\ risk)\ p(aaBB \mid high\ risk)} \tag{2-4}$$

Examining the relative frequency of cases to controls for some multilocus genotype and comparing that frequency to a threshold of one in a balanced dataset is conceptually similar to using a naïve Bayes classifier to categorize genotypes into high and low-risk classes (Hahn and Moore 2004).

MDR has detected genetic interactions contributing to risk in several diseases. Some examples are: sporadic breast cancer (Ritchie et al. 2001), essential hypertension (Moore and Williams 2002; Williams et al. 2004), atrial fibrillation (Tsai et al. 2004), type II diabetes (Cho et al. 2004), coronary artery calcification (Bastone et al. 2004), myocardial infarction (Coffey et al. 2004), schizophrenia (Qin et al. 2005), and amyloid polyneuropathy (Soares et al. 2005). However, these results remain to be replicated in independent datasets.

Association

Association analysis maps genomic space to traits at high resolution by measuring the departure from independence of alleles or genotypes at a single locus with disease status. This is distinct from linkage mapping, which observes within-family transmission of a relatively large genomic region and returns the ratio of probabilities that the region is linked versus unlinked to disease status (Risch and Merikangas 1996). Essentially, linkage analysis asks what location in the genome is relevant to a trait, while association asks what allele or genotype at a locus is relevant. Association studies require a high-density of single-nucleotide polymorphisms (SNPs) or other markers for comprehensive coverage between markers. In contrast, linkage requires many fewer polymorphic sites, but benefits greatly from more alleles at each site.

Case-control population data analysis has been the predominant study design for association testing. These methods were the antecedents of family-based association tests and are relatively easy to perform. Case-control studies are theoretically straightforward and statistically powerful; however, perfect implementation of these techniques is elusive because well-matched controls can be very difficult to ascertain. This is compounded with the often violated assumptions of homogeneous, randomly mating, infinite populations, and spurious signals resulting from potential population stratification and admixture. Population stratification occurs when the frequency of exposures and outcomes differ among subsets of the data, and can lead to false positive and negative findings, as reviewed in (Cardon and Palmer 2003). Admixture occurs when offspring are born to parents from different populations, and thus are not representative of either population. For these reasons, and the difficulties inherent in interaction searches, the many unreplicated association studies are not surprising (Hirschhorn et al. 2002). Additionally, the choice of study design may be determined by the trait that is studied. Early-onset traits may be more straightforward for ascertaining offspring parents and siblings, since appropriate unrelated juvenile controls are difficult to collect. Conversely for late onset diseases it might be easier to collect case-control samples, since parents and siblings of elderly persons are likely unavailable.

Family-based association

Family-based association methods have some properties that are complementary to a case-control experimental design. These methods inherently counter the effects of population stratification by comparing counts of alleles or genotypes transmitted to affected samples to counts of transmissions to unaffecteds and/or counts of untransmitted alleles or genotypes. The advantage of this measurement is that the underlying allele and genotype frequency (exposure) of the control samples are matched to the cases (Cordell and Clayton 2005). Additionally, these designs allow estimation of maternal or paternal-specific genotypic effects, maternal-fetal interaction, and imprinting effects (Cordell, Barratt, and Clayton 2004; Weinberg, Wilcox, and Lie 1998; Weinberg 1999). Two disadvantages are the difficult ascertainment of entire families and the extra genotyping burden of at least one triad per case-control pair (Cordell and Clayton 2005). Considering the advantages above, alongside the existence of many family datasets ascertained for linkage analysis, and the need to use family-based designs for some traits, development of family-based association methods relevant to this project proceeded as follows.

Transmission Disequilibrium Test (TDT)

Because of the challenges inherent in the case-control design, the transmission/disequilibrium test (TDT) was developed (Spielman, McGinnis, and Ewens 1993). The TDT is based on transmission of alleles from parents to affected offspring. Because of this, the test only has power to detect associations in the presence of linkage. The TDT also avoids control sampling bias by using parental controls. The TDT is most powerful under a multiplicative risk model (Clayton 1999). This model specifies that the presence of two disease alleles confers the squared risk ratio of one allele. This restriction indicates that under a dominant, recessive, or any nonmultiplicative model, that the statistic might lose power to localize susceptibility loci. One may infer from a significant TDT result that the tested allele is linked and associated with a susceptibility allele. The TDT does not assume Hardy-Weinberg equilibrium (HWE), and is a valid test of association in stratified populations.

The TDT has some known limitations. The statistic suffers from type I error rate inflation when there are missing parental genotypes (Curtis and Sham 1995) or genotyping error (Gordon et al. 2001; Mitchell, Cutler, and Chakravarti 2003). (Mitchell, Cutler, and Chakravarti 2003) proved mathematically that genotyping error can cause increases in the TDT type I error rate. They showed these errors can cause apparent transmission distortion at markers with alleles of unequal frequency that indicates common allele overtransmission. Additionally, in 79 published studies reviewed by (Mitchell, Cutler, and Chakravarti 2003), the major allele was reported as overtransmitted to affected offspring 39% of the time, whereas case-control studies of the same markers, the common allele was often identified as a protective factor. Other limitations of the

TDT are that the tests are for nuclear families, only a single affected child could be selected from any pedigree, and both parents must be available for genotyping. The latter often prevents TDT use in late-onset disorders. A further shortcoming of the test is that it requires transmissions to be independent under the null hypothesis. Therefore, within a region of linkage in a family with several affected siblings, the requirement would be violated and potentially result in elevated type I error. This is because the alleles transmitted within the family would not be independent due to the physical constraint of linkage and commensurate lack of independent assortment. Removal of these samples results in the loss of data and power. Thus the TDT is not capable of analyzing more complex family structure than the trio.

Several similar tests have been designed to alleviate this (Boehnke 1986; Horvath and Laird 1998; Spielman and Ewens 1998). One alternative is the Sib-TDT (S-TDT) of Spielman and Ewens. This technique allows multiple siblings and multiple alleles to be tested with good power compared to other tests (Monks, Kaplan, and Weir 1998). S-TDT compares marker allele frequencies in affected and unaffected siblings. Only a single DSP is required per family for the test, but power can be increased by the addition of more unaffected sibs.

Other family-based association methods

The next generation of family-based tests of association suffered from fewer invalid scenarios than the TDT. They use information from affected and unaffected offspring, and empirically estimate the variance of the underlying distribution of genotypes to normalize the statistic. Some of these are the pedigree disequilibrium test (PDT) (Martin et al. 2000), the genotype pedigree disequilibrium test (genoPDT)(Martin et al. 2003), and the family-based test of association (FBAT) (Horvath, Xu, and Laird 2001).

The pedigree disequilibrium test (PDT) (Martin et al. 2000), was developed for use in large general pedigrees with diverse structure. The fundamental measure of the test is the difference in allelic transmissions to affected and unaffected subjects and transmitted and untransmitted counts within families. The PDT is a valid measure of association in complex pedigrees, meaning that the type I error rate does not increase, even when multiple affected family members are used. This is because in the PDT the independent measures contributing to the normalized statistic, T, are entire pedigrees not individual transmissions. Like the TDT, the PDT does not assume HWE and is not affected by stratification. The PDT has been used in several recent association studies (Deak et al. 2005; Ermakov et al. 2006; Vyshkina et al. 2005) and continues to gain popularity in the field.

The TDT and PDT compare allelic transmissions to determine whether linkage and association exists between alleles at a disease locus and the alleles of a marker locus. The genoPDT uses genotypes as the unit of observation. For autosomal loci, an individual has two alleles that may act additively, alone, or interact to influence risk. When this occurs, it may be that neither allele is solely accountable for disease susceptibility. Allele-based methods do not observe joint-influence effects on risk. This is a crucial observation necessary to detect association in disease models featuring genotypic interaction.

Family-based tests of association modeling genotypic risk were proposed (Schaid and Sommer 1993; Weinberg, Wilcox, and Lie 1998) which examined likelihood of marker data for trios or DSPs parameterized in terms of two genotypic risk ratios. Power studies showed that these genotype-based tests can be more powerful than TDT for recessive or dominant disease models, depending on the degree of dominance and allele frequencies (Schaid and Sommer 1994; Weinberg, Wilcox, and Lie 1998). Unfortunately, these methods are not valid tests of linkage and association for multiple sibships (Weinberg, Wilcox, and Lie 1998), meaning that type I errors representing linkage alone would occur at a higher rate than expected.

The genoPDT was developed to address this. The genoPDT is a simple modification of the PDT procedure, where genotypes are observed instead of alleles. The genoPDT is a valid test of association in the presence of linkage in nuclear or extended pedigrees (Martin et al. 2003). The geno-PDT has been employed in several genetic studies of human disease, for instance assisting in the search for susceptibility loci in autism (Ma et al. 2005; Skaar et al. 2005).

The fusion of the two methods, MDR and genoPDT, provide an opportunity to explore epistatic associations in family data. This is a previously unavailable modality for the study of human disease. The multifactor dimensionality reduction pedigree disequilibrium test (MDR-PDT) (Martin et al. 2006) is the first method to conduct family-based indirect association epistasis research. The MDR-PDT is designed to detect indirect association through Linkage Disequilibrium (LD) between tested loci and epistatic disease model loci in diverse family structures. It functions by calculating the genoPDT statistic for the multilocus genotypes for an interaction between some number of loci. Models are evaluated based on the value of a summed statistic and tested for significance by calculating an empiric null distribution by permutation testing.

Methods

Multifactor dimensionality reduction (MDR) algorithm

MDR is a method designed to detect and test for multilocus effects in case-control data. The algorithm for conducting an MDR analysis is as follows:

1. An exhaustive list of all possible interactions within the orders specified by the user is generated. MDR randomly splits the data into k portions, for use in k-fold cross-validation. Cross validation (CV) functions optimally between 5 and 10 intervals, with lower values of k optimized for computation time (Motsinger and Ritchie 2006).

2. In k-1/k of the data, the samples are stratified by multilocus genotype and the counts of cases and controls within each strata are tabulated.

3. Using the information from step 2, the ratios of cases/controls in each class among all multilocus genotypes within a combination of loci are established

4. A variable denoting high or low risk is established for each genotype of the multilocus combination. The decision rule for this coding is determined either by a simple threshold of 1 for the ratio of cases to controls, or by comparing the ratio of cases to controls for each strata to the ratio in the data of fully genotyped individuals for the model under consideration, if Balanced Accuracy (BA) is used for imbalanced data, where counts of cases and controls differ. Balanced Accuracy is the arithmetic mean of sensitivity and specificity, and has been shown to be a superior fitness metric when data are imbalanced (Velez et al. 2007).

5. The ratios are combined to form one variable summarizing risk with two levels for each multilocus comparison. Balanced accuracy is computed and used to select models from each order of comparison, or number of loci, for testing.

6. Steps 2-5 are repeated for all models generated in step 1.

7. The model with the highest BA is tested in the remaining 1/k of the data to determine the model's ability to predict outcomes in independent datasets. For k CV intervals, k models will be tested in test sets.

8. This procedure is repeated k times. Minimized average prediction error and maximized cross-validation consistency over the k-fold cross-validation procedure are used to select the final model. If these two criteria support different models, then the model with fewest loci is selected, according to the principle of statistical parsimony (Ritchie et al. 2001; Ritchie, Hahn, and Moore 2003).

9. A permutation test is conducted to evaluate significance. (described below)

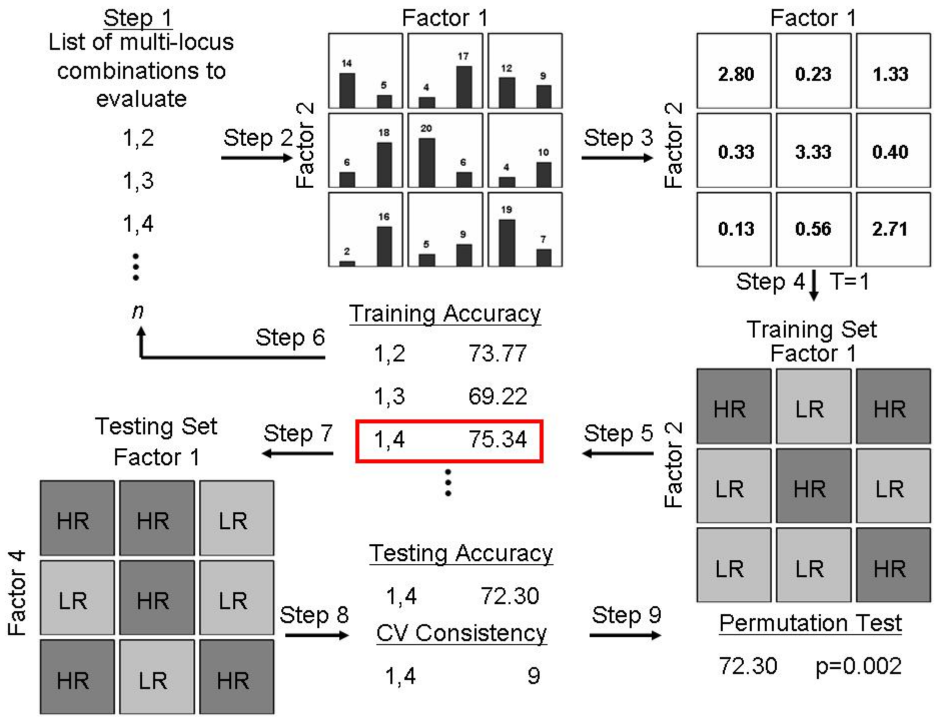


Figure 2-1. Multifactor dimensionality reduction (MDR) algorithm.

To estimate the statistical significance of the result, permutation testing is employed. To estimate the empirical null distribution of results, affection status is randomized according to the original proportions in the dataset. This disrupts associations that may exist between predictor and outcome variables. The MDR procedure is performed as above on the permuted data. This procedure of generating permuted data and subsequent analysis is repeated 1000 times. The actual result is compared to the distribution of ordered results from the permutations to determine significance. A significant result suggests a main or joint effect on risk of genotypes at tested loci.

Power studies with simulated data have shown that MDR has good power to detect epistatic disease models in the presence of 5% genotyping error and 5% missing data. The method loses some power in the presence of 50% phenocopy, and is not robust to 50% locus heterogeneity (Ritchie, Hahn, and Moore 2003). However, recent work with heterogeneity has revealed that this result may lead to undue pessimism since MDR can find some models functioning in complex diseases (Ritchie et al. 2007). Power is lowered slightly by 10,000 noise variables, and type I error is stable at the nominal rate for at least 100-variable simulated datasets (Edwards et al. 2008b). This is in contrast to parametric methods, such as logistic regression, due to strict significance criteria imposed by multiple comparisons correction. Because MDR uses population data, stratification and admixture lowering power and inflating the type I error rate are also a concern.

Pedigree disequilibrium test (PDT)

The PDT is an addition to the lineage of transmission disequilibrium tests, begun by (Spielman, McGinnis, and Ewens 1993) with the TDT. The composition of the TDT is:

$$TDT = \frac{\left(\sum_{i=1}^{h} Y_i\right)^2}{\sum_{i=1}^{h} Y_i^2}$$

(2-5)

Considering a biallelic marker locus, where Y_i is a random variable defined as Y_i = (M_1 transmitted allele – M_1 not transmitted allele) for i = 1,…, h heterozygous parents to affected offspring. The TDT is conducted on affected trios only and is a test of association in the presence of linkage. The null hypothesis of the TDT is that there is association but no linkage to disease at the tested allele. The alternate hypothesis is there is linkage and association to disease at the tested allele. The TDT assumes transmissions from heterozygous parents to affected offspring are independent. The TDT is not a valid test of linkage and association when multiple affected offspring are considered in a region of linkage, as a significant result might only represent linkage and not association (Weinberg, Wilcox, and Lie 1998). For this reason, the TDT is suboptimal for the purpose of following linkage analysis where diverse pedigrees have been ascertained, as it requires that a single affected trio be selected from each pedigree.

Use of the TDT for multiple loci has been considered by (Spielman, McGinnis, and Ewens 1993), using a Bonferroni correction for multiple tests. (McIntyre and Weir 1997; McIntyre et al. 2000) showed use of this correction is appropriate when markers are not associated or linked; however, when markers are in linkage disequilibrium, TDTs at different markers correlate, resulting in a conservative correction due to nonindependence of tests. This results in a loss of power if the Bonferroni correction is used.

The PDT originally was published as an averaged statistic (Martin et al. 2000), where deviations from independent assortment were cataloged by the statistic D.

$$D = \frac{1}{n_T + n_S}\left(\sum_{j=1}^{n_T} X_{Tj} + \sum_{j=1}^{n_S} X_{Sj}\right)$$

(2-6)

Where within an informative nuclear family, at some multiallelic locus, random variables X_T from trio families are defined as (#times allele is transmitted to cases –

#times the allele is not transmitted to cases) and X_S as (#times allele is transmitted to affected sib – #times allele is not transmitted to unaffected sib). Define a random variable D as the sum of X values for a pedigree divided by the number of X values. Each informative pedigree provides one D statistic. An informative pedigree is either a nuclear family with at least one affected child with both parents genotyped at the locus with one parent heterozygous, a DSP with different genotypes at the locus with or without parental genotypes, or an extended pedigree with at least one informative nuclear family or DSP (Martin et al. 2000; Martin, Bass, and Kaplan 2001; Martin et al. 2003). The PDT statistic is constructed by the equation:

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i^2}}$$

(2-7)

T is a random variable given as the ratio of the sum of Ds in the numerator and the square root of the sum of squared Ds is the variance estimate in the denominator. Under the null hypothesis: E(X) = 0, E(D_i) = 0, E(T) = 0, and Var(T) = 1 (Martin et al. 2000).

Under the null hypothesis, the estimate of the statistic variance is the sum of squared Ds. The variance of T is also the variance of the sum of all D_i. The variance of the sum of D statistics is equal to the sum of the variances of the D statistics. This is an application of the principle that independent variances are additive with regard to their contribution to the variance of the aggregate of independent parts. Under the null hypothesis E(D_i) = 0, and E(ΣD_i) = 0, D is a family's transmission deviation score with excess transmissions to cases as the fundamental measurement. The variance of a given D statistic is given by the square of that D. The expected value of the sum of squared independent D statistics estimates the variance of the sum of all D_i and the T statistic (Equation 2-7), which under the null hypothesis has a standard normal N(0,1) distribution.

$$Var\left(\sum_{i=1}^{N} D_i\right) = \sum_{i=1}^{N} Var(D_i) = E\left(\sum_{i=1}^{N} D_i^2\right)$$

(2-8)

Because of these properties, under the null hypothesis the squared **T** statistic is equivalent to a chi-square with one degree of freedom. This makes significance assessment straightforward for a tested allele.

In a simulated power study, PDT was superior to Sib-TDT in 5,000 replicates of 250 three-generation families for 6 models of varying genetic effect and 5,000 replicates of 500 nuclear families with 2 or 5 sibs for 2 models each. PDT was found to be relatively robust to affected misclassification rates as high as 50% in 5,000 replicates of 150 extended pedigrees with at least one affected sibling for 2 models. Type I error rates were at the appropriate level for 6 genetic models in 5,000 replicates of 250 pedigrees and also for varying sample sizes with 10,000 replicates of 25, 50, 100, and 250 pedigree datasets for one model (Martin et al. 2000).

This version of the statistic was found to introduce a possible bias under some genetic models (Martin, Bass, and Kaplan 2001). The bias arises when the critical assumption of the PDT, E(T) = 0 under the null hypothesis, is violated in some scenarios. For some family structures this assumption is violated and the type I error rate could increase. (Martin, Bass, and Kaplan 2001) provides the example of a fully penetrant dominant biallelic model with no phenocopies, rare disease allele d1, and common wild-type allele d2, such that there is one segregating copy per pedigree of the disease allele. In the example, extended three generation pedigrees are sampled with 6 equally likely transmission patterns. Note that all potentially biased pedigrees have affected grandparent (GP2), parent (P2), and offspring (O). Otherwise, there is at most one affected triad per pedigree. Consider a biallelic marker locus (M1, M2) with minor allele M1 such that only one founder is a heterozygote for M1, with equal probability of heterozygosity among the three founders (GP1, GP2, P1). The possible transmission patterns and PDT calculations are detailed below.

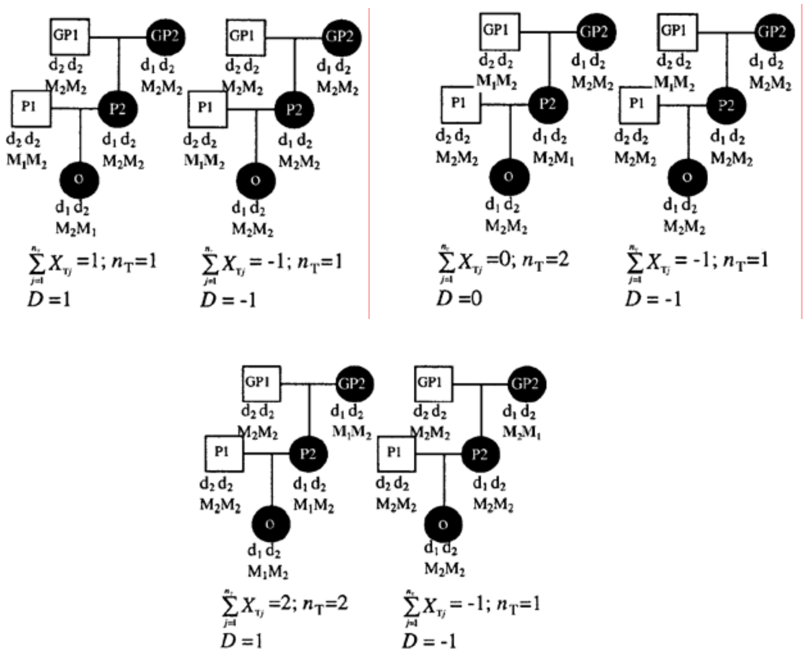


Figure 2-2. Transmission patterns and pedigree disequilibrium test calculations from (Martin, Bass, and Kaplan 2001).

E(D) for these pedigrees is -1/6, causing E(T) ≠ 0 under the null hypothesis, which violates an assumption and increases the type I error. This bias is corrected by the statistic:

$$D = \left[\sum_{j=1}^{n_T} X_{Tj} + \sum_{j=1}^{n_S} X_{Sj} \right]$$

(2-9)

This is the form of the statistic employed by the genoPDT and the MDR-PDT. This statistic has E(D) = 0 and E(T) = 0 for the pedigrees above. The T calculation is unchanged. This formulation is driven by the concept that weight is added for each informative transmission per pedigree, allowing a pedigree to affect the statistic in proportion to its information contribution. Pedigree analysis using statistics based on sums of random variables were also proposed in (Abecasis, Cookson, and Cardon 2000; Martin, Kaplan, and Weir 1997; Rabinowitz and Laird 2000; Teng and Risch 1999).

In a power and type I error study of six genetic models in (Martin, Bass, and Kaplan 2001), PDT-sum had more power for each model than PDT-old with appropriate type I error rates.

The genoPDT is a simple modification of the PDT procedure, where genotypes and not alleles are observed. The genoPDT is applicable to nuclear or extended pedigrees, and can either be used to test a particular genotype, or a global test of all genotypes at a locus simultaneously (Martin et al. 2003). As with the PDT, the statistic is N(0,1) under the null hypothesis, meaning that for a given genotype's T^2, significance can be assessed by comparison to a chi-squared distribution with one degree of freedom. The global statistic is the averaged T^2 statistic across genotypes at a locus multiplied by g-1/g genotypes and compared to a chi-squared with g-1 degrees of freedom. It was reported in (Kaplan, Martin, and Weir 1997; Martin, Kaplan, and Weir 1997), that this statistic converges asymptotically to a chi-square with g-1 degrees of freedom under the null hypothesis when genotype frequencies are equal. For other frequencies, the distribution is approximate, but performed well in simulated data (Kaplan, Martin, and Weir 1997).

To examine power and type I error for geno-PDT in (Martin et al. 2003), simulation studies were conducted with recessive, dominant, and additive modes of inheritance, high and low prevalence, and with equal marker and disease allele frequencies. To estimate type I error, 5000 datasets of 200, 100, and 50 families were simulated and tested for global and individual genotypes. Type I error rates were close to the nominal level for most tests except for those with less than 3% minor allele frequency, for which the global test is conservative. For power, genoPDT was compared to PDT in 1000 replicates of 200 nuclear families and was found to be more powerful than PDT for all but additive models.

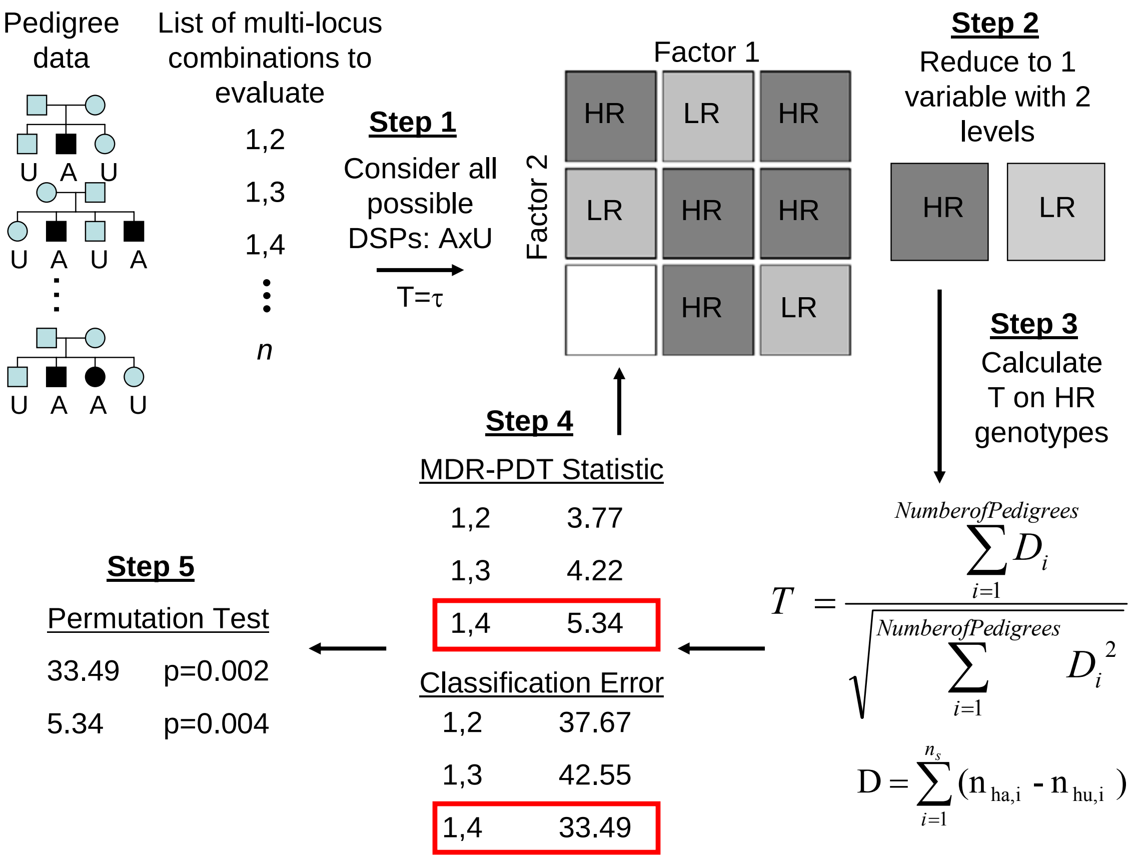
Multifactor dimensionality reduction pedigree disequilibrium test (MDR-PDT)

The MDR-PDT is an approach that uses the genoPDT statistic within the MDR algorithm (Martin et al. 2006). MDR is a nonparametric procedure designed and optimized to find plausible hypotheses about multi-dimensional epistatic models in discrete population or DSP data, allowing the inference of a main or joint effect on disease risk at or near tested variables. The null hypothesis of MDR is that there are no main or joint effects on disease risk of any factor considered. The alternate hypothesis is that there is a main or joint effect on risk among the data considered. The assumptions of MDR are that data have two outcome groups, cases and controls are unstratified with regard to allele and genotype frequency, and are independent or DSP-matched.

