CONFRONTING COMPLEXITY: A COMPREHENSIVE STATISTICAL AND COMPUTATIONAL STRATEGY FOR IDENTIFYING THE MISSING LINK BETWEEN GENOTYPE AND PHENOTYPE

By

Tricia Ann Thornton-Wells

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

Neuroscience

December, 2006

Nashville, Tennessee

Approved:

Professor Jonathan L. Haines

Professor Michael P. McDonald

Professor Jason H. Moore

Professor Marylyn D. Ritchie

To my wonderful parents, Mary and John, who have always believed in me and
have encouraged and enabled me to do whatever I have dreamed

To my amazing husband, Bryce, who is infinitely supportive and
is always poised to provide external motivation when I need it most

To my beautiful baby boy, Greyson, who provides me with joy and
a healthy perspective on life and work

and

To my great uncle, Morgan Freeman, and my great aunt, Sara Campisi,
whose affliction with Alzheimer Disease was a primary motivating factor for this work

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Neuroscience Graduate Program (T32 MH64913), the Department of Biomedical Informatics, the National Library of Medicine Training Grant Fellowship (LMO7450-01), or my advisor, Professor Jonathan L. Haines. I would also like to thank Marylyn Ritchie, Jason Moore and Jonathan Haines for their tremendous help with interpretation of results.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Each member of my Dissertation Committee has provided professional guidance and taught me a great deal about scientific research. I would especially like to thank Michael P. McDonald, Ph.D., the chairman of my committee. I would like to thank Jackie Bartlett for her assistance in learning and applying the traditional genetic analysis methods and Scott Dudek for programming the fuzzy *k*-modes clustering algorithm. I also like to thank fellow graduate students Will Bush, Todd Edwards, Sharon Liang, Alison Motsinger and David Reif for their support and assistance with hashing out ideas and implementing programs.

I am particularly indebted to several individuals, who have been supportive of my interdisciplinary research goals and have provided exceptional mentoring with regard to my career development. Those persons are: Jonathan L. Haines, Ph.D., Director of the Center for Human Genetics Research, Jason H. Moore, Ph.D., Director of the Computational Genetics Laboratory at Dartmouth College, and Elaine Sanders-Bush, Ph.D., Director of the Vanderbilt Brain Institute.

No one has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, John and Mary Thornton, whose love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my loving husband, Bryce, and my incredible son, Greyson, who provide continuous support and motivation. I should also acknowledge my second, and as-yet-unborn son, who has in his own way provided considerable motivation for the timely completion of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Alzheimer Disease
ARI_{HA}	Hubert-Arabie Adjusted Rand Index
BBN	Bayesian Belief Network
CART	Classification and Regression Trees
CPM	Combinatorial Partitioning Method
HWE	Hardy-Weinberg Equilibrium
LOAD	Late-Onset Alzheimer Disease
MARS	Multivariate Adaptive Regression Splines
MDR	Multifactor Dimensionality Reduction
(O)MIM	(Online) Mendelian Inheritance in Man
OSA	Ordered Subset Analysis
QTL	Quantitative Trait Locus
RPM	Restricted Partitioning Method
SNP	Single Nucleotide Polymorphism
THO	Trait Heterogeneity Only
THL	Trait Heterogeneity with Locus Heterogeneity
THG	Trait Heterogeneity with Gene-Gene Interaction
THB	Trait Heterogeneity with Both Locus Heterogeneity and Gene-Gene Interaction

CHAPTER I

INTRODUCTION

Like many common diseases with a genetic basis, the etiology of late-onset Alzheimer disease (LOAD) is complex. Evidence suggests that LOAD is a heterogeneous trait with multiple susceptibility loci and possibly gene-gene interactions involved. While there are existing methods that can address specific components of this etiology, ultimately, the real power of these methods lies in our ability to marry them into a comprehensive approach to genetic analysis, so that their relative strengths and weaknesses can be balanced and a range of alternative hypotheses can be investigated. Thus, I propose a two-stage, multi-pronged approach to the problem of genetic analysis of LOAD in which heterogeneity is first addressed by dissecting-out more homogeneous subsets of the data and then main effects and gene-gene interactions are investigated in each of these subsets.

The theoretical basis for such an approach to the analysis of complex genetic diseases is presented in Chapter II. Definitions and examples of heterogeneity and interactions that complicate genetic analysis are presented. Existing methods for detecting heterogeneity and interactions are reviewed, and gaps in methodology are discussed.

Chapter III presents a simulation study in which the performance of three clustering methods is compared in the task of uncovering trait heterogeneity in simulated data. A novel data simulation algorithm is introduced. The best of the three clustering

methods—Bayesian Classification—is chosen and its applicability to real data (based on its false positive and false negative rates) is investigated.

Chapter IV details an extension of this simulation study in which the implementation of the Bayesian Classification method is modified to improve performance under a wider range of conditions realistic for genetic studies. False positive and false negative rates under these conditions are also investigated.

Chapter V presents an application of the proposed two-stage comprehensive analysis to a late-onset Alzheimer disease dataset. Analysis of heterogeneity is performed using the Bayesian Classification clustering method. Main effect analysis is performed in cluster subsets. For the case-control dataset, the Pearson chi-square test of independence is applied, and for the family-based dataset, two-point linkage analysis, the Pedigree Disequilibrium Test and the Family-Based Association Test are utilized. Interaction analysis is performed using the Multifactor Dimensionality Reduction method. Logistic regression is used to explore the structure of predictive MDR models found significant by permutation testing. Results of these integrated analyses are interpreted, and limitations of the study design and analysis methods are discussed.

In Chapter VI, the entirety of the research comprising this dissertation is put into perspective, discussing the lessons learned and the immediate future directions for this work. New directions for future studies of neurogenetic diseases are also discussed and suggestions are made as to the focus of future research efforts, given current and forthcoming phenotyping technology, such as neuroimaging.

CHAPTER II

BACKGROUND

Adapted from:

Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. Trends in Genetics 20: 640-647, 2004.

"If the only tool you have is a hammer, you tend to see every problem as a nail."

Abraham Maslow, American psychologist, founder of humanistic psychology

"The difficulty lies, not in the new ideas, but in escaping the old ones."

John Maynard Keynes, English economist

Complex Human Genetic Disease

Over the past few decades, most of the success in the field of statistical genetics has come from identifying genes with substantial main (i.e., independent; non-interactive) effects on the disease process. Most statistical tools enabling this success were developed for and are primarily effective in the analysis of simple, Mendelian diseases such as Huntington disease, cystic fibrosis, and early-onset Alzheimer disease. Molecular biologists and geneticists alike now acknowledge that the most common human diseases with a genetic component are likely to have very complex etiologies. However, despite this belief, statistical geneticists continue using primarily traditional

methodologies to attack this complex problem. Traditional statistical methods of genetic analysis, such as linkage and association, have failed to consistently replicate findings of main effect genes, even though they may explain a majority of the genetic effect of a complex disease. For example, over 115 late-onset Alzheimer disease candidate genes have been tested and have generated a positive main effect, but all except apolipoprotein E (APOE) have failed to be consistently replicated (Pericak-Vance MA and Haines JL, 2002). Among the possible reasons for this failure are false positives due to population stratification and true differences in genetic etiology between study populations (Hirschhorn JN et al., 2002). Advances in statistical and computational genetic methodology simply have not kept pace with the advance of available sources of data. There have been a few attempts to address complexity directly, including the development of nonparametric tools, but these have generally limited application. One example is the transmission disequilibrium test that led to the discovery of the insulin receptor gene as a risk factor for diabetes (Spielman RS et al., 1993).

Going forward, statistical geneticists must not only acknowledge but also directly confront the numerous complicating factors that can be involved in complex genetic diseases and that present significant challenges for traditional statistical methods. Only a small fraction of the human genetics literature specifically reports on investigations of such complexity. It is, perhaps, daunting to consider multiple complicating factors, such as locus heterogeneity, trait heterogeneity, and gene-gene interactions (see Figures 1 and 2). However, these must be addressed if we are to have any chance of understanding the genetic legacy of disease left to us by our forebears.

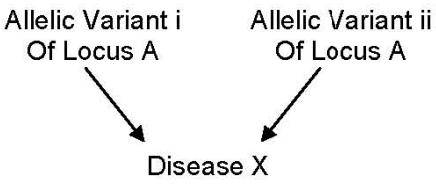
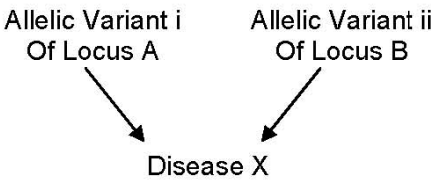
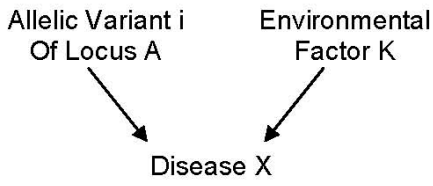
	Allelic Heterogeneity	Locus Heterogeneity	Phenocopy
Definition	when two or more alleles of a single locus are independently associated with the same trait	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	the presence of a disease phenotype that has a non-genetic (random or environmental) basis
Diagram			
Example 1	**Retinitis Pigmentosa** (RP, OMIM# 268000) - in the RHO gene (OMIM#180380), which accounts for 30-40% of autosomal dominant RP, over 100 distinct mutations have been found (Rivolta C et al., 2002; http://www.sph.uth.tmc.edu/RetNet)	**Retinitis Pigmentosa** (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model; still more have been associated with RP under autosomal dominant and X-linked disease models (Rivolta C et al., 2002; http://www.sph.uth.tmc.edu/RetNet)	**Parkinson Disease** (PD, OMIM# 168600) - individuals taking the illicit drug meperidien are sometimes exposed to its by-product MPTP, which causes the destruction of dopaminergic neurons (Langston JW et al., 1984; Langston JW and Ballard P, 1984) and produce the PD phenotype
Example 2	**Cystic Fibrosis** (CF) - over 1000 mutations in the CFTR gene (OMIM# 602421) have been associated with CF (Kulczycki LL et al., 2003)	**Tuberous Sclerosis** (TS, OMIM# 191100) - out of families informative for linkage analysis, half have mutations in the TSC1 gene (located at 9q34), and the other half have mutations in the TSC2 gene (located at 16p13) (Povey S et al., 1994 ; Young J and Povey S, 1998)	**Epilepsy** (OMIM#600669) - traumatic brain injury can result in posttraumatic epileptic seizures occurring within 24 hours or up to several years after the injury (Frey LC, 2003)

Figure 1. Heterogeneity-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples

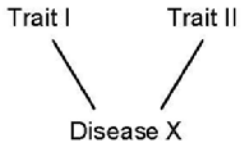
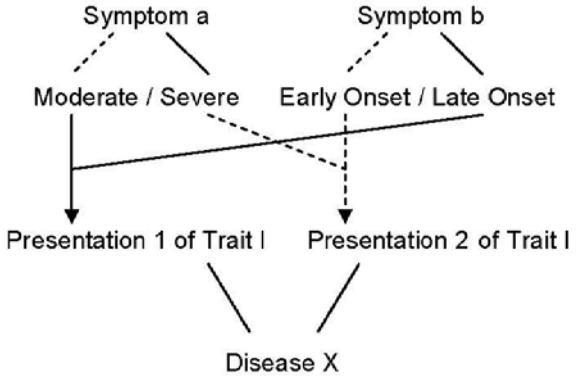
	Trait Heterogeneity	Phenotypic Variability
Definition	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	variation in the degree, or severity, or age of onset of symptoms exhibited by persons who actually have the same trait or disease process
Diagram		
Example 1	**Autosomal Dominant Cerebellar Ataxia** (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms (Harding AE, 1993; Rosenberg RN, 1995), and different genetic loci have been associated with the different subtypes (Devos D et al., 2001)	**Holoprosencephaly** (HPE, OMIM# 236100) - craniofacial abnormalities associated with HPE can range from a single middle incisor to cyclopia
Example 2	**Autism** (OMIM# 209850) - parents and other relatives of autistic individuals often exhibit one or two, but not all three, of the requisite autistic symptomatologies, suggesting autism may be the co-occurrence of three distinct traits (Tager-Flusberg H and Joseph RM, 2003) using subset analysis, some success has been achieved identifying genes associated with one of the three symptomatologies but not as strongly with the broader autistic phenotype (Bradford Y et al., 2001; Shao Y et al., 2002)	**Tuberous Sclerosis** (TS, OMIM# 191100) - the severity of such TS symptoms as mental retardation, kidney disease and facial angiofibroma differ across affected individuals (Lendvay TS and Marshall FF, 2003)

Figure 1, continued. Heterogeneity-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples

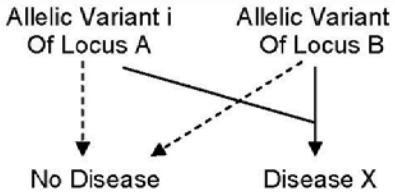
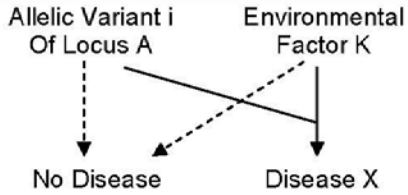
	Gene-Gene Interaction	Gene-Environment Interaction
Definition	when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects	when a DNA variation interacts with an environmental factor, such that their combined effect is distinct from their independent effects
Diagram		
Example 1	**Hirschsprung Disease** (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants (Carrasquillo MM et al., 2002)	**Bovine Spongiform Encephalopathy** (BSE) - all human cases of BSE, which is commonly known as "mad cow disease" and is transmitted through consumption of contaminated feef, were in individuals who were homozygous for the met129 polymorphism in the PRNP gene (OMIM# 176640). In a cluster of unreported cases of BSE in which individuals had been exposed to contaminated brain electrodes, all but one individual was 129met/met; the remaining person was heterozygous for the polymorphism and had a more protracted course than the others (Aguzzi A and Weissman C, 1996; Collinge J et al., 1991)
Example 2	**Creutzfeldt-Jackob Disease** (CJD, OMIM# 123400) and Fatal Familial Insomnia (OMIM# 176640.0010) - the Met129Val polymorphism and Asp178Asn mutation in the PRNP gene (OMIM# 176640) interact, such that when the val129 polymorphism is on the same chromosome as the asn178, the phenotype is fatal familial insomnia (Doh-ura K et al., 1989; Owen F et al., 1990; Collinge J et al., 1991; Palmer MS et al., 1991)	**Unipolar Depression** (OMIM# 608516) - individuals with one or two copies of the short allele of the serotonin transporter (5-HT T) promoter polymorphism were up to two times more likely to develop depressive symptoms, diagnosable depression and suicidality after stressful life events, than individuals homozygous for the long allele, suggesting a gene-environment interaction (Caspi A et al., 2003)

Figure 2. Interaction-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples

Categorization and Analytical Approaches

Each of the factors presented in Figures 1 and 2 complicate statistical analysis in one of two ways—either by creating heterogeneous, or competing, disease models (Figure 1), or else by creating a multifactorial, interacting disease model (Figure 2). The challenge for modeling the relationship between genetic and environmental risk factors (independent variables) and disease endpoints (dependent variables) is different for these two categories. Of course, what exacerbates the complexity is that none of these competing and interacting models are mutually exclusive. Various combinations of (genetic and/or trait) heterogeneity and interactions might be important in any given disease of interest. Thus, to dissect these factors, we must assemble a toolbox of both tried-and-true and newly constructed genetic analysis methodologies, which together can be used to discover the true underlying etiologies of complex traits.

Many complicating factors can be addressed proactively by a well-considered study design. This is perhaps one of the best investments researchers can make to maximize their ability to discover complex genetic disease models. Because the causally complex relationship between the genotype and phenotype is the object of genetic studies, it is important to collect accurate and abundant phenotypic data. In the absence of phenotypic data, there is not even the option of looking for a mapping between genotype and potential clinical subtypes, which could help identify a case of genetic heterogeneity. Established guidelines or protocols concerning data collection should be followed and such data should be made available to others in an accessible format, so as to facilitate future meta-analysis. Information regarding the exposure to potential environmental risk factors should be collected whenever logistically and economically

feasible. Even with the best study design with regard to data collection, an ill-advised or incomplete analysis of the data can still yield disappointing, if not incorrect, results. Thus, we advocate a comprehensive approach to account for both the heterogeneity and the interaction models of disease.

Heterogeneity

For this category of factors, there are multiple independent (predictor) variables or else multiple dependent (outcome) variables that complicate the analysis by creating a heterogeneous model landscape. In the case of allelic or locus heterogeneity or phenocopy, multiple predictor variables (e.g. multiple alleles, multiple loci and/or environmental risk factors) are present, some of which might be unmeasured or unobserved and, therefore, unavailable for inclusion in the disease model. In the case of trait heterogeneity or phenotypic variability, multiple outcome variables are present, which cannot or have not been distinguished based on the available phenotypic information.

Perhaps the most straightforward of the methods for addressing heterogeneity is sample stratification (Figure 3). This method subdivides subjects based on any number of genetic, demographic, clinical or environmental factors to create more homogeneous subsets of the data. The premise of this method is that there are two or more underlying disease models, which are conditional on the factor on which the data are being stratified. For example, one genetic model might be associated with disease in the absence of a specific environmental risk factor; however, when that environmental factor is present, a different set of genetic factors are involved. Using different levels of the stratifying

	Abbreviated Description	Example
Sample Stratification	Manual sorting by covariate(s)	Stratification by age-of-onset led to the confirmation of APP as a risk factor for early-onset Alzheimer Disease (Goate A et al., 1991)
M Test	Manual sorting by covariate(s) + testing difference in recombination fractions across families	Locus heterogeneity across French families with Usher Syndrome was confirmed by the M-test (Larget-Piet D et al., 1994)
Beta Test	Very similar but slightly more powerful statistical test than the M test	Locus heterogeneity between the F9 and fra(x) loci was confirmed with a near maximal result on the Beta test (Risch N, 1988; Brown WT et al., 1987)
Admixture Test	Estimates population admixture present in the sample and evaluates probability of heterogeneity	Evidence for linkage was significantly increased in the 9q32-34 region of interest for tuberous sclerosis when significant heterogeneity was shown with admixture test and only linked families (with 80% posterior probability) were examined (Haines JL et al., 1991)
Ordered Subset Analysis (OSA)	Ordering of families by covariate(s) and calculation of maximum cumulative lod score	Evidence for linkage was significantly increased in two regions of interest for macular degeneration risk genes (14q13 and 6q14) using intraocular pressure and body mass index as covariates in OSA (Schmidt S et al., 2004)
Cluster Analysis	Clustering of individuals to produce subgroups with high intraclass similarity and low interclass similarity	Clustering by pedigree-specific genomic markers led to identification of linkage region of interest not found after sample stratification by self-reported ethnicity (Grigull J et al., 2001)
Latent Class Analysis and Factor Analysis	Similar to cluster analysis, except that "latent" or underlying variables are derived from relationships among known covariates	Factor analysis derived an "insistence on sameness (IS)" factor from the Autism Diagnostic Interview-Revised, which was successfully used as a covariate in OSA to narrow a region of interest for autism susceptibility gene (Shao Y et al., 2003)

Figure 3. Summary of analytical approaches to heterogeneity

factor (e.g. different degrees of environmental exposure), one could perform further analyses, such as logistic regression (discussed in the following section). The main limitation of sample stratification is a reduction in sample size within each stratum and thus a reduction in power.

Some statistical methods that test the hypothesis of locus heterogeneity include the M test (Morton, 1955), the β test (Risch N, 1988) and the Admixture test (Figure 3) (Ott, 1992; Smith CAB, 1963). Each of these methods is solely applicable to family-based data on which linkage analysis is performed. The M test uses *a priori* stratification of subjects based on discrete (or discretized) covariates, such as gender, ethnicity or clinical subtype, and tests for a difference in recombination fractions across the different subsets of families. The β test is a similar but slightly more powerful statistical test than the M test, owing to a difference in their null distributions used to determine statistical significance. The admixture test does not require *a priori* stratification but instead estimates (using maximum likelihood) the degree of admixture present in the sample from two-point or multi-point lod scores between marker and disease loci. It then uses these estimates to evaluate the relative probabilities of linkage with and without heterogeneity. Thus, the M and β tests evaluate a more specific hypothesis, and as a result, have more power than the admixture test. The admixture test also lacks sensitivity and can only account for, not resolve, the underlying heterogeneity.

A more recently developed method to address heterogeneity is the ordered subset analysis (OSA; Figure 3) (Hauser et al., 1998; Hauser et al., 2004). In OSA, a continuous or ordinal covariate, such as blood lipid levels or disease age of onset, is used to rank order families, and then a cumulative lod score is iteratively calculated after each family

is added (in order) to the sample until the cumulative lod score begins to decrease. Thus, those families included in the linkage analysis all provide support for linkage, and the subset of chosen families is more homogeneous with respect to the covariate and, therefore, hopefully, more genetically homogeneous than the whole dataset.

Other methods aimed at producing more homogeneous subsets of the data include cluster analysis, latent class analysis and factor analysis (Figure 3). Unlike the aforementioned statistical tests for heterogeneity that only incorporate linkage analysis, the following methods can also be applied to case-control datasets because they are not tied to any particular statistical analysis of the subsets. There are hundreds of different cluster analysis methods, which operate based on different heuristics and fitness metrics, making them appropriate for particular types of data (continuous versus discrete, low- versus high-dimensional, and so on). They all attempt to produce clusters with high intraclass similarity and/or low interclass similarity and have varying degrees of success. Cluster analysis has been widely used for analyzing DNA and protein microarray data (Slonim DK, 2002) and to find more homogeneous subgroups based on genetic background (Mountain JL and Cavalli-Sforza LL, 1997).