The genoPDT statistic measures transmission disequilibrium of genotypes to affected offspring in general pedigrees, allowing the inference that a tested locus is in LD with a susceptibility locus. The null hypothesis of the genoPDT is that linkage is present with no association to disease among the tested variables. The alternate hypothesis is that both linkage and association are present between measured genotypes and disease variants. The assumptions of the genoPDT are two outcome groups, that each pedigree is independent of other pedigrees in the data, linkage is present between tested loci and a disease variant, and a standard normal test statistic distribution under the null hypothesis.

These two methods form the MDR-PDT, a technique designed to discover loci in LD with susceptibility loci that act alone or jointly on risk in diverse family architectures. MDR-PDT allows the inference that the best model is composed of genetic factors that are related to trait variation. The null hypothesis of the MDR-PDT is that there is no association with disease among variables considered, and that no tested factors act alone or epistatically with other tested factors to influence the tested trait. The alternate hypothesis is that there is linkage and association among tested variables with variants which may act alone or jointly to increase risk. The assumptions of MDR-PDT are the same as those for the genoPDT.

The MDR-PDT is a within-family measure of indirect or direct association between genotype and disease. As described above, the genoPDT statistic functions within the framework of the MDR algorithm (Figure 2-1). The genotypes are the same as those for which MDR would find the ratio of affecteds to unaffecteds. All possible DSPs are generated for each sibship and pooled. This pool of cases and controls are considered to determine which genotypes are high and low risk. The T statistic is calculated for the pooled high-risk genotypes for each interaction. Let this T be the MDR-PDT statistic. The models are ordered and evaluated by the T statistics. A permutation test is applied to estimate an empiric null distribution for significance assessment. The steps of MDR-PDT follow (Figure 2-3).



$$T = \frac{\sum_{i=1}^{NumberofPedigrees} D_i}{\sqrt{\sum_{i=1}^{NumberofPedigrees} D_i^2}}$$

$$D = \sum_{i=1}^{n_s} (n_{ha,i} - n_{hu,i})$$

Figure 2-3. Multifactor dimensionality reduction pedigree disequilibrium test (MDR-PDT) algorithm.

1. All possible DSPs are generated within each sibship (affected times unaffected) and pooled.

2. Each genotype is determined to be high or low risk by comparing the ratio of cases to controls from the pooled DSPs to a threshold τ, such as τ = 1, which indicates positive or negative association with affected status.

3. Statistics for high-risk genotypes are calculated using the D in Eq. 5 with the function in Eq. 3. This is the MDR-PDT statistic for this model. A CE is also calculated for the model, as in the training set for MDR.

4. The procedure repeats for every combination of K loci, calculating an MDR-PDT statistic and CE for each, choosing the largest MDR-PDT statistic as the final result.

5. A permutation test is performed to determine the distribution of the null hypothesis, to which the result from step 4 is compared for significance assessment.

The permutation test is conducted to discover the significance of an n-locus test. Formally, it is performed to test the null hypothesis that the pooled n-locus high-risk genotype statistic is independent of disease status.

The test is based on the idea that under the null hypothesis there is no association between status and genotype. Therefore, any combination of affected sibs is equally likely for a given sibship. In general, for a sibship with *s* siblings (*a* affected, *u* unaffected), there are s!/(a!u!) equally likely permutations. To conduct a test, one of these permutations is randomly selected for each sibship. Each pedigree of real data has a null counterpart of the same size in the pseudosample with same genotypes but randomized status. For families with parental genotypes and a single affected offspring, the permutation test uses transmitted and untransmitted parental alleles in a pseudocontrol rather than genotypes of discordant siblings. The status of the original case and the pseudocontrol created from the untransmitted parental alleles are then permuted for the test. Steps 1-4 are performed to find the value of a maximum MDR-PDT statistic for n-loci and the associated CE from the empirical null distribution. Repeated many times, the null distributions of the maximum MDR-PDT statistic and CE for an n-locus comparison are approximated. The p-value of the result from the original unpermuted data can be estimated by the proportion of the null distribution that exceeds it. This is an empiric p. The test based on the permutation procedure should have the correct type I error, even for sparse data. If the model is significantly associated with disease, it should have a statistic larger than those arising by chance from permuted data.

In simulation-based power studies (Martin et al. 2006), the same 2-locus models used to evaluate MDR (Ritchie, Hahn, and Moore 2003) were tested and compared between MDR and MDR-PDT. Simulations of three pairs of 2-locus models with minor allele frequencies of 0.5, 0.25, and 0.1, respective heritabilities of 5%, 2% and 0.8%, and equal marginal penetrances were evaluated. Examples are presented with 50% phenocopy error, where in triads, 50% of offspring were simulated with random genotypes, and in DSPs, 50% of affected sibs were randomly simulated. 100 replicate 200-triad datasets of each type were analyzed. Power was also investigated for DSPs vs. larger discordant sibships with 2 affecteds and 1 unaffected (AAU). 100 replicate datasets of each type were simulated for each model with and without 50% phenocopy error.

Performance of MDR-PDT was as good as or better than MDR with 10-fold cross validation for trios compared to the transmitted/untransmitted DSPs without 50% phenocopy error. This difference in power increases as the heritability of the model falls, with MDR-PDT being stronger relative to MDR for lower heritability models. For trios simulated with phenocopy error, MDR was slightly more powerful. In the comparison of MDR-PDT between AAU and DSP sibships, the AAU sibships were at least as powerful in every circumstance, and were most often more powerful with and without phenocopy error. This illustrates the gains in power experienced by MDR-PDT for larger pedigrees.

Type I error rate was evaluated for 1, 2 , and 3-locus tests at p = 0.05 (Martin et al. 2006). Samples of 200 triads or 200 AAU sibships with no parental data were simulated with random loci. Best models were chosen by the algorithm and evaluated by the permutation test. Type I errors were close to nominal, ranging from 0.042 to 0.054. This result proves the validity of the permutation procedure as a test for an n-locus result.

These developments described in Chapter II led directly to the methodological and applied specific aims described in the subsequent chapters. The development of family-based tests of association for general pedigrees, and methods developed to detect multilocus associations led directly to the analysis of pedigree data to search for epistasis. The methodological extensions provided in Chapter V are the result of observations made during those analyses.

CHAPTER III

HYPOTHESIS AND SPECIFIC AIMS

Hypothesis: The MDR-PDT algorithm will have improved utility upon the conclusion of my aims by the integration of cross validation and further extensions developed over the course of evaluating power.

Specific aim 1: Implement cross-validation into the MDR-PDT algorithm.

Cross-validation is a procedure in which the data are randomly split into nearly same-size fractions and analysis is performed on each fraction to discover consistently supported models. The purpose of CV is to minimize Type-I error by using a measure of consistency across the analyses and assist in selection of best models. This approach is part of the conventional MDR algorithm, and has been experimentally demonstrated to be effective.

Specific aim 2: Conduct simulation-based power and alpha studies.

Power and Type I error of the MDR-PDT without CV has been assessed previously. Data will be simulated for a variety of circumstances where epistatic disease models, heritabilities, odds ratios, minor allele frequencies, and numbers of interacting loci are varied to rigorously test the algorithm with CV.

Specific aim 3: Implement the likelihood ratio test of significance into MDR-PDT.

The specificity of MDR-PDT is low when strong main effects are present in the data. Fitting saturated conditional logistic regression models for MDR-PDT models from real and permuted data allows the null hypothesis of no interaction to be specifically tested. This approach will take advantage of the speed of MDR-PDT and provide a single valid p-value for the correct null hypothesis. This technique will be applied to the data simulated in AIM 2.

Specific aim 4: Apply MDR-PDT to a real dataset.

MDR-PDT has been applied to an Alzheimer's disease (AD) dataset in previous studies. We have also analyzed a larger AD dataset using MDR-PDT, APL, and the parametric geno-PDT. The results of these analyses will are presented here.

CHAPTER IV

GENERATING LINKAGE DISEQUILIBRIUM PATTERNS IN DATA SIMULATIONS USING GENOMESIMLA

Overview

Genome-Wide association (GWA) studies are becoming a common tool for the exploration of the genetic components of common disease. The analysis of such large scale data presents unique analytical challenges, including problems of multiple testing, correlated independent variables, and large multivariate model spaces. These issues have prompted the development of novel computational approaches. Thorough, extensive simulation studies are a necessity for methods development work to evaluate the power and validity of novel approaches. Many data simulation packages exist, however, the resulting data is often overly simplistic and does not compare to the complexity of real data; especially with respect to LD. To overcome this limitation, we have developed genomeSIMLA. GenomeSIMLA is a forward-time population simulation method that can simulate realistic patterns of LD in both family-based and case-control datasets. In this manuscript, we demonstrate how LD patterns of the simulated data change under different population growth curve parameter initialization settings. These results provide guidelines to simulate GWA datasets whose properties resemble the HapMap.

Adapted from: Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, Martin E, Ritchie MD. Generating linkage disequilibrium patterns in data simulations using genomeSIMLA. Lecture Notes in Computer Science, in press (2007).

Introduction

The initial success of the human genome project is the nearly complete characterization of the consensus human sequence (Finishing the euchromatic sequence of the human genome 2004; Lander et al. 2001). This has greatly increased our ability to describe the structure of genes and the genome and to better design experiments. Perhaps of even more importance for disease gene studies is the HapMap data (The International HapMap Project 2003). This vast pool of characterized common differences between individuals greatly increases our ability to perform targeted or GWA studies by using the measured patterns of LD as a foundation for single nucleotide polymorphism (SNP) selection and data interpretation. SNPs are single base changes in DNA that vary across individuals in a population at a measurable frequency. GWA studies interrogate hundreds of thousands of SNPs throughout the entire human genome to map disease susceptibility or drug response to common genetic variation.

LD is the nonrandom association of alleles at multiple SNPs. This association can be quantified by the squared Pearson's product-moment correlation coefficient (r^2). Also available is a related measure, D', which is the proportion of the maximum possible r^2 given a difference in allele frequencies. The r^2 value gives an indication of the statistical power to detect the effect on disease risk of an ungenotyped SNP, whereas D' is indicative of past recombination events.

Advances that increase the complexity of data simulations will permit investigators to better assess new analytical methods. GenomeSIMLA (an extension of (Dudek et al. 2006)) was developed for the simulation of large-scale genomic data in population based case-control or family-based samples. It is a forward-time population simulation algorithm that allows the user to specify many evolutionary parameters to control evolutionary processes. GenomeSIMLA simulates patterns of LD representative of observed human LD patterns through realistic processes of mating and recombination. This tool will enable investigators to evaluate the sampling properties of any statistical method that is applied to large-scale data in human populations. We describe the algorithm and demonstrate its utility for future genetic studies with GWA.

Background

Multiple technologies now allow a GWA design to be implemented by genotyping between 500,000 and 1.8 million SNPs with high fidelity and low cost. It is conceivable that technological advances will lead to whole genome sequencing in the not too distant future that will involve generating 10-20 million base pair variations per individual. In a GWA approach, a dense map of SNPs is genotyped and alleles, genotypes, or haplotypes are tested directly for association with disease. Estimates suggest that with 500,000 SNPs, ~50-75% of the common variation in the genome is captured (de Bakker et al. 2005). Recent studies have shown that the precise extent of coverage is dependent on study design, population structure, and allele frequency (Barrett and Cardon 2006). Regardless, GWA is by far the most detailed and complete method of genome interrogation currently possible. GWA has successfully detected association with genetic variation in several common diseases including breast cancer (Easton et al. 2007; Hunter et al. 2007), type II diabetes (Saxena et al. 2007; Scott et al. 2007; Welcome Trust Case-Control Consortium 2007; Zeggini et al. 2007), obesity (Lyon et al. 2003), myocardial infarction (McPherson et al. 2007) and others (Welcome Trust Case-Control Consortium 2007).

GenomeSIM and SIMLA

GenomeSIM (Dudek et al. 2006) was developed for the simulation of large-scale genomic data in population based case-control samples. It is a forward-time population simulation algorithm that allows the user to specify many evolutionary parameters and control evolutionary processes. SIMLA (or SIMulation of Linkage and Association) (Bass, Martin, and Hauser 2004; Schmidt et al. 2004) is a simulation program that allows the user to specify varying levels of both linkage and LD among and between markers and disease loci.

SIMLA was specifically designed for the simultaneous study of linkage and association methods in extended pedigrees, but the penetrance specification algorithm can also be used to simulate samples of unrelated individuals (e.g., cases and controls). We have combined genomeSIM as a front-end to generate a population of founder chromosomes. This population will exhibit the desired patterns of LD that can be used as input for the SIMLA simulation of disease models. Particular SNPs may be chosen to represent disease loci according to desired location, correlation with nearby SNPs, and allele frequency. Using the SIMLA method of disease modeling, up to six loci may be selected for main effects and all possible 2 and 3-way interactions as specified in (Marchini, Donnelly, and Cardon 2005) among these 6 loci are available to the user as elements of a disease model. Once these loci are chosen the user specifies disease prevalence, a mode of inheritance for each locus, and relative risks of exposure to the genotypes at each locus. An advantage of the SIMLA approach to the logistic function is it can simulate data on markers that are not independent, yet yield the correct relative risks and prevalence. Many simulation packages using a logistic function for penetrance specification do not have this capability. Modeling of purely epistatic interactions with no detectable main effects, as in genomeSIM, is also supported separately and can simulate 2-way, 3-way, up to n-way interactions. Purely epistatic modeling allows the user to specify a model odds ratio, heritability, and prevalence for disease effects. Thus, the marriage of genomeSIM and SIMLA has allowed for the simulation of large scale datasets with realistic patterns of LD and diverse realistic disease models in both family-based and case-control data.

Alternative genetic data simulation packages

Several genetic data simulation packages are currently available. SIMLINK (Boehnke 1986; Ploughman and Boehnke 1989), SIMULATE, and SLINK (Weeks, Ott, and Lathrop G.M 1990) will simulate pedigrees from an existing dataset. Coalescent-based methods (Kingman 1982) have been used for population based simulation in genetic studies; however, standard approaches which are extremely efficient in simulating short sequences, are not successful for long sequences. GENOME is a novel coalescent-based whole genome simulator developed to overcome previous limitations (Liang, Zollner, and Abecasis 2007). HAP-SAMPLE uses the existing Phase I/II HapMap data to resample existing phased chromosomes to simulate datasets (Wright et al. 2007). In recent years, forward-time population simulations have been developed including easyPOP (Balloux 2001), FPG (Hey 2005), FREGENE (Hoggart et al. 2007), and simuPOP (Peng and Kimmel 2005). All of the existing simulation packages have strengths and weaknesses. The motivation for developing genomeSIMLA is to achieve the ability to simulate: 1) realistic patterns of LD in human populations, 2) GWA datasets in both family and case-control study designs, 3) single or multiple independent main effects, and 4) purely epistatic gene-gene interactions in efficient, user friendly software. Existing simulation packages can do one or more of these, but few are able to succeed in all areas.

Methods

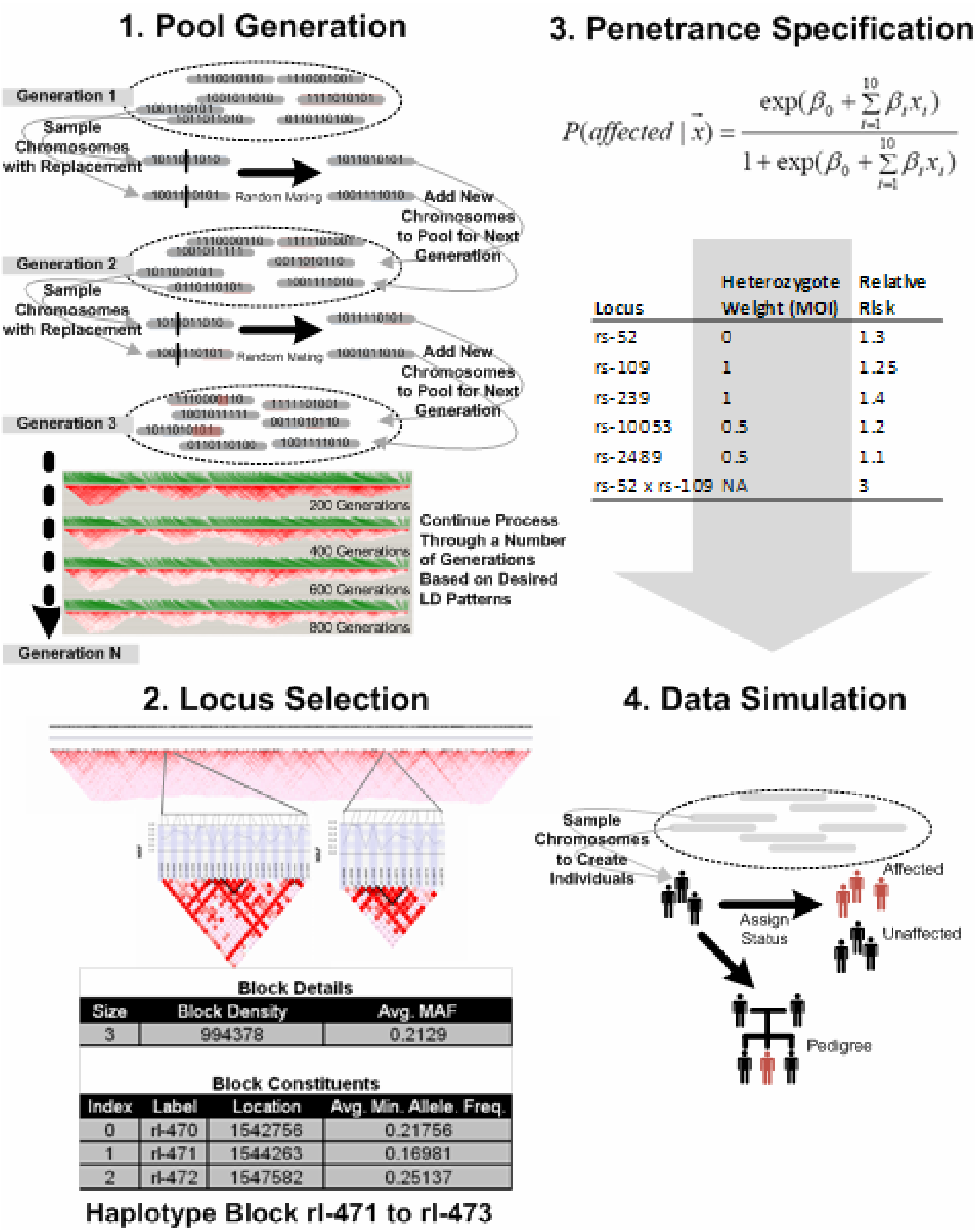
GenomeSIMLA

GenomeSIMLA generates datasets using a forward-time population simulator which relies on random mating, genetic drift, recombination, and population growth to allow a population to naturally obtain LD features. An initial population (or pool of chromosomes) is generated using allele frequencies and positions for a set of desired SNPs or random allele frequencies for real or synthetic SNP locations. Recombinant gametes are created based on intermarker recombination probabilities calculated using Kosambi (accounting for recombination interference) or Haldane (accounting for multiple events between polymorphisms) genetic mapping functions. Recombination probability between two polymorphisms is determined by the Kosambi or Haldane function of the map distance based on a 1 centimorgan per 1 million bases genetic map. The number of crossover events for a pair of parental chromosomes to generate gametes is a random Poisson variable where the expected number of events is the sum of all intermarker recombination probabilities for the chromosome. The two resulting gametes, one from each parent, are then combined to create a new individual. The mapping approximation of 1 million bases per centimorgan is applied here; however, other values could be applied to simulate population-specific genetic maps or recombination hotspots.

The random mating and recombination process continues on the pool of chromosomes for a set number of generations to generate realistic patterns of LD and produce sufficient numbers of chromosomes for drawing datasets. After the pool of chromosomes has developed suitable LD and grown to a useful size, datasets can be drawn by randomly sampling chromosomes with replacement to create nonredundant individuals. Disease-susceptibility effects of multiple genetic variables can be modeled using either the SIMLA logistic function (Bass, Martin, and Hauser 2004; Schmidt et al. 2004) or a purely epistatic multi-locus penetrance function (Moore et al. 2004) found using a genetic algorithm. These individuals are either mated to yield pedigrees, for family-based datasets, or are evaluated by a logistic function or a purely epistatic penetrance function of their genotypic exposures to determine disease status for case-control datasets.

Figure 4-1 illustrates the general steps involved in producing a simulated dataset. As a first step, genomeSIMLA establishes the size of the genome based on the user specified parameters. The total number of SNPs is unlimited except by hardware considerations. We are currently able to simulate at least 500K SNPs. The simulator generates the number of SNPs, recombination fraction, and allele frequencies within user specified margins or boundaries. GenomeSIMLA then generates an initial population (or pool of chromosomes) based on the genome established in the previous step. For each SNP in the genome, the simulator randomly assigns an allele to each chromosome based on the allele frequencies of the SNP. A dual-chromosome representation is used for creating individuals to allow for an efficient representation of the genome and for crossover between chromosomes during the mating process. The genotype at any SNP can be determined simply by adding the values of the two chromosomes at that position. As a result, the genotypes range from 0 to 2 at any SNP.

The initial population forms the basis for the second generation in the simulation. For each cross, four chromosomes are randomly selected with replacement to create two individuals to be the parents for a member of the new generation. Each parent contributes one haploid genome to the child. GenomeSIMLA creates the gametic genotype by recombining the parent's chromosomes. The total number of chromosomes in the pool can be constant or follow a population growth model (linear, exponential, or logistic). This will determine the number of mating/crossover events that occur. GenomeSIMLA continues through a specified number of generations depending on the desired LD patterns.


1. Pool Generation
Generation 1
Sample Chromosomes with Replacement
Random Mating
Add New Chromosomes to Pool for Next Generation
Generation 2
Sample Chromosomes with Replacement
Random Mating
Add New Chromosomes to Pool for Next Generation
Generation 3
200 Generations
400 Generations
600 Generations
800 Generations
Continue Process Through a Number of Generations Based on Desired LD Patterns
Generation N
2. Locus Selection
Haplotype Block rl-471 to rl-473
3. Penetrance Specification
4. Data Simulation
Sample Chromosomes to Create Individuals
Assign Status
Affected
Unaffected
Pedigree

$$P(affected \mid \vec{x}) = \frac{\exp(\beta_0 + \sum_{i=1}^{10} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{10} \beta_i x_i)}$$

Locus	Heterozygote Weight (MOI)	Relative Risk
rs-52	0	1.3
rs-109	1	1.25
rs-239	1	1.4
rs-10053	0.5	1.2
rs-2489	0.5	1.1
rs-52 x rs-109	NA	3

Block Details

Size	Block Density	Avg. MAF
3	994378	0.2129

Block Constituents

Index	Label	Location	Avg. Min. Allele. Freq.
0	rl-470	1542756	0.21756
1	rl-471	1544263	0.16981
2	rl-472	1547582	0.25137

Figure 4-1. Simulator overview. This figure demonstrates the steps of the genomeSIMLA algorithm for simulating data as described in the text. In summary, the process of simulating data is as follows:

1. Develop the chromosome pool using either artificial intermarker distances and recombination or positions from real data. Set the parameters of the population growth to fit the desired LD properties.

2. Select loci to be the disease susceptibility loci in the simulation. Loci can be searched for using built-in search tools allowing the user to screen loci based on allele frequency, block size, and position.
3. Specify the disease model. Either multiple loci with main effects and interactions among them or purely epistatic effects can be modeled
4. Simulate data by either drawing individuals for case-control data or founders for family data.

To create datasets, chromosomes are sampled from the pool with replacement and affection status is assigned based on the user-specified penetrance table or logistic function. Samples are drawn until the desired number of cases and controls are accumulated. In family-based simulation, founders are drawn and offspring are created using the same methods as applied in the pool generation steps. The penetrance function is applied to the parents and offspring to determine status and the resulting pedigrees are retained in a dataset if the study ascertainment criteria are met. Otherwise the pedigrees are discarded and the founder chromosomes are allowed to be drawn again.

GenomeSIMLA: implementation

Performance on desktop grade hardware and interpretable results reporting were main goals of software development. Users can simulate data on modern desktop hardware with at least three Gigahertz processors and two Gigabytes of RAM and have their datasets within 24-48 hours for many parameter settings; though the exact time will be dependent upon the particular growth curve used and the desired chromosome pool size. To achieve these goals we focused on memory requirements, threading, and LD plotting.

C++ allows us to utilize memory with minimal overhead; however, retaining 100,000 chromosomes of 500,000 SNPs each is not a trivial task. To maintain this within the limits of a modern desktop machine, we represent each chromosome as a binary string. Also, unless otherwise specified, genomeSIMLA will only have a single chromosome pool in memory. One drawback of using the binary string for a chromosome is that we are limiting genomeSIMLA to biallelic data. By retaining a single pool in memory, our memory requirements fall reasonably under 2 gigabytes of RAM.

We have implemented two different threading mechanisms to allow users to take full advantage of the hardware available to them. When using genomeSIMLA in 32bit environments, there are at most 4 Gigabytes of memory available to the system. To accommodate users with multiple processors running 32bit operating systems, we allow specification of the number of threads per chromosome. This incurs a minimal memory increase but can speed the calculations up considerably. However, when running genomeSIMLA under 64bit, we allow for configurations to specify any number of chromosomes be managed simultaneously. This is limited by available hardware and process time scales almost linearly with the number of processors available.

To address our reporting needs, we implemented our own LD plotter. Existing LD plotting software, such as haploview, could not accommodate whole chromosome data. As a result, genomeSIMLA is capable of generating whole chromosome LD plots similar to those generated by haploview. Calculating whole-genome LD statistics on large chromosomal pools is a computationally intensive process. To reduce computation time, LD statistics can be optionally calculated on a sample of the entire pool.