Latent class analysis and factor analysis have a goal similar to cluster analysis but instead of directly clustering or classifying data based on known covariates, such as the scores of different items on a psychological or physical functioning test, these two methods try to derive 'latent' or underlying variables, such as summary scores of various test items, from relationships among the known covariates. These latent variables are then used to classify or stratify the data. Latent class analysis has been applied to phenotypic data for several diseases, including attention deficit hyperactivity disorder

(Neuman RJ et al., 1999), Alzheimer's disease (Neuman RJ et al., 2000), autism (Pickles A et al., 1995) and schizophrenia (Sham PC et al., 1996).

It should be noted that all of the methods discussed previously, with the exception of the admixture method, depend on covariate data, whether these be known genetic risk factors, demographic data, phenotypic data or endophenotypes. Not only must such information be available but also these covariates must actually be relevant to, or be surrogates for, the existing heterogeneity. If the data are incomplete, the performance of many of these methods for dissecting heterogeneity suffers and attempts to correct this problem by imputing data can introduce spurious associations. In the absence of such relevant, complete data, we are left with seemingly few options of how to proceed when we suspect heterogeneity to have a role.

To overcome some of these problems it might be advantageous to adapt the same basic principles of the aforementioned methods to the more complex data. For instance, although clustering methods have been heavily utilized for microarray data, few studies have looked into clustering genotypic data from association-based studies to identify multilocus patterns that characterize particular subsets of the data. Some clustering methodologies appropriate for such discrete data include hypergraph clustering (Han EH et al., 1997a), Bayesian classification (Hanson R et al., 1991) and fuzzy *k*-modes clustering (Huang Z and Ng MK, 1999).

Interactions

Gene–gene and gene–environment interactions are two complex genetic factors (Figure 2) that create a rugged model landscape for statistical analysis. There is clear and

convincing evidence that gene–gene interactions, whether synergistic or antagonistic, are not only possible but also are probably ubiquitous (Moore JH, 2003; Tong AH et al., 2004). Similarly, gene–environment interactions are likely to be discovered if properly investigated. Thus, it is crucial that complex genetic datasets be properly interrogated for possible underlying interactions.

Analytically it can be difficult to distinguish between heterogeneity and interactions. Many of the methods that address heterogeneity might be equally applicable to uncovering interactions. For instance, the discovery of linkage to a particular locus in only one subset of data produced by sample stratification could be indicative of heterogeneity, or it could be indicative of an interaction between the locus and the covariate used to stratify the data. However, there is also an entirely different set of methods that are particularly well suited to discovering interactions (but not heterogeneity; Figure 4).

One traditional approach still widely used today is regression. In particular, logistic regression is used when the outcome variable is discrete, for example, disease status (i.e. you either have the disease or you do not) (Figure 4). Logistic regression enables direct modeling of the mathematical relationship of genetic and other risk factors to disease status. However, this 'workhorse' suffers from the curse of dimensionality, meaning that as the distribution of data across numerous combinations of factors becomes sparse, the parameter estimates become unreasonably biased, particularly when the ratio of sample size to independent variables is below ten to one (Concato et al., 1993; Moore JH and Williams SM, 2002; Peduzzi P et al., 1996). Thus, when considering a combination of loci, one or more of which have low minor allele frequencies, the number

	Description	Example
Logistic Regression	Mathematical modeling of relationship of discrete genetic and other risk factors to disease status	Three two-locus Interactions among three folate-related genes were found by logistic regression to increase risk of neural tube defects (ReltonCL et al., 2004)
Focused Interaction Testing Framework (FITF)	Likelihood ratio tests of interaction are performed after prescreening with chi-square goodness-of-fit test	A significant multilocus effect between the NQO1, MPO and CAT genes involved in the oxidative stress pathway were found to be associated with asthma in multiple independent datasets (Millstein J et al., 2005)
S Sum Statistic (Set Association analysis)	Selects the 'best' set of SNPs, whose summary statistic is statistically significant	A significant nine-SNP interaction among 62 candidate genes was found to be associated with restenosis after angioplasty (Hoh J et al., 2001; Zee RY et al., 2002)
Linear Regression	Mathematical modeling of relationship between continuous outcome variable(s) and genetic risk factors to disease status	The nonmodulating hypertension phenotype was found to be associated with an interaction among the angiotensinogen, angiotensin-converting enzyme, and aldosterone synthase genes (Kosachunhanun N et al., 2003)
Multivariate Adaptive Regression Splines (MARS)	Generalization of stepwise linear regression particularly suited for high-dimensional problems with many independent variables	An interaction between two inflammation-related genes—the P-selectin and interleukin-4 genes—was found by MARS to be associated with ischemic stroke (Cook NR et al., 2004)
Classification and Regression Trees (CART)	Iteratively subdivides data to build a hierarchical classification model	Evidence of linkage to cardiovascular disease traits was strengthened in behaviorally-distinct subgroups constructed by CART (Costello TJ et al., 2003)
Bayesian Belief Network	Probabilistic reasoning system that builds a hierarchical model of interactions	A network of SNPs and microsatellites in candidate genes was found to predict cervical cancer with reasonable specificity (Horng JT et al., 2004)
Combinatorial Partitioning Method (CPM)	Utilizes data reduction to investigate gene-gene interactions	CPM found evidence for non-additive effects of the ACE and PAI-1 genes, in addition to additive effects found by linear regression, in the prediction of plasma PAI-1 levels (Moore JH et al., 2002)
Restricted Partition Method (RPM)	Modification to CPM which heuristically restricts the search for high-level interactions	This new method was used to analyze 10 quantitative measures and 10 SNPs in candidate genes related to irinotecan metabolism but did not identify any significant associations (Culverhouse R et al., 2004)
Multifactor Dimensionality Reduction (MDR)	Utilizes data reduction to investigate gene-gene interactions	An interaction between two SNPs in the DNA repair gene *XPD* and smoking status was found by MDR to be associated with bladder cancer (Andrew et al., 2006)
Artificial Neural Networks	Utilizes pattern recognition to find models for disease risk with multiple gene-gene interactions	Analysis of Type 1 diabetes mellitus data reproduced previously reported results showing highest lod scores for the IDDM1 and IDDM2 loci (Lucek P et al., 1998)

Figure 4. Summary of analytical approaches to interactions

of individuals with certain multilocus genotype combinations will be so small (or perhaps equal to zero), that one cannot reasonably estimate, or generalize to the population, what is the disease risk for that combination of genotypes. Missing or incomplete data can also create or exacerbate the problem of sparse data. In addition, many standard approaches to implementing logistic regression, such as forward stepwise regression, require significant main effects to be modeled before including interaction effects between factors. This is a major methodological limitation for situations where each locus has relatively small main (non-interactive) effects but more substantial interactive effects because none of those interactive effects would ever be considered.

A more recently developed statistical method for evaluating gene–gene interactions is the focused interaction testing framework (FITF) (Millstein et al., 2006). This method is applicable to case-control data and uses likelihood ratio tests on increasingly greater orders of interaction between genes. To reduce the number of interactions tested, a prescreening step is applied in which a goodness-of-fit chi-square statistic is used to detect association among candidate genes in the pooled case-control data. Multiple testing is addressed by controlling false discover rates. This method is reported to have better power to detect interactions than Multifactor Dimensionality Reduction (MDR, discussed below) when the genes involved have recessive, dominant or additive effects (Millstein et al., 2006). However, the reported difference in power may be attributable to the particular implementation of MDR, which differs from that recommended by MDR's authors, and to a disconnect between how the methods determine the success of an analysis of simulated data.

Another recently developed method for gene-gene interactions is the S-sum statistic, which is designed to overcome the curse of dimensionality and the multiple-testing problems by reducing any number of independent variable statistics into one sum statistic and then using permutation testing to correct for an experiment-wise Type I error rate, which is the probability of concluding that there is an effect when one does not actually exist (Hoh J et al., 2001; Ott J and Hoh J, 2003). 'Set association' analysis is the authors' term for the application of the S statistic to SNP marker data from candidate genes or regions (Figure 4). This method selects the 'best' set of *n* number of single nucleotide polymorphisms (SNPs), whose S_n statistic is statistically significant, leading to the inference that the entire set of SNPs might be interacting in some way to increase disease risk, or else that they are all contributing independently to disease risk. However, because the summed statistics are all single-marker statistics, set association analysis does not look at any specific (non-additive) interactive effects among markers and would be likely to miss nonlinear or antagonistic types of gene–gene interactions. This method has successfully identified a set of seven SNPs, which together were associated with restenosis incidence ($P < 0.0001$) and explained over 11% of the overall variance (Zee RY et al., 2002). In theory the S statistic can be used with any number of test statistics on discrete or continuous data, but its applications and limitations are still being evaluated (Wille A et al., 2003).

When the outcome variable is continuous, as is the case for a quantitative trait locus (QTL), such as serum prolactin levels, linear regression can be used to model the relationship between risk factors and QTL status (Figure 4). However, linear regression faces the same limitations logistic regression does regarding parameter estimation and

modeling interactions. Cheverud and Routman (Cheverud JM and Routman EJ, 1995) developed an alternative parameterization of gene–gene interactions based on its effects on genetic variance components (additive, dominance and interaction); however, it is limited to evaluating only two loci at a time and all possible genotypes must be present in the sample.

Multivariate adaptive regression splines (MARS) (Cook NR et al., 2004; Friedman J, 1991) is a generalization of stepwise linear regression that is particularly suited for high-dimensional problems in which many independent variables might be modeled. MARS is also similar to classification and regression trees (CART) (Cook NR et al., 2004; Morgan JN and Sonquist JA, 1963; Province MA et al., 2001; Shannon WD et al., 2001), which iteratively subdivide data to build a hierarchical classification model. A Bayesian belief network (BBN) (Good IJ, 1961) is a probabilistic reasoning system that builds a topological (but necessarily hierarchical) model of interactions (joint probabilities) (Figure 4). BBN, CART and MARS all suffer from the same problem of sequential conditioning that can plague many other regression-based methods, which makes it difficult to discover interactions (especially higher-order interactions) among predictor variables, depending on the strength of their individual (or lower-order interaction) effects. The binary nature of CART further limits its ability to model any additive interaction. Still, the most troubling limitation that plagues all these methods is their inability to model, much less discover, nonlinear interactions.

Two types of computational methods—data reduction and pattern recognition—that come from the computer science field offer the potential for uncovering such nonlinear interactions, with increased tolerance for missing or incomplete data (Figure 4).

Nelson *et al.* (Nelson MR et al., 2001) developed a combinatorial partitioning method (CPM) that utilizes data reduction to investigate gene–gene interactions. CPM has shown success in building multilocus models with nonlinear interactions to explain and predict variability in plasma triglyceride (Nelson MR et al., 2001) and plasma plasminogen activator inhibitor 1 levels (Moore JH et al., 2002). Culverhouse *et al.* (Culverhouse R et al., 2004) developed a modification of the CPM method, the restricted partition method (RPM), which heuristically restricts the exhaustive search used in CPM and thereby reduces its computational load for evaluating interactions. Multifactor dimensionality reduction (MDR) is one data reduction method developed specifically for genotypic data that has been successful at finding gene–gene interactions in both simulated data (Hahn LW and Moore JH, 2004; Hahn et al., 2003; Ritchie MD et al., 2001; Ritchie MD et al., 2003) and real data (Ashley-Koch et al., 2006; Cho YM et al., 2004; Ma et al., 2005; Qin et al., 2005; Ritchie et al., 2001; Tsai CT et al., 2004; Williams SM et al., 2004). Artificial neural networks perform pattern recognition and have been applied to genotypic data with varied success (Lucek P et al., 1998; Marinov M and Weeks D, 2001; McCulloch W and Pitts W, 1943; Sherriff A and Ott J, 2001). However, recent work has improved the reliability of artificial neural networks through their optimization by evolutionary computation (EC) algorithms (Fogel GB and Corne DW, 2002), which use a computational search methodology uniquely suited for rugged model landscapes (Ritchie et al., 2003b). One limitation of these computational methods is the potential difficulty of interpreting the biological implications of the resulting predictive models (Moore JH and Ritchie MD, 2004; Moore and Williams, 2002).

Retooling for the Future

None of the aforementioned methodologies is superior in all respects for the range of complicating factors that might be present in any given dataset. Given the relative shortcomings of our current analyses in complex diseases, we need to extend greatly the range of available analytical tools. There is a crucial need for extensive reevaluation of existing methodologies for complex diseases, as well as for massive efforts in new method development. It is important that empirical studies be conducted to compare and contrast the relative strengths and weaknesses of methods on specific types of problems. For example, although cluster analysis has shown promise in numerous other scientific and mathematical fields, its use with genetic, particularly discrete genotypic data, has not been adequately explored. Similarly, artificial neural networks modified with evolutionary computation have great potential for discovering nonlinear interactions among genes and environmental factors. However, work is still ongoing to evaluate its limitations with regard to the heritability and effect sizes that can be detected.

Ultimately, the real power of existing and yet-to-be-developed methods lies in our ability to marry them into a comprehensive approach to genetic analysis, so that their relative strengths and weaknesses can be balanced and few alternative hypotheses are left uninvestigated. We propose routinely taking a two-step approach to analysis because no single method adequately investigates heterogeneity and interaction issues simultaneously. For example, clustering or ordered subset analysis can be used first to uncover genotypic and/or phenotypic heterogeneity and to subdivide the data into more homogeneous groups. Then in a second step, specific tests of interactions, such as the S sum statistic approach or the multifactor dimensionality reduction method can be used to

investigate gene–gene or gene–environment interactions within each of the homogenized subgroups. This is still not a perfect approach, but it is an important improvement over the more common alternative of a single-pronged approach to analysis.

Such a combined strategy must be the future of genetic statistical analysis. We must harness our knowledge and experience of existing methods even as we open our minds to newly fashioned techniques and approaches. By thus 'retooling' our analyses, we provide the best opportunity for uncovering the genetic basis of common human disease.

CHAPTER III

A COMPARISON OF CLUSTERING METHODS

Adapted and expanded from previous work completed for

Masters Thesis in Biomedical Informatics (2005) and published as follows:

Thornton-Wells TA, Moore JH, Haines JL. Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. BMC Bioinformatics 7:204, 2006.

Background

For over 30 years, cluster analysis has been used as a method of data exploration (Anderberg MR, 1973). Clustering is an unsupervised classification methodology, which attempts to uncover 'natural' clusters or partitions of data. It involves data encoding and choosing a similarity measure, which will be used in determining the relative 'goodness' of a clustering of data. No one clustering method has been shown universally effective when applied to the wide variety of structures present in multidimensional datasets. Instead, the choice of suitable methods is dependent on the type of target data to be analyzed. Clustering has been utilized widely for the analysis of gene expression (e.g., DNA microarray) data; however, its application to genotypic data has been limited (Slonim DK, 2002).

Most traditional clustering algorithms use a similarity metric based on distance that may be inappropriate for categorical data such as genotypes. Newer methods have been developed with categorical data in mind and include extensions of traditional methods

and application of probabilistic theory. Three such methods were chosen (as discussed in a subsequent section) to compare in the task of discovering trait heterogeneity using multilocus genotypes—Bayesian Classification (Hanson R et al., 1991), Hypergraph-Based Clustering (Han EH et al., 1997a), and Fuzzy *k*-Modes Clustering (Huang Z and Ng MK, 1999)—all of which are appropriate for categorical data.

Methods

Data Simulation

To compare the performance of clustering methodologies in the task of uncovering trait heterogeneity in genotypic data, datasets were needed in which such heterogeneity was known to exist. Since there are no well-characterized real datasets available that fit this description, a simulation study was needed. Genetic models that contained two binary disease-associated traits, such that there is trait heterogeneity among 'affected' individuals, were used. In addition, some of the models incorporate locus heterogeneity, a gene-gene interaction, or both. Figure 5 depicts the structure of the four genetic models used to simulate the genotypic data.

Four prevalence levels were simulated for each genetic model: (1) fifteen percent, which is characteristic of a common disease phenotype such as obesity (Flegal KM et al., 1998), (2) five percent, which is characteristic of a relatively common disease such as prostate cancer (Narod SA et al., 1995), (3) one percent, which is

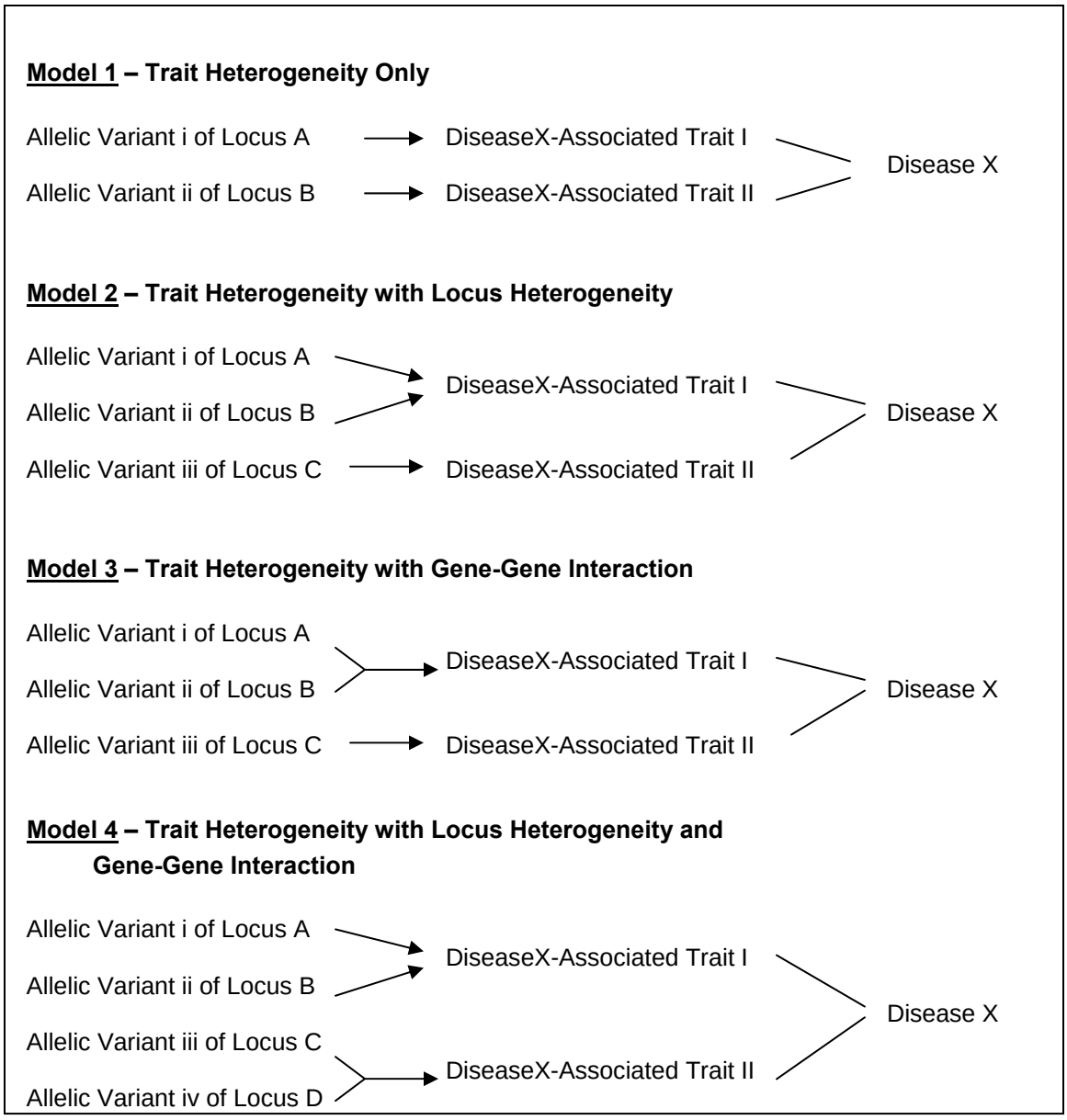


Figure 5. Structure of Genetic Models Used for Data Simulation

characteristic of a less common disease such as schizophrenia (Schultz S and Andreasen N, 1999), and (4) one tenth of one percent, which is characteristic of a more uncommon disease such as multiple sclerosis (Kurtzke JF, 1991). Three realistic levels of sample

size were simulated for each model: 200, 500 and 1000 affected individuals. Finally, four levels of non-functional loci were simulated: 0, 10, 50 and 100. The inclusion of non-functional loci adds a random noise effect that is present in real candidate gene studies in which the functional locus or loci are among many more suspected but actually non-functional loci. All loci, including the functional loci, were simulated to have equal biallelic frequencies of 0.5.

Although the above parameter settings are by no means exhaustive of the biologically plausible situations, the outlined conditions are reasonable and specify 192 different sets of data specifications due to the combinatorial nature of the study design. To have adequate power to detect a difference in performance among clustering methodologies, 100 datasets per set of parameters were simulated, resulting in a total of 19,200 simulated datasets.

For the purposes of simulating these data, a novel data simulation algorithm capable of incorporating these complex genetic factors in an epidemiologically-sound manner was designed and developed (Figure 6). Penetrance is the probability of having a particular phenotype given a specific genotype (single or multilocus). Prevalence, on the other hand, is the percentage of individuals in a population that have a particular phenotype. The penetrance levels of the two simulated disease-associated traits are constrained by the overall prevalence level of the simulated disease. The two traits were simulated to contribute equally to the prevalence of the associated disease (fifty percent trait heterogeneity), such that a small but naturally occurring degree of overlap would be present, representing individuals having both disease-associated traits, instead of just one or the other. These penetrance tables are inputs for the new data simulation algorithm.

Penetrance Function Array: each cell value represents the probability of having the disease-associated trait, given the (multilocus) genotype

Unaffecteds Probability Array: each cell value represents the probability of having the multilocus genotype given that the disease status is unaffected, which is the probability of being negative for all traits, or the joint probability of being negative for each trait, given the genotype frequency (prior probability)

Affecteds Probability Array: each cell value represents the probability of having the multilocus genotype given that the disease status is affected, which is the probability of being positive for at least one trait, which is the same as 1 – probability of being negative for all traits, or 1- joint probability of being negative for each trait, given the genotype frequency (prior probability)

Pseudocode:

1. Allocate two probability arrays, one for Affecteds and one for Unaffecteds, each of size $\prod_{i=1}^{L} \sum_{j=1}^{A_i} j$ where L is the total number of loci and A_i is the number of alleles for locus i.