Growth Curve Parameter Sweep

To develop an understanding of the consequences of different population growth curve parameter settings, we have designed a series of experiments. The hypothesis is that some combination of population growth parameters will emulate the average profile of correlation by distance observed in the HapMap data. We used a generalized logistic curve, or Richards curve, to model realistic population growth (Richards 1959) Equation 4-1. The Richards growth curve consists of five parameters: A -- the initial population size or lower asymptote, C -- the carrying capacity of the population or the upper asymptote, M -- the generation of maximal growth, B – the growth rate, and T – a parameter that determines if the point of maximal growth occurs near the lower or upper asymptote.

$$Y = A + \frac{C}{(1 + Te^{-B(x-M)})^{1/T}} \tag{4-1}$$

This function provides a parameterized theoretical basis for population growth, though real population growth likely has more stochastic variability. To allow variability in population growth, we implemented a jitter parameter that draws a random number from a uniform distribution over a range specified by the user and adds or subtracts that percentage from the population size predicted by the growth curve. For the purposes of the parameter exploration in this study, however, the jitter parameter was set to zero.

We scanned through a wide range of parameters to find population growth profiles providing suitable correlation among genetic variables for data simulation. Since there were five parameters to vary and many possible values for each, we were still limited to a small subset of the possible sets of growth parameters available (Table 4-1). Prior to this study, we performed a number of parameter sweeps to evaluate ranges that were likely to yield interesting and realistic LD patterns (results not shown) in a population of 100,000 chromosomes. For this study, we split the parameter sweep into three scans. In total, 726 combinations of parameter settings were examined for average LD over distance.

Table 4-1. Parameter sweep of population growth parameters for the logistic function: settings for three scans.

Parameters	Scan 1	Scan 2	Scan 3
A - Lower asymptote	500, 750, 1000	100, 150, 200, 250, 300	750, 1000, 1250, 1500
C - Upper asymptote	120k, 500k, 900k	110k, 120k	120k
M - Maximum growth time	305, 315, 325, 335, 345, 355	350, 400, 450	500, 1000, 1500, 2000, 2500, 3000
B - Growth rate	0.005, 0.0075, 0.01	0.018, 0.02, 0.022, 0.025	0.02, 0.025, 0.03, 0.035, 0.04
T - Maximum growth position	0.1. 0.2, 0.3	0.1	0.1
Total parameters	486	120	120

We predict that a common usage of genomeSIMLA software will be to simulate case-control and family-based whole-genome association datasets containing 300,000-500,000 biallelic markers across the genome. These data could be used to evaluate the sampling properties of new or established association methods or techniques to characterize the genetic structure of populations. While genomeSIMLA can simulate data of this magnitude, for this study, we wanted to focus on a single chromosome. Thus, we simulated the 6031 chromosome 22 markers used on the Affymetrix 500K SNP Chip.

To visualize the results of each parameter combination, average R^2 by distance in kilobases was graphed for the simulated data and for the CEPH (Caucasian), Yoruba

(African), and Chinese/Japanese HapMap populations. This representation captures global estimates of correlation by distance across the entire chromosome.

Results

Parameter settings in Scan 1 did not yield LD which was comparable to HapMap samples. A trend was observed among the better fitting models that the parameters C and T always functioned best when set to 120k and 0.1, respectively. Scan 2 examined very small initial populations and more rapid growth to strengthen LD profiles through rapid genetic drift. These unfortunately also resulted in the fixing, or drifting to zero frequency, of many alleles. Scan 3 focused on larger initial populations, late maximum growth, and rapid growth. These simulations were the most successful and resulted in several curves which approximated LD in HapMap samples well. One such example is presented in Figure 4-2.

While not a perfect fit to any population, the curve represents a good approximation of the correlation observed in the data. Of note is the fit in the shorter ranges, since short range LD is more related to the power to detect associations with disease susceptibility loci (Durrant et al. 2004). A sample of the actual LD observed among the markers is presented in Figure 4-3. The goal of this study was to obtain data which on average is similar to HapMap data. Since we initialized the chromosomes with random minor allele frequency and the measure r^2 is sensitive to this parameter, it is not expected that each intermarker correlation will be identical to the value calculated from the HapMap data. However, it can be seen here that the major features and regions of high and low correlation are captured. The growth curve in Figure 4-2 and the LD shown in Figure 4-3 were generated with the following parameters: A=1500, C=120000, M=500, B=0.02, T=0.1. D', an alternate measure of LD, was more difficult to fit than R^2. The curves for the simulated data generally were stronger than those observed for the real data in the short ranges but weaker at long ranges. The reasons for this are unknown but are a topic of further study for genomeSIMLA.

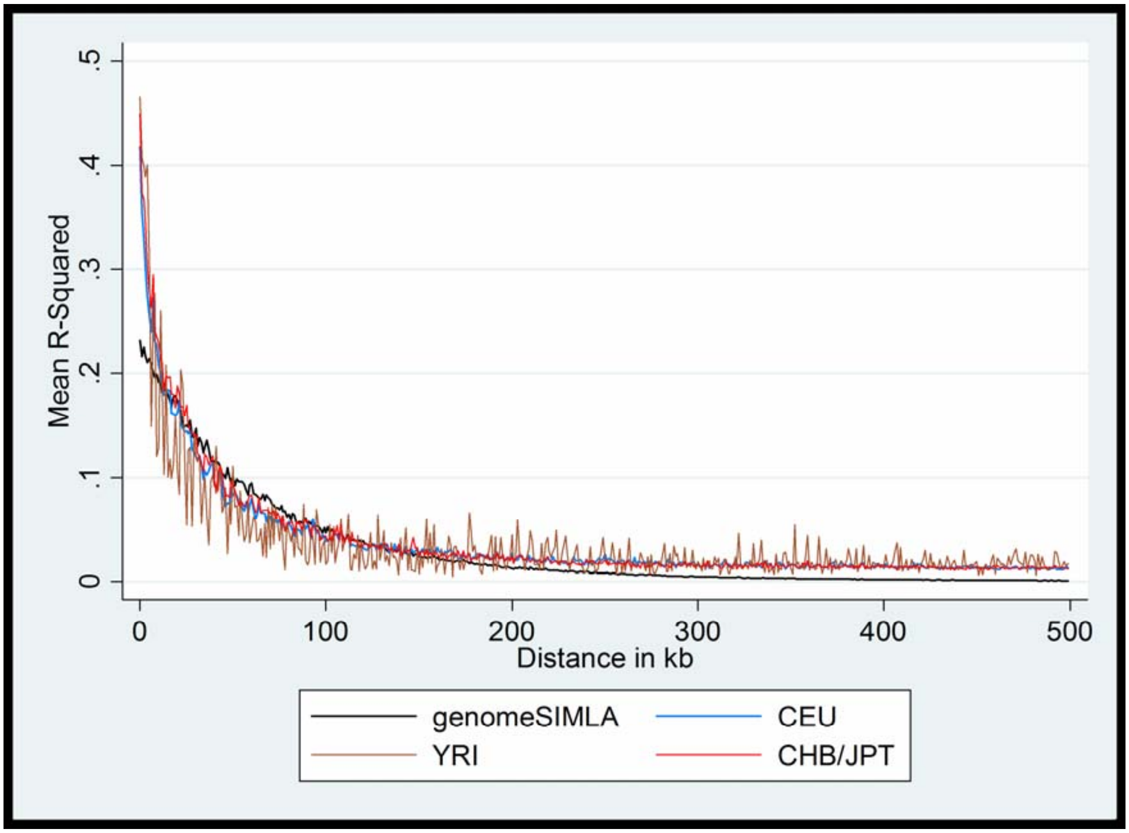


Figure 4-2. Average R^2 by distance (kb) for simulated, CEPH (Caucasian), YRI (Yoruba African), and CHB/JPT (Chinese/Japanese) samples

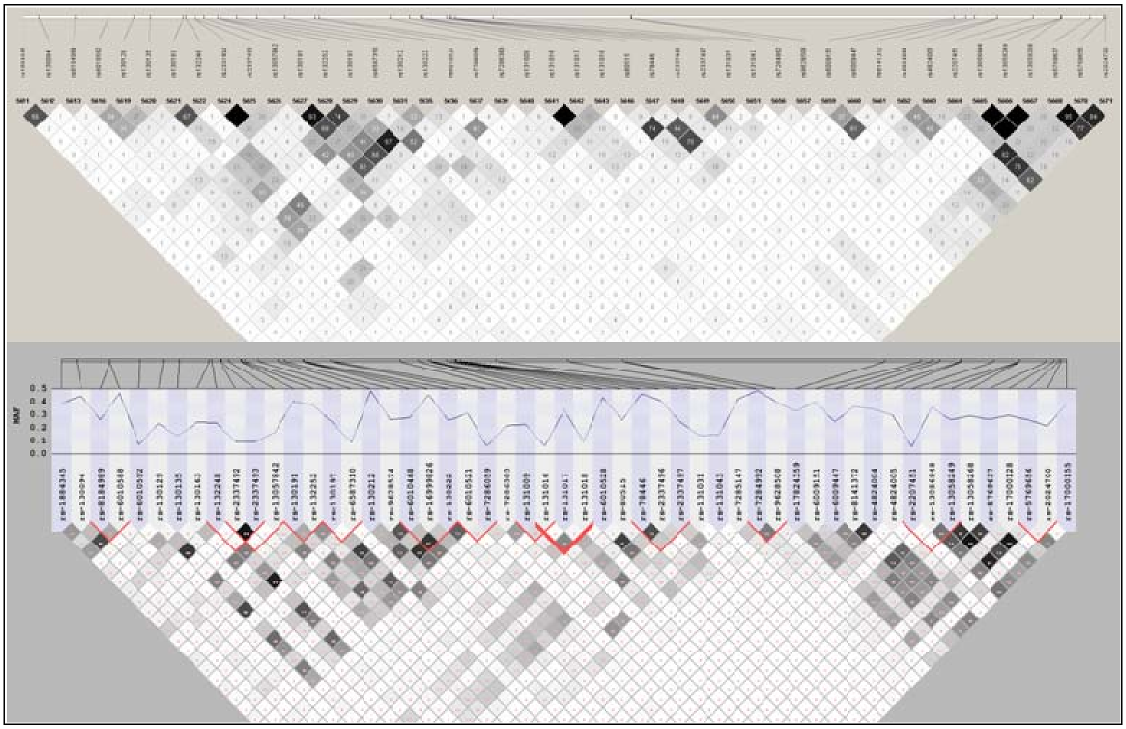


Figure 4-3. Sample of LD from the simulation detailed in Figure 2 of R^2 plots from HapMap CEPH samples (above) and simulated data.

We also measured the time to completion for various size simulations. We examined the markers for the Affymetrix 500K in chromosomes 1 and 22 and the full chip (Table 4-2) for the growth parameters in Figures 4-2 and 4-3. To reduce the time required to scan a growth curve for ideal LD patterns, genomeSIMLA utilizes both sampled and complete LD. When generating sampled LD plots, genomeSIMLA draws LD plots for a small region (1000 SNPs) of each chromosome and limits the participants to a relatively small number (1500).

Table 4-2. Time to completion for pool advancement to 100,000 chromosomes and graphical LD calculation and representation for up to 500,000 SNPs on Dual quad core 2.33 Ghz and 16 GB RAM

Simulation	Processors	LD Calculation	Time
Chr1	1	Sampled	13m 41s
Chr1	1	Complete	88m 45s
Chr1	4	Sampled	5m 41s
Chr1	4	Complete	33m 4s
Chr22	1	Sampled	2m 15s
Chr22	1	Complete	12m 27s
Chr22	4	Sampled	1m 33s
Chr22	4	Complete	4m 30s
500k	4	Sampled	74m 52s
500k	4	Complete	367m 54s
500k	8	Sampled	29m 22s
500k	8	Complete	123m 21s

Discussion

We found that tuning the parameters to emulate the average pattern of correlation in real human populations was difficult. However, some settings we used provided good qualitative fit to the observed real data. Statistical evaluation of these distributions was difficult, since tests on distributions are extremely sensitive and strong ceiling effects were observed. We initialized our chromosome pools with random allele frequency independent data, and only allowed the recombination probabilities to directly mimic those expected from the Kosambi function for the HapMap physical distances. This procedure was a proof of principle that it is neither necessary to directly resample HapMap chromosomes or use computationally inefficient coalescent models to effectively simulate the properties of unobserved samples from real human populations.

One potential reason for the deviations of our simulated data from those observed in the HapMap populations was that genomeSIMLA simulates phased chromosomes with no heterozygote ambiguity. As a result, genomeSIMLA does not employ an Expectation Maximization (EM) algorithm (Excoffier and Slatkin 1995) to phase genotype data. The phased data available from the HapMap is processed from raw genotypes using PHASE (Stephens and Scheet 2005), which is a notably different means of LD calculation. The effects of EM algorithms on the observed average LD by distance when the true LD is known has not been investigated, but will be a topic of further study.

The results we observed here show that genomeSIMLA is an effective platform for simulating large-scale genetic data. Each individual pool was expanded to 100,000 chromosomes before termination, which typically took less than 10 minutes including LD calculation. Additionally, methods other than purely stochastic independent initialization for pools of chromosomes could be used, which could lead to superior data properties and less generations of population growth.

The speed and scale of the genomeSIMLA software is sufficient to provide timely results to investigators conducting power studies for various methods. The software architecture ensures that the user can access all available computational power to do very large whole-genome size studies. The time to completion for various size simulations for single and multiple processors are presented in Table 4-2. Those times include the time required to calculate and provide an interactive graphical interface of LD pictures for locus selection. These times are very fast given the computational task and represent the advanced implementation which is presented here. Demonstrations, manuals, and genomeSIMLA software for Mac, PC, and Linux are available for download at http://chgr.mc.vanderbilt.edu/genomeSIMLA. With this capability, researchers who develop novel methods to analyze genetic data can quickly and accurately estimate the performance and sampling properties of those methods.

CHAPTER V

A GENERAL FRAMEWORK FOR FORMAL TESTS OF INTERACTION AFTER EXHAUSTIVE SEARCH METHODS WITH PRACTICAL APPLICATIONS TO MDR-PDT

Overview

As genetic epidemiology looks beyond mapping single disease susceptibility loci, interest in detecting epistatic interactions between genes has grown. The dimensionality and comparisons required to search the epistatic space and the inference for a significant result pose challenges for testing epistatic disease models.

The Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) was developed to test for multilocus models in pedigree data. In the present study we rigorously tested MDR-PDT with new cross-validation (CV) and omnibus model selection algorithms by simulating a range of heritabilities, odds ratios, minor allele frequencies, and numbers of interacting loci. Additionally, given that the permutation-based hypothesis test of the MDR-PDT does not evaluate effect modification across genotypes and that this property might inflate the Type I error rate for the null hypothesis of no interaction, we chose to implement a regression-based permutation test.

We found that MDR-PDT performs similarly with 5 and 10 fold CV. Also, the sensitivity did not fall a large amount when MDR-PDT selected best models using the omnibus approach compared to the n-locus approach. We also demonstrate that fitting a regression model on the same data as analyzed by MDR-PDT is a biased procedure and is not a valid test of interaction. The regression-based permutation test implemented here conducts a valid test of interaction after a search for multilocus models, and can be used with any method which conducts a search to find a multilocus model representing an interaction.

Introduction

Gene-gene (epistasis) and gene-environment interactions are long-recognized phenomena that have become a focus of the method development arm of genetic epidemiology. Descriptions of interaction have been developing throughout the histories of statistics and genetics. Interaction or effect modification was defined by R.A. Fisher as the deviation from additivity in a mathematical model (Fisher 1918). Sewell Wright related this concept to genetics by stating that the relationship between genotype and phenotype is dependent on dynamic interactive networks of genes and environmental factors (Wright 1932).

The implication of Fisher's view is that a statistically epistatic relationship between multilocus genotypes and phenotype featuring effect modification is not detectable with a linear combination of variables. This describes the regression model with no interaction term. The regression term for an interaction specifically tests the null hypothesis that simultaneous exposure to multiple factors does not significantly increase the link function beyond the sum of average changes in the link function for exposure to each factor. Statistical epistasis is thereby a population-level event measured on average outcomes among samples, as opposed to biological epistasis (Bateson 1909), which is the result of the physical interactions of biomolecules in individuals (Moore and Williams 2002). It has been postulated that variation in biologically epistatic processes among individuals within populations yields epistatic statistical signals (Moore and Williams 2005).

Epistasis can be easily observed at the population level in *S. cerevisiae* (Tong et al. 2001), and these observations have elucidated novel gene functions influencing extensively studied biological processes (Tong et al. 2004; Schuldiner et al. 2005). Epistasis has also been observed in *Drosophila Mercatorum* for the abnormal abdomen phenotype (DeSalle and Templeton 1986; Hollocher et al. 1992; Hollocher and Templeton 1994) as listed in (Templeton 2000). Animal models of Hirschsprung disease in mice have demonstrated effect modification and epistasis among genes (Cantrell et al. 2004). In humans, explorations of genetic statistical epistasis have been performed for several diverse traits with various methods, such as adverse drug reactions (Wilke, Moore, and Burmester 2005), Alzheimer disease (Martin et al. 2006), asthma (Chan et al. 2006; Millstein et al. 2006), autism (Ma et al. 2005; Ashley-Koch et al. 2006), bladder cancer (Andrew et al. 2006), schizophrenia (Morris et al. 2007; Qin et al. 2005), multiple sclerosis (Brassat et al. 2006), irinotecan metabolism (Culverhouse, Klein, and Shannon 2004) and many others. These studies demonstrate that variation in genes that function together in and across biochemical pathways can have unforeseen effects on phenotypes, and that exhaustive searches through these spaces can reveal models that predict gene functions and phenotype outcomes (Segre et al. 2005). As individual susceptibility loci with main effects on various traits are discovered, large amounts of trait variance remain unexplained for many diseases; suggesting statistical epistasis may play a role in many disease phenotypes.

From a method development perspective, the difficulties encountered when searching for replicable statistical epistasis in human populations are essentially two-fold. The first difficulty encountered is multiple comparisons, due to the extremely large space that must be searched to exhaustively catalog all possible interactions for a set of variables. This problem is usually solved with permutation testing, where the entire search is performed on randomized data many times to estimate the distribution of statistics under the null hypothesis of no association after a same-size search. The other main difficulty is the curse of dimensionality (Bellman 1961), and essentially refers to the loss of precision due to excessive data subdivision. In regression, when modeling high-order interactions with insufficient data the curse of dimensionality causes coefficient and standard error inflation, leading to inaccurate estimates of effect sizes and low precision (Hosmer and Lemeshow 2000).

Traditional parametric statistics such as logistic regression (Hosmer and Lemeshow 2000) have limited utility when searching for interactive effects in a large search space, whether searching through genetic loci (Templeton 2000) or environmental exposures (Schlicting and Pigliucci 1998). These methods do not natively adjust for many comparisons or accommodate scenarios with high dimensionality. As the number of predictor variables increases, the number of comparisons necessary to explore the entire statistically epistatic search space expands rapidly.

As discussed in Chapter II, several derivations of the logistic regression procedure have been developed to detect statistical epistasis such as the focused interaction testing framework (FITF) (Millstein et al. 2006), and stepwise logistic regression (Marchini, Donnelly, and Cardon 2005); however, purely interactive effects or those with weak main effects are not likely to be detected, and higher order interactions still suffer from the curse of dimensionality.

Many solutions to the curse of dimensionality have been proposed using data mining and machine learning approaches such as pattern recognition and data reduction. Pattern recognition describes the use of an algorithm to discover data patterns that discern groups from fully dimensionalized data (See Chapter II). These ideas are implemented in MDR (Hahn, Ritchie, and Moore 2003; Hahn and Moore 2004; Moore 2004; Moore et al. 2006; Moore 2007; Ritchie et al. 2001; Ritchie, Hahn, and Moore 2003) for discrete outcomes. MDR in particular maps data from several variables with many dimensions to a single binary variable which can then be used in any other context, such as regression or contingency table statistics, for association analysis.

The data reduction step of MDR is amenable to decreasing the dimensionality of genetic variables for logistic regression or any other subsequent method for modeling the relationship between multilocus genotypes and phenotype (Velez et al. 2007; Moore et al. 2006; Martin et al. 2006). Logistic regression assumes linear relationships between exposures and outcomes, and is also iterative for maximum likelihood estimates, resulting in insensitivity to detect some disease models and computational inefficiency. Therefore, MDR is optimized for sensitivity and computational speed, acknowledging the difficulties inherent in exhaustive searches for epistasis.

The MDR-PDT was developed to perform exhaustive searches for epistasis in pedigree data. As described in Chapter II, MDR-PDT uses all informative offspring when calculating association statistics, and has superior power in family data than MDR.

The initial debut of MDR-PDT did not feature cross validation (CV) or a means to select from among orders of models, two features of MDR. Cross validation is a feature of the original MDR algorithm that is used to find consistent signals in the data and select a single best model from among orders of model. This procedure has been shown to be effective in simulation studies at multiple levels of CV (Motsinger and Ritchie 2006). Here we present an algorithm to perform CV in family data and select a best model from among models of various sizes.

Another element of analyzing real datasets with MDR or MDR-PDT is the structure of the hypothesis test performed on the best models observed by the algorithms. The null hypothesis of the permutation test is no association signal from the loci, as defined by the ability of the high-low risk variable to identify variation in penetrances across multilocus genotypes. MDR and MDR-PDT make no assumptions about the statistical relationship between loci and traits or about the mode of inheritance at individual loci (Ritchie et al. 2001; Ritchie, Hahn, and Moore 2003). As a result, this null hypothesis is very general and is capable of detecting disease models featuring nonlinear effects and purely epistatic effects (Moore 2004). However, this strength can also be a weakness, since this test will not distinguish between groups of noninteracting main effects and interactions among variables with or without main effects. Thus the interpretation of a significant result for an MDR or MDR-PDT model is not clear, and the investigator is left with a constellation of loci and no knowledge about the relationship among them, other than they are associated as a group with the trait. Additionally, we show here that hypothesis tests with a null of no interaction performed post hoc using a regression procedure on a multilocus model on the same data where an MDR or MDR-

PDT analysis was conducted are substantially biased toward the alternate hypothesis. Due to this lack of specificity in the MDR and MDR-PDT procedures and bias in post hoc estimation techniques, a valid test of interaction for an MDR or MDR-PDT model is presented here.

Materials and methods

MDR-PDT

The MDR-PDT procedure and hypothesis test is described in Chapter II (Figure 2-1).

Cross validation

A notable difference between MDR and MDR-PDT from the (Martin et al. 2006) simulation studies is the ability of MDR to choose a single best model when several orders of model, for instance 2-locus and 3-locus, have been considered. In practice this is a very important capability since it allows much larger searches to be performed under a single hypothesis test, thus increasing sensitivity by removing the need for multiple testing corrections across orders of model post-hoc. This capability from MDR is based on the CV procedure.

To implement CV in MDR-PDT, the consideration of how to evenly split the data must be taken. In MDR, the data are case-control, with each individual representing an independent observation and proportion of the data available. In MDR, the data are binned into equal-size bins prior to analysis based on counts of cases and controls, with no regard for missing data, so some bins may be unequal splits for some loci. For MDR-PDT the data are independent pedigrees of various structures and sizes, each contributing different amounts of information to the dataset. The units of information which are used by MDR-PDT are discordant sibling pairs (DSPs) and transmitted/Untransmitted (T/UT) pairs, so the quantification of this information is necessary to evenly bin the pedigree data.

Consider a dataset consisting of pedigrees containing extended sibships of arbitrary size. Let x_{ij} be the number of possible DSPs and T/UT pairs from a sibship sharing both parents in a pedigree, where i indexes sibhips i = (1,2, … , n) in a pedigree j, j = (1,2, … , m). x_i will be found in a sibship by ((Affected sibs x Unaffected sibs) + Affected sibs). Thereby, $x_{ij} = \sum x_i$ for full sibships within a pedigree j. This gives the maximum information available to the statistic for that pedigree. Let $X = \sum_j \sum x_{ij}$ for all j pedigrees be the total such pairs of offspring from all families in the dataset. x_{ij}/X gives the proportion of potential observations from each pedigree compared to the proportion of total possible observations for the entire dataset. To perform CV, randomly split the data by putting intact families into k bins, the value of k specified by the user. Let k_i be the proportion of information from the total data for a bin, given by $k_i = \sum(x_{ij}/X)$ for i = (1, 2, … , k). Set a variance threshold V_x for the variance across values of k_i across bins for the split, where the variance will not exceed V_x. The variance for a split into n bins of the dataset will be $V = \Sigma_k(k_i(X) – 1/k(X))^2/k$. Compare V_x to V. If $V_x < V$, reject the split and repeat the procedure 30 times. Continue until $V_x > V$. If no split provides a satisfactory binning of the data, relax V_x or change the number of bins.

Once the data are split into approximately equal parts, an extension allowing best model selection for MDR-PDT is possible. Each CV interval is used as a test set as in MDR to develop a measure of how well a model will predict disease status in independent samples. This procedure also provides a measure of cross-validation consistency (CVC) for each model found. One issue in MDR-PDT that is not a problem for MDR is the fact that the MDR-PDT statistic that is calculated for the best model is not comparable across orders of models. As a result, this statistic cannot be used to determine across orders of model which is the strongest signal. Because of this, two fitness metrics were employed to select best models. Prediction error (PE) and the matched odds ratio (MOR) were both calculated and power calculations for each were performed. The prediction error is defined as the average classification error from test sets during the CV procedure. The matched odds ratio is calculated by pooling the DSPs or T/UT pairs from test set pedigrees and plotting them in a 2x2 table relating the high/low risk variable to status. The ratio of DSPs and T/UT pairs that are correctly classified to those that are incorrectly classified is a matched odds ratio. The average MOR is calculated from test sets across CV intervals. The omnibus procedure is as follows and is illustrated in Figure 5-1.