2. For each penetrance function p(Status=Affected | Multilocus Genotype)
   ==>Distribute 1-p across relevant cells of Unaffecteds probability array

3. Populate cells of the Affecteds probability array with 1-(cell probability) of corresponding cells of the Unaffecteds probability array

4. For each locus
   ==>Distribute allele frequencies across appropriate cells of both probability arrays

5. Generate the specified number of unaffected individuals from the Unaffecteds probability array

6. Generate the specified number of affected individuals from the Affecteds probability array

7. Determine the status of each disease-associated trait for each affected individual thus…. If the affected individual has a high-risk genotype combination for that disease-associated trait, then that individual is affected for that trait. Otherwise, the individual is unaffected for that disease-associated trait. (By design, each affected individual will be affected at one or more disease-associated traits.)

Figure 6. Novel Data Simulation Algorithm. Simulates trait heterogeneity, locus heterogeneity and gene-gene interactions in an epidemiologically-sound manner. The inputs are penetrance function arrays, which are translated into probability arrays for affecteds and unaffecteds, separately. Then affected and unaffected individuals (with multilocus genotypes) are simulated from those respective arrays.

For one fourth of the models, trait heterogeneity only is involved (not locus heterogeneity or gene-gene interactions), and there is one genetic risk factor for each of the two traits. Each locus acts in a recessive manner, such that affected individuals have both copies of the high-risk allele at the disease-associated "functional" locus (Figure 7). A naturally occurring degree of overlap between the two traits can result, such that some affected individuals have the high-risk genotypes for both traits.

(a)

1A1A	1A1B	1B1B
0	0	x

(b)

2A2A	2A2B	2B2B
0	0	x

Figure 7. Genetic Model THO (Trait Heterogeneity Only)
The penetrance tables for Trait I (a) and Trait II (b) are presented. Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype. For each of the two traits, a Mendelian recessive genetic model is used, in which the trait is penetrant only when two copies of the high risk (B) allele are present. The penetrance (x) is constrained by the desired overall disease prevalence to be simulated (0.001, 0.01, 0.05 or 0.15).

In the second quarter of the datasets, a locus heterogeneity model described by Li and Reich (Li WT and Reich J, 2000) was also simulated (Figure 8b) so that for one of the traits, there are two associated loci, each of which is responsible for roughly half of the individuals affected with the trait. In that locus heterogeneity model, each of the

functional loci acts in a recessive manner, such that the disease-associated genotype for the locus consists of two copies of one high-risk allele. For the other trait, a recessive model was implemented, as described above (Figure 8a). By chance, there might be some affected individuals who have the high-risk genotype from the first trait as well as one of the high-risk genotypes from the second trait.

(a)

1A1A	1A1B	1B1B
0	0	x

(b)

	2A2A	2A2B	2B2B
3A3A	0	0	x
3A3B	0	0	x
3B3B	x	x	x

Figure 8. Genetic Model THL (Trait Heterogeneity with Locus Heterogeneity)
The penetrance tables for Trait I (a) and Trait II (b) are presented. Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype. For Trait I, a Mendelian recessive genetic model is used, in which the trait is penetrant only when two copies of the high risk (B) allele are present. For Trait II, a locus heterogeneity model described by Li and Reich (Li WT and Reich J, 2000) is used, in which the trait is penetrant only when two copies of the high risk allele at one or both loci are present (in this case the B alleles for locus 2 and 3 are high risk).

In the third quarter of the datasets, a gene-gene interaction was simulated for one of the two traits. The "diagonal" gene-gene interaction model, first described by Frankel and Schork (Frankel WN and Schork NJ, 1996) and later by Li and Reich (Li WT and

Reich J, 2000), which is nonlinear and nonadditive in nature, was used (Figure 9b). Under this model, a multilocus genotype is high-risk if it has exactly two high-risk alleles from either of the two associated loci. A multilocus genotype with fewer than or greater than two high-risk alleles is not associated with disease. For the other trait, a recessive model was implemented, as described above (Figure 9a). By chance, there might be some affected individuals who have the high-risk genotype from the first trait as well as one of the high-risk genotypes from the second trait.

(a)

1A1A	1A1B	1B1B
0	0	x

(b)

	2A2A	2A2B	2B2B
3A3A	0	0	x
3A3B	0	0.5x	0
3B3B	x	0	0

Figure 9. Genetic Model THG (Trait Heterogeneity with Gene-Gene Interaction) The penetrance tables for Trait I (a) and Trait II (b) are presented. Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype. For Trait I, a Mendelian recessive genetic model is used, in which the trait is penetrant only when two copies of the high risk (B) allele are present. For Trait II, the "diagonal" genetic model first described by Frankel & Schork (Frankel WN and Schork NJ, 1996) and later by Li and Reich (Li WT and Reich J, 2000) is used. Two loci (2 and 3) are involved, each with two alleles (A and B), and the trait is penetrant only when *exactly* two copies of the high risk allele from either locus are present.

In the fourth quarter of the datasets, one trait is simulated to involve locus heterogeneity (Figure 10a), while the other is simulated to involve the "diagonal" gene-

gene interaction, as described above (Figure 10b). Thus, there are some affected individuals who, by chance, will have one high-risk genotype from the first trait as well as one high-risk genotype from the second trait.

(a)

	1A1A	1A1B	1B1B
2A2A	0	0	x
2A2B	0	0	x
2B2B	x	x	x

(b)

	3A3A	3A3A	3A3A
4A4A	0	0	x
4A4B	0	0.5x	0
4B4B	x	0	0

Figure 10. Genetic Model THB (Trait Heterogeneity with Both Locus Heterogeneity and Gene-Gene Interaction). The penetrance tables for Trait I (a) and Trait II (b) are presented. Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype. For Trait I, a locus heterogeneity model described by Li and Reich (Li WT and Reich J, 2000) is used, in which the trait is penetrant only when two copies of the high risk allele at one or both loci are present (in this case the B alleles for locus 2 and 3 are high risk). For Trait II, the "diagonal" genetic model first described by Frankel & Schork (Frankel WN and Schork NJ, 1996) and later by Li and Reich (Li WT and Reich J, 2000) is used. Two loci (2 and 3) are involved, each with two alleles (A and B), and the trait is penetrant only when *exactly* two copies of the high risk allele from either locus are present.

The input file for each of the clustering methods, which are described below, includes genotype and trait status information. Each row is a single individual. Column headings include unique individual number, trait status (affected for Trait 1, Trait 2, or

both), and all simulated loci. Genotypes for each locus are encoded nominally (not ordinally), such that no genetic model assumptions are incorporated. Loci are numbered, and alleles are lettered. Thus, for a given locus '3' that has two alleles 'A' and 'B', the three possible genotypes are '3A3A', '3A3B', and '3B3B'. A different nomenclature could easily be used, however, since the methods simply treat each genotype as a character string for labeling purposes only and do not attribute any meaning or order to them.

Clustering Methods

There exists a very large number of clustering algorithms and even more implementations of those algorithms. The choice of which clustering methodology to use should be determined by the kind of data being clustered and the purpose of the clustering (Kaufman L and Rousseeuw PJ, 1990). Genotypic data are categorical, which immediately narrows the field of appropriate methods for this study to only a few. Three different clustering methodologies were chosen that are suitable for categorical data and are appealing due to their speed or theoretical underpinnings.

The goal of this cluster analysis is to find a partitioning of the affected individuals based on multilocus genotypic combinations that maps onto the trait heterogeneity simulated in the data. For example, consider a dataset with 10 loci (numbered 1 to 10), each of which has two alleles (A and B), such that at each locus there are three possible genotypes (AA, AB and BB). It is likely that among affected individuals in the dataset, subsets of individuals will share specific genotypes or multilocus combinations of genotypes (such as 2B2B; or 3A3B and 9A9B together), either by chance or because such

combinations are related to genetic background, phenotypic variability, or trait heterogeneity in some way. Thus, a successful clustering would be one in which all the individuals who were simulated to have Trait I end up in one or more clusters that do not have any individuals unaffected for Trait I and all individuals who were simulated to have Trait II end up in one or more distinct clusters that do not have any individuals unaffected for Trait II (Figure 11). (Those individuals, who by chance have both Trait I and Trait II, could be 'correctly' placed in any cluster.) Such a clustering would effectively eliminate the noise present among affected individuals due to trait heterogeneity. In the case where locus heterogeneity is also simulated, an even more successful clustering would be one in which there are two or more Trait II clusters, each of which has only those individuals who have a specific high-risk genotype (e.g., 2B2B from Figure 12) and none that do not.

Bayesian Classification

The first clustering method is Bayesian Classification (Cheeseman P and Stutz J, 1996; Hanson R et al., 1991). The corresponding AutoClass software is freely available from Peter Cheeseman at the NASA Ames Research Center. Bayesian Classification (BC) aims to find the most probable clustering of data given the data and the prior probabilities. In the case of genotypic data, prior probabilities are based on genotype frequencies, which for the purpose of the proposed data simulations are set in accordance with Hardy-Weinberg equilibrium and equal biallelic frequencies of 0.5. The most probable clustering of data is determined from two posterior probabilities. The first involves the probability that a particular individual belongs to its 'assigned' cluster, or

(a)

	Locus										Trait	
Indiv	**1**	**2**	3	4	5	6	7	8	9	10	**1**	**2**
1	**BB**	AB	AB	AB	AB	AA	AB	BB	AB	BB	**X**	
2	AB	**BB**	BB	AB	BB	BB	AB	AB	BB	AB		**X**
3	**BB**	**BB**	AA	AA	AB	AB	AA	AB	BB	AB	**X**	**X**
4	AB	**BB**	AB	AB	AB	AB	BB	AB	AA	AB		**X**
5	**BB**	AB	AA	AB	AA	AB	AA	AB	AA	BB	**X**	
6	**BB**	AB	AB	AB	BB	BB	AB	AA	AB	AB	**X**	
7	**BB**	**BB**	BB	BB	AB	AB	AA	AB	BB	AB	**X**	**X**
8	AB	**BB**	AB	AB	AA	AA	AB	BB	AB	BB		**X**
9	**BB**	AA	AB	AB	BB	AB	AB	AA	AB	AB	**X**	
10	AB	**BB**	AB	BB	AB	AB	BB	AB	AB	AA		**X**
11	AA	**BB**	AA	AA	AA	AB	AA	AB	AB	AB		**X**
12	**BB**	AB	BB	BB	AB	BB	AB	BB	AA	AB	**X**	
13	AB	**BB**	AB	AA	AB	AB	BB	AB	AA	AA		**X**
14	**BB**	AA	AB	AB	BB	BB	AB	AA	AB	AB	**X**	
15	AB	**BB**	BB	BB	AB	AA	AB	BB	AB	AA		**X**

(b)

	Locus										Trait	
Indiv	**1**	**2**	3	4	5	6	7	8	9	10	**1**	**2**
1	**BB**	AB	AB	AB	AB	AA	AB	BB	AB	BB	**X**	
3	**BB**	**BB**	AA	AA	AB	AB	AA	AB	BB	AB	**X**	**X**
5	**BB**	AB	AA	AB	AA	AB	AA	AB	AA	BB	**X**	
6	**BB**	AB	AB	AB	BB	BB	AB	AA	AB	AB	**X**	
9	**BB**	AA	AB	AB	BB	AB	AB	AA	AB	AB	**X**	
12	**BB**	AB	BB	BB	AB	BB	AB	BB	AA	AB	**X**	
14	**BB**	AA	AB	AB	BB	BB	AB	AA	AB	AB	**X**	

(c)

	Locus										Trait	
Indiv	**1**	**2**	3	4	5	6	7	8	9	10	**1**	**2**
2	AB	**BB**	BB	AB	BB	BB	AB	AB	BB	AB		**X**
4	AB	**BB**	AB	AB	AB	AB	BB	AB	AA	AB		**X**
7	**BB**	**BB**	BB	BB	AB	AB	AA	AB	BB	AB	**X**	**X**
8	AB	**BB**	AB	AB	AA	AA	AB	BB	AB	BB		**X**
10	AB	**BB**	AB	BB	AB	AB	BB	AB	AB	AA		**X**
11	AA	**BB**	AA	AA	AA	AB	AA	AB	AB	AB		**X**
13	AB	**BB**	AB	AA	AB	AB	BB	AB	AA	AA		**X**
15	AB	**BB**	BB	BB	AB	AA	AB	BB	AB	AA		**X**

Figure11. Hypothetical Clustering of a THO Dataset
(a) A small dataset consistent with the Trait Heterogeneity Only (THO) genetic model (see Figure 7) is presented. All individuals with the high risk genotype (BB) at locus 1 have Trait I, and all individuals with the high risk genotype (BB) at locus 2 have Trait II. Some individuals have both high risk genotypes and, therefore, both traits.
A successful clustering of this dataset might be one in which there are two clusters (b) and (c), such that one cluster contains only individuals who have Trait I (b) and the other cluster contains only individuals who have Trait II (c).

otherwise stated as the probability of the individual's multilocus genotype, conditional on it belonging to that cluster, with its characteristic genotypes. The second posterior probability involves the probability of a cluster given its assigned individuals, or otherwise stated as the probability of the cluster's characteristic genotypes, conditional on the multilocus genotypes of the individuals assigned to that cluster.

In actuality, individuals are not 'assigned' to clusters in the hard classification sense but instead in the fuzzy sense they are temporarily assigned to the cluster to which they have the greatest probability of belonging. Thus, each individual has its own vector of probabilities of belonging to each of the clusters. The assignment of individuals is also not considered the most important result of the clustering method. A ranked listing is produced of all loci in the dataset with their corresponding normalized "attribute influence" values (ranging between 0 and 1), which provide a rough heuristic measure of relative influence of each locus in differentiating the classes from the overall dataset. Thus, emphasis is placed on the identification of which attributes, or loci, are most important in producing the clustering. This information that can then be used to more directly stratify affected (and/or unaffected) individuals, for instance, by using the top n most influential loci identified, and to enable meaningful interpretation of the clustering result.

Hypergraph Clustering

The second method is Hypergraph Clustering (Han EH et al., 1997a). It has been implemented in the hMETIS software, which is freely available from George Karypis at the University of Minnesota. Hypergraph clustering seeks a partitioning of vertices, such

that intracluster relatedness meets a specified threshold, while the weight of hyperedges cut by the partitioning is minimized. In this case, vertices represent single locus genotypes, hyperedges represent association rules, and hyperedge weights represent the strength of the association rules. For instance, if a specific genotype at one locus co-occurs with a specific genotype at another locus, an association rule linking those two genotypes would be created, and that rule would have a weight equivalent to the proportion of individuals in the dataset that had both of those genotypes. Thus, for our purposes, association rules are multilocus genotype combinations that are found in the dataset. The freely available LPminer program was used to generate the association rules (Seno M and Karypis G, 2001). LPminer searches the database for multilocus genotype combinations that appear together with substantial frequency (above a prespecified "support" percentage) and outputs this info as a list of association rules. hMETIS takes these association rules and uses them to create a hypergraph in which single locus genotypes are vertices and association rules dictate the presence and weight of hyperedges. hMETIS creates a partition of the hypergraph such that the weight of the removed hyperedges is minimized. It achieves this by using a series of phases, somewhat analogous to the stages of a simulated annealing algorithm, in an attempt to avoid making decisions which are only locally (not globally) optimal.

This process results in a partitioning (or clustering) of the genotypes in a dataset. If a single dataset were being analyzed, this information by itself could be sufficiently helpful since it would provide information about which multilocus genotypes appear with such frequency that they characterize groups of individuals. Individuals could be directly stratified using such multilocus combinations (similar to the way attribute influence

values in the Bayesian Classification method could be used). However, for the purpose of comparing the results of Hypergraph Partitioning to those of the other two methods, which produce clusters, or partitions, of individuals (not genotypes), such a partitioning of individuals still needed to be created. Since a given individual could have more than one of the multilocus genotypes specified by different hyperedges in the final partitioning, the partitioning of individuals was not entirely straightforward. Thus, a heuristic was devised such that each individual would be assigned to the partition, or cluster, for which it had the highest percentage of matching genotypes (Figure 12). More specifically, for each cluster, the number of loci represented by one or more genotypes in that cluster was determined (L_c). Then, for each individual, for each cluster, the number of matching genotypes between the cluster and the individual (M_{ic}) was divided by L_c, producing a vector of similarity percentages per individual, similar to the vector of probabilities used by the Bayesian Classification and Fuzzy *k*-Modes Clustering methods. Each individual was then assigned to the cluster with which it had the greatest similarity.

Fuzzy *k*-Modes Clustering

The third clustering method is Fuzzy *k*-Modes Clustering (Huang Z and Ng MK, 1999). *k*-Modes is a trivial extension to categorical data of the popular *k*-means algorithm. In both methods, cluster centroids can be initialized at random or by one of many seeding strategies (Duda RO and Hart PE, 1973), and individuals are assigned to their nearest cluster centroids. Then, cluster centroids are reevaluated based on their newly assigned individuals. For the *k*-means algorithm, the centroid is calculated as the mean vector of genotypes across individuals. However, for nominal data, such means are

(a)

Cluster 1	**Cluster 2**	**Cluster 3**
1B1B	2B2B	7A7B
3A3B		9A9A

(b)

	Locus										Percentage of Matching Genotypes by Cluster		
Indiv	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**
1	**BB**	AB	AB	AB	AB	AA	AB	BB	AB	BB	**100**	0	50
2	AB	**BB**	BB	AB	BB	BB	AB	AB	BB	AB	0	**100**	50
3	**BB**	**BB**	AA	AA	AB	AB	AA	AB	BB	AB	50	**100**	0
4	AB	**BB**	AB	AB	AB	AB	BB	AB	AA	AB	50	**100**	50
5	**BB**	AB	AA	AB	AA	AB	AA	AB	AA	BB	**50***	0	50
6	**BB**	AB	AB	AB	BB	BB	AB	AA	AB	AB	**100**	0	50
7	**BB**	**BB**	BB	BB	AB	AB	AA	AB	BB	AB	50	**100**	0
8	AB	**BB**	AB	AB	AA	AA	AB	BB	AB	BB	50	**100**	50
9	**BB**	AA	AB	AB	BB	AB	AB	AA	AB	AB	**100**	0	50
10	AB	**BB**	AB	BB	AB	AB	BB	AB	AB	AA	50	**100**	0
11	AA	**BB**	AA	AA	AA	AB	AA	AB	AB	AB	0	**100**	0
12	**BB**	AB	BB	BB	AB	BB	AB	BB	AA	AB	50	0	**100****
13	AB	**BB**	AB	AA	AB	AB	BB	AB	AA	AA	50	**100**	0
14	**BB**	AA	AB	AB	BB	BB	AB	AA	AB	AB	**100**	0	50
15	AB	**BB**	BB	BB	AB	AA	AB	BB	AB	AA	0	**100**	50

Figure 12. Example of Post-processing of Hypergraph Clustering Result
Hypergraph clustering produces a clustering of genotypes, instead of individuals. Thus, a clustering of individuals must be induced from this clustering of genotypes. As described in the text, an individual in assigned to the cluster for which it has the highest percentage of matching genotypes. Given the dataset presented in Figure 11(a) and a clustering of genotypes that is presented here (a), a clustering of individuals can be induced (b). For each individual (row), the percentage of matching genotypes that is highlighted indicates to which cluster the individual becomes assigned. Notice that for individual 5, there is a tie between the percentage of matching genotypes for clusters 1 and 3. In such cases, we arbitrarily assign the individual to the lower numbered cluster. Since cluster 3 does not contain any high-risk genotypes, it does not facilitate the goal of creating a clustering that maps to the simulated trait heterogeneity, and in the case of individual 12, it ends up capturing an individual who would preferably be clustered in cluster 1.

not necessarily meaningful, and the *k*-modes algorithm instead determines the centroid as the mode vector of genotypes across individuals. Genotypes are encoded nominally (not ordinally), such that no two genotypes are considered 'closer' than another two, and the 'distance' between an individual and a centroid is calculated as the cumulative number of non-matching genotypes across all loci. After cluster centroids are reevaluated, individuals are again assigned to their nearest centroids, and this process is repeated until the assignment of individuals to clusters does not change. Figure 13 demonstrates the first steps of the *k*-modes clustering, using the same dataset presented in Figures 11 and 12. The straightforward algorithm was developed in the C++ language. The number of clusters (*k*) was prespecified to be 2, 3, 4, 5 or 6. All five possible *k* were run for each dataset. Each cluster centroid was initially set to the values of a randomly selected individual in the dataset being analyzed. Both a 'fuzzy' and a 'hard' version of the *k*-modes algorithm were implemented and tested, and while their results on test datasets were comparable, the fuzzy version did perform slightly better and provided more information, which could be used for interpretation of results. Thus, the fuzzy version was chosen for use in these analyses.

Statistical Analysis

Comparison of Clustering Methods

Each clustering method has its own metric(s) for evaluating the "goodness" of a clustering of data. Since these methods are being tested on simulated data, classification error of a given clustering can be calculated as the number of misclassified individuals

(a)

	Locus									
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1 (1)	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB
2 (5)	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB
3 (12)	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB
4 (15)	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA

(b)

	Locus										Cluster Distance			
Indiv	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**
1	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB	**0**	6	5	5
2	AB	BB	BB	AB	BB	BB	AB	AB	BB	AB	8	8	**6**	6
3	BB	BB	AA	AA	AB	AB	AA	AB	BB	AB	8	**5**	7	8
4	AB	BB	AB	AB	AB	AB	BB	AB	AA	AB	7	**6**	7	7
5	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB	6	**0**	5	10
6	BB	AB	AB	AB	BB	BB	AB	AA	AB	AB	**4**	7	5	8
7	BB	BB	BB	BB	AB	AB	AA	AB	BB	AB	8	6	**5**	6
8	AB	BB	AB	AB	AA	AA	AB	BB	AB	BB	**4**	7	8	4
9	BB	AA	AB	AB	BB	AB	AB	AA	AB	AB	**5**	7	8	8
10	AB	BB	AB	BB	AB	AB	BB	AB	AB	AA	7	8	8	**4**
11	AA	BB	AA	AA	AA	AB	AA	AB	AB	AB	9	**5**	9	8
12	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB	5	7	**0**	5
13	AB	BB	AB	AA	AB	AB	BB	AB	AA	AA	8	7	8	**6**
14	BB	AA	AB	AB	BB	BB	AB	AA	AB	AB	**5**	7	6	8
15	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA	5	10	5	**0**

(c)

	Locus									
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	BB	**AA**	AB	AB	**BB**	AA	AB	**AA**	AB	**AB**
2	BB	**BB**	AA	**AA**	AA	AB	AA	AB	AA	**AB**
3	BB	**BB**	BB	BB	AB	BB	AB	**AB**	**BB**	AB
4	AB	BB	**AB**	BB	AB	**AB**	**BB**	**AB**	AB	AA

Figure 13. Example of *k*-Modes Clustering
In this example, the same dataset presented in Figure 11 is used to demonstrate the different steps involved the *k*-modes clustering algorithm, and *k* was chosen to be 4, such that four clusters will initially be formed. (a) The cluster centroids are seeded by randomly selecting the genotypes of actual individuals in the dataset. The number in parentheses beside the cluster number is the individual used to seed that cluster. (b) Individuals are then compared to each of the cluster centroids, and the number of nonmatching genotypes between each cluster centroid and that individual are recorded. The individual is then assigned to the cluster for which it had the fewest number of nonmatching genotypes (in bold). (c) The next step is to update the cluster centroids based on the individuals now assigned to the clusters. The mode genotype among individuals assigned to a cluster becomes the centroid genotype at that locus. Genotypes that changed from the initialization to the update are shown in bold.

(d)

	Locus										Cluster Distance			
Indiv	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**
1	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB	**4**	9	7	7
2	AB	BB	BB	AB	BB	BB	AB	AB	BB	AB	6	7	**3**	7
3	BB	BB	AA	AA	AB	AB	AA	AB	BB	AB	8	**3**	4	6
4	AB	BB	AB	AB	AB	AB	BB	AB	AA	AB	7	6	6	**3**
5	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB	8	**3**	8	8
6	BB	AB	AB	AB	BB	BB	AB	AA	AB	AB	**2**	8	6	8
7	BB	BB	BB	BB	AB	AB	AA	AB	BB	AB	8	4	**2**	5
8	AB	BB	AB	AB	AA	AA	AB	BB	AB	BB	**5**	8	8	6
9	BB	AA	AB	AB	BB	AB	AB	AA	AB	AB	**1**	7	7	7
10	AB	BB	AB	BB	AB	AB	BB	AB	AB	AA	8	7	6	**0**
11	AA	BB	AA	AA	AA	AB	AA	AB	AB	AB	8	**2**	7	6
12	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB	7	7	**3**	8
13	AB	BB	AB	AA	AB	AB	BB	AB	AA	AA	9	5	7	**2**
14	BB	AA	AB	AB	BB	BB	AB	AA	AB	AB	**1**	8	6	8
15	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA	7	9	5	**4**

Figure 13, continued. Example of *k*-Modes Clustering
(d) After the centroids are updated, the individuals are reevaluated as to which cluster they most closely resemble and are assigned to that cluster. Only individual 4 was assigned to a different cluster than it was previously. Steps (c) and (d) are repeated until no genotypes are changed in any cluster centroid and no individuals' cluster assignments are changed.

divided by the total number of individuals. However, simple classification error has its disadvantages. Firstly, in cases such as this where there is overlap between the known classes, the researcher must make an arbitrary decision as to when individuals who have been simulated to have both traits, not just one or the other, are considered to be misclassified. The decision about error is equally arbitrary when the number of resulting clusters is greater than the number of known classes. For instance, if the individuals belonging to one class were divided into two classes by the clustering algorithm, calculating classification error would require either (1) that none of those individuals be considered incorrectly classified, since they are all in homogenous clusters, or else (2) that all individuals from one of those clusters be considered misclassified. Neither choice

seems to satisfactorily capture the "goodness" of the clustering result. Subsequently, it is not advisable to compare the classification error of two clustering results for which the number of clusters differs.

It is for these reasons alternative cluster recovery metrics were investigated. The Hubert-Arabie Adjusted Rand Index (ARI_{HA}) addresses the concerns raised by classification error and was, therefore, chosen to evaluate the goodness of clustering results from the three clustering methods being compared (Hubert L and Arabie P, 1985). Calculation of the ARI_{HA} involves determining (1) whether pairs of individuals, who were simulated to have the same trait, are clustered together or apart and (2) whether pairs of individuals, who do not have the same trait, are clustered together or apart. The ARI_{HA} is robust with regard to the number of individuals being clustered, the number of resulting clusters, and the relative size of those clusters (Steinley D, 2004). It is also sensitive to the degree of class overlap, which is desirable since it will penalize more for good clusterings that occur by chance than classification error would. When interpreting ARI_{HA} values, 0.90 and greater can be considered excellent cluster recovery, 0.80 and greater is good cluster recovery, 0.65 and greater reflects moderate cluster recovery, and less than 0.65 indicates poor cluster recovery. These values were derived from empirical studies showing observations cut at the 95th, 90th, 85th and 80th percentiles corresponded to ARI_{HA} values of 0.86, 0.77, 0.67 and 0.60 respectively (Steinley D, 2004).

The ARI_{HA} was used as the gold standard measure to compare the performance of the three clustering methods. Three categorical variables were created that could be tested using the nonparametric chi-square test of independence. The ARI_{HA} values were

discretized into a 1 or 0 depending on whether they met or exceeded the cutoff values for excellent, good and moderate cluster recovery, as described above. A chi-square test of independence was performed testing the null hypothesis that the number of clusterings achieving a certain ARI_{HA} value was independent of the clustering method, thereby evaluating whether one method significantly outperformed the others. Five percent was chosen as the significance level (alpha).

Applicability to Real Data

As a reminder, the ultimate goal of this research is to find a clustering method that works well at uncovering trait heterogeneity in real genotypic data. Unlike for the current simulation study, for real data it is not known *a priori* to which clusters individuals belong, otherwise the clustering would not be necessary. Indeed, it is the goal of clustering to uncover natural clusters or partitions of data using the method-specific "goodness" metric as a guide. In preparation for application of a clustering method to real data, after choosing the superior method, that method's internal clustering metrics were analyzed using permutation testing to determine how good a proxy they are for ARI_{HA}.

One hundred permuted datasets per simulated dataset was chosen, which should result in a reasonable approximation of the null distribution but would not put unreasonable strain on resources and time (Good P, 2000). Genotypes were permuted within loci across individuals, such that the overall frequency of genotypes at any one locus was unchanged, but the frequency of multilocus genotypes was altered at random. This created a null sample in which the frequency of multilocus genotypes was no longer

associated with trait status except by chance. The empirically-determined superior clustering method was applied to each permuted dataset and both the internal clustering metric values and the ARI_{HA} were determined. For each set of 100 permuted datasets, the significance of each of the simulated dataset results was determined based on whether they exceeded the values at the significance level in the corresponding null distribution. Ten percent was chosen as the acceptable Type I error rate since these methods serve as a means of data exploration to be followed by more rigorous, supervised analyses on individual clusters of the data. However, the more conventional levels of 0.05 and 0.01 were also evaluated. Finally, the ability of permutation testing to preserve acceptable Type I (false positive) and Type II (false negative) error rates was evaluated at the three specified significance levels.

Results

Descriptive statistics and plots for the Hubert-Arabie Adjusted Rand Index results were produced. Mean ARI_{HA} values for Bayesian Classification, Hypergraph Clustering and Fuzzy *k*-Modes Clustering were 0.666, 0.354 and 0.556, respectively. Confidence intervals around the means were also produced to demonstrate the preciseness of the ARI_{HA} measurements. The results for each method across all datasets are presented in Table 1. Mean ARI_{HA} values differed by genetic model type, with higher scores achieved on Trait Heterogeneity Only (THO) datasets for the Bayesian Classification and Hypergraph Clustering methods (Figure 14).

Table 1. Confidence Intervals around ARI_{HA} Means by Method

			95% Confidence Interval	
Method	**Mean**	**Standard Error**	**Lower End**	**Upper End**
Bayesian	0.666	0.001	0.664	0.667
Hypergraph	0.354	0.001	0.352	0.355
Fuzzy *k*-Modes	0.556	0.001	0.555	0.558

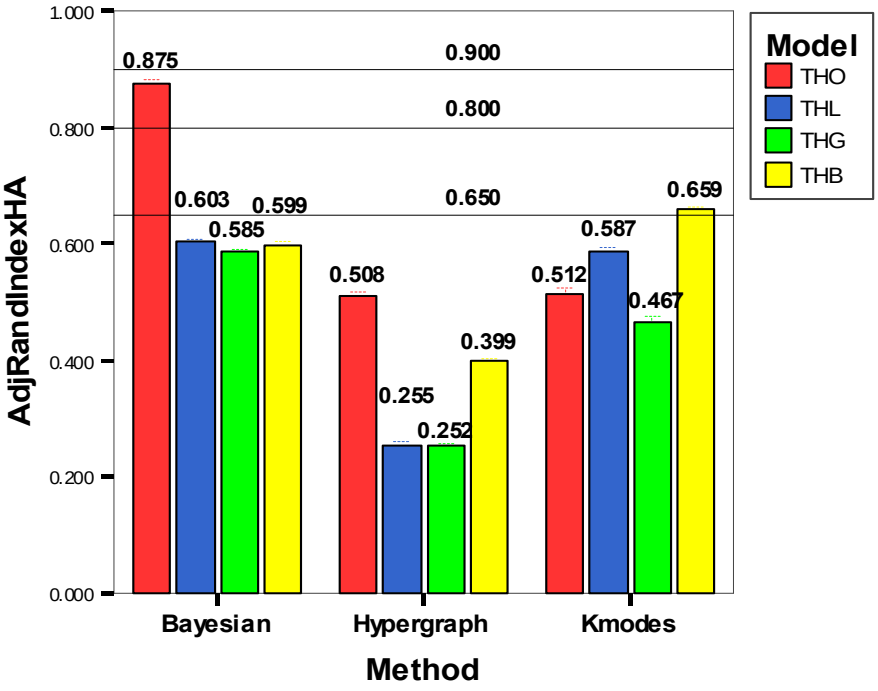


Figure 14. Comparison of ARI_{HA} Means by Method and Model. Bars represent means, and error bars, which are very short and may be difficult to see, represent 95% confidence intervals. Horizontal lines represent thresholds for quality of cluster recovery: 0.90 for excellent recovery, 0.80 for good recovery and 0.65 for moderate recovery.

Results are displayed as percentages by clustering method (Figure 15) and by clustering method and genetic model (Figure 16). A chi-square test of independence was performed testing the null hypothesis that the number of clusterings achieving the

specified ARI_{HA} cutoff value was independent of the clustering method. The three methods performed significantly differently on each of the ARI_{HA} cutoff statistics (Table 2). Bayesian Classification outperformed the other two methods. However, across all the dataset parameters, Bayesian Classification achieved moderate or better recovery on only 48% of the datasets (Figure 15).

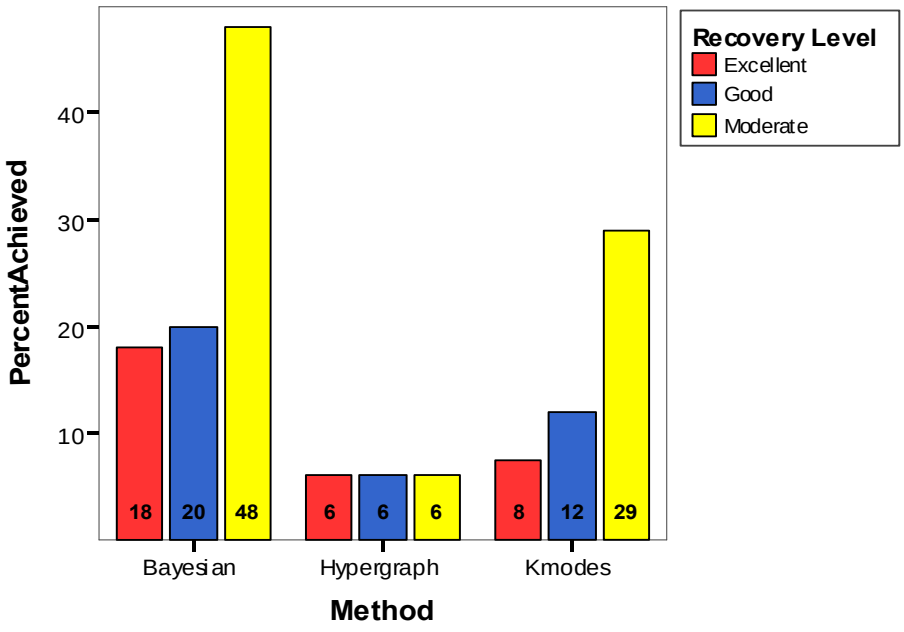


Figure 15. Percentage of Clustering Results Achieving Cluster Recovery Levels by Method

Table 2. Overall Results of Chi-Square Test of Independence. The null hypothesis that the percentage of clustering results achieving the specified cluster recovery level does not differ across clustering methods was tested.

Cluster Recovery Statistic	χ^2	**df**	**p**
%Results achieving Excellent cluster recovery ($ARI_{HA} \geq 0.90$)	1787	2	< 0.001
%Results achieving Good cluster recovery ($ARI_{HA} \geq 0.80$)	1614	2	< 0.001
%Results achieving Moderate cluster recovery ($ARI_{HA} \geq 0.65$)	8565	2	< 0.001

The performance of the three clustering methods across different dataset parameters was evaluated to find particular conditions under which one method consistently achieved good or excellent recovery (not just better recovery than the other two methods). For those datasets simulated under the THO model, Bayesian Classification performed well, with over 73 percent of its resulting clusterings achieving an ARI_{HA} value of 0.90 or greater, indicating excellent recovery (Figure 16). For this subset of the datasets, Bayesian Classification outperformed the other two methods, and again there was a significant difference in performance across the three methods, as measured by a chi-square test of independence on each of the three new ARI_{HA} cutoff statistics (Table 3). Analysis of the other simulation parameters failed to show as great a difference among methods where the 'winning' method performed as well as the Bayesian Classification performed in the THO datasets (data not shown). Thus, this subset of data was chosen for further investigation into the efficacy of using the Bayesian Classification method to uncover trait heterogeneity in **real** data.

The Bayesian Classification method produces two internal clustering metrics for each resulting cluster, or class: (1) class strength, and (2) cross-class entropy. Class strength is a heuristic measure of how strongly each class predicts "its" instances and is reported as the log of class strength. Cross-class entropy is a measure of how strongly the class probability distribution function differs from that of the dataset as a whole. Because each metric is reported per resulting cluster, or class, the average metric value across clusters was calculated and utilized for evaluating cluster fitness. To evaluate the validity of using the Bayesian Classification internal clustering metrics—class strength and cross-class entropy—as a proxy for the ARI_{HA} (since ARI_{HA} is unknown for

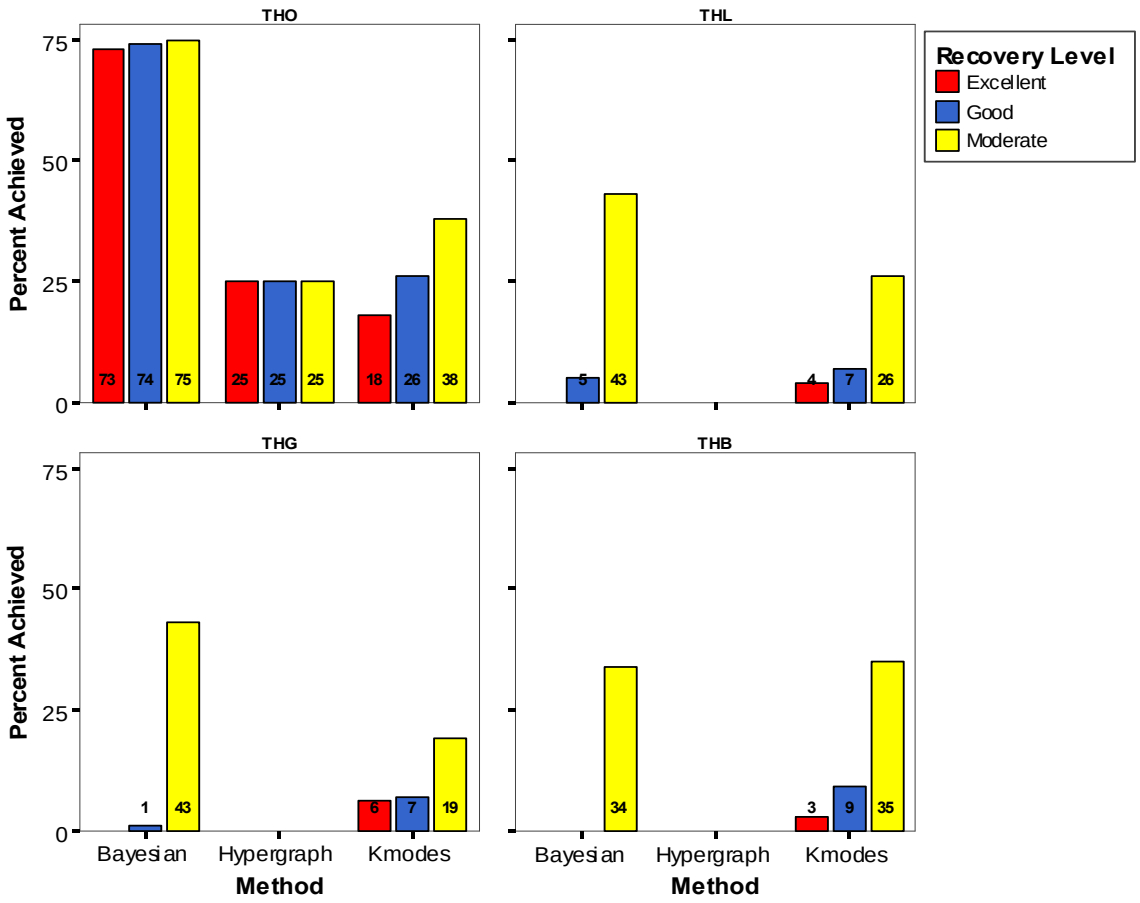


Figure 16. Percentage of Clustering Results Achieving Cluster Recovery Levels by Method and Model

Table 3. Results of Chi-Square Test of Independence for THO Datasets.
The null hypothesis that the percentage of clustering results achieving the specified cluster recovery level does not differ across clustering methods was tested.

Cluster Recovery Statistic	**Model**	χ^2	**df**	**p**
%Results achieving Excellent cluster recovery ($ARI_{HA} \geq 0.90$)	THO	3713	2	< 0.001
%Results achieving Good cluster recovery ($ARI_{HA} \geq 0.80$)	THO	3107	2	< 0.001
%Results achieving Moderate cluster recovery ($ARI_{HA} \geq 0.65$)	THO	2609	2	< 0.001

real data), permutation testing was performed. Resulting p-values for ARI_{HA}, average log of class strength and average cross class entropy were used to calculate false positive and false negative rates at three significance levels of 0.01, 0.05 and 0.10. A clustering result was considered a false positive if it was considered significant according to **either** average log of class strength or average cross class entropy but was not considered significant according to our ARI_{HA} standard. A clustering result was considered a false negative if it was called not-significant according to **both** average log of class strength and average cross class entropy but was considered significant according to ARI_{HA}. Figures 17 and 18 show the false positive and false negative rates, respectively, by alpha level.

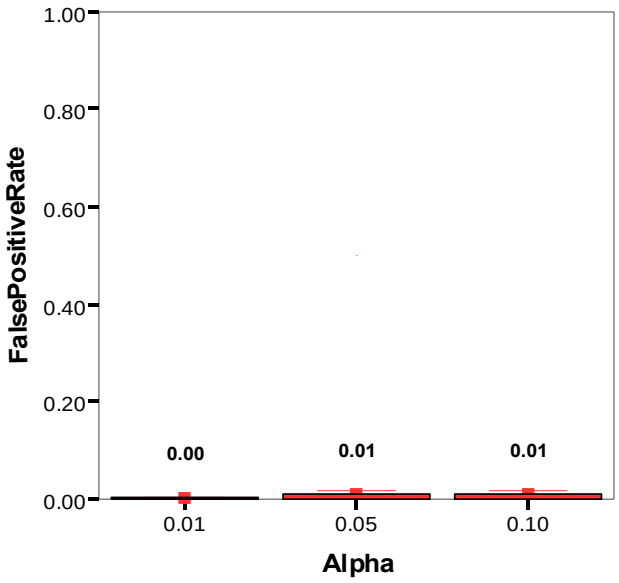


Figure 17. False Positive Rate by Significance Level (Alpha).

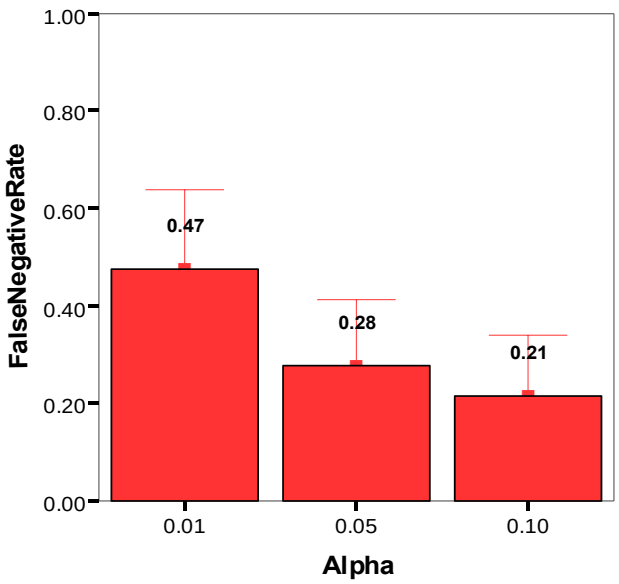


Figure 18. False Negative Rate by Significance Level (Alpha)

The false positive, or Type I, error rate was controlled very well at three percent or less for all three significance levels. The false negative, or Type II, error rate was not controlled as well, however. At the least stringent significance level ($\alpha = 0.10$), the Type II error rate was 18 percent, and at the most stringent level ($\alpha = 0.01$), the rate was 47 percent. Other simulation parameters were examined for their impact on the false negative rate, and Figures 19 and 20 show the false negative rate by alpha level paneled by number of nonfunctional loci and number of affecteds (sample size), respectively. As might be expected, the lowest false negative rates were achieved for datasets with the lowest number of nonfunctional loci (10) and the greatest sample size (1000).