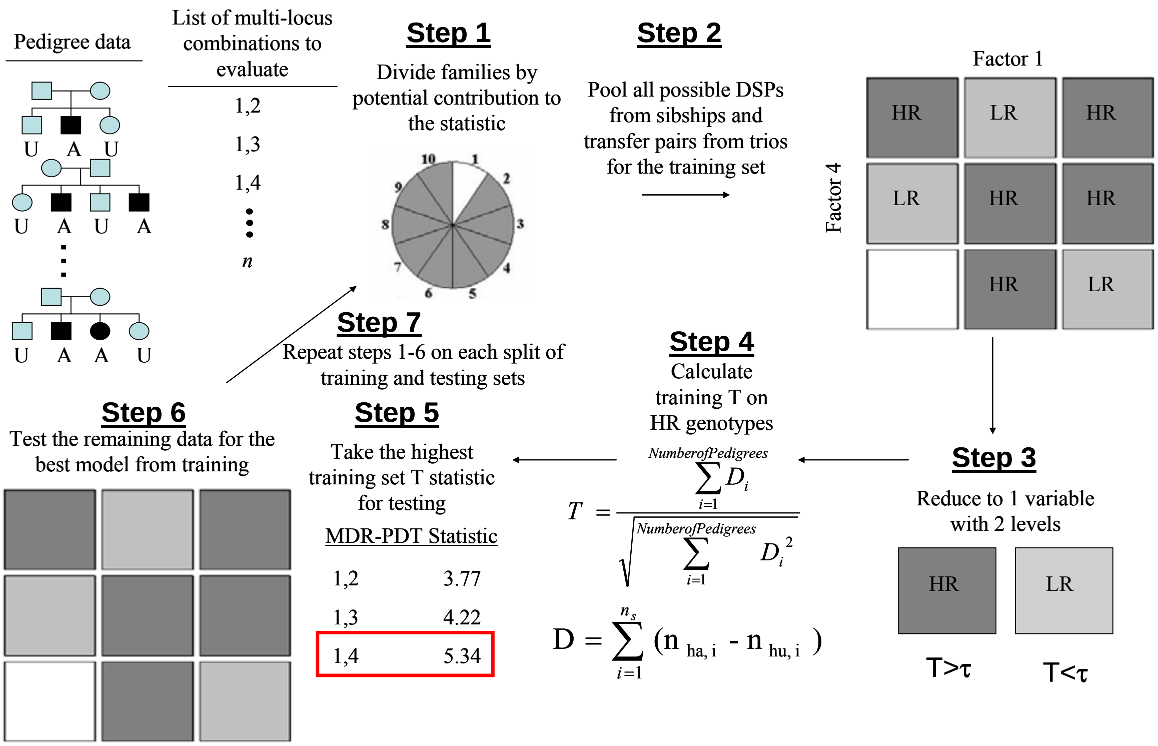

Pedigree data
U A U
U A U A
U A A U
List of multi-locus combinations to evaluate
1,2
1,3
1,4
n
Step 1
Divide families by potential contribution to the statistic
1 2 3 4 5 6 7 8 9 10
Step 2
Pool all possible DSPs from sibships and transfer pairs from trios for the training set
Factor 1
Factor 4
HR LR HR
LR HR HR
HR LR
Step 3
Reduce to 1 variable with 2 levels
HR LR
T>τ T<τ
Step 4
Calculate training T on HR genotypes
$T = \frac{\sum_{i=1}^{NumberofPedigrees} D_i}{\sqrt{\sum_{i=1}^{NumberofPedigrees} D_i^2}}$
$D = \sum_{i=1}^{n_s} (n_{ha,i} - n_{hu,i})$
Step 5
Take the highest training set T statistic for testing
MDR-PDT Statistic
1,2 3.77
1,3 4.22
1,4 5.34
Step 6
Test the remaining data for the best model from training
Step 7
Repeat steps 1-6 on each split of training and testing sets

Figure 5-1. MDR-PDT algorithm with CV

1. Data are split into K equal parts

2. All possible DSPs and T/UT pairs are generated within each sibship (affected times unaffected) and pooled within K-1/K of the data. This is a training set.

3. Each genotype is determined to be high or low risk by comparing the genoPDT statistic (Martin et al. 2003) from the pooled DSPs and T/UT pairs to a threshold τ, such as τ = 0, which indicates positive or negative association with affected status.

4. Statistics for high-risk genotypes are calculated using the MDR-PDT statistic (Martin et al. 2006).

5. The procedure repeats for every combination of loci within the order range specified, calculating an MDR-PDT statistic for each, choosing the largest MDR-PDT statistic from each order as the best model.

6. Matched odds ratios or prediction error is calculated from the testing set for each best model of each order using the high-low risk levels established during training.

7. Steps 1-6 are repeated in the other splits of the data, so that each CV interval is used as a test set. Where the same model is observed in multiple training sets, a measure of CVC is observed. To select the best from among all models found in training, CVC is considered first, then the tiebreaker is decided using the average PE or MOR from test sets.

A permutation test is performed to determine the distribution of the null hypothesis of no association. The result from step 7 is compared to this distribution for significance assessment.

Regression test of interaction

To conduct a formal test of interaction among the variables in a model resulting from an exhaustive search, the size of the search must be accounted for when determining the critical value of the statistic for significance. Otherwise when comparing the statistic to the parametric value for significance, the test is not valid and is strongly biased toward the alternate hypothesis. To accomplish a valid test, a straightforward extension to the MDR-PDT or MDR algorithm is implemented. For simplicity, MDR-PDT will be referred to here, but this method is also applicable to MDR results.

Where a best two through N-locus model is found by the MDR-PDT omnibus algorithm, the genotypes at the model loci are determined to be high or low risk by individual assessment of each model locus by MDR-PDT. This binary coding for genotypes is then used to fit saturated full and reduced conditional logistic regression equations with multiplicative interaction terms assessing effect modification for simultaneous exposure to high-risk genotypes for each model locus. The interaction term corresponding to the best model from the MDR-PDT search is removed for the reduced model, leaving any nested interaction terms in place, and the likelihoods of each model are recorded. A likelihood ratio statistic is calculated for this interaction term. Then the data are permuted as usual, and MDR-PDT chooses the best two through N-locus model for each permutation. The regressions are fit as in the original data and the resulting likelihood ratio statistics from each permutation are sorted from largest to smallest. The statistic from the real data is then compared to this distribution for significance assessment.

Simulations

Power and Type I error of the MDR-PDT without CV has been measured in (Martin et al. 2006). GenomeSIMLA (Edwards et al. 2008a) software has been developed by merging the software packages genomeSIM (Dudek et al. 2006) and SIMLA (Bass, Martin, and Hauser 2004; Schmidt et al. 2004) (Chapter IV) to simulate pedigree data with purely epistatic penetrance tables.

Epistatic models were simulated with a genetic algorithm, modified from (Moore et al. 2004), for 2 and 3 loci, minor allele frequency of 0.2 or 0.4, and heritability of 0.005, 0.01 0.03, 0.05 or 0.1 There were a total of 20 genetic models, each of which were simulated as 100 20-locus datasets with 100, 500, and 1000 pedigrees (Table 5-1). The odds ratio in this table is the average ratio of odds between high-risk and low-risk cells, where high-risk cells are those multilocus genotypes for which the penetrance equals or exceeds the prevalence for the model. All penetrance tables used to simulate genetic data are presented in the Appendix.

Pairs of noninteracting model loci were simulated in 500 and 2000 20-marker DSP pedigrees. The effect sizes of the model loci were simulated at relative risks 1.5, 2, 4, and 6. The model loci were independent and had a dominant model for the minor allele with relative frequency of 0.2. All loci in the simulations were independent to provide conservative estimates of power due to increased data noise. It is expected that data with extensive correlation among non-model loci would provide fewer spurious signals relative to independent loci since correlated loci would tend to behave similarly to one another in epistatic models, thereby effectively reducing the number of independent non-model variables. This is analogous to the principles underlying the multiple testing correction method of (Nyholt 2004)

Table 5-1. Models examined in the simulation study.

Loci	Minor allele frequency	Heritability	Odds Ratio
2	0.2	0.005	1.1
2	0.2	0.01	1.26
2	0.2	0.03	1.53
2	0.2	0.048	1.79
2	0.2	0.09	3
2	0.4	0.005	1.15
2	0.4	0.01	1.28
2	0.4	0.03	1.56
2	0.4	0.05	1.79
2	0.4	0.1	2.85
3	0.2	0.005	1.19
3	0.2	0.01	1.36
3	0.2	0.03	1.58
3	0.2	0.05	2.1
3	0.2	0.1	3.2
3	0.4	0.005	1.21
3	0.4	0.01	1.32
3	0.4	0.03	1.52
3	0.4	0.05	2.23
3	0.4	0.12	3.5

Type I error of Conditional Logistic Regression with correction for sharing among multiple affected siblings in regions of linkage (Siegmund et al. 2000), MDR-PDT with CV, with and without the LR procedure were estimated by simulating 1000 20-marker 500-DSP datasets with no penetrance function specified and random minor allele frequency.

To estimate the type I error rate of regression following an MDR-PDT search, the best 2-locus model was chosen from each null dataset using MDR-PDT. Two loci were also chosen at random from each dataset. The genotypes at each model locus were then classified as high or low-risk using MDR-PDT. This coding was then used in the subsequent regression for each model, where a likelihood ratio statistic was calculated for each interaction term and compared to a chi-squared, one degree of freedom distribution for significance assessment.

To estimate the Type I error rate for MDR-PDT with CV and the regression procedure following an MDR-PDT search, each of the 1000 null datasets were permuted 100 times to determine whether the best model from the original null dataset exceeded the 5th largest value from the 100 permutations, corresponding with an alpha of 0.05. Where a null dataset yielded a statistic that equaled or exceeded the 5th largest permutation, a Type I error occurred and was scored.

Results

Type I error of regression after MDR-PDT

The Type I error rate for the experimental scenario where random pairs of loci were chosen and followed by fitting full and reduced regression models with and without the interaction term was 0.048. This is very close to the nominal rate and serves as a negative control to demonstrate that there are not other biases present in the regression procedure. The Type I error rate of the likelihood ratio statistic for the regression interaction term corresponding to MDR-PDT two-locus models when compared to chi-squared with one degree of freedom was 0.39. Therefore, testing effect modification using logistic regression in the same dataset where an MDR-PDT result was found is a biased procedure. We later show that this bias can be remedied by permutation testing the interaction term statistic (shown below).

Type I error of MDR-PDT in the presence of independent main effects

When strong main effects are present in a dataset, MDR-PDT might find, test, and reject the null hypothesis for models consisting of these loci. MDR-PDT does not recognize that these models do not represent effect modification. If the null is rejected, these findings are type I errors with regard to the null hypothesis of no interaction. However, the null hypothesis of MDR-PDT is a general null of no association, leading to rejection for such scenarios. This situation is more severe as effect and sample size grow (Tables 5-2a, b). The false positive rate when evaluating these models for the regression extension in the absence of effect modification is zero, demonstrating that the specificity, defined as 1-(false positive rate) is extremely superior to that of the conventional permutation test. This evaluation was performed as a power study and not a type I error study, meaning that the power reported is the power to detect the model loci as the best model and reject the null from a permutation test, rather than the rate at which any model might reject the null. Therefore, the regression-based test of interaction remedies this problem at the testing stage of the algorithm.

Table 5-2a. Type I errors with regard to the null hypothesis of no interaction occur in MDR-type algorithms for 500 DSP families.

Relative Risk	MDR-PDT power prediction error	MDR-PDT power matched odds ratio	Regression power prediction error	Regression power matched odds ratio
1.5	0	0	0	0
2	7	0	0	0
4	21	2	0	0
6	29	4	0	0

Table 5-2b. Type I errors with regard to the null hypothesis of no interaction occur in MDR-type algorithms for 2000 DSP families.

Relative Risk	MDR-PDT power prediction error	MDR-PDT power matched odds ratio	Regression power prediction error	Regression power matched odds ratio
1.5	4	0	0	0
2	17	1	0	0
4	54	24	0	0
6	58	50	0	0

Type I error of the regression permutation test

The regression-based permutation test was conducted in 1000 500-DSP datasets with no penetrance function. One hundred permutations were performed for each null dataset due to long computation times. The Type I error of the procedure was 0.058 at an alpha rate of 0.05. This value was not significantly different from 0.05. The Type I error rate for the MDR-PDT with CV alone was 0.052.

MDR-PDT N-locus vs. omnibus

These results show that for a variety of models, the CV procedure and omnibus model selection criteria, as described in Figure 5-1, function well (Figures 5-2a-e). The sensitivity of five and ten-fold CV are very similar across all the simulated scenarios. Also, the sensitivity of the PE and MOR metrics were very similar. Compared to the n-locus search, where only interactions of the order present in the simulated model were sought, the n-locus with CV performed almost as well as or better as n-locus searches without CV. The omnibus search, where two and three-locus models were examined, tended to lose some sensitivity; however, this can be explained by the larger number of comparisons performed for those searches.

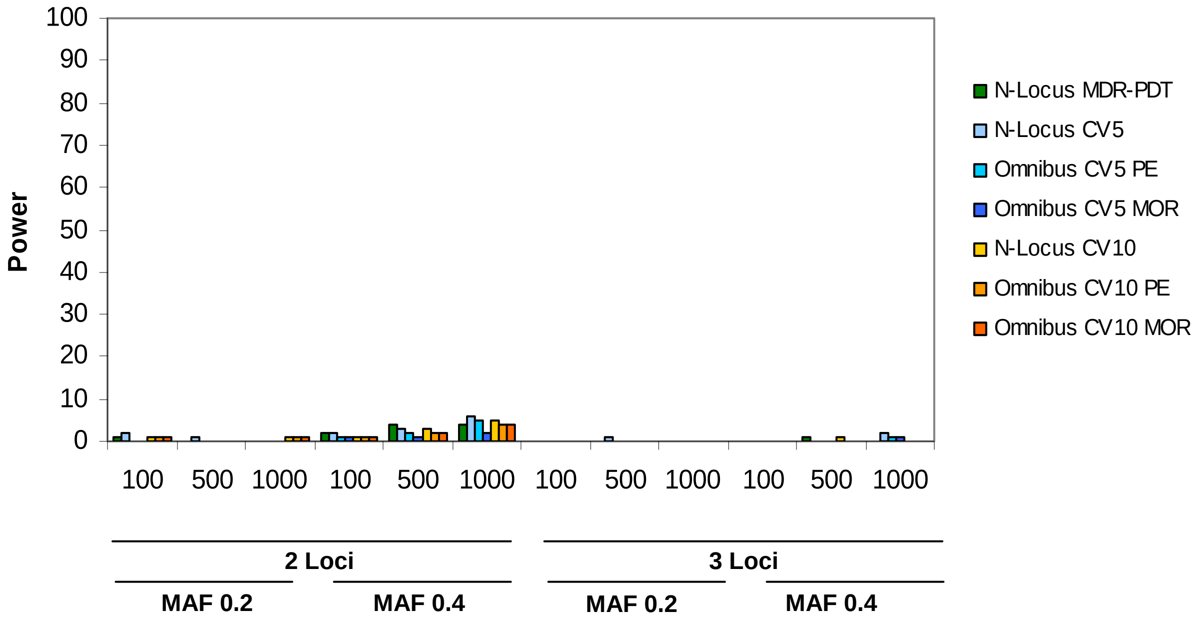


Figure 5-2a. MDR-PDT with N-locus searches with and without cross-validation versus the omnibus procedure for broad-sense heritability 0.005.

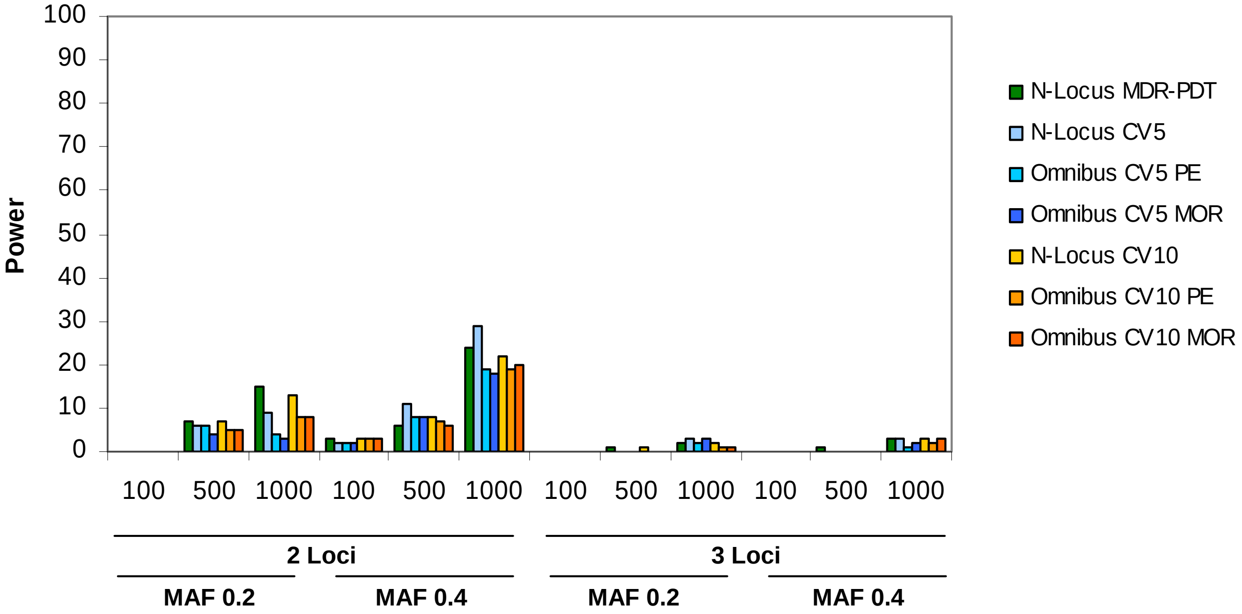


Figure 5-2b. MDR-PDT with N-locus searches with and without cross-validation versus the omnibus procedure for broad-sense heritability 0.01.

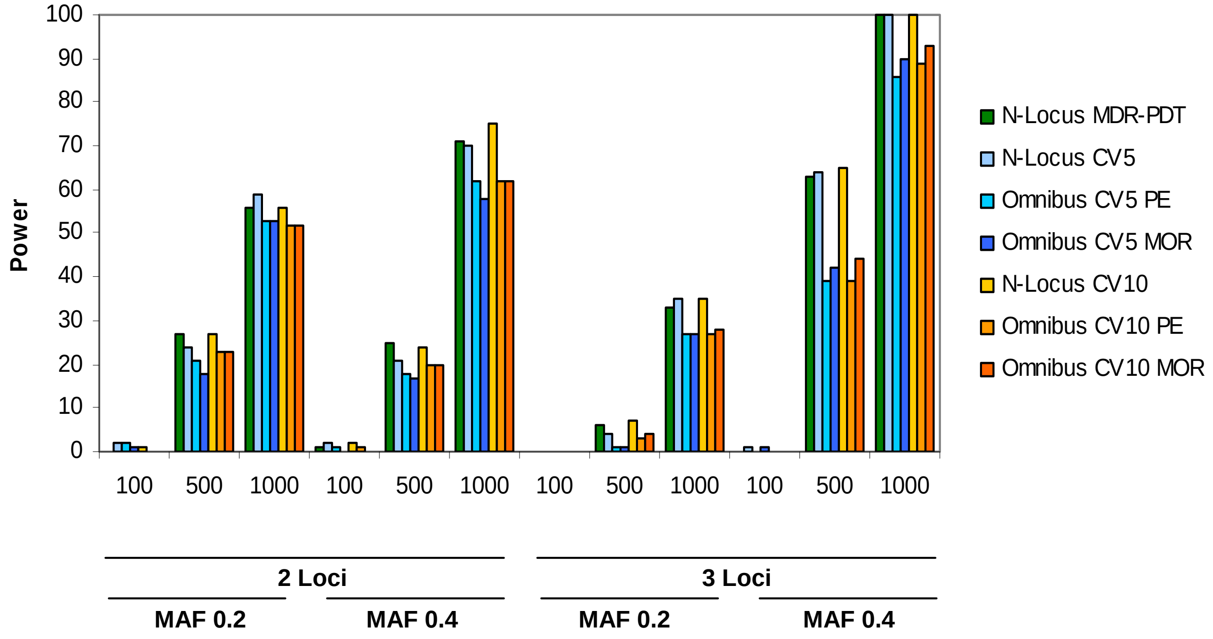


Figure 5-2c. MDR-PDT with N-locus searches with and without cross-validation versus the omnibus procedure for broad-sense heritability 0.03.

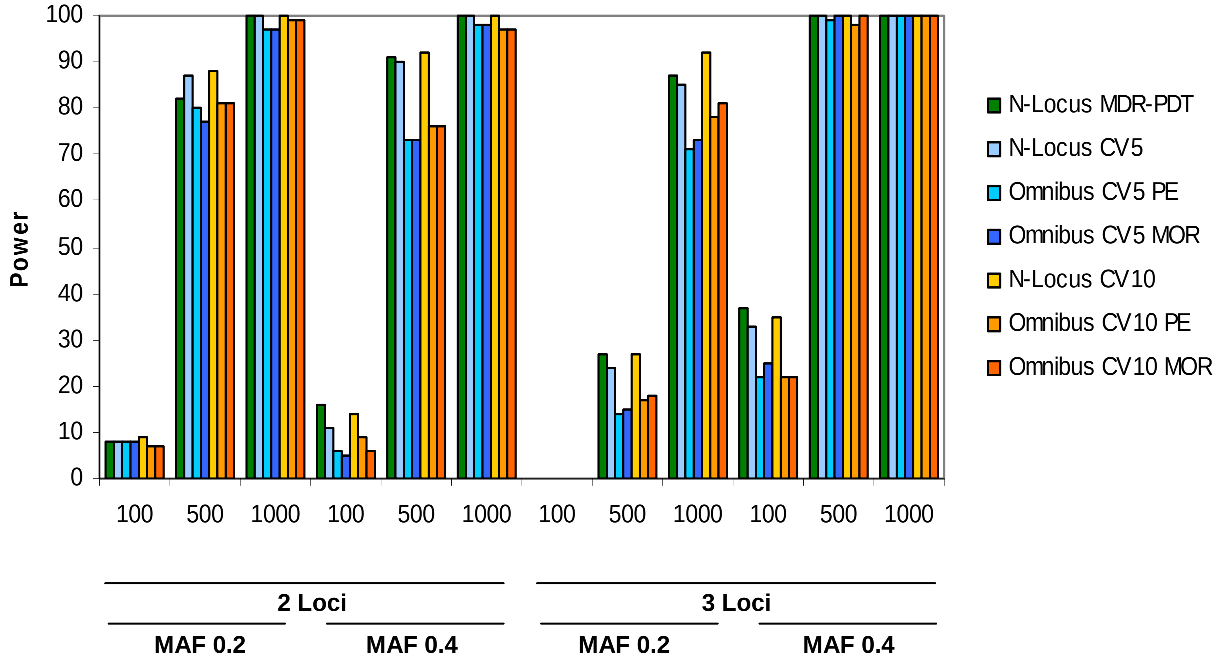


Figure 5-2d. MDR-PDT with N-locus searches with and without cross-validation versus the omnibus procedure for broad-sense heritability 0.05.

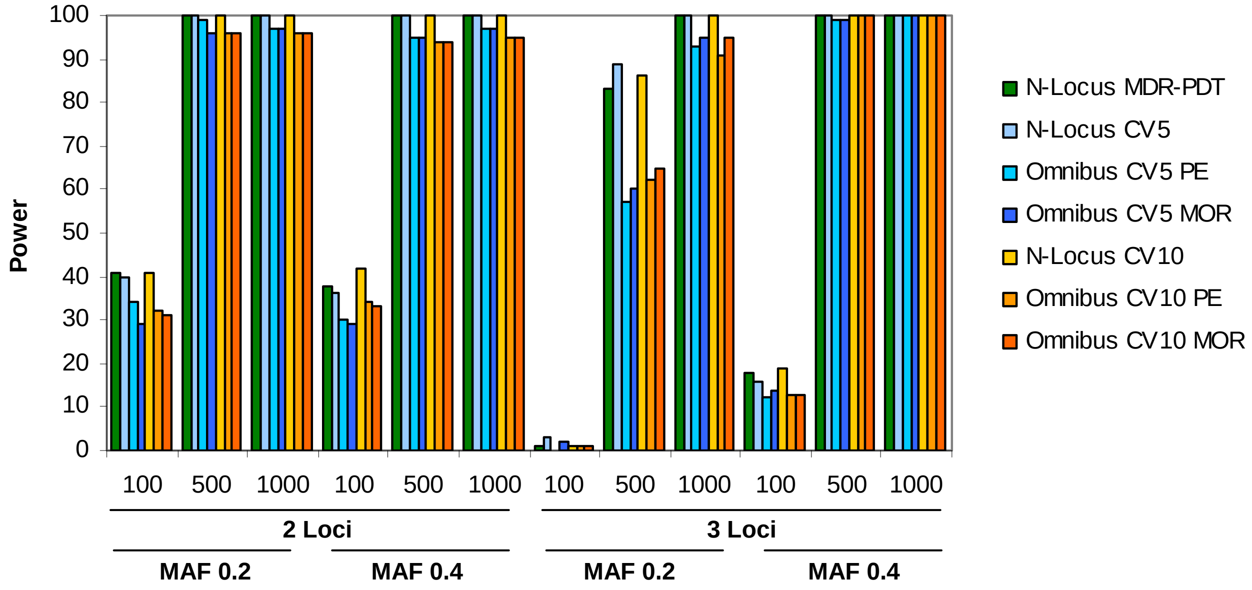


Figure 5-2e. MDR-PDT with N-locus searches with and without cross-validation versus the omnibus procedure for broad-sense heritability 0.1.

MDR-PDT omnibus permutation test vs. regression permutation test

These experiments compare the power of permutation testing of either the PE or MOR fitness metrics after CV with the MDR-PDT omnibus procedure with model selection by either PE or MOR followed by the regression-based permutation test. Only the models from the N-locus vs. omnibus experiments with broad-sense heritabilities of 0.03 or larger were tested. The results from those experiments are presented in Figures 5-3a and 5-3b. These results show that the regression-based permutation test is less powerful than the MOR or PE fitness functions. This loss of power was anticipated, since another level of model evaluation was added to the algorithm with more specificity than the previous metrics. For some models, the regression test is competitive, but never more powerful. It is also notable for having power in multilocus models displaying no marginal main effect.