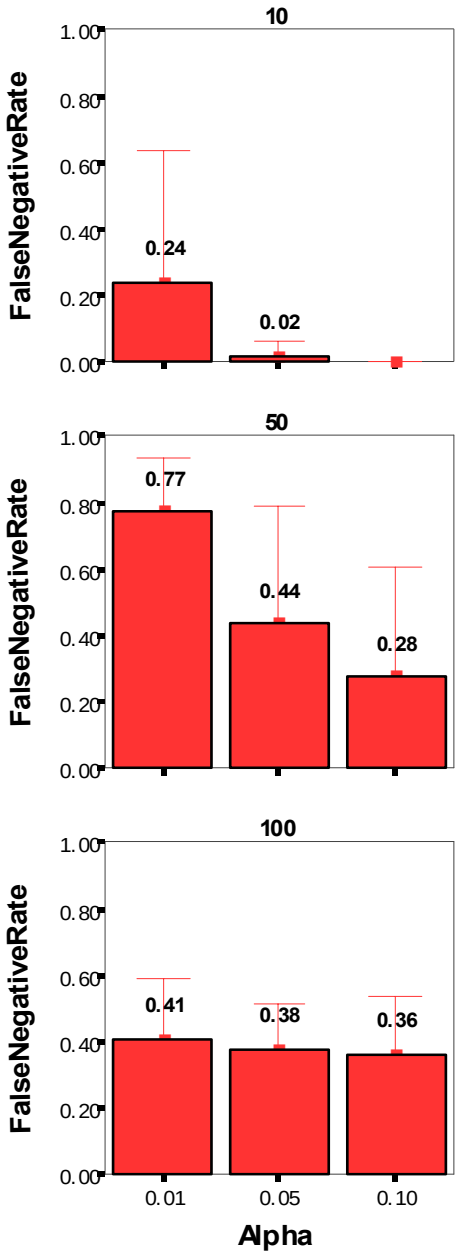


Figure 19. False Negative Rate by Significance Level (Alpha), Paneled by Number of Nonfunctional Loci. These rates are across all genetic models (THO, THL, THG and THB).

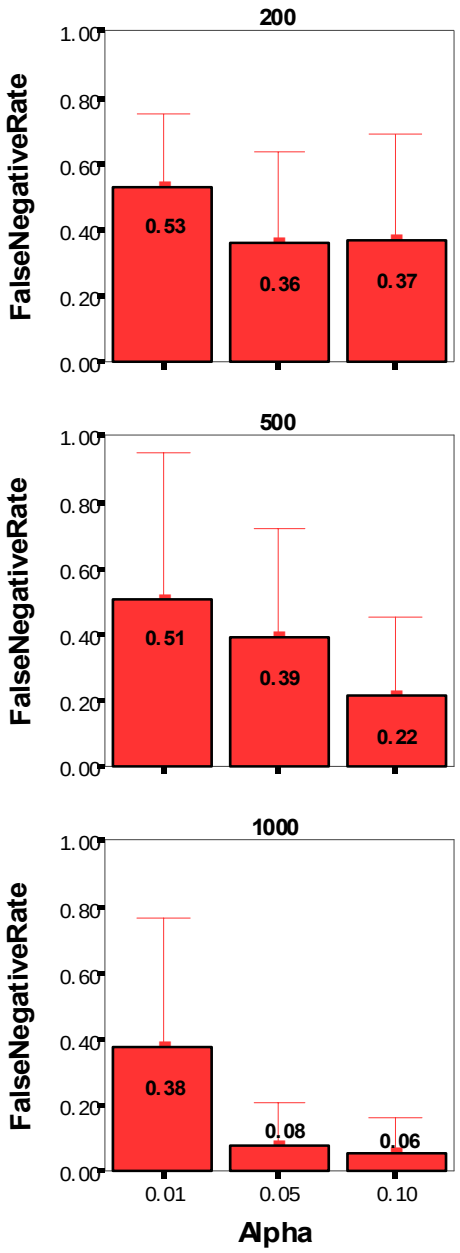


Figure 20. False Negative Rate by Significance Level (Alpha), Paneled by Number of Affecteds (Sample Size). These rates are across all genetic models (THO, THL, THG and THB)

Discussion

Data Simulation

The new data simulation algorithm produced complex genotypic datasets that included trait heterogeneity, locus heterogeneity and gene-gene interactions. Most existing simulation software that attempt to simulate heterogeneity do so by allowing the user to specify what portion of the dataset is to be simulated under one model versus another, and the resulting individuals are simply combined into one dataset. In the new algorithm, however, the disease penetrance models, which were used to simulate the data, were constructed so that overall prevalence levels were controlled, allowing naturally occurring overlaps, in which some individuals would have both traits (and their associated multilocus genotypes) by chance. This is important because it more closely simulates the natural variation one would expect under the "common disease, common variant" hypothesis in which there is very little if any selective pressure against alleles that increase disease risk only slightly or only in combination with other susceptibility alleles at the same or distinct loci (Cargill et al., 1999; Chakravarti, 1999; Reich and Lander, 2001; Risch and Merikangas, 1996). This novel data simulation algorithm should prove very useful for future studies of other proposed genetic analysis methods for complex diseases.

Comparison of Clustering Methods

The Bayesian Classification method outperformed the other two methods across most dataset parameter combinations, with the exception of the most complex

model (THB) on which Fuzzy *k*-Modes Clustering performed best. When the results were further examined to find a set of parameters for which one or more methods performed well, Bayesian Classification achieved excellent recovery for 73% of the datasets with the THO model (Figure 16) and achieved moderate recovery for 56% of datasets with 500 or more affecteds and for 86% of datasets with 10 or fewer nonfunctional loci (Figures 21 and 22). Neither Hypergraph Clustering nor Fuzzy *k*-Modes Clustering achieved good or excellent cluster recovery even under a restricted set of conditions (data not shown).

Bayesian Classification was obtained as closed-source software, for which there are numerous parameters that can be optimized, as discussed in Chapter IV. Initial parameter settings were chosen as recommended by the authors based on the type of data being analyzed. However, it is possible that alternative settings may yield better results. For example, for datasets with the more complex genetic models, greater numbers of nonfunctional loci and smaller sample sizes, the maximum number of classification trials and/or the maximum number of classification cycles per trial may need to be longer, and those parameters concerned with convergence rate and stopping criteria may need to be changed to delay convergence. If improvements in performance could be achieved with reasonable time and resource tradeoffs, such changes would certainly be desirable. Further investigation of this matter is discussed in Chapter IV.

It was disappointing that Hypergraph Clustering did not perform very well under most conditions, despite its intuitive appeal as a method that would find frequently-occurring multilocus genotypic patterns. The Hypergraph Clustering method has been reported to work well with very large variable sets (on the order of thousands), which

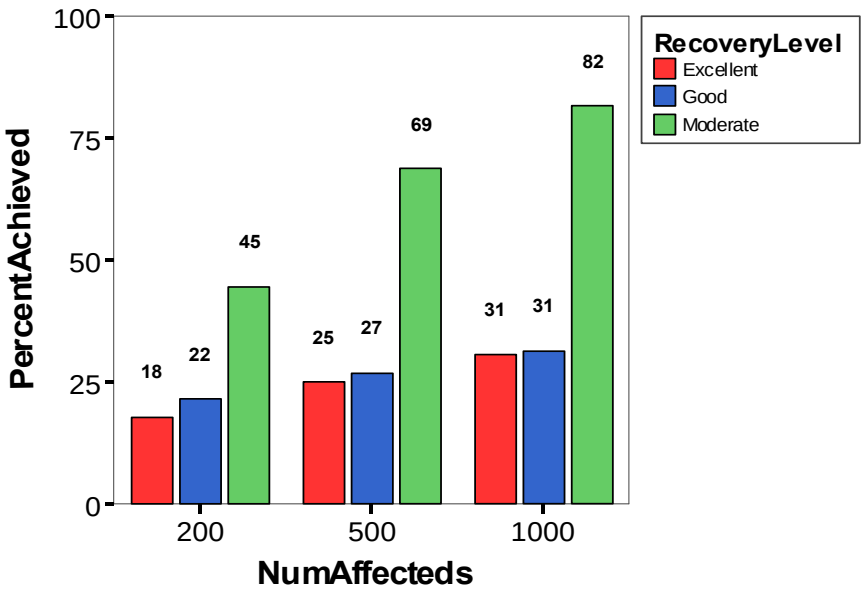


Figure 21. Percentage of Bayesian Classification Clustering Results Achieving Cluster Recovery Levels by Number of Affecteds (Sample Size)

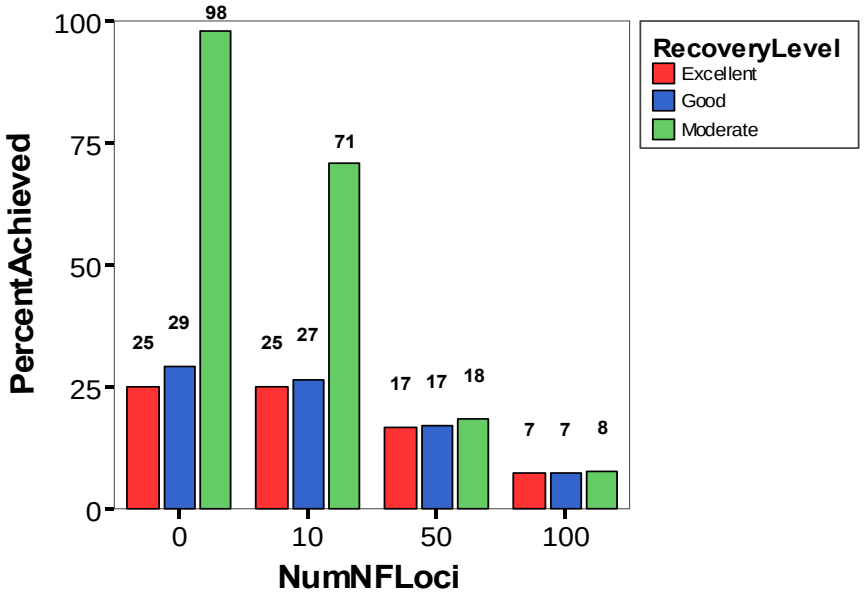


Figure 22. Percentage of Bayesian Classification Clustering Results Achieving Cluster Recovery Levels by Number of Nonfunctional Loci

have complex patterns for which large numbers of clusters (10-20+) were relevant (Han EH et al., 1997b). However, there has been no examination of the method's performance on smaller variable sets. Thus, it is possible that the restricted patterns present in our multilocus genotypic data were too simple and sparse and that the method is simply tuned to search for more complex patterns. Also, we were required to devise a translation of the resulting partitioning of genotypes into a clustering of individuals. We tested several such translations and implemented the best process out of several tested. Oftentimes, even when the method correctly chose the functional genotypes to be in different partitions, too many other nonfunctional genotypes were also chosen, which meant that the difference between an individual's likelihood of belonging to one cluster versus another was too small, making the choice of cluster assignment almost arbitrary.

The Fuzzy *k*-Modes Clustering method performed comparably to Bayesian Classification for the more complex datasets and was much less computationally intensive. It has been widely reported that the performance of *k*-means algorithms is highly variable depending on the method of seeding the initial cluster centroids (Duda RO and Hart PE, 1973). While we used the recommended method of selecting individuals from the dataset to serve as the initial cluster modes, we perhaps could have achieved better results if we implemented an additional step to ensure that the initial centroids were substantially dissimilar to each other. This is supported by evidence that when the Fuzzy *k*-Modes Clustering resulted in only one cluster (effectively no partitioning of the data), the initial centroids were very similar and the method had converged early so that individuals had equal probability of belonging to any of the

clusters. In such cases, the individual was arbitrarily assigned to the first cluster, thereby leading to all other clusters being empty.

As expected, the simpler the model, the better the performance of the three clustering algorithms, with the exception that the Hypergraph Clustering and Fuzzy *k*-Modes Clustering methods performed somewhat better on the THB (Trait Heterogeneity with Both locus heterogeneity and gene-gene interaction) datasets than they did on the THL (Trait Heterogeneity with Locus heterogeneity) and THG (Trait Heterogeneity with Gene-gene interaction) datasets. Likewise, in general, the fewer the nonfunctional loci and the larger the sample size, the better the performance.

Applicability to Real Data

To determine the efficacy of using the Bayesian Classification method on real data, the reliability of its internal clustering metrics at finding good clusterings was evaluated. Using the combination of the average log of class strength and the average cross class entropy to determine significance, the false positive rate was controlled very well, at three percent or less for all three significance levels. The false negative rate was acceptably low (18 percent) for the less stringent significance levels of 0.10. However, it was high (47 percent) for the most stringent significance level of 0.01. Thus, if a clustering of data were called significant according to permutation testing using either the average log of class strength or the average cross class entropy, one could be quite confident that the result were real. Typically geneticists prefer to accept a higher false positive rate to increase power; however, there is indeed a trade-off between these two types of error. Valuable time and resources can be spent on follow-up studies, and it can

be very detrimental to pursue leads that do not have a good chance of yielding new information about the disease under study. Therefore, we would recommend the Bayesian Classification method for use in the first stage of a comprehensive analysis strategy to detect heterogeneity and then main effects and interactions, with the caveat that a negative result should be interpreted carefully and may indicate that other methods for detecting heterogeneity should be considered as well.

CHAPTER IV

FURTHER EVALUATION OF BAYESIAN CLASSIFICATION

Background

The Bayesian Classification method is effective at uncovering trait heterogeneity in simulated genotypic data while preserving very low false positive rates and reasonably low false negative rates. However, these results were for the simplest of simulated genetic models and may not generalize to more complicated models. This chapter will present an extension of the previous work in which the Bayesian Classification method is modified to improve its performance under a wider set of simulation conditions. As discussed in Chapter III, it is possible that the parameter settings used in the initial data simulation study were not appropriate for the more complex genetic models. The goal of this study is to test different parameter settings to make improvements in performance for the more complex models without compromising performance for the simplest ones.

In addition, false positive and false negative rates will be determined for a wider range of simulation conditions. It is possible that even though the method performance decreases for these more complex models, the false positive rate will remain well-controlled, such that positive results are very trustworthy, in which case the method would still be useful. Conversely, along with decreased performance, an increased false positive rate (decrease in power) would prevent reasonable conclusions from being drawn about its results and thus render the method's use inadvisable. Thus, determining how the method behaves under these wider set of conditions will allow us to have more

confidence in our inferences about results from application of Bayesian Classification to real data.

Methods

Modification of Parameter Settings

The Bayesian Classification method software has over 30 different parameter settings that can be modified by the user to tweak the method's application. Six parameters were chosen as being likely to affect method performance on more complex data patterns since they affect how and what kind of search is performed in looking for the best clustering of the data. They determine initial search conditions, the type of search performed (i.e., what types of stopping criteria are used), and what values those stopping criteria impose. The six chosen parameters include: (1) start_j_list, (2) max_n_tries, (3) try_fn_type, (4) halt_range, (5) halt factor, and (6) max_cycles.

Table 4 shows the settings for each of these six parameters used in the initial simulation study and in the current extension to that study. Only one parameter setting was modified at a time, so that the effect of that particular setting change could be evaluated in comparison to the initial settings. For each of the new modified parameter settings, Bayesian Classification was applied to all 19,200 datasets that were simulated according to specifications detailed in Chapter III.

One decision the search algorithm must make is what the optimal number of clusters is for the data. The start_j_list parameter specifies a list of numbers that are the initial quantity of clusters the search algorithm tries when optimizing this value. This list

guides the search but does not restrict the algorithm, since it will also try other values that it deems likely to produce more optimal results. For the problem of detecting heterogeneity, we are primarily interested in clustering results where the number of classes is ten or less; therefore, the default start_j_list was modified accordingly (see Table 4).

Table 4. Bayesian Classification Parameter Settings in Simulation Studies. Note that try_fn_type is listed twice since all three search strategies were tried—'converge_search_3' initially and both 'converge' and 'converge_search_4' in the current simulation study.

Parameter	Initial Setting	Modified Setting
start_j_list	2,3,5,7,10,15,25	10,9,8,7,6,5,4,3,2,1
max_n_tries	50	100
try_fn_type	converge_search_3	converge
try_fn_type	converge_search_3	converge_search_4
halt range	0.5	0.75
halt_factor	0.0001	0.001
max_cycles	200	500

The max_n_tries parameter specifies a limit on the number of times the algorithm will produce a clustering of the data. Thus, the higher the value of this parameter, the longer the search will last and, in theory, the better the likelihood that the algorithm will find a globally optimal solution. The max_n_tries parameter was increased from 50 to 100, thereby doubling the maximum number of attempts at finding the optimal solution. Larger values were tested on a few datasets, but the computation time was not feasible, given the large volume of simulated datasets to be evaluated. Ideally, on a real dataset,

one would set this parameter to the default of 0, allowing unlimited numbers of tries at reaching the optimal solution, within the constraints of other search parameter settings.

The try_fn_type parameter specifies one of three search strategies ('converge', 'converge_search_3' and 'converge_search_4') the algorithm may use in searching for an optimal solution. The three strategies use different types of stopping criteria based on convergence measures. The default setting is the 'converge' algorithm, which is thought to perform better on a wide variety of problems than the other two algorithms (Taylor W et al., 2002). The authors indicate that the two alternative search algorithms may perform better on some problems but will perform substantially worse on others. Since this was one of the most critical parameters, we tried both alternative algorithms.

The halt_range and halt_factor parameters affect the convergence rate and, conversely, the number of cycles the search strategy will use. Increasing these values decreases the convergence rate. Therefore, we increased each of them, in turn. The halt_range parameter was increased from 0.5 to 0.75. The halt_factor was increased by a factor of ten from 0.0001 to 0.001. Higher values were also tested but were found to be cost-prohibitive in run time.

The max_cycles parameter specifies an upper limit on the number of cycles the search will perform while the convergence criteria have not been met. The default value of 200 was increased to 300. Higher values were tested on a small number of datasets but were found to increase computation time beyond reasonable limits, given the large volume of datasets being evaluated.

Applicability to Real Data

Using the best group of parameter settings, as determined by the aforementioned simulations, permutation testing to determine false positive and false negative rates was performed on datasets under a wider set of data simulation conditions. Data simulated under the Trait Heterogeneity with Locus Heterogeneity (THL) and Trait Heterogeneity with Gene-Gene Interaction (THG) models, as described in Chapter III, were evaluated, where the prevalence was 15 percent, the number of nonfunctional loci was either 10 or 100 and the sample size was either 500 or 1000. In the interest of time and computational resources, only the first 50 (out of 100) replicates of each set of conditions were used, and for each of the replicates, 500 permutated datasets were created, resulting in 200,000 datasets that were analyzed.

Results

Figures 23-25 show how method performance differed with each parameter setting modification, as measured by the percentage of clustered datasets achieving moderate (Figure 23), good (Figure 24) or excellent (Figure 25) cluster recovery according to the Hubert-Arabie Adjusted Rand Index. There was essentially no improvement in method performance for either model for each of the modified parameter settings, and in fact, modifications in two parameters (start_j_list and try_fn_type) led to decreases in performance. Thus, we concluded that the initial parameter settings were the best we had discovered and that those settings should be used going forward.