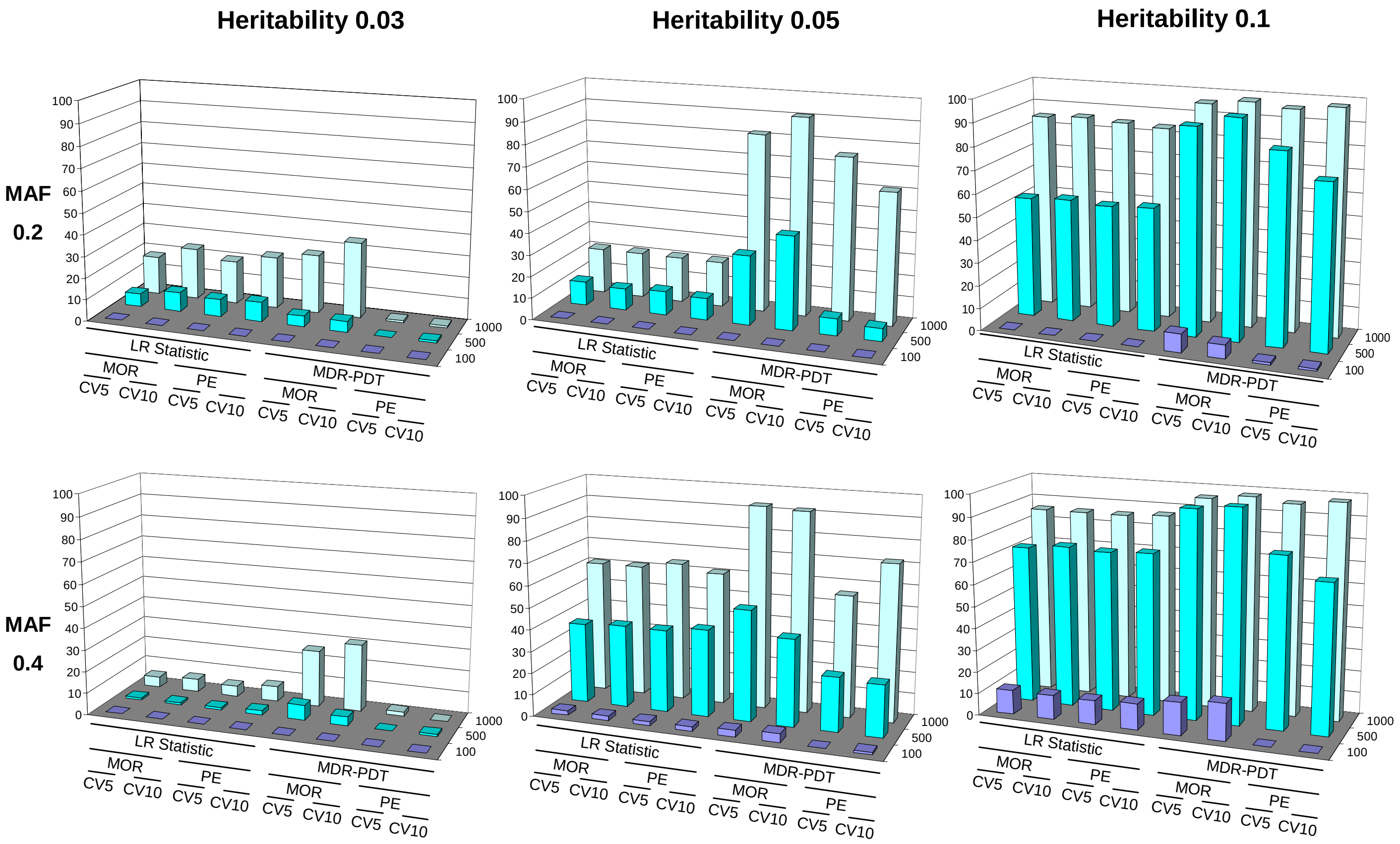


Figure 5-3a. Power of MDR-PDT with either PE or MOR with or without the regression extension to detect simulated two-locus models

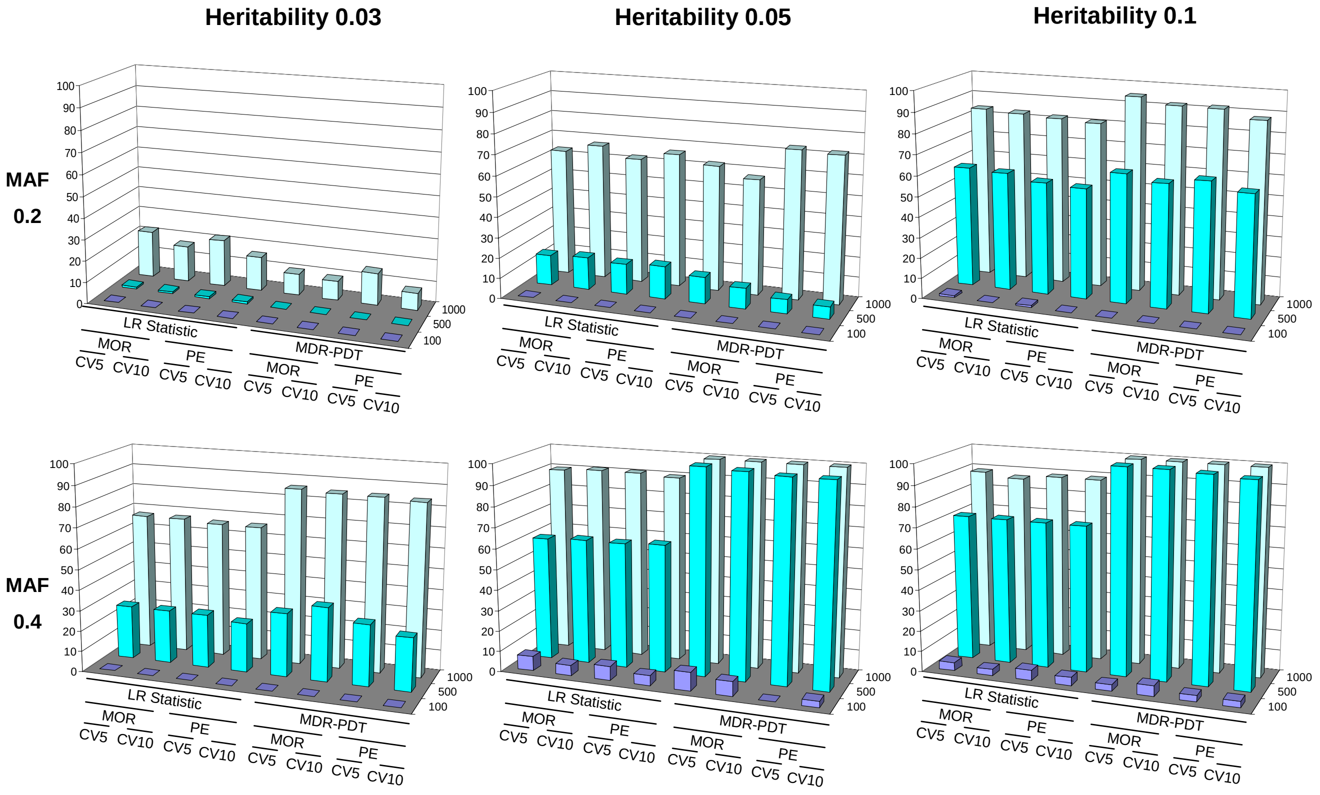


Figure 5-3b. Power of MDR-PDT with either PE or MOR with or without the regression extension to detect simulated three-locus models

Discussion

We have introduced three extensions to the MDR-PDT: an algorithm for binning families evenly for CV, a means of selecting from among orders of multilocus models, and a test of effect modification which has much higher specificity than the previous method of hypothesis testing.

In general, the regression-based testing approach is less powerful than the MDR-PDT omnibus permutation test. However, the power for the MDR-PDT regression test is reasonable considering the many sources of error which may lead to incorrect inferences when using methods that search for interactions. It is more likely in real data that a result that rejects the null hypothesis of the regression-based test will replicate in an independent sample with methods looking for interactions than a result which rejects the general null of no association. Knowing the relationships between the variables in a model and the trait of interest is crucial to prediction and understanding the roles of factors in complex disease. Therefore, we advocate tests that provide high specificity and interpretability, at the expense of some statistical power.

We also investigated the bias introduced in parametric statistics when tests are performed for MDR-PDT models in the same data as the models were found. We found that such procedures are very biased to the alternate hypothesis and lead to many Type I errors. The distribution of the null must be adjusted for the size of the search conducted for the test to be valid; otherwise strong bias is introduced into results.

We developed a new algorithm for splitting families into CV intervals, and a new means of selecting best models from among several orders of model. We showed these methods are effective in simulated data and do not greatly decrease the sensitivity of

MDR-PDT. This approach is philosophically identical to the original MDR algorithm. We also developed a valid test of the null hypothesis of no interaction, and showed it has reasonable performance in scenarios where there are negligible main effects. It is expected this method would perform better where the interacting main effects were not very small, since the method uses the binary risk variable from MDR-PDT at a single locus to encode genotypes for regression. The primary method for searching epistatic spaces with regression, FITF (Millstein et al. 2006), explicitly requires that main effects be present in multilocus models to be detected. We do not have this constraint, and so a broader class of models may be detected. Additionally, use of MDR-PDT to constructively induct model locus genotypes to the high and low-risk variable for regression substantially reduces the dimensionality of modeling from 3^n to 2^n and the sparseness that arises due to the curse of dimensionality.

Regression in general offers a flexible framework for testing generalized associations between variables. Part of the strength of the regression modeling approach is the specificity with which hypotheses may be tested. However, in the context of modeling interactions from a large space of possible multilocus models, this can also be a weakness. The possible ways to model interactions, encode genotypes, and correct for multiple comparisons make regression cumbersome in epistasis searches. Here, we offer a nonparametric framework for detecting multilocus models, encoding genotypes with constructive induction, specifically modeling interactions, and adjusting null distributions of interaction test statistics for the size of the search conducted.

Some future directions of this work will include extending this test of effect modification to case-control data for MDR. Additionally, more sensitive methods for interaction detection than regression, such as generalized estimating equations (Liang and Zeger 1986; Zeger and Liang 1986), which may improve sensitivity while preserving specificity (Hancock et al. 2007). We also will incorporate the ability to include covariates in the regression models to adjust for potential confounding.

This approach is very flexible, and could be adapted to any method searching for epistasis using a permutation test. For instance, one could fit linear regression models for RPM (Culverhouse, Klein, and Shannon 2004) multilocus models for quantitative traits. The approach might be applied to exotic computational methods such as genetic programming neural networks (Motsinger et al. 2006), which use computer learning and evolution principals to search for models. Regardless of the means to search for interactions and the test used to provide a specific test of the null hypothesis of no effect modification across genotypes, this framework incorporates the qualities of methods designed to accommodate the curse of dimensionality and multiple comparisons with the specificity of parametric modeling methods.

CHAPTER VI

AN ASSOCIATION ANALYSIS OF ALZHEIMER CANDIDATE GENES REVEALS A RISK HAPLOTYPE IN *ACE* AND MULTILOCUS ASSOCIATION BETWEEN *ACE, A2M* AND *LRRTM3*

Overview

Alzheimer 's disease (AD) is the most common form of progressive dementia in the elderly. It is a neurodegenerative disorder characterized by the neuropathologic findings of intracellular neurofibrillary tangles and extracellular amyloid plaques that accumulate in vulnerable brain regions. AD etiology has been studied by many groups, but since the discovery of the *APOE* ε4 allele, no further genetic variation has been mapped conclusively with the late-onset form of the disease. In this study, we examined genetic association with late-onset Alzheimer's susceptibility in 738 Caucasian families and an independent case-control dataset exploring 11 candidate genes. In addition to tests for main effects and haplotype analyses, the Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) was used to search for single-locus effects and 2-locus and 3-locus gene-gene interactions associated with AD in the family data. We observed significant haplotype effects in *ACE* in both family and case-control samples. *ACE* was also part of a significant 2-locus and 3-locus MDR-PDT joint effects model with Alpha-2-Macroglobulin (*A2M*), which mediates the clearance of Aβ, and Leucine-Rich Repeat Transmembrane 3 (*LRRTM3*), a nested gene in Alpha-3 Catenin (*CTNNA3*) that binds Presenilin 1. These genes are related to amyloid beta clearance; thus this constellation of effects might constitute an axis of susceptibility for late-onset

AD. This consistent result between independent data sets of families and unrelated cases and controls is strong evidence in favor of *ACE* as a susceptibility locus for AD.

Introduction

Alzheimer's disease (AD) (OMIM 104300, 104310) is the most common form of progressive dementia in the elderly. More than 4 million Americans are afflicted with this debilitating disorder and many studies have been conducted to elucidate an etiology. The discovery of the *APOE ε4* risk factor demonstrated that genetic analysis can be successful in complex disease research. However, between 42% and 68% of cases do not carry the *ε4* allele (Henderson et al. 1995; Lucotte et al. 1994; Ritchie et al. 1996; Hardy, Myers, and Wavrant-De 2004). Additional environmental and genetic factors likely play a role in Alzheimer's susceptibility. Some of these putative factors are explored in the current study. Genes known to interact with presenilins, amyloid beta (Aβ) clearance, and cardiovascular disease are surveyed due to their known or hypothesized biological relevance.

One gene that has been associated with the early onset form of AD is amyloid precursor protein. Duplications in this gene have been associated with the disease (Rovelet-Lecrux et al. 2006). Amyloid precursor proteins and presenilins influence autosomal dominant, early-onset disease due to altered Amyloid Protein Precursor processing, leading to Aβ deposition (Goate 2006; Hardy 1997; Levy-Lahad et al. 1995; Rogaev et al. 1995; Sherrington et al. 1995). Variation in these genes has not been shown to influence late-onset susceptibility, which is far more prevalent. They do, however, provide insight into the pathophysiology of the disorder. The *ε4* allele of *APOE* causes increased risk of AD, while the *ε2* allele is protective (Chartier-Harlin et al. 1994; Corder et al. 1993). The mechanism by which *APOE ε4* influences risk of AD is unknown, but is likely related to Aβ processing (Bales et al. 1999).

Fourteen years after the discovery of *APOE*, single-locus approaches by many groups have not discovered any additional candidates consistently associating with late-onset AD. This inability to unravel the mechanism underlying the trait, given steadily increasing ascertainment and genotyping capability, illustrates the difficulty of finding AD genes.

One of these candidate genes is angiotensin converting enzyme (*ACE*) (Alvarez et al. 1999; Kehoe et al. 1999; Scacchi et al. 1998). *ACE* functions in several biological systems that may lead to AD, such as the cardiovascular and Aβ pathways (Hu et al. 2001). *ACE* is part of the renin-angiotensin system regulating homeostasis (Reid 1992). *ACE* plasma concentrations are increased in persons bearing a 287bp deletion in intron 16 of the gene (Rigat et al. 1990). This *ACE* I/D is also a risk locus for cardiovascular disease (Hessner et al. 2001; Malik et al. 1997), which may share some common etiological factors with AD (Breteler et al. 1994; Hofman et al. 1997). *ACE* may also promote Aβ degradation, providing a more direct link to AD (Hemming and Selkoe 2005; Hu et al. 2001).

Some other genes where previous associations have been observed but inconsistently replicated are alpha-2-macroglobulin (*A2M*) (Blacker et al. 1998) and alpha-T-catenin (*CTNNA3*), and a nested gene leucine-rich repeat transmembrane protein 3 (*LRRTM3*) (Ertekin-Taner et al. 2003; Martin et al. 2005). *A2M* has protease inhibitor activity (Bergqvist and Nilsson 1979), and mediates the clearance of Aβ deposits. The marker rs3832852, a 5bp insertion/deletion in intron 18, has been studied in many samples and demonstrates a significant main effect in our analyses (Supplementary Table 1). *CTNNA3* binds to beta catenin which then interacts with presenilin 1, variants in which are associated with early-onset AD. There have also been previous reports of linkage to late-onset AD at the *CTNNA3*/*LRRTM3* locus, in addition to *CTNNA3* association with Aβ-42 levels in late-onset AD families (Ertekin-Taner et al. 2000; Ertekin-Taner et al. 2003). *ACE*, *A2M* and *LRRTM3* were also found as the best multilocus model by Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (Martin et al. 2006) analysis of our family sample.

The study of gene-gene interactions has become a common element of current AD candidate gene research. The presence of gene-gene interactions explaining AD risk could explain why the search for AD loci since the *APOE* discovery has been relatively fruitless (Ioannidis 2007). Such interactive effects can exist without the presence of substantial main effects, making detection with single-locus analysis unlikely (Hirschhorn et al. 2002). Such a scenario requires the evaluation of all participating sites for detection, or proxy sites in linkage disequilibrium (LD) with those mutations. New methods for the analyses of large interaction search spaces are now available and were applied here for family and case-control data (Martin et al. 2006; Ritchie et al. 2001; Ritchie, Hahn, and Moore 2003). Due to strong biological and epidemiological evidence of a genetic etiology for AD but lack of consistent single-locus findings, AD would appear to be an ideal trait to begin a search for epistasis among past candidates.

The goal of this study is to explore effects explaining AD through single-locus analysis, haplotypes, and epistatic gene-gene interactions among these variants.

Table 6-1. Family data details and ascertainment

Family Type	**Total Families**	**CAP Families**	**NIMH Families**	**IU Families**	**Discordant sibling pairs**	**Affected Relative Pairs**
Multiplex	580	87	349	124	1111	1153
Singleton	158	78	3	29	161	0

Materials and methods

Study population

Family data: The data for this study consisted of genotypes in both a family sample and an independent case-control sample. The family sample has been described elsewhere (Martin et al. 2005) and contains 738 families collected through three ascertainment groups: the Collaborative Alzheimer Project (CAP: The Joseph and Kathleen Bryan ADRC and the Center for Human Genetics at Duke University, the Center for Human Genetics Research at Vanderbilt University Medical Center, and the University of California at Los Angeles Neuro-psychiatric Institute); National Institutes of Mental Health (NIMH); and the National Cell Repository for AD at Indiana University Medical Center (IU). The family sample is described in Table 6-1. The singleton dataset contains 158 families with one sampled affected family member and any number of unaffected siblings. The multiplex dataset contains 580 families with at least two sampled affected family members.

All affected individuals met the NINDS/ADRDA criteria for probable or definite AD. Unaffected relatives from the CAP and NIMH sites showed no signs of dementia upon examination. Unaffected individuals from IU were classified based on self report.

The mean (SD) age at onset (AAO) in affected individuals was 72.31 (9.09) years, and the mean (SD) age at examination (AAE) was 74.82 (11.02) years.

Case-control data: The case-control dataset consisted of 296 unrelated cases and 566 unrelated controls independent of the family data. The average age of exam (standard deviation) for cases was 79.02(6.76), and controls was 73.63(6.30). The average age of onset (SD) for cases was 71.78(7.82). The age of onset in cases and age of examination in controls were not significantly different. Unrelated cases were determined to be affected by examination based on the same criteria as the cases in the family data. Priority for selection was given to cases where age of onset was known, Parkinson's disease (PD) was not present, depression status was known, and documentation proving AD was available. Unaffected controls required unaffected status confirmed by examination, no first-degree relatives with AD, no PD, otherwise no dementia, and adequate DNA for genotyping. Unrelated cases and controls were collected at the Center for Human Genetics at Duke University and the Center for Human Genetics Research at Vanderbilt University Medical Center. Also ascertained in the case control data was hypertension status, which was measured by survey as having ever being diagnosed with hypertension.

Table 6-2. Gene and SNP information for Alzheimer's candidate genes

Gene	SNP rs#	Position (bp)	Chromosome Band	Alleles Major/Minor	Role
PZP	rs10842971	9194563	12p13.31	A/T	Coding exon
	rs3213831	9208040	12p13.31	T/C	Coding exon
	rs2277413	9209051	12p13.31	C/T	Coding exon
	rs3213832	9212768	12p13.31	C/T	Coding exon
	rs12230214	9238059	12p13.31	C/G	Coding exon
A2MP	rs16918212[B]	9276225	12p13.31	C/A	Not annotated
	rs34362[B]	9276692	12p13.31	C/T	Not annotated
	rs17804080[B]	9279277	12p13.31	C/T	Not annotated
LRP1	rs1799986	55821533	12q13.3	C/T	Coding exon
	rs1800127	55825349	12q13.3	C/T	Coding exon
	rs1800174	55846076	12q13.3	G/A	Intron (boundary)
	rs1800181	55864555	12q13.3	C/T	Intron (boundary)
	rs2075699	55871411	12q13.3	C/T	Coding exon
	rs1800154	55875926	12q13.3	C/T	Coding exon
	rs1800165	55877493	12q13.3	T/C	Intron (boundary)
	rs11172124	55881222	12q13.3	G/A	Intron (boundary)
	rs9669595	55881333	12q13.3	G/A	Intron
	rs7956957	55889082	12q13.3	G/C	Promoter
CTNNA3	rs1786927	67352267	10q21.3	G/A	Intron
	rs2126750	67507709	10q21.3	T/A	Intron
	rs7911820	67534145	10q21.3	G/T	Intron
	rs12357560	67534187	10q21.3	T/C	Intron
	rs7070570	67534610	10q21.3	A/G	Intron
	rs7074454	67534965	10q21.3	T/C	Intron
	rs6480140	67538887	10q21.3	A/C	Intron
	rs997225	67952976	10q21.3	G/A	Intron
LRRTM3	rs1925583	68349950	10q21.3	G/T	Promoter
	rs942780	68406547	10q21.3	A/G	Intron
	rs1925617	68434823	10q21.3	T/G	Intron
NCSTN	rs6668576	157130094	1q23.2	T/C	Intron
	rs10494342	157130193	1q23.2	T/G	Intron
	rs2038781	157130457	1q23.2	G/C	Intron
	rs12239747[B]	157134138	1q23.2	A/G	Coding exon
	rs7528638[B]	157136976	1q23.2	C/G	Intron (boundary)
	rs6427515	157138184	1q23.2	C/T	Intron
	rs4656256	157144092	1q23.2	A/G	Promoter
COG2	rs3789662	227135608	1q42.2	A/G	3' UTR
	rs7536290	227143437	1q42.2	A/G	3' UTR
AGT	rs3789670	227150449	1q42.2	C/T	Intron
	rs2478545	227150856	1q42.2	C/T	Intron
	rs4762[B]	227152712	1q42.2	G/A	Coding exon
	rs2148582	227156534	1q42.2	T/C	Intron (boundary)
	rs5051[B]	227156607	1q42.2	C/T	Promoter

	rs5050[B]	227156621	1q42.2	T/G	Promoter
	rs1326886	227166495	1q42.2	A/G	Promoter
A2M	rs3832852	9137444	12p13	CCATA/del	Splice Site
	rs1800433[A]	9123618	12p13	A/G	Coding Exon
APOE	ε2, ε3, ε4	50101007	19q13.31	ε2, ε3, ε4	-
ACE	rs4291	58907926	17q23.3	A/T	Promoter
	rs4295	58910030	17q23.3	G/C	Intron (boundary)
	rs4311	58914495	17q23.3	C/T	Intron (boundary)
	rs4329	58917190	17q23.3	A/G	Intron
	rs4646994	58919636	17q23.3	del/ins	Intron
	rs4343	58919763	17q23.3	G/A	Coding exon
	rs4353	58924154	17q23.3	A/G	Intron
	rs4978	58927493	17q23.3	T/C	Coding exon

[A] genotyped only in family data
[B] genotyped only in case-control data

Genotyping methods

The list of SNPs selected for this study is shown in Table 6-2. The rationale for including each gene in the list of candidates is detailed in Table 6-3. The SNPs were designed to be genotyped on the Applied Biosystems, Taqman 7900HT allelic discrimination system and were either custom (Assay by Design) or inventoried (Assay on Demand) assays. All genotyping reactions were run according to the standard genotyping methods as outlined by Applied Biosystems protocols and were performed on 3ng of genomic DNA per reaction. All SNPs were held to a minimum genotyping efficiency of 95%. Quality control was performed on the SNPs by using matched pairs of quality control samples placed within and between the 384 well plates. Laboratory technicians were blinded to the matching pattern, affection status, and pedigree information.

Table 6-3. Names and roles of candidate genes in AD

Gene Symbol	Gene Name	Candidate Rationale
PZP	Pregnancy Zone Protein	closely related to *A2M*, maps to the same region
A2MP	Alpha 2 Macroglobulins of pregnancy	*PZP* analog, may participate in cardiovascular remodeling
LRP1	Low density lipoprotein receptor-related protein 1	APOE receptor related protein
CTNNA3	Catenin, Alpha-3	alpha-T-catenin binds beta-catenin; beta-catenin interacts with PSEN1
LRRTM3	Leucine rich repeat transmembrane 3	nested gene in *CTNNA3*
NCSTN	Nicastrin	forms complex with PS1 and 2
COG2	Component of oligomeric golgi complex 2	essential component of intracellular protein trafficking
AGT	Angiotensinogen	enzymatic target of ACE
A2M	Alpha 2 Macroglobulin	mediates the clearance of A-Beta, the main component of amyloid beta deposits
APOE	Apolipoprotein Epsilon	validated AD association, VLDL transport
ACE	Angiotensinogen converting enzyme	associations found previously between *ACE* and AD

Family data

PDT: The PDT is described in Chapter II.

MDR-PDT: The MDR-PDT is described in Chapter II.

The presence of a statistically significant signal from MDR-PDT is not necessarily an interaction in the formal sense. This result may be due to a strong main effect, group of main effects, a nested interaction, a combination of any of these, or an actual interaction representing effect modification across genotypes. Such an interaction should be formally tested using conditional logistic regression.

Tag SNPs in family data were chosen using tagger, a function within the haploview software package (Barrett et al. 2005; Gabriel et al. 2002) for the MDR-PDT analyses to remove redundant variables from the data, which reduce the power of MDR-PDT. An r^2 threshold of 0.8 and LOD of 3 were used to choose tag SNPs in order to eliminate nearby markers with very similar information and maximize power for MDR-PDT analysis. The abridged data contained 32 of the original 47 markers.

Conditional logistic regression: For single-locus effect size evaluation, the referent group was the major allele homozygote. The adjustment of Siegmund et al. (Siegmund et al. 2000) was implemented to correct confidence intervals for familial correlation in regions of linkage.

Association in the presence of linkage (APL): The Association in the Presence of Linkage (APL) statistic (Chung, Hauser, and Martin 2006; Martin et al. 2003) was employed to measure haplotype associations in family data. APL measures the difference in the number of copies of an allele or haplotype in affected offspring from the expected number of copies under the null hypothesis of no association conditional on parental genotypes. APL uses nuclear families with at least one affected offspring. When parental genotypes are missing, they are inferred using the expected probabilities of consistent parental mating types. APL correctly adjusts for correlated transmissions to multiple affected siblings by estimating IBD probabilities. The probability IBD 0, 1, 2 and the haplotype frequency are estimated by EM algorithm (Clark 1990; Excoffier and Slatkin 1995; Long, Williams, and Urbanek 1995).

To estimate the variance of the APL statistic, a bootstrapping approach is used (Chung, Hauser, and Martin 2006). Bootstrap samples are taken with replacement across families, forming same-size pseudosamples consisting of replicates of some families and missing others at random. The variance of the APL statistic calculated for all pseudosamples is the estimated sampling variance for the statistic. This variance can be used to test the null hypothesis of no association allowing for the presence of linkage.

Case-control data analysis

Chi-square and Fisher's exact tests: To test for association of sex with genotypes in the case-control data chi-squared or Fisher's exact tests of differences between frequencies of alleles and genotypes between sexes were performed in controls at each marker. This test should detect where sampling error has distorted the distribution of alleles or genotypes by sex at autosomal markers. Since there is a difference in prevalence by sex in AD, such a scenario in the data could cause confounding. If the genotype frequency tests were significant at the 0.05 level, then sex-stratified chi-squared or Fisher's exact tests of Hardy-Weinberg Equilibrium (HWE) and association with disease at alleles and genotypes were performed.

Sex stratification, single site allele and genotype frequency and association, and HWE analyses in controls were performed using Powermarker statistical software (Zaykin et al. 2002). Where the number of observations for a cell from the 3x2 table stratifying the data by genotype and status was five or less, Fisher's exact test was used to assess HWE and association with AD.

MDR: MDR (Ritchie et al. 2001) was used to search for interactions in the case-control data. MDR exhaustively screens all possible interactions and ranks results by the signal detected by balanced accuracy and cross-validation consistency in case-control data to find models with the most potential to be real interactions. MDR performs well across many genetic simulation scenarios where purely epistatic relationships existed between status and a set of variables with an absence of main effects (Ritchie, Hahn, and Moore 2003).

In the case-control data, all models including *APOE* were highly significant by the permutation test. *APOE* was deleted and the analysis was run again. Tags were chosen using the same criteria as in the family data.

Logistic regression: To estimate single-locus effect sizes in parallel with the family data, effect sizes in case-control data were estimated using logistic regression using the major allele homozygote as the referent group (Stata Corp 2005).

Haplotype analysis: Haplotype analyses for case-control data were performed using the haplo.cc and haplo.glm functions in Haplo.Stats (Schaid et al. 2002). A 3-marker sliding window was run to identify associations among correlated sets of markers. Full haplotypes were tested and haplotype exposure odds ratios were estimated using the most frequent haplotype as the referent group.

Bioinformatic tools: The website SNPer (Riva and Kohane 2004) and Entrez PubMed were used to collect information on candidate genes and genotyped markers. Online Inheritance in Man (OMIM) (Online Mendelian Inheritance in Man 2006) was used to collect information about the phenotype and candidate genes. The Alzgene database at www.Alzgene.org (Bertram et al. 2007) was also used to collect information about AD association studies.

Multiple testing was accounted for depending on the type of analysis. MDR and MDR-PDT both inherently correct for the search conducted with permutation testing. Multiple tests of main effects were corrected using Nyholt's method SNPSpD (Nyholt 2004) with the modification of (Li and Ji 2005). The effective number of tests for the 47 markers that were in both datasets was 28.9 for the founders from family data and 29.4 for the controls from the case-control data, showing the similarity of correlation among these independent samples. To commit as few type I errors as possible, the threshold for significance from the case-control dataset were applied to the results, to yield a threshold for significance of 0.0017.

For purposes of assessing significance where two tests have been performed for the same null hypothesis in independent samples, we used Fisher's method (Fisher 1950) to merge p-values from the same SNP in different samples. We then compared the merged p-value to the threshold for significance given the effective number of independent tests established by SNPSpD. This threshold is determined by the Sidak correction for multiple tests (Sidak 1967) which is slightly more liberal than the Bonferroni correction but provides the exact correction necessary to return the experiment-wise error rate to the desired level.

Results

Family data single-locus results

For the family dataset, single-locus associations were examined with allele and genotype PDT statistics. These results are presented in Table 6-4. Either alleles or genotypes at 17 markers in 8 candidate genes were significantly associated with AD in the families without correction for multiple testing. Of these, only *APOE* survives a conservative Bonferroni correction for tests at all markers. *PZP* SNP rs12230214 (C/G), a nonsynonymous L/V change located in exon 11 (allele p = 0.18, genotype p = 0.05). Two *LRP1* SNPs, rs9669595 (A/G), located in intron 65 (allele p = 0.02, genotype p = 0.04), and rs7956957 (G/C), located in intron 78 (allele p = 0.08, genotype p = 0.02). Two *CTNNA3* intron 13 SNPs, rs7911820 (G/T) (allele p = 0.02, genotype p = 0.03) and rs7074454 (C/T) (allele p = 0.01, genotype p = 0.02) and 1 intron 14 SNP, rs12357560 (T/C) (allele p = 0.06, genotype p = 0.03). One LRRTM3 intron 7 SNP, rs1925617 (T/G) was significantly associated with AD (allele p = 0.65, genotype p = 0.01). One NCSTN intron 2 SNP, rs2038781 (G/C) was significantly associated (allele p = 0.05, genotype p = 0.04). rs3832852 (ins/del), a 5-base insertion in *A2M* that spans the upstream splice site for exon 18 (allele p = 0.01, genotype p = 0.01). The *APOE* allele ε4 was highly significantly associated with AD in both allele and genotype tests (allele p < 0.001, genotype p < 0.001). The remaining seven markers significantly associated with disease at alleles and genotypes were all found in *ACE*. The *ACE* markers were: rs4291 (A/T), 239 base pairs upstream of exon 1 (allele p = 0.02, genotype p = 0.01); rs4295 (G/C), an intron 2 marker (allele p = 0.07, genotype p = 0.03); rs4311 (C/T), an intron 9 marker (allele p = 0.1, genotype p = 0.03); rs4646994 (del/ins), a 287bp indel in intron 16 (allele p = 0.02, genotype p = 0.07); rs4343 (A/G), a synonymous coding SNP in exon 16 (allele p = 0.01, genotype p = 0.03); rs4353 (A/G), a marker in intron 19 (allele p = 0.04, genotype p = 0.04); and rs4978 (C/T), a synonymous coding SNP in exon 23 (allele p = 0.01, genotype p = 0.01).

Table 6-4. Single-locus analysis of AD candidate genes in family data

Gene	rs#	Risk Allele	P-Value				
			[1]Allele-PDT	[2]Genotype-PDT			[3]Global-PDT
				11[4]	12	22[5]	
PZP	rs10842971	A	0.167	**0.305**	0.886	0.26	0.458
	rs3213831	C	0.585	0.503	**0.651**	**0.894**	0.448
	rs2277413*	T	0.077	0.032	**0.087**	**0.871**	0.082
	rs3213832	T	0.216	0.215	**0.216**	**1**	0.359
	rs12230214*	G	0.182	0.066	**0.05**	0.663	0.085
LRP1	rs1799986	T	0.605	0.9	0.69	**0.239**	0.595
	rs1800127	C	0.421	**0.421**	0.421	**1**	0.421
	rs1800174	A	0.362	0.685	0.614	**0.389**	0.679
	rs1800181	C	0.111	0.378	0.472	**0.146**	0.321
	rs2075699	T	0.199	0.559	0.421	**0.194**	0.41
	rs1800154	C	0.071	**0.164**	0.855	0.198	0.298
	rs1800165	C	0.243	0.686	0.426	**0.218**	0.463
	rs11172124	A	0.67	**0.812**	0.443	**0.487**	0.687
	rs9669595*	A	0.023	0.223	0.313	**0.038**	0.103
	rs7956957*	G	0.08	0.883	0.07	**0.022**	0.057
CTNNA3	rs1786927	G	0.9	**0.956**	**0.943**	0.9	0.992
	rs2126750	A	0.095	**0.103**	0.87	0.287	0.281
	rs7911820*	G	0.016	**0.029**	0.297	0.206	0.083
	rs12357560*	C	0.06	0.028	**0.029**	**1**	0.041
	rs7070570	G	0.078	0.052	**0.076**	**0.864**	0.099
	rs7074454*	C	0.005	**0.019**	0.295	0.112	0.048
	rs6480140	A	0.553	0.134	0.061	**0.327**	0.107
	rs997225	A	0.825	**0.593**	0.475	**0.655**	0.717
LRRTM3	rs1925583	T	0.763	**0.585**	0.634	**0.928**	0.355
	rs942780	G	0.054	**0.096**	0.368	0.237	0.19
	rs1925617*	T	0.652	**0.01**	0.001	**0.092**	0.001
NCSTN	rs6668576	T	0.34	**0.37**	0.693	0.67	0.683
	rs10494342	G	0.555	0.553	**0.555**	**1**	0.792
	rs2038781*	G	0.052	**0.035**	0.026	**0.317**	0.031
	rs6427515	T	0.815	0.814	**0.814**	**1**	0.964
	rs4656256	G	0.396	0.425	**0.535**	**0.726**	0.683
COG2	rs3789662	G	0.532	**0.523**	0.589	0.865	0.784
	rs7536290	A	0.819	0.966	**0.731**	0.413	0.768
AGT	rs3789670	C	0.457	0.437	**0.441**	**1**	0.671
	rs2478545	C	0.867	**0.833**	0.863	**1**	0.976
	rs2148582	T	0.839	**0.295**	0.195	**0.451**	0.328
	rs1326886	G	0.378	**0.236**	0.157	**0.336**	0.236
A2M	rs3832852*	del	0.002	0.001	**0.004**	**0.544**	0.002
	rs1800433	A	0.315	**0.131**	0.2	**1**	0.271
APOE	ε2, ε3, ε4*	ε4	<0.001	<0.001	**<0.001**	**<0.001**	<0.001
ACE	rs4291*	A	0.015	**0.012**	0.1	0.392	0.038
	rs4295*	G	0.068	**0.028**	0.054	0.864	0.058

	rs4311*	T	0.106	0.963	0.054	**0.026**	0.056
	rs4329	G	0.092	0.324	0.708	**0.113**	0.298
	rs4646994*	ins	0.017	0.072	**0.857**	**0.074**	0.116
	rs4343*	A	0.01	0.076	**0.974**	**0.027**	0.069
	rs4353*	G	0.04	0.255	0.581	**0.044**	0.152
	rs4978*	C	0.008	0.115	0.593	**0.012**	0.048

[1] Allele-PDT- Association test for alleles
[2] Genotype-PDT-Association test for genotypes, bold numbers indicate genotypes overrepresented in cases
[3] Global-PDT- Global test of association for all genotypes at the locus
[4] Major allele homozygote
[5] Minor allele homozygote
*Nominally significant association observed at this marker

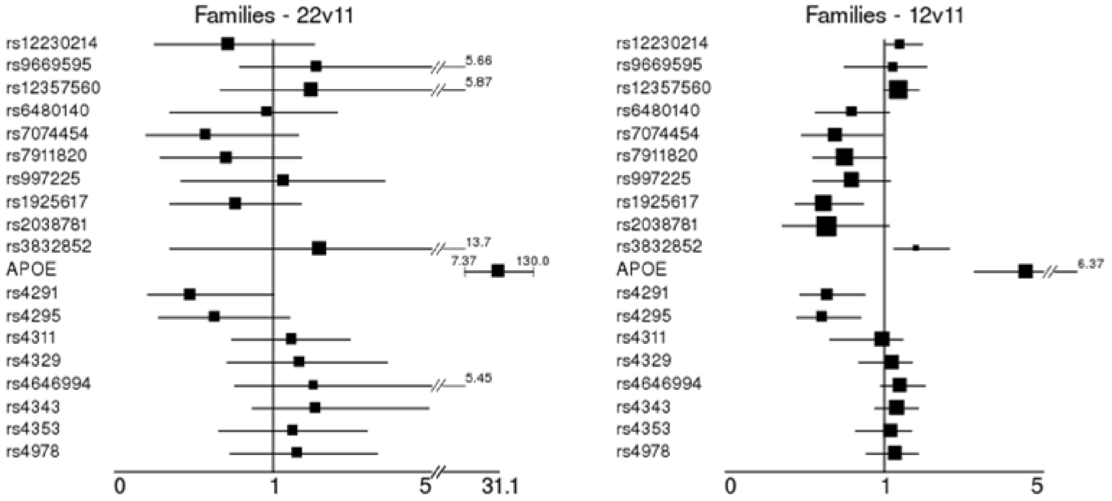


Figure 6-1. Odds ratio estimates of effect size from family data using conditional logistic regression and the correction of Siegmund et al**.** All estimates are for markers that were significantly associated with AD at alleles or genotypes in either family or case-control data. Estimates are for the homozygote major allele (11) as the referent group versus the other 2 genotypes (22[1], 12[2]).
[1] Homozygous minor allele vs. homozygous major allele - 22 genotype compared to 11 (referent group)
[2] Heterozygote vs. homozygous major allele -12 genotype compared to 11 (referent group)

Conditional logistic regression was run to estimate the effect sizes observed in the family sample among those markers that were significant at either alleles or genotypes in families or case-control samples. Of note in these estimates are those estimates for markers that were significantly associated in the case-control sample (Table 6-5, and described below), thus attempting to remedy the bias encountered when effect size estimation and association detection are performed on the same data. The major allele homozygote was used as the referent group for these analyses. These results are detailed in Figure 6-2. *APOE* had a very strong effect in these data for the ε4 homozygote (OR = 31.1 95% CI = 7.37-130) and the ε4 heterozygote (OR = 4.57 95% CI = 3.28-6.57). Other than *APOE*, seven significant single-locus genotype effects in five genes were observed in the family data. The *PZP* marker rs12230214 (OR = 1.42, 95% CI = 1.04-1.95) had a statistically significant effect for the CG heterozygote. Two *CTNNA3* markers showed a significant effect: rs12357560 (OR 1.37, 95% CI = 1-1.87) for the TC heterozygote and rs7074454 (OR 0.69, 95% CI = 0.48-0.99) for the TC heterozygote. The *LRRTM3* marker rs1925617 (OR 0.619, 95% CI = 0.44-0.87) had a significant effect estimated for the TG heterozygote. The *A2M* marker rs3832852 (OR = 1.81, 95% CI = 1.25-2.64) had a significant effect estimate for the splice site deletion heterozygote. Two markers in *ACE* had significant effect estimates. They were rs4291 (OR = 0.48, 95% CI = 0.21-1.0) for the A allele homozygote and (OR = 0.64, 95% CI = 0.47-0.88) for the AT heterozygote, and rs4295 (OR = 0.62, 95% CI = 0.45-0.85) for the GC heterozygote.

Table 6-5. Single-locus analysis of AD candidate genes in case-control data

Gene	rs#	Minor Allele	Minor Allele Relative Freq.	P-value		
				[1]HWE (controls)	[2]Allele	[3]Genotype
PZP	rs10842971	T	0.316	0.278	0.231	0.184
	rs3213831	C	0.423	0.156	0.325	0.464
	rs2277413	T	0.302	0.718	0.853	0.921
	rs3213832	T	0.057	1	0.696	0.641
	rs12230214	G	0.282	0.007	0.551	0.288
A2MP	rs16918212	A	0.1	0.462	1	0.739
	rs34362	T	0.271	0.07	0.718	0.417
	rs17804080	T	0.098	0..311	0.919	0.458
LRP1	rs1799986	T	0.158	0.86	0.527	0.566
	rs1800127	T	0.021	0.616	0.538	0.534
	rs1800174	A	0.25	0.496	0.512	0.851
	rs1800181	T	0.241	0.151	0.502	0.779
	rs2075699	C	0.242	0.475	0.663	0.529
	rs1800154	T	0.318	0.362	0.231	0.499
	rs1800165	C	0.254	0.489	0.826	1
	rs11172124	A	0.253	0.6	0.503	0.511
	rs9669595	A	0.33	0.682	0.508	0.798
	rs7956957	C	0.35	0.666	0.6	0.861
CTNNA3	rs1786927	A	0.409	0.471	0.569	0.851
	rs2126750	A	0.348	0.528	0.205	0.439
	rs7911820	T	0.377	0.616	0.168	0.389
	rs12357560	C	0.232	0.599	0.909	0.922
	rs7070570	G	0.272	0.403	0.997	0.766
	rs7074454	C	0.375	0.679	0.149	0.356
	rs6480140*	C	0.367	0.377	0.613	0.008
	rs997225*	A	0.209	0.709	0.033	0.082
LRRTM3	rs1925583	T	0.467	0.443	0.964	0.633
	rs942780	G	0.195	0.621	0.724	0.78
	rs1925617	G	0.46	0.203	0.408	0.435
NCSTN	rs6668576	C	0.466	0.291	0.854	0.91
	rs10494342	G	0.053	1	1	0.917
	rs2038781	C	0.043	0.579	0.549	0.501
	rs12239747	G	0.052	1	0.707	0.928
	rs7528638	G	0.053	0.654	0.556	0.766
	rs6427515	T	0.068	0.717	0.902	1
	rs4656256	G	0.19	0.363	0.054	0.123
COG2	rs3789670	T	0.103	0.807	0.192	0.278
	rs7536290	G	0.152	0.066	0.943	0.885
AGT	rs3789670	T	0.103	0.807	0.192	0.278
	rs2478545	T	0.213	0.106	0.46	0.452
	rs4762	A	0.123	0.017	0.597	0.528
	rs2148582	C	0.399	0.302	0.159	0.262

	rs5051	T	0.394	0.219	0.128	0.253
	rs5050	G	0.163	0.043	0.471	0.299
	rs1326886	G	0.084	0.353	0.124	0.285
A2M	rs3832852	del	0.15	0.403	0.878	0.608
APOE	ε2, ε3, ε4*	ε4	0.087	0.95	<0.001	<0.001
ACE	rs4291	T	0.378	0.611	0.541	0.324
	rs4295	C	0.381	0.6	0.71	0.233
	rs4311	T	0.492	0.125	0.191	0.437
	rs4329	G	0.435	0.608	0.103	0.211
	rs4646994	ins[4]	0.443	0.751	0.065	0.105
	rs4343*	A	0.433	0.513	0.048	0.144
	rs4353	G	0.438	0.349	0.082	0.191
	rs4978	C	0.447	0.948	0.201	0.389

[1] HWE-Hardy Weinberg Equilibrium
[2] Allele-chi-square test comparing allele frequencies of Alzheimer cases to controls
[3] Genotype-chi-square test comparing genotype frequencies of Alzheimer cases to controls
[4] Minor allele at rs4646994 is a 287bp insertion
*Nominally significant association observed at this marker

Case-control data single locus results

The results of tests at single loci from the case-control data are in Table 6-5. Three markers in 2 genes significantly deviated from HWE in controls. One was the *PZP* marker rs12230214, minor allele frequency (MAF): 0.28 (p = 0.01). The *AGT* markers rs5050, MAF: 0.16 (p = 0.04) and rs4762, MAF: 0.123 (p = 0.01) also significantly deviated from HWE.

Allele and genotype frequency differences among controls between males and females were significant at 4 markers in 4 genes. These tests were conducted to make observations regarding potential confounding by sex where sampling error had caused association of autosomal alleles and genotypes with sex in controls. Such spurious associations in the data might lead to confounding since there is an association between sex and AD.

PZP marker rs12230214 (allele p = 0.01, genotype p = 0.05), *LRP1* marker rs1800127 (allele p = 0.03, genotype p = 0.03), *LRRTM3* marker rs942780 (allele p = 0.01, genotype p = 0.02), *A2M* marker rs3832852 (allele p = 0.01, genotype p = 0.02), all had significantly different frequencies by sex in controls at both alleles and genotypes. Each of these markers was tested separately in males and females for HWE and allele and genotype frequency differences between cases and controls. Among these tests, significant deviations from HWE were found in control females for *PZP* marker rs12230214 (p = 0.02).

Statistically significant single-locus differences in allele or genotype frequency between cases and controls were observed at 3 markers in 2 genes. One marker in *ACE* was significantly associated with disease at alleles. rs4343 (A/G) MAF: 0.46, a synonymous SNP in exon 16 (allele p = 0.05, genotype p = 0.14). Two markers in *CTNNA3* were significantly associated with disease. The markers rs6480140 (A/C) MAF: 0.37, a SNP in intron 14 (allele p = 0.61, genotype p = 0.01) and rs997225 (A/G), a SNP in intron 10 (allele p = 0.03, genotype p = 0.08). The *APOE* marker MAF: 0.09, (allele p < 0.001, genotype p < 0.001) was very strongly associated with disease. Again only APOE survived a conservative Bonferroni correction for multiple tests at all loci.

Merged results

Merged p-values for the family and case-control single-locus tests were analyzed using Fisher's method (Fisher 1950). This approach allows for the evidence against the null hypothesis across tests to be combined into a single statistic for each null hypothesis. Global p-values were used from the family-based tests on genotypes. The results of this analysis are presented in Table 6-6.

Table 6-6. Merged p-values from case-control and family data using Fisher's method

Gene	rs #	Allele p-value	Genotype p-value
PZP	rs10842971	0.164	0.293
	rs3213831	0.506	0.534
	rs2277413	0.245	0.271
	rs3213832	0.435	0.568
	rs12230214	0.331	0.115
LRP1	rs1799986	0.683	0.703
	rs1800127	0.563	0.560
	rs1800174	0.498	0.895
	rs1800181	0.217	0.597
	rs2075699	0.399	0.548
	rs1800154	0.084	0.432
	rs1800165	0.523	0.820
	rs11172124	0.704	0.719
	rs9669595	0.064	0.288
	rs7956957	0.194	0.197
CTNNA3	rs1786927	0.855	0.987
	rs2126750	0.096	0.382
	rs7911820*	0.019	0.143
	rs12357560	0.213	0.162
	rs7070570	0.276	0.271
	rs7074454*	0.006	0.087
	rs6480140*	0.706	0.007
	rs997225	0.125	0.225
LRRTM3	rs1925583	0.961	0.560
	rs942780	0.166	0.431
	rs1925617*	0.618	0.004
NCSTN	rs6668576	0.649	0.917
	rs10494342	0.882	0.959
	rs2038781	0.130	0.080
	rs6427515	0.961	0.999
	rs4656256	0.104	0.292
COG2	rs3789670	0.335	0.550
	rs7536290	0.972	0.942
AGT	rs3789670	0.301	0.500
	rs2478545	0.765	0.802
	rs2148582	0.402	0.297
	rs1326886	0.190	0.249
A2M	rs3832852*	0.013	0.009
APOE	ε2, ε3, ε4*	<0.001	<0.001
ACE	rs4291*	0.047	0.066
	rs4295	0.195	0.072

	rs4311	0.099	0.115
	rs4329	0.054	0.237
	rs4646994*	0.009	0.066
	rs4343*	0.004	0.056
	rs4353*	0.022	0.132
	rs4978*	0.012	0.093

* Nominally significant association observed at this marker

Several markers in *ACE* were nominally significant at alleles and trending at genotypes. CTNNA3 marker rs7074454 was also significant at alleles and trending at genotypes. A2M SNP rs3832852 was nominally significant at alleles and genotypes. Again, Only *APOE* survived a Bonferroni correction for multiple tests.

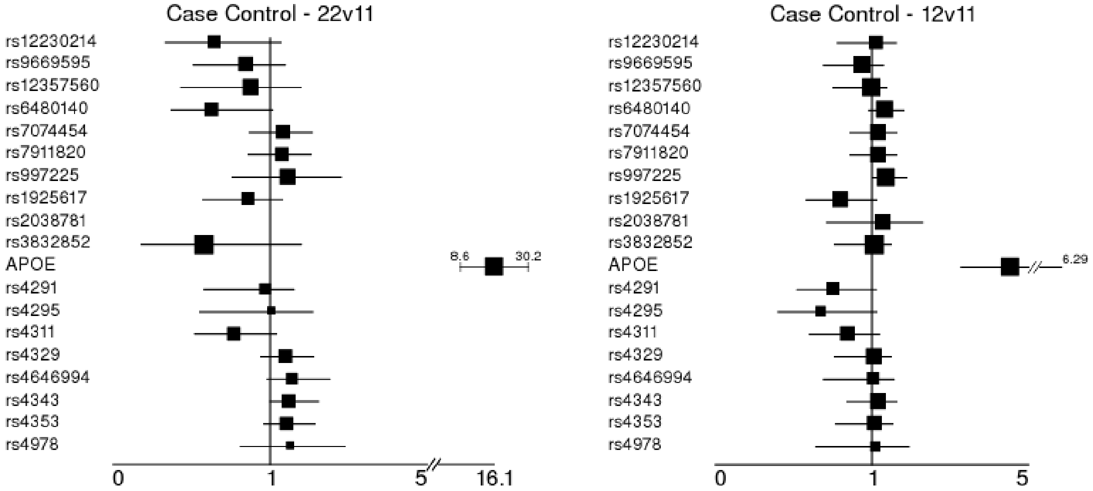


Figure 6-2. Odds ratio effect size estimates for significant single-locus associations from case-control data**.** All estimates are for markers that were significantly associated with AD at alleles or genotypes in either family or case-control data. Estimates are for the homozygote major allele (11) as the referent group versus the other 2 genotypes (22[1], 12[2]).

[1] Homozygous minor allele vs. homozygous major allele - 22 genotype compared to 11 (referent group)

[2] Heterozygote vs. homozygous major allele - 12 genotype compared to 11 (referent group)

To estimate the effect of each significant finding from family or case-control data, odds ratios and 95% confidence intervals using the major allele homozygote as the referent group were estimated in the case-control data using logistic regression from the STATA statistical software package. Since the markers associating with AD did not significantly differ in genotype frequency by sex, no adjustment for confounding by sex was performed. Also, since no difference was detected between age of onset and controls, no adjustment for age was performed. These results are presented in Figure 6-3. Of note in these results are those loci which demonstrated association in the family sample. Significant effects were detected at *APOE* ε4 homozygotes (OR 16.1 95% CI = 8.6-30.2), *APOE* ε4 heterozygotes (OR 4.55 95% CI = 3.28-6.29), the CTNNA3 SNP rs997225 GA heterozygote (OR 1.39 95% CI = 1.02-1.89) and *ACE* SNP rs4343 for the minor allele homozygote, (OR 1.49 95% CI = 1.0-2.23). Markers in *ACE* were also assessed for confounding by hypertension status. After forcing hypertension status into the models the OR point estimates did not change substantially, indicating that hypertension was not a confounder for those variables.

Haplotype results

Haplotype analysis was performed across all candidate markers in pairwise LD as defined by a D' of 0.95 or greater in the family data with APL using a 3-locus sliding window. These tests identified overlapping 3-locus haplotypes in the *ACE* gene that were significantly associated with AD in the family data set. Results of this procedure are in Table 6-7a. These results suggest a consistent signal of association with disease on a common haplotype background throughout these *ACE* markers. This signal is from a chromosome containing an array of minor alleles at each of these markers. This diffuse association signal is detectable at each individual marker, but this phenomenon is also observed in the case-control data, which makes the family result worthy of note. Also, the p-values observed at these overlapping 3-locus haplotypes are smaller than those for most of the single-locus statistics.