False positive and false negative rates were calculated based on permutation testing results on the THL and THG genetic model datasets, as specified above. Overall,

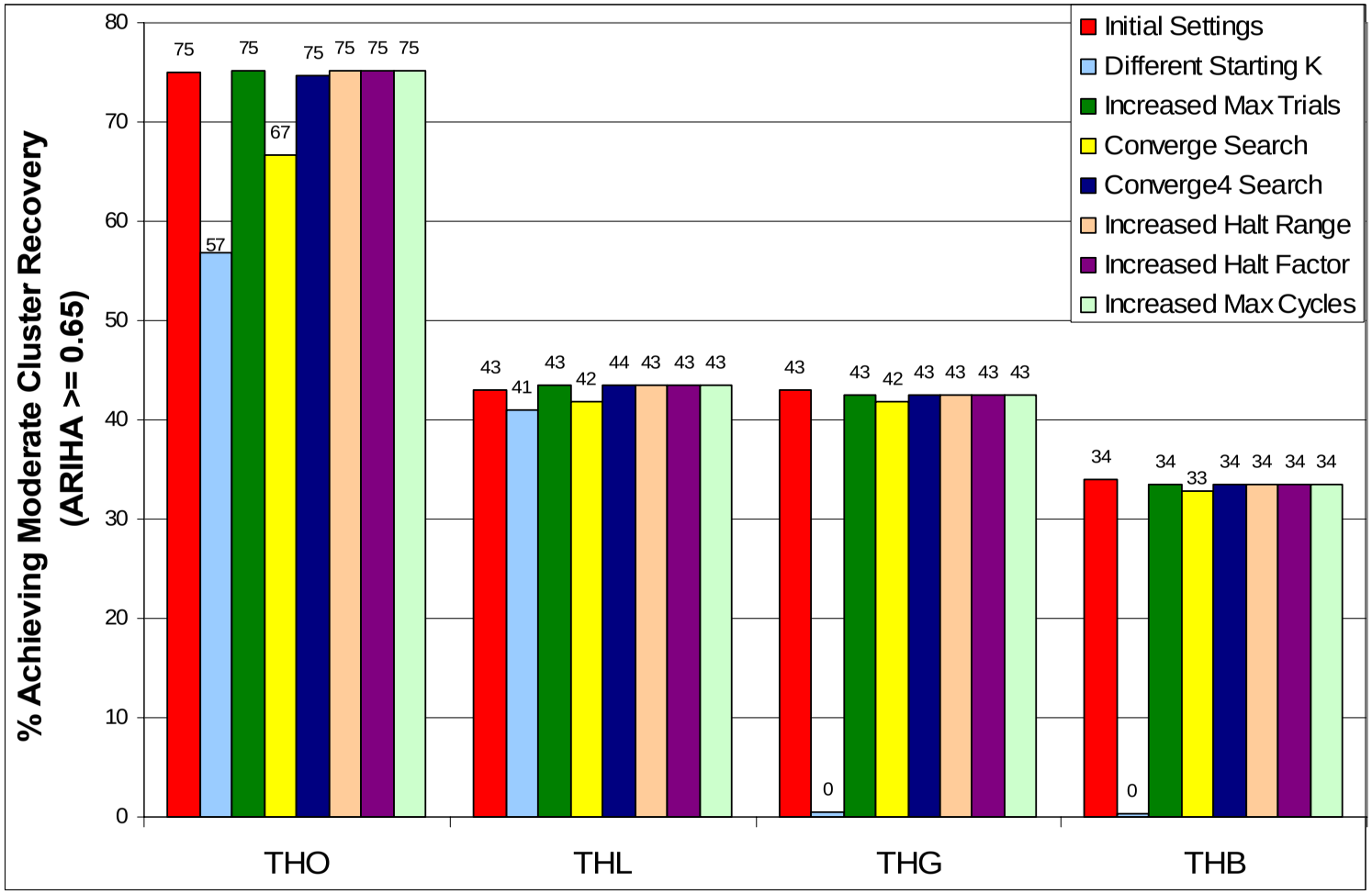


Figure 23. Moderate Cluster Recovery across Modified Parameter Settings. Method performance is measured as the percentage of clustered datasets achieving moderate cluster recovery, as determined by having a Hubert-Arabie Adjusted Rand Index value of >= 0.65.

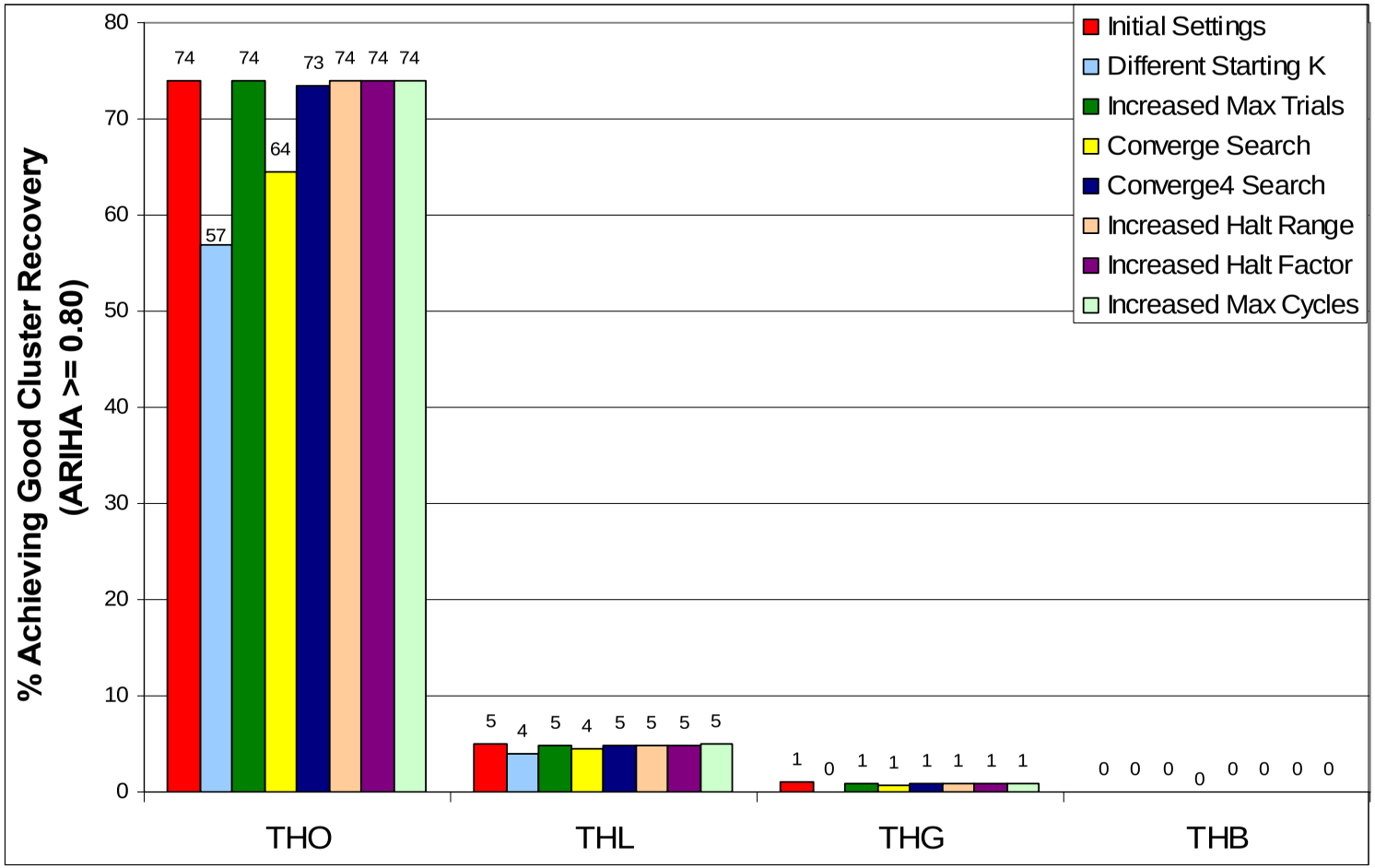


Figure 24. Good Cluster Recovery across Modified Parameter Settings. Method performance is measured as the percentage of clustered datasets achieving good cluster recovery, as determined by having a Hubert-Arabie Adjusted Rand Index value of >= 0.80.

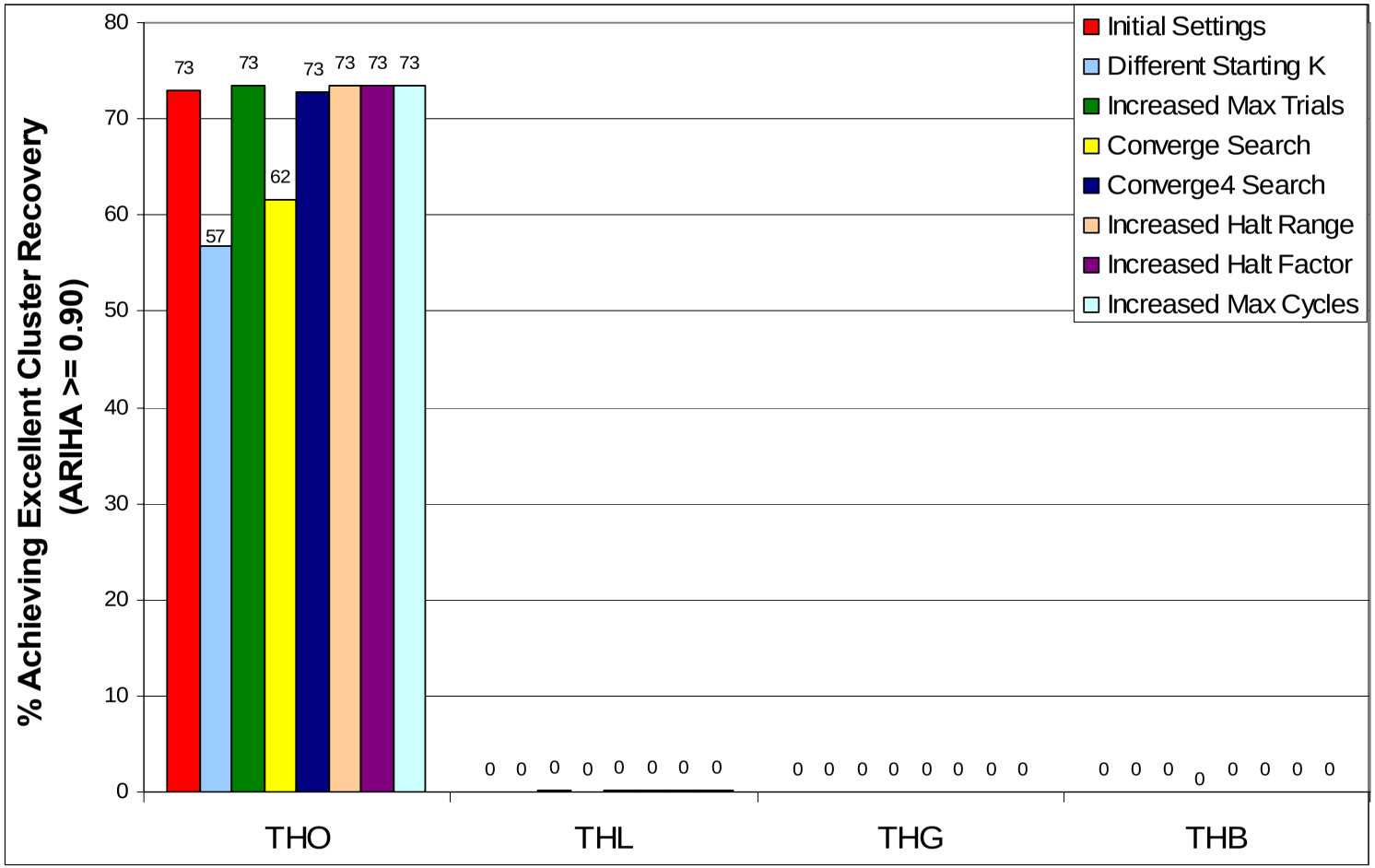


Figure 25. Excellent Cluster Recovery across Modified Parameter Settings. Method performance is measured as the percentage of clustered datasets achieving excellent cluster recovery, as determined by having a Hubert-Arabie Adjusted Rand Index value of >= 0.90.

false positive rates were still well-controlled, although less so for the THG datasets, where ten percent of the clustering results determined significant by the Bayesian Classification internal clustering metrics (alpha = 0.01) were actually not significant according to the Hubert-Arabie Adjusted Rand Index (ARI_{HA}) (Figure 26). Conversely, false negative rates were better for THG datasets than they were for the THL or the previously-evaluated THO datasets. At the most liberal alpha of 0.10, only sixteen percent of the clustering results deemed not significant by the internal clustering metrics were actually significant by the ARI_{HA} (Figure 26).

A more detailed breakdown of this same data is presented in Figure 27 showing how false positive and false negative rates track with the number of significant results by internal clustering metric and by ARI_{HA}, for each set of simulation conditions, where alpha is ten percent. Note that the vast majority of clustering results are significant by ARI_{HA} across all sets of conditions and that high error rates are very specific to certain sets of simulation conditions. False positive rates were at or below five percent for all sets of simulation conditions. Even in the worst case, for datasets simulated under the more complex THG model, with 10 (versus 100) nonfunctional loci, clusterings results still yielded false positive rates between 11 and 12 percent, very close to alpha of ten percent. False negative rates were near zero for most sets of conditions. However, for datasets containing 100 (versus 10) non-functional loci, the false positive rates ranged from two to 94 percent, with the highest rates for datasets with 500 (versus 1000) affecteds.

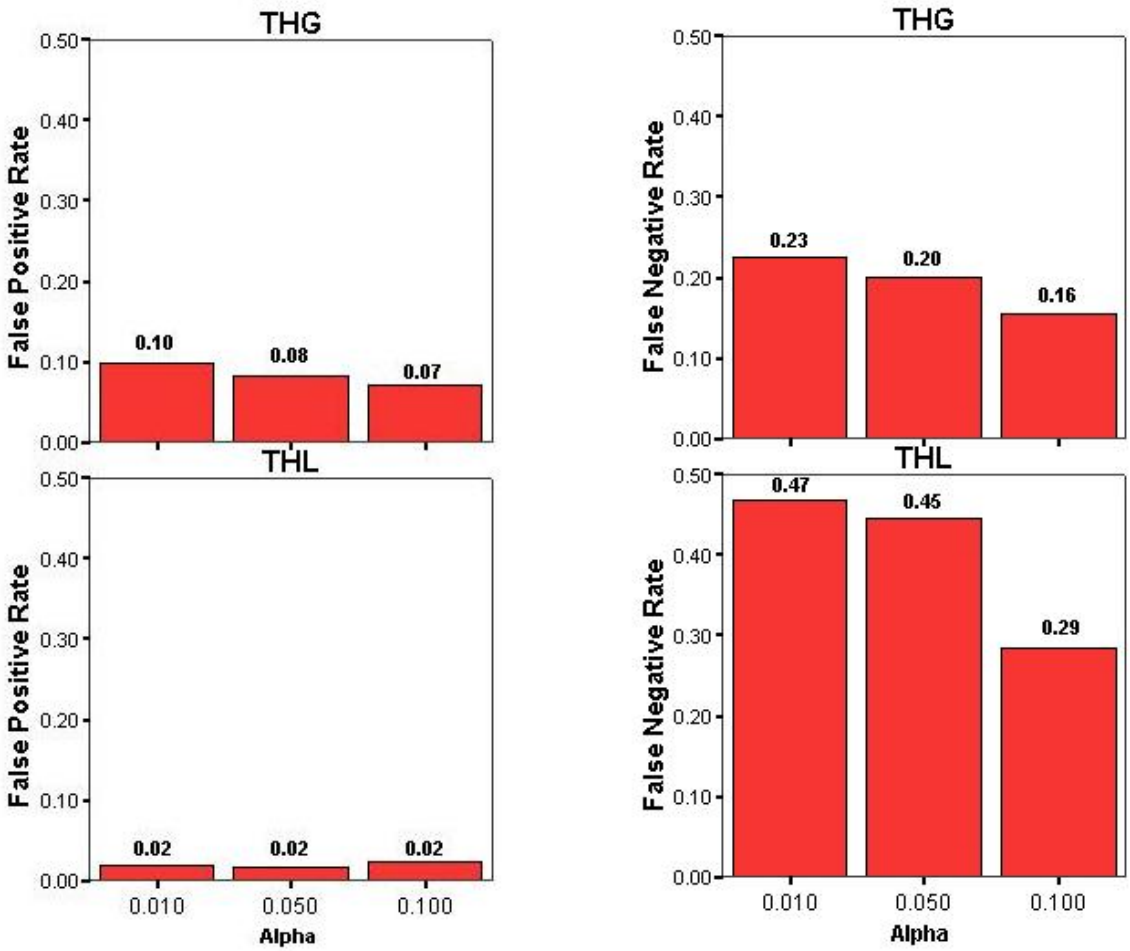


Figure 26. Error Rates for THG and THL Genetic Model Results. False positive rates are shown in the first column and false negative results are shown in the second column. Row one shows results for the Trait Heterogeneity with Gene-Gene Interaction genetic model datasets. Row two shows results for the Trait Heterogeneity with Locus Heterogeneity genetic model datasets. The three bars represents results at the significance levels (alpha) of one percent, five percent and ten percent.

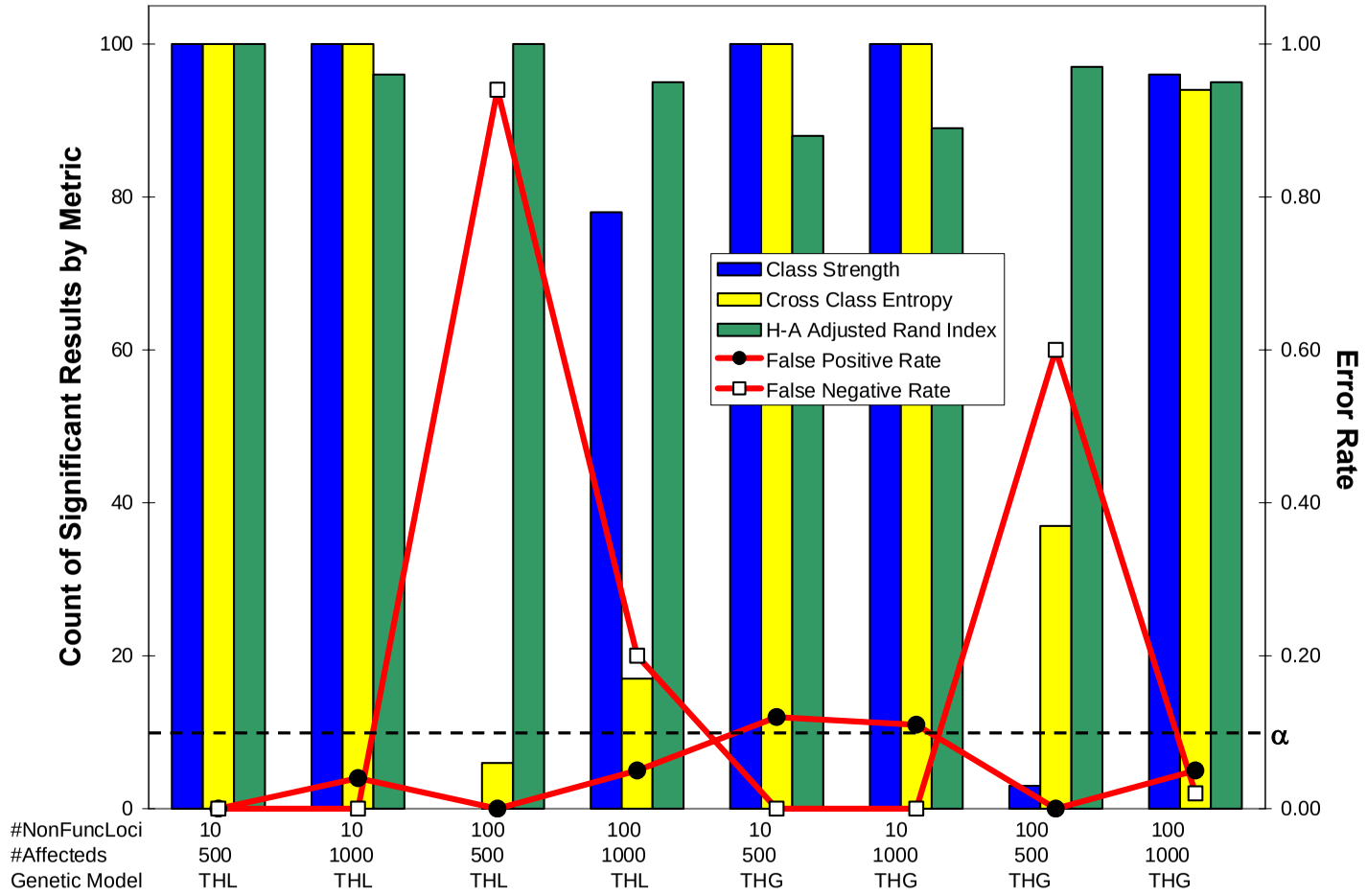


Figure 27 Permutation Testing Results at Alpha of Ten Percent. For each cluster of bars and line points, the x-axis specifies the dataset conditions for which the results are shown. Bars represent counts of the clustering results determined significant by the specified metric (see legend). Lines represent error rates as a fraction of one.

Discussion

Attempts to improve upon the performance of the Bayesian Classification method at the task of detecting heterogeneity in genotypic datasets were unsuccessful. The parameter settings chosen in the initial simulation study detailed in Chapter III, which were based on the methods' authors' recommendations and the characteristics of the simulated data, turned out to be as good as or better than any setting modifications applied in the current study.

Extended permutation testing of a wider set of simulation conditions provided insight into how the method's two internal clustering metrics—class strength and cross class entropy—compare to the gold standard of the Hubert-Arabie Adjusted Rand Index. The internal clustering metrics were biased based on the dataset characteristics of sample size and, especially, number of non-functional loci. This is important to keep in mind when the Bayesian Classification method is applied to real data, in which the underlying pattern of inheritance and presence (or degree) of heterogeneity and gene-gene interactions are unknown. These results suggest that one can place a high degree of confidence in a positive (significant) result based on permutation testing of the method's internal clustering metrics, but less so when the number of nonfunctional loci in the dataset is fairly low, since a dataset containing a gene-gene interaction may have a slightly inflated false positive rate under these conditions.

Interpretation of a negative (not significant) result is more difficult. Under the majority of conditions simulated, the false negative rate is very well controlled and the power to detect the underlying heterogeneity present in the data is high. However, there are large fluctuations in false negative rates due primarily to differences in the number of

results considered significant by the internal clustering metrics (Figure 27). For a dataset with a high number of non-functional loci (100) and a moderate sample size (500), the false negative rate may approach 100 percent, eliminating all power to detect heterogeneity when it exists, which is of course discouraging. Further simulation studies exploring the "breakpoint" or slope of the false negative rates between the two extremes of the current simulation conditions may further aid in interpretation of negative results.

Even though the current simulation study found that most results were significant by permutation testing (using ARI_{HA}), recall that performance as measured by the percentage of results achieving 'excellent', 'good' or 'moderate' cluster recovery was low (less than 45 percent of datasets achieving moderate cluster recovery) for the more complex datasets (Chapter III). One aspect of this issue is that the internal clustering metrics used by Bayesian Classification are biased under certain dataset conditions (discussed above). It is also possible that the null distribution we created, in which the relationship within multilocus genotypes was disrupted, was not the most appropriate choice for the question we were asking and was, therefore, leading to erroneous conclusions. The goal of the permutation testing is to test whether the clustering results, with their corresponding average class strength and average cross-class entropy values, have uncovered structure unlikely to be present (by chance) in data that has no real (functional) underlying structure. Perhaps we should permute only the genotypes of the known functional loci or of the loci with the highest influence values. In real data, since the functional loci are unknown as such, we would only be able to use influence values as a guide to choosing which loci to permute. This would disrupt the relationship(s) among loci already identified by the clustering algorithm as being the strongest, but it would

leave any other, presumably weaker multilocus genotype patterns in tact. Thus, if those patterns are sufficiently strong, that the clustering (on the original, unpermuted data) would not stand out as being significantly different from what could be found in the permuted data. However, this would shift the bias even more in a conservative direction, increasing the false negative rate, which we are interested in reducing.