Table 6-7a. Significant family and case-control data haplotype association in ACE

APL haplotypes 3-Marker Scan		Haplotype	Relative Frequency	P-value	
Gene	**Markers**			**Haplotype**	**Global**
ACE	rs4291 – rs4295	AG	0.495	0.383	0.672
	rs4311 – rs4329 – rs4646994[1]	TGI	0.460	0.045	0.290
	rs4329 – rs4646994[1] – rs4343	GIA	0.450	0.013	**0.020**
	rs4646994[1] – rs4343 – rs4353	IAG	0.450	0.004	**0.004**
	rs4343 – rs4353 – rs4978	AGC	0.450	0.003	**0.020**

[1] 289 base pair Alu repeat: D major allele (289bp absent), I minor allele (289bp present)

Table 6-7a. Significant family and case-control data haplotype association in ACE

Haplo.stats 3-Marker Scan		Haplotype	Relative Frequency		OR	CI	P-value
Gene	**Markers**		**Cases**	**Controls**			
ACE	rs4291 - rs4295	AG	0.65	0.61	1.13	0.89-1.45	0.338
	rs4311 - rs4329 - rs4646994[1]	TGI	0.48	0.43	1.21	0.98-1.48	0.06
	rs4329 - rs4646994[1] - rs4343	GIA	0.48	0.43	1.22	1-1.49	**0.046**
	rs4646994[1] - rs4343 - rs4353	IAG	0.48	0.43	1.23	1.01-1.51	**0.048**
	rs4343 - rs4353 - rs4978	AGC	0.48	0.43	1.22	1.01-1.50	**0.049**
	rs4311 - rs4329 - rs4646994 - rs4343 - rs4353 - rs4978	TGIAGC	0.48	0.42	1.22	1.01-1.50	**0.041**

[1] 289 base pair Alu repeat: D major allele (289bp absent), I minor allele (289bp present)

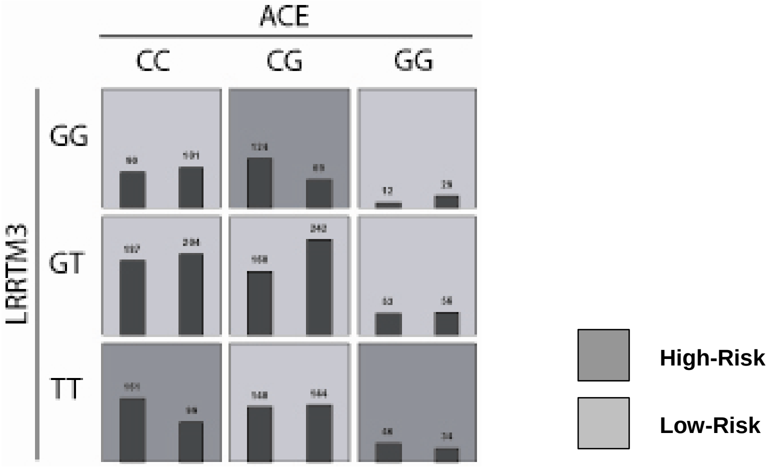
In the family data, the *ACE* gene contained several significant markers and overlapping associated haplotypes. No other regions in the family sample contained significant haplotypes. To follow up this observation, and to validate the haplotype findings in *ACE* from the family data, the Haplo.Stats software package was used to estimate haplotype frequencies and test haplotype associations in *ACE* in case-control data. The results of both sets of tests in family and case-control data are presented in Tables 6-7a and 6-7b. A sliding window scan of the markers in *ACE*, analogous to that performed in the family data, was conducted among markers in strong LD ($r^2 > 0.9$, D' > 0.95) in both datasets. This scan yielded an odds ratio and 95% confidence interval for all haplotypes versus the most common haplotype, and a chi-squared test for haplotype frequency differences between cases and controls. Every 3-locus haplotype in *ACE* between rs4311 and rs4978 had a chi-squared p-value < 0.05. The 2-locus haplotype including rs4291 and rs4295 was not significant in either dataset. The 6-locus haplotype including rs4311, rs4329, rs4646994, rs4343, rs4353, and rs4978 had an OR estimate very close to 1.2 and 95% confidence intervals at approximately 1.0-1.5, which was very similar to those estimates for the 3-locus sliding window through that region. This indicates that chromosomes in this area of the gene tend to be either all major or minor alleles with little recombination in two primary haplotypes. Also of interest is the similarity of haplotype frequency estimates between the family and case-control data. The consistent estimated susceptibility haplotype frequency and pattern of significance strongly suggests a main effect is present in this gene.

MDR-PDT results

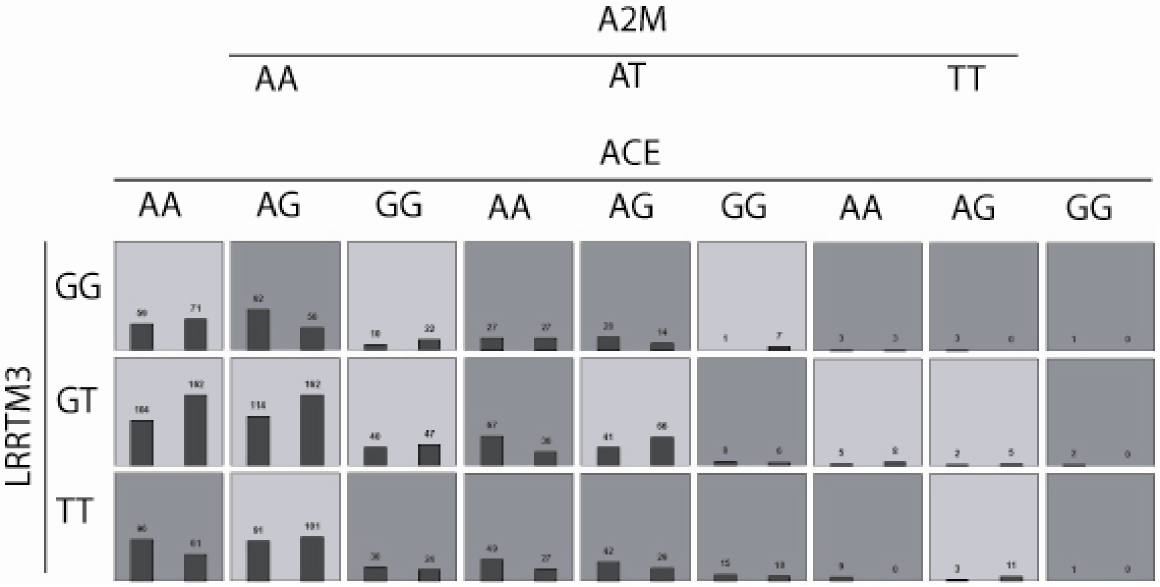
Having explored single-locus main effects at alleles, genotypes, and haplotypes, we began a search for multi-locus signals significantly associated with disease using the MDR-PDT in family data and MDR in case-control data. The MDR-PDT models are presented in Table 6-8 and Figures 6-3a and 6-3b. MDR and MDR-PDT were run with all markers and every model including the *APOE* marker was highly significant by the permutation test. To remove the very strong *APOE* signal from the data, the *APOE* marker was then removed from the data and another search for interactions was conducted among tags. Haplotype tag SNPs were chosen using the haploview software function tagger ($r^2 = 0.8$, LOD = 3), and the best models were found by MDR and MDR-PDT. The best two and 3-locus models from the full data without *APOE* contained the same markers as those chosen from the tag SNP data for the MDR-PDT. This indicated that the signal observed at these models was detected by MDR-PDT, but the known issue of power loss with increasing numbers of markers caused the failure to reject. No MDR model was significant by the permutation test. The best MDR model was a 3-locus model including LRP1 SNP rs1800165, PZP SNP rs3213831, and PZP SNP rs10842971 (CVC 2/5, PE 43.41, p-value = 0.34). Two significant signals were found by MDR-PDT. The 2-locus model included rs1925617 in *LRRTM3* and rs4295 in *ACE* (MDR-PDT statistic $p < 0.001$). The 3-locus model included rs1925617 in *LRRTM3*, rs4291 in *ACE*, and rs1800433 in *A2M* (MDR-PDT statistic $p < 0.001$).

Table 6-8. Summary of MDR-PDT results

# of Loci	Best Model For Each Interaction	SNPs	T Statistic	Classification Accuracy (%)
1	[*LRRTM3*]	rs1925617	3.32	54.32
2	[*LRRTM3-ACE*]	rs1925617-rs4291	4.49	56.89
3	[*LRRTM3-ACE-A2M*]	rs1925617-rs4291-rs1800433	5.65	59.58



6-3a. MDR-PDT two locus model. Summary of multilocus interactions between *LRRTM* and*, ACE*. Each multifactorial cell is labeled as “high risk” or “low risk”. For each multifactorial combination, empirical distributions of cases *(left bar in cell)* and controls *(right bar in cell)* are shown. The classification accuracy for the 2-locus model is 56.89% (p-value <0.001), with a t-statistic of 4.49 (p-value <0.001).



6-3b. Summary of multilocus interactions between *LRRTM3, ACE* and *A2M*. Each multifactorial cell is labeled as "high risk" or "low risk". For each multifactorial combination, empirical distributions of cases *(left bar in cell)* and controls *(right bar in cell)* are shown. The classification accuracy for the 3-locus model is 59.58% (p-value < 0.001) with an MDR-PDT-statistic of 5.65 (p-value < 0.001).

Discussion

These results highlight the *ACE* gene as a risk factor in AD. In both family and case-control samples, significant associations were observed when considering *ACE* haplotypes. Notably in the case-control samples, only one single-locus test was marginally significant at rs4343 for the test on genotypes, but the haplotype tests on specific, overlapping sets of alleles were significant. The p-values from the family data haplotype analysis were also smaller than those from the single-locus analysis, suggesting more signal was detected when considering the entire region. This finding strongly indicates that a genetic background exists in *ACE* in the Caucasian population that is associated with this array of alleles and AD. This replication in both data types strongly supports the hypothesis that the *ACE* gene may harbor real risk variants for AD.

Some of the evidence for *ACE* association was found across many studies in recent meta-analyses of association results (Lehmann et al. 2005). Haplotype associations have also been previously observed in *ACE* for AD in five independent case-control samples (Kehoe et al. 2003). We observed significant haplotype association with different markers than those studies in case-control and family data, further supporting the hypothesis of association of the gene to late-onset AD.

The *ACE* intron 16 I/D polymorphism, a marker in the associated haplotype, has been previously reported to associate with AD. This association has a plausible biological explanation, since *ACE* degrades Aβ peptide in vitro (Hu et al. 2001), and the insertion allele results in decreased plasma levels of the *ACE* protein (Hemming and Selkoe 2005; Rigat et al. 1990; Tiret et al. 1992).

Overall, association results in the *ACE* gene for Caucasians have been inconsistent in past studies. In 31 case-control association studies in Caucasians with markers in *ACE* reviewed by (Bertram et al. 2007), there were 12 positive findings, 15 negative findings and 4 trends suggesting association between *ACE* and AD. The sample sizes were larger in studies where positive associations were observed (1-sided t-test p-value for cases = 0.04, p-value for controls = 0.03, p-value for overall sample size = 0.02), suggesting that differential power might explain some of the previous inconsistency. Additionally, associations of AD with *A2M* have been inconsistent, where in Alzgene.org there have been six positive associations, two trends and 35 negative associations observed in case-control Caucasian samples for markers in that gene. However, in family data there were three positive, one trend, and two negative associations. Most of those studies were performed on rs3832852. There was not a significant sample size difference across study outcomes for sample sizes. For *CTNNA3*, there were seven negative and two positive associations in case-control samples, but in family samples four of five studies found variants associating with AD. Evidence exists for associations at each of these genes, and interactions among them may explain some of the previous inconsistency.

Interactions likely are relevant to genetic epidemiology of AD, and the MDR-PDT represents a unique capability to search for such effects in family data. MDR, the analogous approach in case-control data, has been used to find several interactive effects for various phenotypes. MDR has been used to detect genetic interactions contributing to risk in several diseases. Some examples are: sporadic breast cancer (Ritchie et al. 2001), essential hypertension (Moore and Williams 2002; Williams et al. 2004), atrial fibrillation (Tsai et al. 2004), type II diabetes (Cho et al. 2004), coronary artery calcification (Bastone et al. 2004), myocardial infarction (Coffey et al. 2004), schizophrenia (Qin et al. 2005), and amyloid polyneuropathy (Soares et al. 2005). MDR-PDT has been shown in simulation to have better power than MDR when families are large, as in these data (Martin et al. 2006).

The family and case-control analyses detailed in this manuscript were conducted in parallel with the exception of the haplotype analysis, which was done in families and then on the *ACE* markers specifically in the case-control data. This was done to compare results from either dataset and produce stronger initial findings for future replication. The ACE haplotype results were somewhat unexpected, since the individual markers did not show much association signal in the case-control sample, yet the haplotype analysis revealed the association.

Single-locus analysis in the family sample yielded many more significant results than in the case-control samples. The family dataset was relatively large with more than 4,500 individuals, compared to 860 from the case-control data. More power was available to detect relatively subtle effects in families than in case-control samples. The estimated effects of associated genotypes in the family sample were similar to the estimates from the case-control sample for most of the interesting markers which were detected in either sample. This supports the reliability and generalizability of these estimated values for these markers.

To maximize sensitivity, uncorrected p-values are presented, with the exception of MDR-PDT which adjusts within each k-locus test. All the tests performed in this analysis were considered significant at the 0.05 level. The rate at which significant results corroborate one another when the null is true is the product of the alpha levels, or for these analyses 0.0025, so there is some inherent protection from Type I error in this study design. In total, 259 tests were performed to detect association during this analysis. If the null hypothesis were true for every test performed, and if all tests were completely independent, which these are not due to LD and variable sharing among models, then the expectation would be 12.95 95% CI [6.08, 19.83] Type I errors throughout these results at the 0.05 alpha level. There are 32 p-values that are 0.05 or less from these 259 tests (z-score = 5.44, p = $2.6x10^{-8}$), suggesting that association signals exist in these data and that the null hypothesis is not true at all tested variables. Five significant tests corroborated each other between the samples, all in *ACE*. This far exceeds the expectation of 0.26 such events across all pairs of tests. The strict assumptions of this z-score test are not met here;

however, neither are the assumptions of the Bonferroni correction or the false discovery rate procedure (Benjamini and Hochberg 1995) which assume independent tests.

Future directions for these investigations into the mechanism underlying late-onset AD should include further ascertainment and genotyping in the *ACE* gene, as well as functional studies targeting potential molecular etiologies involving *ACE*. Plasma levels of *ACE* and putative downstream targets relevant to AD should be measured to make observations regarding coordinate regulation, feedback systems, and continuous measurements further explaining this pattern of association. The presence of *A2M* and *CTNNA3*/*LRRTM3* SNPs in the MDR-PDT model also point to Aβ accumulation as a factor predicting late-onset AD, as these genes all relate to Aβ clearance. Perhaps these variants and others related to Aβ clearance with similarly subtle effects are sufficient to cause disease for late-onset cases among more aged persons, whereas other more acute genetic lesions elsewhere, such as in the presenilins, might precipitate an earlier onset. It may be that we have already discovered the main causes of AD in the constellation of weak main effects that have been observed to date. The attributable risk of the *ACE* haplotype alone explains about 16.6% of late-onset AD cases among those exposed to the haplotype. For the entire Caucasian population, the population attributable risk for the haplotype explains about 8%, or 320,000 late-onset Alzheimer's cases. While this fraction is small, it has been replicated here and elsewhere, implicating this locus as not only contributing modest risk, but also supporting the biological hypothesis regarding plasma concentrations of *ACE* and their relationship to Aβ concentrations.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions

New developments in genotyping technology have provided genetic epidemiologists with novel tools with which to study traits. These developments were made without certain knowledge of how best to use platforms with hundreds of thousands of polymorphisms in human populations. Applying these tools effectively for mapping genomic locations to traits can proceed in several ways. The most conceptually straightforward approach is one where financial cost is not a consideration and enough samples are collected to provide sufficient statistical precision for current association methods to overcome whatever multiple testing corrections are applied to the analysis. Such an inelegant approach would only be available for traits that pose the most serious threats to public health and that were relatively common and straightforward to ascertain. Even under these hypothetical circumstances, some aspects of trait mapping may yet elude investigators. The search for epistatic disease models is an example of an analysis problem that poses conspicuous challenges that have not yet been thoroughly explored in simulation for GWA data. Haplotype association mapping is also a difficult issue, both for reasons of computation required and for result visualization and inference. It is for all these commonly acknowledged reasons that sophisticated software such as genomeSIMLA as presented in Chapter IV should be developed to assist in statistical method development. Thereby more efficient statistical methods, GWA analysis software packages, study designs, and specific analytical issues can be evaluated for sampling properties in realistic data scenarios. Traits that might not receive such grandiose support as the major public health problems might become accessible to study with GWA with improved methodological approaches, the overall cost of human genetics studies decreased, and the rate of success increased.

We observed significant association in the Alzheimer’s family and case-control samples at the *ACE* locus (Appendix Table 1). This association is commonly accepted by many familiar with the genetic epidemiology of Alzheimer’s disease, although without the strength of evidence or certainty of the *APOE* locus. Additionally, we observed a significant multilocus model including three genes involved in Aβ processing, which included *ACE*. While no biological network of genes has been directly implicated in Alzheimer’s susceptibility, it is likely that this pathway is important given these associations and the histopathological hallmarks of AD. Future studies of AD should continue to apply multilocus methods to accumulate evidence for networks of genes, rather than restricting searches to single loci.

During the analysis of the Alzheimer’s data with MDR-PDT and MDR, a phenomenon occurred that demanded further study. All models that included the *APOE* locus were highly significant by the permutation test. Since it is not likely that effects at all these loci undergo modification in the presence of simultaneous exposure with *APOE* variants, alternative explanations were sought. Upon contemplation of this scenario and the structure of the null hypothesis for the permutation test for MDR and MDR-PDT, it became apparent that strong main effects cause rejection of the null hypothesis. Since these methods are designed to find interactions in particular, this lack of specificity would cause the reporting of Type I errors in the literature with regard to finding genetic effect modifiers. This issue was recognized in the original MDR-PDT literature, and the means used to determine whether signals were the result of epistasis was to fit a regression model *post hoc* and assess the interaction term corresponding to the best model from MDR or MDR-PDT. We showed here with simulation that this procedure is not valid when both analyses are performed in the same data. We also show with simulation that the specificity of the MDR-PDT permutation test decreases as main effects and sample size increase. The solution we propose in Chapter V is to fit full and reduced regression models to data using constructive induction to encode genotypes as binary variables. The likelihood ratio statistic for the interaction term corresponding to the best model from MDR-PDT is recorded for comparison to the empirical null from the subsequent permutation test. By using this measure, the null hypothesis is changed to no interaction, and we show with simulation that specificity in the presence of independent main effects is returned to 100%. Some disadvantages of this method are that the power relative to the original permutation procedure is reduced, the MDR-PDT is still as liable as ever to find the strong main effects as the best model, and the computation time is somewhat increased. However, the null will not be rejected for inappropriate scenarios, which was a major shortcoming of the previous approach.

Future Directions

GenomeSIMLA: The work presented here is preliminary to many future projects. For genomeSIMLA, some future projects are to perform a methodological comparison of several current GWA simulation methods. The computation time and properties of the data produced would be evaluated for similarity to known human samples. The flexibility of trait modeling and embedded tools for data evaluation would also be a featured part of the comparison. Additional improvements to the algorithm would be to incorporate population stratification and admixture capability, population bottlenecks, new mutation, multiple-generation family data, more disease susceptibility loci, quantitative traits, and selection. We feel that the tools we have developed thus far provide a good basis from which to implement all these extensions. We most likely will also incorporate the coalescent modeling software GENOME as an optional front-end instead of the forward-time option. This option would be named GENOMESIMLA for those who prefer coalescent-based simulations. The benefit of the modular design of the software we have developed thus far is that we can easily incorporate other approaches for use with our modeling and visualization capability.

Regression test of significance: For the regression-based hypothesis test for MDR-PDT, an immediate future direction is to implement the analogous methodology for MDR to be used with case-control data. A project that would result from that work would be the comparison of the MDR-logistic to FITF. In such a simulation study, models featuring interaction with and without main effects would be simulated in large-scale data simulated with genomeSIMLA. We anticipate that these results will show that MDR-logistic has power to detect interactions with no main effects while FITF does not. We also anticipate that this study will show that MDR-logistic has better power than FITF when the number of loci in an interaction is large. The improved power for large models is because of the use of constructive induction to encode genotypes for regression with MDR-logistic and the resulting dimensionality reduction. FITF does not enjoy this benefit. For instance, for an MDR-logistic model with 3 loci, there are 2^3, or 8, levels of exposure, for FITF, there are 3^3, or 27. For 4 locus models, there are 2^4, or 16, for MDR-logistic and 3^4, or 81, for FITF. For large numbers of loci, we expect the power of MDR-logistic to be superior due to the use of permutation testing. FITF uses an FDR correction for multiple tests that assumes all models and loci are independent. This conservative correction becomes more statistically inefficient relative to the permutation test as the search space increases. Of course, it is already known that FITF has no power to detect models with no main effects, a shortcoming we shall be quick to point out.

Some further extensions of the effect modification tests after searches for interactive models might be to augment methods featuring neural networks, such as grammatically evolving neural networks with methods which test for interaction. Also, other faster and more powerful methods to test for effect modification will be sought.

GWA analysis: Another area of interest is the analysis of GWA data. Currently, MDR and MDR-PDT are too slow to run permutation tests in these data. This is because there are 125.99 billion 2-locus models in a dataset with 500,000 markers. However, even if permutation testing were possible in these data, it is likely that the critical values for significance after such an immense search will be so extreme that there will be no power to reject the null in a hypothesis test without huge sample sizes. The compromise I propose is to simply split the data in half, run MDR on the first half to find interesting models, and then conduct hypothesis tests with a null of no interaction with regression in the other half. This may initially seem an unreasonable compromise, but it carries several notable advantages. The first advantage is multiple testing. Since the two samples are independent, there is no need to adjust the regression results for the search by MDR. If only one model is tested with regression, then the parametric alpha of 0.05 could be applied to declare significance. Another advantage not available with the current permutation testing strategy is the possibility of finding multiple interactions with a single MDR run. If for instance the alternate hypothesis of interaction is true for several models, and they represent several of the top ten MDR models, then there is a good chance to reject these nulls at an alpha of 0.005 with even half of a moderate to large sample. Finally, the most important advantage to this approach is feasibility and speed. Currently, it takes about 4 days on 160 processors to conduct a single search for 2-locus models with MDR in 500 cases and 500 controls. One thousand permutations would take 10.95 years to perform on 160 processors, and perhaps a million permutations or more would be a more reasonable approximation for the tails of the null distribution for such an immense search space. This procedure could be performed in a few days, and would have more power than the impossible permutation procedure.

Other future directions: Finally, I have interest in adapting the spectral decomposition methods of Nyholt to interaction searches. This method could be a means of circumventing permutation testing altogether. It would not be feasible for GWA data, but it could be useful for MDR-PDT, which has an especially slow permutation test. Essentially, a correlation matrix would be computed for the interaction terms for all possible interactions, and Nyholt's procedure run. This should adjust the threshold for significance for the formal test of interaction for the entire search space by determining the effective number of independent tests performed, as in SNP data with LD. While not as statistically efficient as permutation testing, this could be done almost instantly, as opposed to a permutation test which might take a week or longer. If significant, a permutation test would be run to obtain a more accurate estimate of significance.

Ultimately, methods which both simulate epistasis and detect epistasis with analysis will be important features of any comprehensive effort to explain genetic susceptibility to disease. These advances provided in this dissertation are incremental steps toward that goal. More efficient statistics and more advanced simulations, combined with sound study design and thoughtful analysis will continue to reveal the genes, and by extension the mechanisms that lead to complex genetic disease.

APPENDIX

Penetrance tables

Penetrance tables in order from Table 5-1

Table 5-1. Models examined in the simulation study.

Model	Loci	Minor allele frequency	Heritability	Odds Ratio
1	2	0.2	0.005	1.1
2	2	0.2	0.01	1.26
3	2	0.2	0.03	1.53
4	2	0.2	0.048	1.79
5	2	0.2	0.09	3
6	2	0.4	0.005	1.15
7	2	0.4	0.01	1.28
8	2	0.4	0.03	1.56
9	2	0.4	0.05	1.79
10	2	0.4	0.1	2.85
11	3	0.2	0.005	1.19
12	3	0.2	0.01	1.36
13	3	0.2	0.03	1.58
14	3	0.2	0.05	2.1
15	3	0.2	0.1	3.2
16	3	0.4	0.005	1.21
17	3	0.4	0.01	1.32
18	3	0.4	0.03	1.52
19	3	0.4	0.05	2.23
20	3	0.4	0.12	3.5

Model 1

aabb	0.151
aabB	0.155
aaBB	0.165
aAbb	0.153
aAbB	0.161
aABB	0.082
AAbb	0.182
AAbB	0.046
AABB	0.530

Model 2

aabb	0.157
aabB	0.139
aaBB	0.184
aAbb	0.144
aAbB	0.186
aABB	0.010
AAbb	0.136
AAbB	0.105
AABB	0.793

Model 3

aabb	0.514
aabB	0.587
aaBB	0.604
aAbb	0.604
aAbB	0.436
aABB	0.369
AAbb	0.467
AAbB	0.643
AABB	0.907

Model 4

aabb	0.238
aabB	0.315
aaBB	0.000
aAbb	0.271
aAbB	0.099
aABB	0.788

AAbb	0.305
AAbB	0.155
AABB	0.000

Model 5

aabb	0.315
aabB	0.525
aaBB	0.183
aAbb	0.505
aAbB	0.066
aABB	0.807
AAbb	0.342
AAbB	0.487
AABB	0.045

Model 6

aabb	0.193
aabB	0.205
aaBB	0.247
aAbb	0.203
aAbB	0.199
aABB	0.210
AAbb	0.247
AAbB	0.205
AABB	0.103

Model 7

aabb	0.171
aabB	0.155
aaBB	0.105
aAbb	0.160
aAbB	0.152
aABB	0.137
AAbb	0.087
AAbB	0.150
AABB	0.309

Model 8

aabb	0.417
aabB	0.361
aaBB	0.713
aAbb	0.454
aAbB	0.489
aABB	0.352
AAbb	0.488
AAbB	0.484
AABB	0.315

Model 9

aabb	0.201
aabB	0.156
aaBB	0.016
aAbb	0.157
aAbB	0.150
aABB	0.131
AAbb	0.017
AAbB	0.131
AABB	0.529

Model 10

aabb	0.140
aabB	0.496
aaBB	0.511
aAbb	0.425
aAbB	0.345
aABB	0.325
AAbb	0.724
AAbB	0.164
AABB	0.192

Model 11

aabbcc	0.203
aabbcC	0.203
aabbCC	0.283
aabBcc	0.203
aabBcC	0.207
aabBCC	0.211

aaBBcc	0.257
aaBBcC	0.229
aaBBCC	0.029
aAbbcc	0.208
aAbbcC	0.201
aAbbCC	0.214
aAbBcc	0.195
aAbBcC	0.214
aAbBCC	0.000
aABBcc	0.221
aABBcC	0.065
aABBCC	0.065
AAbbcc	0.269
AAbbcC	0.209
AAbbCC	0.004
AAbBcc	0.221
AAbBcC	0.076
AAbBCC	0.038
AABBcc	0.000
AABBcC	0.059
AABBCC	0.050

Model 12

aabbcc	0.144
aabbcC	0.145
aabbCC	0.207
aabBcc	0.146
aabBcC	0.167
aabBCC	0.160
aaBBcc	0.250
aaBBcC	0.198
aaBBCC	0.023
aAbbcc	0.153
aAbbcC	0.169
aAbbCC	0.138
aAbBcc	0.156
aAbBcC	0.175
aAbBCC	0.012
aABBcc	0.030
aABBcC	0.000
aABBCC	0.077
AAbbcc	0.279
AAbbcC	0.016
AAbbCC	0.031

AAbBcc	0.134
AAbBcC	0.030
AAbBCC	0.119
AABBcc	0.003
AABBcC	0.028
AABBCC	0.422

Model 13

aabbcc	0.452
aabbcC	0.505
aabbCC	0.488
aabBcc	0.497
aabBcC	0.322
aabBCC	0.246
aaBBcc	0.442
aaBBcC	0.659
aaBBCC	0.793
aAbbcc	0.463
aAbbcC	0.363
aAbbCC	0.641
aAbBcc	0.440
aAbBcC	0.689
aAbBCC	0.418
aABBcc	0.140
aABBcC	0.687
aABBCC	0.583
AAbbcc	0.566
AAbbcC	0.292
AAbbCC	0.225
AAbBcc	0.377
AAbBcC	0.650
AAbBCC	0.116
AABBcc	0.436
AABBcC	0.418
AABBCC	0.385