There is also the question of how good we need the clustering results to be. Is moderate cluster recovery (ARI_{HA} >= 0.65) good enough to enable our statistical methods to find main effects and/or gene-gene interactions that were previously masked by heterogeneity? Is an ARI_{HA} of only 0.50 or even 0.35 good enough? To answer that question, we would need to perform main effect and gene-gene interaction tests on the simulated data before and after clustering and determine the power to detect the effect in the before and after datasets. If a clustering result with a certain ARI_{HA} leads to a substantial increase in the power to detect an effect in the data, then the method is working well, for our purposes. If, instead, only a clustering result achieving good cluster recovery (ARI_{HA} >= 0.80) aids in the detection of effects obscured by heterogeneity, then it is indeed very important that the relationship between ARI_{HA} and statistical significance based on permutation testing be well-understood and, if necessary, that the permutation testing procedure be modified to enable clearer interpretation of results.

CHAPTER V

APPLICATION OF TWO-STAGE ANALYSIS APPROACH TO LATE-ONSET ALZHEIMER DISEASE DATA

Background

Alzheimer's disease (AD; MIM: 104300) is a neurodegenerative disorder characterized clinically by a decline in two or more areas of cognition, one of which is usually episodic memory, in the absence of acute causes (Pericak-Vance MA and Haines JL, 2002). Presenting symptoms range from memory impairment to visuospatial disorientation, language impairment, depression and psychotic episodes. This range of symptoms suggests extensive cortical damage largely in the hippocampus but also in posterior-parietal areas, temporal-parietal systems or even frontal lobe areas (Fox NC and Rossor MN, 2000; Perry and Hodges, 2000; Roses, 1997; Small et al., 2000). While gross sensory and motor abnormalities generally rule out AD, some moderate disturbances similar to those seen with Parkinson Disease (PD), such as tremor, rigidity and bradykinesia, may instead suggest a distinct subtype—AD with Parkinson Disease (Brown et al., 1998; Chen et al., 1991; Mayeux et al., 1985; Molsa et al., 1984; Perry et al., 1997). While AD can occur as early as the third decade of life (Cruts et al., 1995), it most commonly occurs after the sixth decade. The age of onset for late-onset Alzheimer disease (LOAD) is generally defined to be after age 60 or 65 but extends into the ninth decade (Pericak-Vance MA and Haines JL, 2002). The prevalence of AD was estimated to be 13.5 million worldwide and 4.5 million in the United States in 2000, with

projections for 2005 up to 21.2 million worldwide (Hebert et al., 2003; Katzman R and Fox P, 1999).

AD is defined pathologically by the presence of two abnormalities in the cerebral cortex. The first is senile plaques that have an amyloid beta (Aβ) protein core, and the second is neurofibrillary tangles, which contain the microtubule-associated protein tau (Goedert M, 1999; Wisniewski et al., 1993). It remains controversial whether the plaques and tangles are themselves pathogenic or whether they are merely "tombstones" of other pathogenic processes (Glabe C, 2000). Only a weak link between plaque load and severity of illness has been found, while the load of neurofibrillary tangles may be more strongly correlated with severity (Guillozet et al., 2003; Mufson et al., 1999). Also, both plaques and tangles have been found in normal older adults, leading many to suggest that these abnormalities are secondary effects arising from the true pathological mechanisms underlying AD. In addition, Lewy bodies, which contain fibrils of aggregated, insoluble alpha-synuclein (McKeith et al., 2004), have been observed in up to 20% of AD cases in the substantia nigra (which is characteristic of PD) and elsewhere in the brain (Ditter and Mirra, 1987; Growden, 1995; McKeith et al., 1996). A growing body of literature suggests substantial overlap among AD, dementia with Lewy bodies, and Parkinson Disease (Pericak-Vance MA and Haines JL, 2002). It is possible that the developments of Aβ plaques, neurofibrillary tangles and Lewy bodies have common physiological pathways. However, it is also possible each one of these features (plaques, tangles and Lewy bodies) is a distinct trait, with its own etiology, which would mean that AD is a heterogeneous trait that would be better defined as the coincident state of having the trait for plaques and the trait for tangles. Likewise, AD with PD could then be better

described as the concomitance of the three traits for plaques, tangles and Lewy bodies. Such dissection and categorization of AD is speculative and controversial but not without support.

AD has a strong, albeit complex, genetic component, as evidenced by recent family-based studies reporting sibling recurrence risk ratios between 4 and 5, indicating that a sibling of a person with LOAD is 4-5 times more likely to develop LOAD than someone in the general population (Breitner et al., 1988; Hirst et al., 1994; Sadovnick et al., 1989). Also, twin studies show a concordance rate of 0.49 for monozygotic twins versus 0.18 for dizygotic twins (Bergem, 1994). This demonstrates that there is an almost 3 fold increased risk of developing AD for siblings that share all, versus (on average) half, of their genes with their affected twin. Still, the fact that the monozygotic concordance rate is far from 100 percent suggests that other factors, including environment, are likely involved. In addition, segregation analyses of LOAD show a complex genetic etiology with multiple genes and environmental factors involved (Daw et al., 1999; Daw et al., 2000; Pericak-Vance MA and Haines JL, 2002; Rao et al., 1994; van Duijn et al., 1993). Some environmental risk factors under investigation include head trauma, plasma homocysteine levels and non-steroidal anti-inflammatory drugs, the last of which is purported to have a protective effect (Andersen et al., 1995; Breitner et al., 1995; Mayeux et al., 1995; Roberts et al., 1994; Seshadri et al., 2002).

The only known gene conferring risk for LOAD is apolipoprotein E (APOE). It is estimated that at least fifty percent of the genetic effect of LOAD remains unexplained (Daw et al., 2000; Roses AD et al., 1995; Slooter et al., 1998). Over 115 LOAD candidate genes have been tested and have generated a positive main effect, but all except

APOE have failed to be consistently replicated (Pericak-Vance MA and Haines JL, 2002) (Figure 28). While the initial reports may have been false positive findings, alternatively, these inconsistencies could be indicative of heterogeneity and/or environmental interactions across the entire phenotype. Reported differences of incidence and prevalence between ethnic and gender groups are also indicative of interactions with environment and/or genetic background. The possibility of gene-gene interactions has been explored only superficially (Pericak-Vance MA and Haines JL, 2002).

Late Onset Alzheimer Disease is just one example of a complex disease, in which traditional statistical methods of analysis such as linkage and association have failed to identify main effect genes. Among the possible reasons for this failure are false positives due to population stratification and true differences in genetic etiology between study populations (Hirschhorn JN et al., 2002). In addition, while a small number of supervised computational methods exist for discovering gene-gene interactions, the power of these methods drops dramatically when locus or trait heterogeneity is present (Ritchie et al., 2003a). Current statistical approaches for detecting heterogeneity, such as the admixture test (Ott J and Hoh J, 2003; Smith, 1963), are neither sensitive nor powerful and can merely account for, not resolve, any underlying heterogeneity (see Chapter II).

It is possible that phenotypic data could be utilized to improve the performance of these methods in the face of locus or trait heterogeneity by facilitating heuristic stratification of data. For instance, age of onset data was used to stratify AD patients, leading to the detection of association with the apolipoprotein E4 allele in late-onset and sporadic cases (Saunders et al., 1993; Strittmatter et al., 1993). However, for most diseases, particularly neurological diseases, little detailed phenotypic data has been

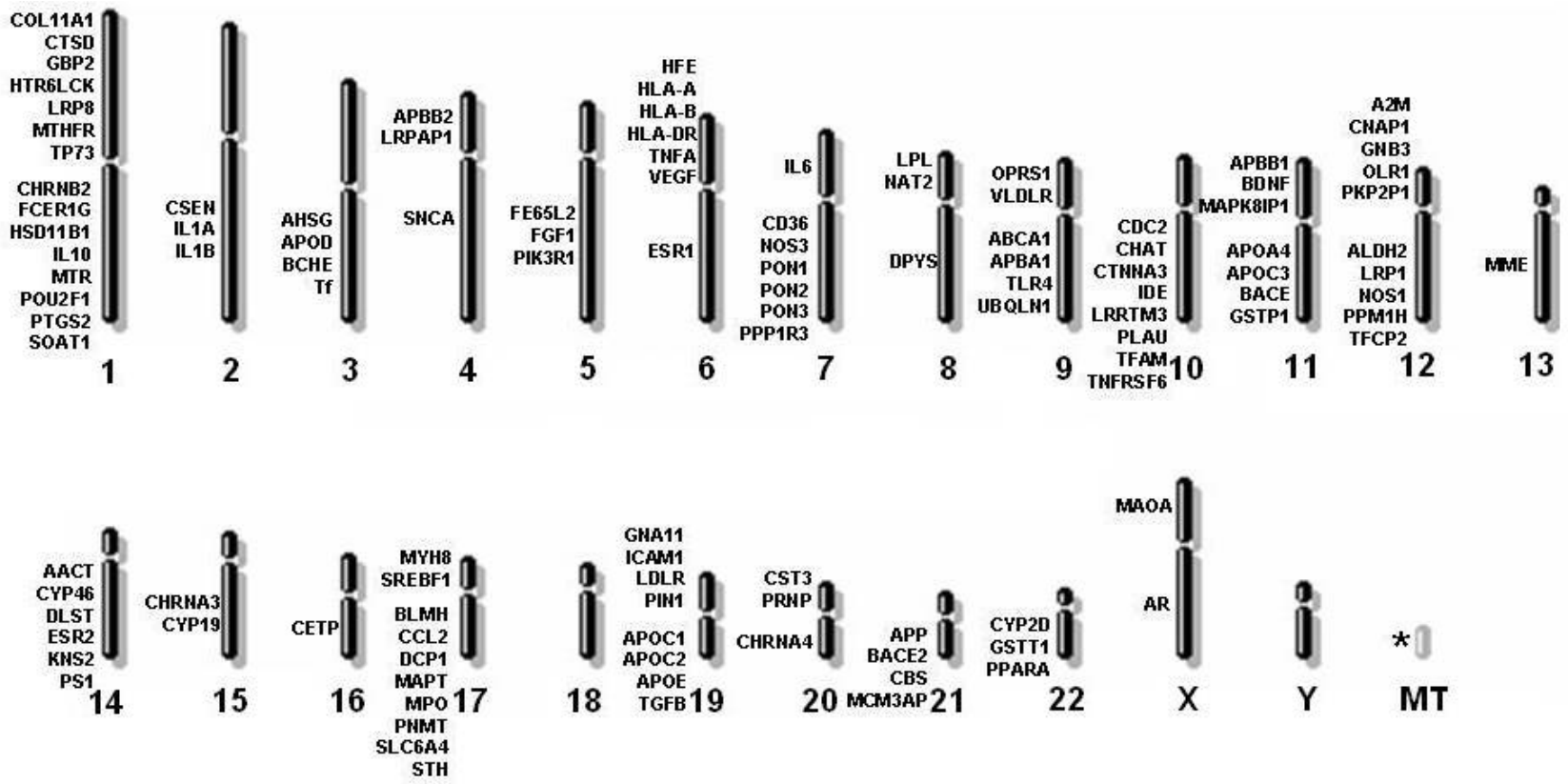


Figure 28 Candidate Genes for Late-Onset Alzheimer Disease. Each of these 116 genes shown next to their respective chromosome has received at least one positive report for linkage or association with LOAD in the literature.

consistently collected in combination with genotypic data. Postmortem histological data are rare for cases, even rarer for controls, and neuroimaging can be expensive and challenging with mentally ill patients. It is for these reasons that an unsupervised method, such as the Bayesian Classification method investigated in Chapters III and IV, which does not rely on phenotypic data, would be valuable to mine potentially heterogeneous genotypic data as a means of data stratification and hypothesis generation.

In Chapter II, a comprehensive two-step approach to analysis was proposed in which heterogeneity is first addressed and then main effects and interactions are subsequently investigated in the more homogeneous subsets discovered in the first stage. In this chapter, an application of this two-stage approach to a LOAD dataset is presented in which cluster analysis is first used to uncover heterogeneity and to subdivide the data into more homogeneous groups. Then in the second stage, traditional linkage and association tests are used to detect main effects and a computational data reduction method is used to investigate gene–gene interactions within each of the subgroups.

Methods

Specifics of Late-Onset Alzheimer Disease Dataset

The late-onset Alzheimer Disease dataset includes samples obtained by (1) Dr. Jonathan L. Haines at Vanderbilt University, Dr. Pericak-Vance at Duke University and Dr. Gary Small at UCLA of the Collaborative Alzheimer Project (the CAP dataset), (2) the Indiana Alzheimer Disease Center National Cell Repository (the IU dataset), and (3) the National Institute of Mental Health Alzheimer Disease Genetics Initiative dataset (the

NIMH dataset). Although the NIMH and IU datasets represent a rich resource for generating hypotheses, they are in use by multiple groups (including CAP). In contrast, the CAP dataset represents an independent set of families that can, therefore, be used to confirm and extend initial findings.

All subjects are Caucasian Americans. Written consent was obtained from all participants in agreement with protocols approved by the institutional review board at each contributing institution. Alzheimer Disease was diagnosed according to the NINCDS-ADRDA criteria (McKhann et al., 1984). Age of onset was recorded as the age at which the first symptoms were noted by the participant or family member. Only subjects with an age of onset of 65 or greater were included in this late-onset dataset.

Markers previously genotyped in over 25 candidate genes and a region of interest (ROI) on chromosome 10 were included in the dataset. The data were then 'cleaned' to remove markers and subjects with high percentages of missing data. This was an iterative process that resulted in a dataset with 148 markers in the chromosome 10 ROI and in 22 candidate genes residing on eight different chromosomes. All chosen markers were genotyped in at least 90 percent of included subjects (Figures 29 and 30), and all chosen subjects were genotyped for greater than 85 percent of the included markers (Figures 31 and 32).

Most of the functional candidate genes chosen here are purported to have some role in LOAD through their involvement in the processing of amyloid precursor protein (APP; MIM: 104760), the secretion of its product, Aβ, and/or the phosphorylation of tau or regulation of microtubules within neurons. Table 5 lists alphabetically the 22 genes

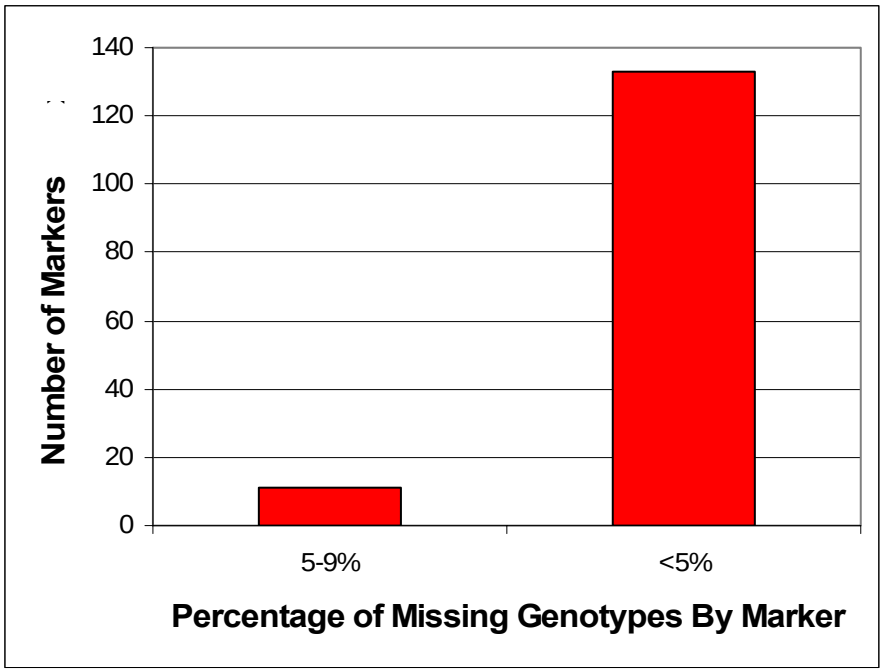


Figure 29. Family-Based Data: Percentage of Missing Genotypes by Marker

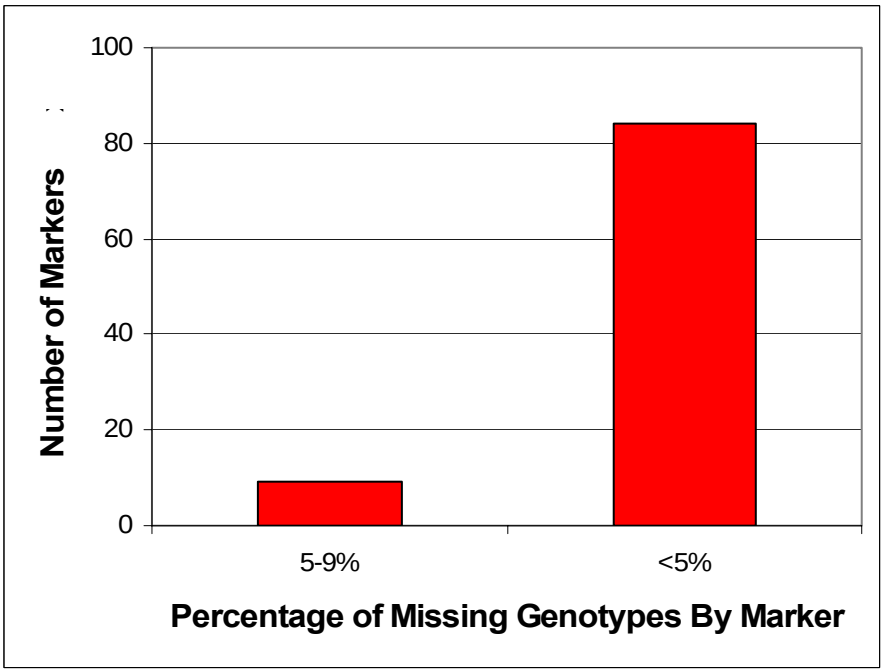


Figure 30. Case-Control Data: Percentage of Missing Genotypes by Marker

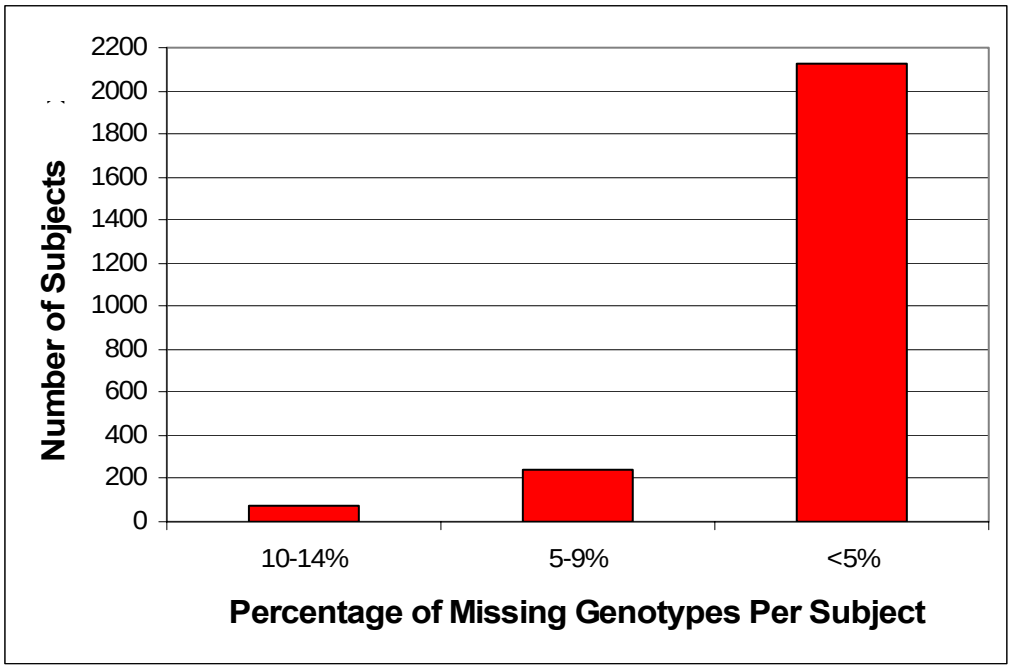


Figure 31. Family-Based Data: Percentage of Missing Genotypes by Subject

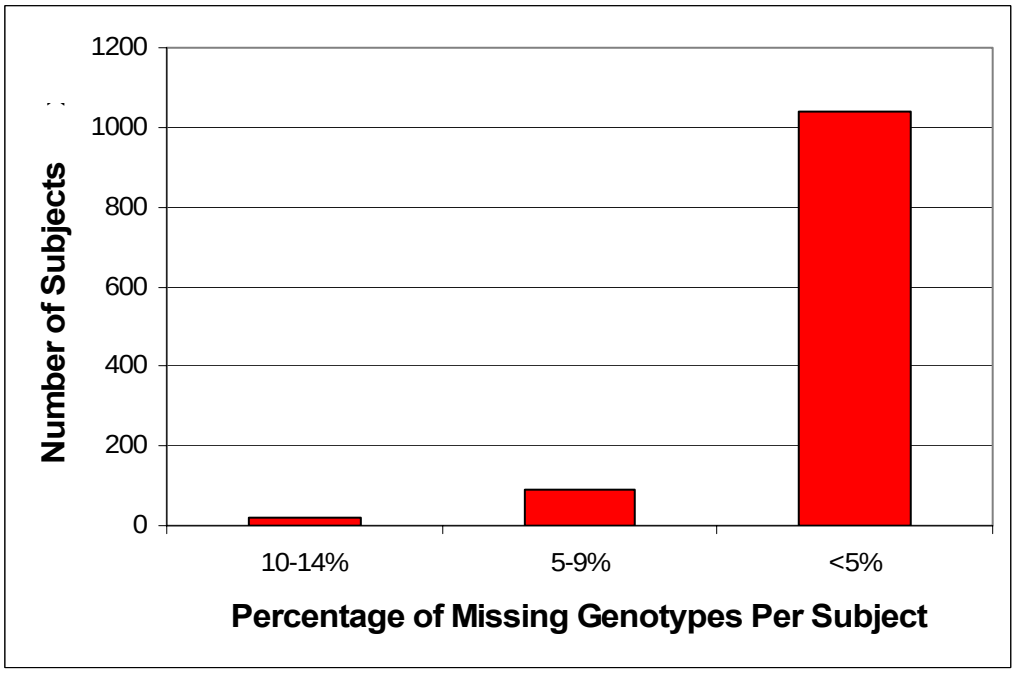


Figure 32. Case-Control Data: Percentage of Missing Genotypes by Subject

genotyped in one or both of the samples, along with their full names and identification numbers in the Mendelian Inheritance in Man (MIM) and Entrez Genome databases of the National Center for Biotechnology Information (NCBI).