Model 14

aabbcc	0.153
aabbcC	0.151
aabbCC	0.147
aabBcc	0.159
aabBcC	0.095

aabBCC	0.261
aaBBcc	0.000
aaBBcC	0.758
aaBBCC	0.006
aAbbcc	0.158
aAbbcC	0.104
aAbbCC	0.189
aAbBcc	0.129
aAbBcC	0.347
aAbBCC	0.016
aABBcc	0.002
aABBcC	0.002
aABBCC	0.002
AAbbcc	0.331
AAbbcC	0.003
AAbbCC	0.000
AAbBcc	0.083
AAbBcC	0.004
AAbBCC	0.037
AABBcc	0.010
AABBcC	0.009
AABBCC	0.032

Model 15

aabbcc	0.418
aabbcC	0.570
aabbCC	0.231
aabBcc	0.297
aabBcC	0.426
aabBCC	0.722
aaBBcc	0.259
aaBBcC	0.791
aaBBCC	0.707
aAbbcc	0.533
aAbbcC	0.033
aAbbCC	0.430
aAbBcc	0.532
aAbBcC	0.557
aAbBCC	0.536
aABBcc	0.487
aABBcC	0.249
aABBCC	0.192
AAbbcc	0.202
AAbbcC	0.589

AAbbCC	0.433
AAbBcc	0.858
AAbBcC	0.154
AAbBCC	0.623
AABBcc	0.292
AABBcC	0.258
AABBCC	0.316

Model 16

aabbcc	0.176
aabbcC	0.175
aabbCC	0.152
aabBcc	0.184
aabBcC	0.174
aabBCC	0.188
aaBBcc	0.173
aaBBcC	0.186
aaBBCC	0.092
aAbbcc	0.177
aAbbcC	0.181
aAbbCC	0.169
aAbBcc	0.181
aAbBcC	0.173
aAbBCC	0.179
aABBcc	0.186
aABBcC	0.172
aABBCC	0.158
AAbbcc	0.181
AAbbcC	0.182
AAbbCC	0.119
AAbBcc	0.168
AAbBcC	0.184
AAbBCC	0.170
AABBcc	0.110
AABBcC	0.168
AABBCC	0.544

Model 17

aabbcc	0.156
aabbcC	0.157
aabbCC	0.133
aabBcc	0.149

aabBcC	0.149
aabBCC	0.157
aaBBcc	0.149
aaBBcC	0.151
aaBBCC	0.114
aAbbcc	0.165
aAbbcC	0.142
aAbbCC	0.143
aAbBcc	0.155
aAbBcC	0.164
aAbBCC	0.140
aABBcc	0.124
aABBcC	0.132
aABBCC	0.141
AAbbcc	0.158
AAbbcC	0.147
AAbbCC	0.123
AAbBcc	0.132
AAbBcC	0.141
AAbBCC	0.084
AABBcc	0.108
AABBcC	0.149
AABBCC	0.704

Model 18

aabbcc	0.519
aabbcC	0.419
aabbCC	0.467
aabBcc	0.726
aabBcC	0.519
aabBCC	0.414
aaBBcc	0.499
aaBBcC	0.372
aaBBCC	0.689
aAbbcc	0.305
aAbbcC	0.678
aAbbCC	0.541
aAbBcc	0.506
aAbBcC	0.476
aAbBCC	0.627
aABBcc	0.425
aABBcC	0.643
aABBCC	0.441
AAbbcc	0.786

AAbbcC	0.519
AAbbCC	0.624
AAbBcc	0.472
AAbBcC	0.420
AAbBCC	0.238
AABBcc	0.519
AABBcC	0.569
AABBCC	0.802

Model 19

aabbcc	0.383
aabbcC	0.659
aabbCC	0.449
aabBcc	0.602
aabBcC	0.233
aabBCC	0.485
aaBBcc	0.457
aaBBcC	0.756
aaBBCC	0.222
aAbbcc	0.635
aAbbcC	0.249
aAbbCC	0.374
aAbBcc	0.517
aAbBcC	0.641
aAbBCC	0.411
aABBcc	0.154
aABBcC	0.475
aABBCC	0.531
AAbbcc	0.510
AAbbcC	0.560
AAbbCC	0.657
AAbBcc	0.189
AAbBcC	0.346
AAbBCC	0.864
AABBcc	0.387
AABBcC	0.747
AABBCC	0.597

Model 20

aabbcc	0.198
aabbcC	0.712
aabbCC	0.719

aabBcc	0.455
aabBcC	0.312
aabBCC	0.431
aaBBcc	0.094
aaBBcC	0.360
aaBBCC	0.746
aAbbcc	0.483
aAbbcC	0.273
aAbbCC	0.387
aAbBcc	0.527
aAbBcC	0.514
aAbBCC	0.294
aABBcc	0.523
aABBcC	0.403
aABBCC	0.019
AAbbcc	0.639
AAbbcC	0.105
AAbbCC	0.580
AAbBcc	0.213
AAbBcC	0.440
AAbBCC	0.386
AABBcc	0.701
AABBcC	0.877
AABBCC	0.626

Appendix Table 1. Significant associations observed in Chapter VI

Gene	rs number	risk allele	family allele	family genotype	CC allele	CC genotype	Fisher's allele	Fisher's genotype
PZP	rs2277413*	T	0.077	0.082	-	-	-	-
PZP	rs12230214*	G	0.182	0.085	-	-	-	-
LRP1	rs9669595*	A	0.023	0.103	-	-	-	-
LRP1	rs7956957*	G	0.08	0.057	-	-	-	-
CTTNA3	rs7911820*	G	0.016	0.083	-	-	-	-
CTTNA3	rs12357560*	C	0.06	0.041	-	-	-	-
CTTNA3	rs7074454*†	C	0.005	0.048	-	-	0.006	0.087
CTTNA3	rs6480140§†	C	-	-	0.613	0.008	0.706	0.007
CTTNA3	rs997225§	A	-	-	0.033	0.082	-	-
LRRTM3	rs1925617*‡†	T	0.652	0.001	-	-	0.618	0.004
NCSTN	rs2038781*	G	0.052	0.031	-	-	-	-
A2M	rs1800433‡	A	-	-	-	-	-	-
A2M	rs3832852*†	del	0.002	0.002	-	-	0.013	0.009
APOE	ε2, ε3, ε4*†	ε4	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
ACE	rs4291*‡	A	0.015	0.038	-	-	-	-
ACE	rs4295*	G	0.068	0.058	-	-	-	-
ACE	rs4311*	T	0.106	0.056	-	-	-	-
ACE	rs4646994*†	ins	0.017	0.116	-	-	0.009	0.066
ACE	rs4343*§†	A	0.01	0.069	0.048	0.144	0.004	0.056
ACE	rs4353*†	G	0.04	0.152	-	-	0.022	0.132
ACE	rs4978*†	C	0.008	0.048	-	-	0.012	0.093

*Significant at alleles or genotypes in family sample

§Significant at alleles or genotypes in case-control sample

‡Part of Significant MDR-PDT models

†Significant by Fisher's merged p-value statistic

References

1. (2003) The International HapMap Project. *Nature* 426 (6968):789-796.

2. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011):931-945.

3. Online Mendelian Inheritance in Man. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) . 4-20-2006.
   Ref Type: Electronic Citation

4. Abecasis GR, Cookson WO, and Cardon LR (2000) Pedigree tests of transmission disequilibrium. *Eur.J.Hum.Genet.* 8 (7):545-551.

5. Alvarez R et al (1999) Angiotensin converting enzyme and endothelial nitric oxide synthase DNA polymorphisms and late onset Alzheimer's disease. *J.Neurol.Neurosurg.Psychiatry* 67 (6):733-736.

6. Andrew AS et al (2006) Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis* 27 (5):1030-1037.

7. Ashley-Koch AE et al (2006) An analysis paradigm for investigating multi-locus effects in complex disease: examination of three GABA receptor subunit genes on 15q11-q13 as risk factors for autistic disorder. *Ann.Hum.Genet.* 70 (Pt 3):281-292.

8. Bales KR et al (1999) Apolipoprotein E is essential for amyloid deposition in the APP(V717F) transgenic mouse model of Alzheimer's disease. *Proc.Natl.Acad.Sci.U.S.A* 96 (26):15233-15238.

9. Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *J.Hered.* 92 (3):301-302.

10. Barrett JC and Cardon LR (2006) Evaluating coverage of genome-wide association studies. *Nat.Genet.* 38 (6):659-662.

11. Barrett JC et al (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21 (2):263-265.

12. Bass MP, Martin ER, and Hauser ER (2004) Pedigree generation for analysis of genetic linkage and association. *Pac.Symp.Biocomput.*:93-103.

13. Bastone L et al (2004) MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. *Hum Hered.* 58 (2):82-92.

14. Bateson W (1909) Mendel's Principles of Heredity. Cambridge University Press: Cambridge.

15. Bellman RE (1961) Dynamic Programming. Princeton University Press: Princeton, NJ.

16. Benjamini Y and Hochberg Y (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J.R.Statist Soc.B* 57 (1):289-300.

17. Bergqvist D and Nilsson IM (1979) Hereditary alpha 2-macroglobulin deficiency. *Scand.J.Haematol.* 23 (5):433-436.

18. Bertram L et al (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat.Genet.* 39 (1):17-23.

19. Blacker D et al (1998) Alpha-2 macroglobulin is genetically associated with Alzheimer disease. *Nat.Genet.* 19 (4):357-360.

20. Boehnke M (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. *Am.J.Hum.Genet.* 39 (4):513-527.

21. Brassat D et al (2006) Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. *Genes Immun.* 7 (4):310-315.

22. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall: New York, NY.

23. Breteler MM et al (1994) Cardiovascular disease and distribution of cognitive function in elderly people: the Rotterdam Study. *BMJ* 308 (6944):1604-1608.

24. Cantrell VA et al (2004) Interactions between Sox10 and EdnrB modulate penetrance and severity of aganglionosis in the Sox10Dom mouse model of Hirschsprung disease. *Hum.Mol.Genet.* 13 (19):2289-2301.

25. Cardon LR and Palmer LJ (2003) Population stratification and spurious allelic association. *Lancet* 361 (9357):598-604.

26. Chan IH et al (2006) Gene-gene interactions for asthma and plasma total IgE concentration in Chinese children. *J.Allergy Clin.Immunol.* 117 (1):127-133.

27. Chartier-Harlin MC et al (1994) Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. *Hum.Mol.Genet.* 3 (4):569-574.

28. Cho YM et al (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47 (3):549-554.

29. Chung RH, Hauser ER, and Martin ER (2006) The APL test: extension to general nuclear families and haplotypes and examination of its robustness. *Hum.Hered.* 61 (4):189-199.

30. Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol.Biol.Evol.* 7 (2):111-122.

31. Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am.J.Hum.Genet.* 65 (4):1170-1177.

32. Coffey CS et al (2004) 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.

33. Cook NR, Zee RY, and Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat.Med* 23 (9):1439-1453.

34. Cordell HJ, Barratt BJ, and Clayton DG (2004) Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet.Epidemiol.* 26 (3):167-185.

35. Cordell HJ and Clayton DG (2005) Genetic association studies. *Lancet* 366 (9491):1121-1131.

36. Corder EH et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261 (5123):921-923.

37. Culverhouse R, Klein T, and Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 27 (2):141-152.

38. Curtis D and Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. *Am.J.Hum.Genet.* 56 (3):811-812.

39. de Bakker PI et al (2005) Efficiency and power in genetic association studies. *Nat.Genet.* 37 (11):1217-1223.

40. Deak KL et al (2005) SNPs in the neural cell adhesion molecule 1 gene (NCAM1) may be associated with human neural tube defects. *Hum.Genet.* 117 (2-3):133-142.

41. DeSalle R and Templeton AR (1986) The molecular through ecological genetics of abnormal abdomen. III. Tissue-specific differential replication of ribosomal genes modulates the abnormal abdomen phenotype in Drosophila mercatorum. *Genetics* 112 (4):877-886.

42. Dudek SM et al (2006) Data simulation software for whole-genome association and other studies in human genetics. *Pac.Symp.Biocomput.*:499-510.

43. Durrant C et al (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am.J.Hum.Genet.* 75 (1):35-43.

44. Easton DF et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.2007.Jun.28.*:1087-1093.

45. Edwards TL et al (2008a) Generating Linkage Disequilibrium Patters in Genetic Data Simulations using genomeSIMLA. *Lecture Notes in Computer Science* In Press.

46. Edwards TL et al (2008b) Exploring the power of Multifactor Dimensionality Reduction in Large Studies or the Presence of Genetic Heterogeneity. *Human Heredity*.

47. Ermakov S et al (2006) Variation in femoral length is associated with polymorphisms in RUNX2 gene. *Bone* 38 (2):199-205.

48. Ertekin-Taner N et al (2000) Linkage of plasma Abeta42 to a quantitative locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Science* 290 (5500):2303-2304.

49. Ertekin-Taner N et al (2003) Fine mapping of the alpha-T catenin gene to a quantitative trait locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Hum.Mol.Genet.* 12 (23):3133-3143.

50. Excoffier L and Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol.Biol.Evol.* 12 (5):921-927.

51. Fisher R (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans.R.Soc.Edin.* 52:399-433.

52. Fisher RA (1950) Statistical methods for research workers. 11 ed. Hafner: New York.

53. Gabriel SB et al (2002) The structure of haplotype blocks in the human genome. *Science* 296 (5576):2225-2229.

54. Goate A (2006) Segregation of a missense mutation in the amyloid beta-protein precursor gene with familial Alzheimer's disease. *J.Alzheimers.Dis.* 9 (3 Suppl):341-347.

55. Gordon D et al (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am.J.Hum.Genet.* 69 (2):371-380.

56. Hahn LW and Moore JH (2004) Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico.Biol.* 4 (2):183-194.

57. Hahn LW, Ritchie MD, and Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19 (3):376-382.

58. Hancock DB et al (2007) Methods for interaction analyses using family-based case-control data: conditional logistic regression versus generalized estimating equations. *Genet.Epidemiol.* 31 (8):883-893.

59. Hardy J (1997) Amyloid, the presenilins and Alzheimer's disease. *Trends Neurosci.* 20 (4):154-159.

60. Hardy J, Myers A, and Wavrant-De VF (2004) Problems and solutions in the genetic analysis of late-onset Alzheimer's disease. *Neurodegener.Dis.* 1 (4-5):213-217.

61. Hemming ML and Selkoe DJ (2005) Amyloid beta-protein is degraded by cellular angiotensin-converting enzyme (ACE) and elevated by an ACE inhibitor. *J.Biol.Chem.* 280 (45):37644-37650.

62. Henderson AS et al (1995) Apolipoprotein E allele epsilon 4, dementia, and cognitive decline in a population sample. *Lancet* 346 (8987):1387-1390.

63. Hessner MJ et al (2001) Age-dependent prevalence of vascular disease-associated polymorphisms among 2689 volunteer blood donors. *Clin.Chem.* 47 (10):1879-1884.

64. Hey J (2005) A computer program for forward population genetic simulation.

65. Hirschhorn JN et al (2002) A comprehensive review of genetic association studies. *Genet.Med.* 4 (2):45-61.

66. Hofman A et al (1997) Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study. *Lancet* 349 (9046):151-154.

67. Hoggart CJ et al (2007) Sequence-level population simulations over large genomic regions. *Genetics*.

68. Hollocher H and Templeton AR (1994) The molecular through ecological genetics of abnormal abdomen in Drosophila mercatorum. VI. The non-neutrality of the Y chromosome rDNA polymorphism. *Genetics* 136 (4):1373-1384.

69. Hollocher H et al (1992) The molecular through ecological genetics of abnormal abdomen. IV. Components of genetic variation in a natural population of Drosophila mercatorum. *Genetics* 130 (2):355-366.

70. Horvath S and Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am.J.Hum.Genet.* 63 (6):1886-1897.

71. Horvath S, Xu X, and Laird NM (2001) The family based association test method: strategies for studying general genotype--phenotype associations. *Eur.J.Hum.Genet.* 9 (4):301-306.

72. Hosmer DW, Lemeshow S (2000) Applied logistic regression. 2nd ed ed. Wiley: New York.

73. Hu J et al (2001) Angiotensin-converting enzyme degrades Alzheimer amyloid beta-peptide (A beta); retards A beta aggregation, deposition, fibril formation; and inhibits cytotoxicity. *J.Biol.Chem.* 276 (51):47863-47868.

74. Hunter DJ et al (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat.Genet.2007.Jul.*:870-874.

75. Ioannidis JP (2007) Non-replication and inconsistency in the genome-wide association setting. *Hum.Hered.* 64 (4):203-213.

76. Kaplan NL, Martin ER, and Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. *Am.J.Hum.Genet.* 60 (3):691-702.

77. Kehoe PG et al (2003) Haplotypes extending across ACE are associated with Alzheimer's disease. *Hum.Mol.Genet.* 12 (8):859-867.

78. Kehoe PG et al (1999) Variation in DCP1, encoding ACE, is associated with susceptibility to Alzheimer disease. *Nat.Genet.* 21 (1):71-72.

79. Kingman J (1982) The coalescent. *Stochastic Processes Appl* 13:235-248.

80. Kooperberg C and Ruczinski I (2005) Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 28 (2):157-170.

81. Kooperberg C et al (2001) Sequence analysis using logic regression. *Genet Epidemiol* 21 Suppl 1:S626-S631.

82. Lander ES et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822):860-921.

83. Lehmann DJ et al (2005) Large meta-analysis establishes the ACE insertion-deletion polymorphism as a marker of Alzheimer's disease. *Am.J.Epidemiol.* 162 (4):305-317.

84. Levy-Lahad E et al (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 269 (5226):973-977.

85. Li J and Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95 (3):221-227.

86. Liang KY and Zeger SL (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 73 (1):13-22.

87. Liang L, Zollner S, and Abecasis GR (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics* 23 (12):1565-1567.

88. Long JC, Williams RC, and Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am.J.Hum.Genet.* 56 (3):799-810.

89. Lou XY et al (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am.J.Hum.Genet.* 80 (6):1125-1137.

90. Lucotte G et al (1994) Association of apolipoprotein E allele epsilon 4 with late-onset sporadic Alzheimer's disease. *Am.J.Med.Genet.* 54 (3):286-288.

91. Lyon HN et al (2003) The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. *PLoS.Genet.2007.Apr 27.*:e61.

92. Ma DQ et al (2005) Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. *Am.J.Hum.Genet.* 77 (3):377-388.

93. Malik FS et al (1997) Renin-angiotensin system: genes to bedside. *Am.Heart J.* 134 (3):514-526.

94. Marchini J, Donnelly P, and Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat.Genet* 37 (4):413-417.

95. Martin ER et al (2003) Genotype-based association test for general pedigrees: the genotype-PDT. *Genet.Epidemiol.* 25 (3):203-213.

96. Martin ER, Bass MP, and Kaplan NL (2001) Correcting for a potential bias in the pedigree disequilibrium test. *Am.J.Hum.Genet.* 68 (4):1065-1067.

97. Martin ER et al (2005) Interaction between the alpha-T catenin gene (VR22) and APOE in Alzheimer's disease. *J.Med.Genet.* 42 (10):787-792.

98. Martin ER, Kaplan NL, and Weir BS (1997) Tests for linkage and association in nuclear families. *Am.J.Hum.Genet.* 61 (2):439-448.

99. Martin ER et al (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am.J.Hum.Genet.* 67 (1):146-154.

100. Martin ER et al (2006) A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet.Epidemiol.* 30 (2):111-123.

101. McIntyre LM et al (2000) Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genet.Epidemiol.* 19 (1):18-29.

102. McIntyre LM and Weir BS (1997) Hardy-Weinberg testing for continuous data. *Genetics* 147 (4):1965-1975.

103. McPherson R et al (2007) A common allele on chromosome 9 associated with coronary heart disease. *Science* 316 (5830):1488-1491.

104. Millstein J et al (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. *Am.J.Hum.Genet.* 78 (1):15-27.

105. Mitchell AA, Cutler DJ, and Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am.J.Hum.Genet.* 72 (3):598-610.

106. Monks SA, Kaplan NL, and Weir BS (1998) A comparative study of sibship tests of linkage and/or association. *Am.J.Hum.Genet.* 63 (5):1507-1516.

107. Moore JH (2004) Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert.Rev.Mol.Diagn.* 4 (6):795-803.

108. Moore JH (2007) Genome-wide analysis of epistasis using mutifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zu X, Davidson I (eds) Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data. IGI Press: Hershey.

109. Moore JH et al (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J.Theor.Biol.* 241 (2):252-261.

110. Moore JH et al (2004) Routine Discovery of High-Order Epistasis Models for Computational Studies in Human Genetics. *Applied Soft Computing* 4:79-86.

111. Moore JH and Williams SM (2002) New strategies for identifying gene-gene interactions in hypertension. *Ann.Med.* 34 (2):88-95.

112. Moore JH and Williams SM (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27 (6):637-646.

113. Morris DW et al (2007) Dysbindin (DTNBP1) and the Biogenesis of Lysosome-Related Organelles Complex 1 (BLOC-1): Main and Epistatic Gene Effects Are Potential Contributors to Schizophrenia Susceptibility. *Biol.Psychiatry*.

114. Motsinger AA et al (2006) GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics* 7:39.

115. Motsinger AA and Ritchie MD (2006) The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. *Genet.Epidemiol.* 30 (6):546-555.

116. Nelson MR et al (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11 (3):458-470.

117. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am.J.Hum.Genet.* 74 (4):765-769.

118. Peng B and Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics.*

119. Ploughman LM and Boehnke M (1989) Estimating the power of a proposed linkage study for a complex genetic trait. *Am.J.Hum.Genet.* 44 (4):543-551.

120. Qin S et al (2005) 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 (7):807-814.

121. Rabinowitz D and Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum.Hered.* 50 (4):211-223.

122. Reid IA (1992) Interactions between ANG II, sympathetic nervous system, and baroreceptor reflexes in regulation of blood pressure. *Am.J.Physiol* 262 (6 Pt 1):E763-E778.

123. Richards F (1959) A flexible growth function for empirical use. *Journal of Experimental Botany* 10:290-300.

124. Rigat B et al (1990) An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J.Clin.Invest* 86 (4):1343-1346.

125. Ripley B (1996) Pattern recognition and neural networks. Cambridge University Press: Cambridge.

126. Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273 (5281):1516-1517.

127. Ritchie K et al (1996) Characteristics of Alzheimer's disease patients with and without ApoE4 allele. *Lancet* 348 (9032):960.

128. Ritchie MD et al (2007) Genetic heterogeneity is not as threatening as you might think. *Genet.Epidemiol.* 31 (7):797-800.

129. Ritchie MD, Hahn LW, and Moore JH (2003) 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 (2):150-157.

130. Ritchie MD et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69 (1):138-147.

131. Riva A and Kohane IS (2004) A SNP-centric database for the investigation of the human genome. *BMC.Bioinformatics.* 5:33.

132. Rogaev EI et al (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 376 (6543):775-778.

133. Rovelet-Lecrux A et al (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat.Genet.* 38 (1):24-26.

134. Saxena R et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316 (5829):1331-1336.

135. Scacchi R et al (1998) DNA polymorphisms of apolipoprotein B and angiotensin I-converting enzyme genes and relationships with lipid levels in Italian patients with vascular dementia or Alzheimer's disease. *Dement.Geriatr.Cogn Disord.* 9 (4):186-190.

136. Schaid DJ et al (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am.J.Hum.Genet.* 70 (2):425-434.

137. Schaid DJ and Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am.J.Hum.Genet.* 53 (5):1114-1126.

138. Schaid DJ and Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. *Am.J.Hum.Genet.* 55 (2):402-409.

139. Schlicting, Pigliucci (1998) Phenotypic evolution: A reaction norm perspective. Sinauer Associates, Inc.

140. Schmidt M et al (2004) Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: environmental covariates, gene-gene and gene-environment interaction. *Stat.Appl.Genet.Mol.Biol.2005.*:Article15.

141. Schuldiner M et al (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 123 (3):507-519.

142. Scott LJ et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316 (5829):1341-1345.

143. Segre D et al (2005) Modular epistasis in yeast metabolism. *Nat.Genet.* 37 (1):77-83.

144. Sherrington R et al (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 375 (6534):754-760.

145. Sidak Z (1967) Rectangular Confidence Regions for Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* 62 (318):626-&.

146. Siegmund KD et al (2000) Testing linkage disequilibrium in sibships. *Am.J.Hum.Genet.* 67 (1):244-248.

147. Skaar DA et al (2005) Analysis of the RELN gene as a genetic risk factor for autism. *Mol.Psychiatry* 10 (6):563-571.

148. Soares ML et al (2005) Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. *Hum.Mol.Genet.* 14 (4):543-553.

149. Spielman RS and Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am.J.Hum.Genet.* 62 (2):450-458.

150. Spielman RS, McGinnis RE, and Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am.J.Hum.Genet.* 52 (3):506-516.

151. Stata Corp (2005) Stata Statistical Software: Release 9 StataCorp LP, College Station, TX.

152. Stephens M and Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am.J.Hum.Genet.* 76 (3):449-462.

153. Tahri-Daizadeh N et al (2003) Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Res* 13 (8):1952-1960.

154. Templeton A (2000) Epistasis and complex traits. In: Wade M, Brodie BI, Wolf J (eds) Epistasis and Evolutionary Process. Oxford University Press: London.

155. Teng J and Risch N (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* 9 (3):234-241.

156. Tiret L et al (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. *Am.J.Hum.Genet.* 51 (1):197-205.

157. Tong AH et al (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294 (5550):2364-2368.

158. Tong AH et al (2004) Global mapping of the yeast genetic interaction network. *Science* 303 (5659):808-813.

159. Tsai CT et al (2004) Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 109 (13):1640-1646.

160. Velez DR et al (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet.Epidemiol.* 31 (4):306-315.

161. Vyshkina T et al (2005) Genetic variants of Complex I in multiple sclerosis. *J.Neurol.Sci.* 228 (1):55-64.

162. Weeks DE, Ott J, and Lathrop G.M (1990) SLINK: A general simulation paorgram for linkage analysis. *American Journal of Human Genetics* 47:A204.

163. Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am.J.Hum.Genet.* 65 (1):229-235.

164. Weinberg CR, Wilcox AJ, and Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am.J.Hum.Genet.* 62 (4):969-978.

165. Welcome Trust Case-Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.2007.Jun.7.*:661-678.

166. Wilke RA, Moore JH, and Burmester JK (2005) Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. *Pharmacogenet.Genomics* 15 (6):415-421.

167. Williams SM et al (2004) Multilocus analysis of hypertension: a hierarchical approach. *Hum.Hered.* 57 (1):28-38.

168. Wright FA et al (2007) Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* 23 (19):2581-2588.

169. Wright S (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the 6th International Congress of Genetics* 1.

170. Zaykin DV et al (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum.Hered.* 53 (2):79-91.

171. Zeger SL and Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42 (1):121-130.

172. Zeggini E et al (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316 (5829):1336-1341.