The family-based dataset, derived from all three ascertainment sources, consists of 654 families with 1422 subjects with possible, probable or definite LOAD and 744 cognitively normal elderly individuals. Of these families, 328 contain a total of 1279 discordant sibling pairs (DSPs), in which one sibling is affected with LOAD and the other is unaffected. For this sample, there are 138 markers genotyped in 22 genes on 8 chromosomes, plus the ROI on chromosome 10. The CAP dataset also includes a clinic-based unrelated case-control sample of 451 cases with possible, probable or definite LOAD and 699 cognitively normal elderly controls who were either spouses of AD patients or subjects recruited from outpatient clinics at the participating institutions. For this case-control sample, there are 93 markers genotyped in 19 genes on eight chromosomes. Across the family-based and case-control samples, there are 82 markers in common, covering 18 genes on 8 chromosomes and the ROI on chromosome 10. Table 6 lists all markers genotyped, giving their chromosomal location and noting whether they are genotyped in the family-based dataset and/or the case-control dataset. One marker, labeled '1920', is actually a combination of two adjacent single nucleotide polymorphisms—rs2456777 and rs2456778—that could not be distinguished by the Taqman probe used for genotyping (Liang X et al., 2006).

Table 5. Genes Covered by Markers Genotyped in One or Both Samples

Symbol	Location	Name	MIM ID	Gene ID
A2M	12p13.3-p12.3	Alpha-2-macroglobulin	103950	2
A2MP	12p13.3-p12.3	Alpha-2-macroglobulin pseudogene	-	3
ACE	17q23.3	Angiotensin 1 converting enzyme (petidyl-dipeptidase A)	106180	1636
AGT	1q42-q43	angiotensinogen	106150	183
APOE	19q13.2	Apolipoprotein E	107741	348
CDC2	10q21.1	Cell division cycle 2	116940	983
COG2	1q42.2	component of oligomeric Golgi complex 2	606974	22796
GAPDH	12p13	Glyceraldehydes-3-phosphate dehydrogenase	138400	2597
GAPDHS	19q13.1	Glyceraldehydes-3-phosphate dehydrogenase, spermatogenic	609169	26330
IDE	10q23-q25	Insulin degrading enzyme	146680	3416
LIPC	15q21-q23	Lipase, hepatic	151670	3990
LRP1	12q13-q14	Low-density lipoprotein receptor-related protein 1	107770	4035
LRRTM3	10q21.3	Leucine-rich repeat transmembrane neuronal 3 protein	-	347731
LTA	6p21.3	Lymphotoxin alpha (TNF superfamily, member 1)	153440	4049
OLR1	12p13.2-p12.3	Oxidized density lipoprotein (lectin-like) receptor 1	602601	4973
PLAU	10q24	Urokinase-type plasminogen activator	191840	5328
PPM1H	12q14.1-q14.2	Protein phosphotase 1H (PP2C domain containing)	-	57460
PZP	12p13-p12.2	Pregnancy-zone protein	176420	5858
TNF	6p21.3	Tumor necrosis factor (TNF superfamily, member 2)	191160	7124
TNFRSF6 / FAS	10q24.1	Necrosis factor receptor superfamily member 6	134637	355
UBQLN1	9q21.2-q21.3	Ubiquilin 1	605046	29979
VR22 / CTNNA3	10q22.2	Catenin (cadherin-associated protein), alpha 3	607667	29119

Table 6. Markers Genotyped in Family-Based and Case-Control Samples. Chromosomal location is given according to NCBI dbSNP Human Build 126. Markers with no gene listed were chosen to cover the region of interest on chromosome 10.

Chrom	Gene	Marker	NCBI Location	Present in Fam	Present in CC
1	COG2	rs3789662	227135608	X	X
1	AGT	rs7536290	227143437	X	X
1	AGT	rs3789670	227150449	X	X
1	AGT	rs2478545	227150856	X	X
1	AGT	rs4762	227152712	X	X
1	AGT	rs2148582	227156534	X	X
1	AGT	rs5051	227156607	X	X
1	AGT	rs1326886	227166495	X	X
6	LTA	rs1799724	31650461	X	X
6	TNF	rs1800750	31650942	X	X
6	LTA	rs1800629	31651010	X	X
6	LTA	rs361525	31651080	X	X
6	TNF	rs4645843	31652541	X	X
9	UBQLN1	rs7866234	83508371		X
9	UBQLN1	rs2781003	83508569	X	X
9	UBQLN1	rs2781002	83508579		X
9	UBQLN1	rs12344615	83510749		X
9	UBQLN1	rs2780995	83520722		X
9	UBQLN1	rs10868038	83521233	X	X
9	UBQLN1	rs11140213	83531038	X	X
10		rs10826594	29623140	X	
10		rs1023207	32134896	X	
10		rs1319013	33583935	X	
10		rs1148247	35536952	X	
10		rs6482044	37892393	X	
10		rs6593491	42585568	X	
10		rs1890739	45074179	X	
10		rs1806797	48357923	X	
10		rs7097397	49695402	X	
10		rs14327	51735896	X	
10		rs1904018	53523252	X	
10		rs4998401	55575412	X	
10		rs4935648	57804443	X	
10		rs10763551	59943904	X	
10	CDC2	1920	61896492	X	X
10	CDC2	rs7919724	62165848	X	X
10	CDC2	rs2448341	62205963	X	X
10	CDC2	rs2448347	62215148	X	X
10		rs7090884	63632032	X	
10		rs1935	64597829	X	
10		rs7089698	65054573	X	

Table 6, continued. Markers Genotyped in Family-Based and Case-Control Samples.

Chrom	Gene	Marker	NCBI Location	Present in Fam	Present in CC
10	VR22	4783	67208785	X	X
10	VR22	rs1786927	67352267	X	X
10	VR22	rs2126750	67507709	X	
10	VR22	rs4745886	67530329	X	X
10	VR22	rs7911820	67534145	X	X
10	VR22	rs7070570	67534610	X	
10	VR22	rs7074454	67534965	X	
10	VR22	rs10822719	67535076	X	X
10	VR22	rs6480140	67538887	X	
10	VR22	rs922347	67652964	X	X
10	VR22	rs4463744	67778486	X	X
10	VR22	rs2441718	67806967	X	X
10	VR22	rs2939947	67808364	X	X
10	VR22	rs2456737	67825340	X	X
10	VR22	rs4746606	68061108	X	X
10	VR22	rs7909676	68104803	X	X
10	LRRTM3	rs1001016	68347044	X	X
10	LRRTM3	rs12769870	68347401	X	X
10	LRRTM3	rs1925583	68349950	X	X
10	LRRTM3	rs2394314	68350254	X	X
10	LRRTM3	rs1925577	68358439	X	
10	LRRTM3	rs10762122	68386380	X	X
10	LRRTM3	rs942780	68406547	X	X
10	LRRTM3	rs1925617	68434823	X	X
10	LRRTM3	rs1925622	68439644	X	X
10	LRRTM3	rs1925632	68469620	X	X
10	LRRTM3	rs1952060	68472940	X	X
10	LRRTM3	rs2147886	68488649	X	X
10	LRRTM3	rs2251000	68494777	X	X
10	LRRTM3	rs2764807	68498938	X	X
10	LRRTM3	rs10762136	68513538	X	X
10	VR22	rs11593235	68546044	X	X
10	VR22	rs10997591	68671884	X	X
10	VR22	rs7903421	68951738	X	X
10	VR22	rs3096244	69080192	X	X
10		rs870801	71599752	X	
10		rs1227047	73104105	X	
10	PLAU	rs1916341	75341168	X	X
10	PLAU	rs2227564	75343107	X	X
10	PLAU	rs2227566	75343737	X	X
10	PLAU	rs2227568	75343885	X	X
10	PLAU	rs4065	75346470	X	X
10		rs1898071	77477033	X	
10		rs1439042	80374264	X	
10		rs1336439	82822237	X	
10		rs11816558	84709583	X	
10		rs3750686	87198514	X	

Table 6, continued. Markers Genotyped in Family-Based and Case-Control Samples.

Chrom	Gene	Marker	NCBI Location	Present in Fam	Present in CC
10	TNFRSF6	rs1800682	90739943	X	X
10	TNFRSF6	rs1324551	90741496	X	X
10	TNFRSF6	rs2031612	90756960	X	X
10	TNFRSF6	rs2296600	90760419	X	X
10		rs4933194	92501347	X	
10	IDE	rs2251101	94201284	X	X
10	IDE	rs1832196	94258314	X	X
10	IDE	rs7076966	94315491	X	X
10	IDE	rs4646954	94323807	X	X
10	IDE	rs3758505	94324758	X	X
10	IDE	rs7099761	94325779	X	X
10	IDE	rs1544210	94477781	X	X
10		rs701865	95371763	X	
10		rs4372378	97234998	X	
10		rs2039826	99516658	X	
10		rs2255901	101629786	X	
10		rs3127242	103303589	X	
10		rs7084783	105314160	X	
10		rs2058980	107379174	X	
10		rs10509859	109803462	X	
12	GAPD	rs7307229	6513864		X
12	GAPD	rs3741916	6514252		X
12	GAPD	rs3741918	6514517		X
12	GAPD	rs1060621	6514957		X
12	GAPD	rs1060620	6514983		X
12	GAPD	rs1060619	6515042		X
12	A2M	rs1800433	9123618	X	
12	A2M	rs3832852	9137444	X	
12	PZP	rs10842971	9194563	X	X
12	PZP	rs3213831	9208040	X	X
12	PZP	rs2277413	9209051	X	X
12	PZP	rs3213832	9212768	X	X
12	PZP	rs12230214	9238059	X	X
12	A2MP	rs16918212	9276225	X	X
12	A2MP	rs34362	9276692	X	X
12	A2MP	rs17804080	9279277	X	X
12	OLR1	rs1050283	10203556	X	
12	LRP1	rs1799986	55821533	X	
12	LRP1	rs1800127	55825349	X	
12	LRP1	rs1800174	55846076	X	
12	LRP1	rs1800181	55864555	X	
12	LRP1	rs2075699	55871411	X	
12	LRP1	rs1800154	55875926	X	
12	LRP1	rs1800165	55877493	X	
12	LRP1	rs11172124	55881222	X	
12	LRP1	rs9669595	55881333	X	
12	LRP1	rs7956957	55889082	X	

Table 6, continued. Markers Genotyped in Family-Based and Case-Control Samples.

Chrom	Gene	Marker	NCBI Location	Present in Fam	Present in CC
12	PPM1H	rs2029721	61435611	X	X
15	LIPC	rs6078	56621285	X	X
15	LIPC	rs6083	56625302	X	X
17	ACE	rs4291	58907926	X	X
17	ACE	rs4295	58910030	X	
17	ACE	rs4311	58914495	X	
17	ACE	rs4329	58917190	X	
17	ACE	rs4646994	58919636	X	X
17	ACE	rs4343	58919763	X	X
17	ACE	rs4353	58924154	X	
17	ACE	rs4978	58927493	X	
19	GAPDS	rs4806173	40716765	X	X
19	GAPDS	rs12984928	40721692		X
19	APOE	rs440446	50101007	X	X

Statistical Analysis

A comprehensive, two-stage approach to analysis was performed in which heterogeneity was first investigated in the dataset and then main effects and gene-gene interactions were investigated among the resulting subsets or clusters of data. Although all of the markers in the dataset had been previously tested for main effects and some even for interactions, this testing was performed at different time points over the past 10 years and, therefore, the samples on which they were tested vary to different degrees from the sample being analyzed in the current study. It is for this reason that a preliminary analysis of the complete datasets was performed prior to the two-stage analysis, using all the main effect and interaction-detection methods proposed for the subsets of data.

Analysis of deviations from Hardy-Weinberg equilibrium (HWE) and linkage equilibrium were tested using the Haploview program (Barrett et al., 2005) on the complete case-control and family-based datasets. Hardy-Weinberg Equilibrium

stipulates the expected ratio of individuals in a population who have each of a marker's possible genotypes, based solely on that marker's allele frequencies. Deviations from HWE in a sample could be indicative of genotyping error or a violation of one HWE's assumptions—random mating, no selection, no mutation, no migration and infinite or large sample size. Alternatively, it could be evidence for association. Linkage disequilibrium (LD) is the statistically observed (population) phenomenon of two or more segments of DNA being observed together more often that would be expected by chance. When LD exists between two or more markers, there is essentially one signal or effect coming from those markers. If one or more of the markers in LD exhibit an association with disease, it could be any one of those markers (or another variant not genotyped in the dataset that is also in LD with one or more of these markers) that is the functionally relevant one.

The Bayesian Classification method (Cheeseman P and Stutz J, 1996; Hanson R et al., 1991), previously investigated in simulation studies described in Chapters III and IV, was used to detect heterogeneity. For the family-based and case-control data, separately, the affected individuals in the dataset were subjected to cluster analysis, and the resulting clustering created subsets, which were more homogeneous than the complete dataset. Each cluster subset was then recombined with the entire group of unaffected individuals from the respective dataset for subsequent analysis of main effects and interactions.

For the family-based data, two-point heterogeneity lod score (HLOD) linkage analysis using FASTLINK and HOMOG (Ott, 1999) and two methods for detecting main effect association—the family-based association test (FBAT) (Horvath et al., 2001) and

the pedigree disequilibrium test (PDT) (Martin et al., 2000; Martin et al., 2001)—were performed. Linkage analysis tests whether a marker and a disease locus co-segregate within families (according to a specific genetic model), in violation of Mendel's laws, which would suggest that the disease susceptibility allele is at or near the marker in question. Both recessive and dominant disease models are tested, and the maximum heterogeneity lod score, which is the highest lod score found for either model under the range of full range of possible theta values, is reported. Tests for allelic association are nonparametric and detect deviations in the expected frequency of a marker allele with respect to disease status, which would suggest that the disease susceptibility allele is, or is in linkage disequilibrium with, the marker in question. The FBAT for allelic association uses data from discordant sibpairs and from nuclear families (decomposing extended pedigrees, if present, into nuclear families), whereas the PDT can use data from discordant sibpairs, from nuclear families, and from intact extended pedigrees (without decomposition and accounting for intrafamilial correlation). For the case-control data, a chi-square test of independence was used to detect main effect associations. In each case, a genotype-based model was tested in which the distribution of cases to controls at each of the possible genotypes was compared.

For both the family-based and the case-control datasets, the multifactor dimensionality reduction (MDR) method was used to detect gene-gene interactions (Hahn et al., 2003; Ritchie et al., 2001). MDR is a nonparametric data reduction computational method that performs an exhaustive search of the data space, looking for combinations of genetic markers and/or environmental factors whose genotypes or levels, when reduced to a single risk variable with two levels—high- and low-risk—predict disease status.

Using 5-fold cross validation, we measured the average balanced prediction accuracy (across the five cross-validation intervals) of every possible combination the best one-, two- and three-way MDR models. Accuracy is a function of the percentage of true positives (TP) and true negatives (TN), defined as TP/(TP+FN) (Moore et al., 2006). Because each of the datasets tested were unbalanced to some degree—meaning that the number of affecteds differed substantially from the number of unaffecteds—the metric 'balanced accuracy' was actually used, along with an adjusted threshold for determining risk status. The adjusted threshold further corrects for the imbalance in the data by comparing the ratio of affecteds to unaffecteds with the particular multilocus genotype being considered to the ratio in the overall dataset. For each of the one-locus, two-locus and three-locus combinations, the 'best' MDR model was chosen as the one with the best average balanced prediction accuracy. All 'best' MDR models were evaluated for statistical significance using permutation testing with 1000 permutations.

For each of the 'best' two- and three-marker MDR models achieving prediction accuracy of 55 percent or greater, the markers in those MDR models were used in logistic regression analyses to further characterize the underlying statistical models. Logistic regression can determine the structure of the model, in terms of whether markers are influencing or predicting disease status primarily through independent (main) effects or through interactions with each other. One can also obtain odds ratios from logistic regression, which are helpful in interpreting these models. For the case-control data, a logistic regression analysis was performed in SPSS, and for the family-based data, a multivariate logistic regression method, which controls for intrafamily correlation, was

implemented in SAS (Martin ER et al., 2005; Siegmund et al., 2000) and applied to all discordant sibpairs.

Results

Analysis of Complete Datasets

Linkage analysis of the complete family-based dataset detected the known effect of APOE (HetLOD = 7.963) and other marginal linkage scores (HetLOD between 1 and 1.5) for one marker in AGT and four markers in VR22. The FBAT detected the known association of APOE (χ^2=86.989, df=2, p<0.001) as well as two substantial effects (χ^2=13.876, df=1, p<0.001 and χ^2=9.085, df=1 p=0.003) and one marginal effect (χ^2=4.343, df=1, p=0.037) in ACE, and five other marginal effects (χ^2>4.2, p<0.05) in LRRTM3, PLAU and A2MP. The PDT detected the known association with APOE (χ^2=98.388, df=2, p<0.001), two other substantial effects—one in A2M (χ^2=6.772, df=1, p=0.009) and one in ACE (χ^2=7.104, df=1, p=0.008)—and 10 other marginal effects (χ^2> 4.5, p<0.05). Table 7 presents results from all three tests on all markers showing statistically significant effects (p<0.05) according to at least one test. Analysis using the chi-square test of independence on the complete case-control dataset detected the known association with APOE (χ^2=171.62, df=5, p<0.001) and seven other marginal effects in CDC2, VR22, LRRTM3 and GAPDH (χ^2>6.2, p<0.05) (Table 8).

MDR gene-gene interaction analysis was performed on both the complete family-based and complete case-control datasets. Since MDR works by comparing the ratio of affected to unaffected individuals but does not account for intrafamilial correlations, for

Table 7. Main Effect Analysis Results for Complete Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	2-Pt Linkage Max HetLOD	PDT Chi-Square	PDT p-value	FBAT ChiSquare	FBAT p-value
1	AGT	rs5051	1.033	0.114	0.736	0.010	0.921
10	VR22	rs7070570	1.366	1.416	0.702	0.248	0.619
10	VR22	rs2441718	1.407	0.312	0.577	1.771	0.183
10	VR22	rs2456737	1.038	1.143	0.285	2.835	0.092
10	VR22	rs7909676	1.068	4.540	0.033	2.682	0.101
10	LRRTM3	rs1925622	0.302	2.849	0.091	4.285	0.038
10	LRRTM3	rs1925632	0.140	2.052	0.152	5.283	0.022
10	LRRTM3	rs2764807	0.097	3.556	0.059	4.462	0.035
10	PLAU	rs2227568	0.000	5.170	0.023	3.446	0.063
10	PLAU	rs4065	0.000	3.152	0.076	4.987	0.026
10		rs4933194	0.052	4.676	0.031	0.886	0.347
12	A2M	rs3832852	0.011	6.772	0.009	1.587	0.208
12	A2MP	rs34362	0.047	0.904	0.342	4.673	0.031
12	LRP1	rs1800154	0.000	4.017	0.045	2.145	0.143
12	LRP1	rs9669595	0.003	4.599	0.032	1.939	0.164
12	LRP1	rs7956957	0.000	4.059	0.044	2.343	0.126
17	ACE	rs4291	0.000	7.104	0.008	13.876	< 0.001
17	ACE	rs4295	0.000	3.23	0.072	9.085	0.003
17	ACE	rs4646994	0.000	5.481	0.019	3.056	0.080
17	ACE	rs4343	0.000	4.516	0.034	3.689	0.055
17	ACE	rs4353	0.000	4.887	0.027	3.405	0.065
17	ACE	rs4978	0.000	6.503	0.011	4.343	0.037
19	APOE	rs440446	7.963	98.388	< 0.001	86.989	< 0.001

Table 8. Main Effect Analysis Results for Complete Case-Control Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	Pearson's ChiSquare	Pearson's p-value
10	CDC2	rs2448347	6.581	0.037
10	VR22	rs1786927	7.035	0.030
10	VR22	rs2441718	8.553	0.014
10	VR22	rs2456737	6.222	0.045
10	LRRTM3	rs942780	7.586	0.023
10	LRRTM3	rs1925617	6.465	0.039
12	GAPD	rs1060621	7.188	0.027
19	APOE	rs440446	171.62	< 0.001

family-based data, only discordant sibpairs (DSPs) are used in the analysis. Two datasets were created—the first with only one randomly chosen DSP per family (designated '1DSP') and the second with all individuals who are part of one or more DSPs in a family (designated 'AllDSPs'). MDR detected the main effect of APOE in all three datasets (Case-Control, 1DSP and AllDSPs) by choosing APOE as the best one-locus models with perfect (5 of 5) cross-validation consistency and by including APOE in the best two- and three-locus models as well, all of which were statistically significant ($p < 0.05$) (Table 9). To give MDR the opportunity to detect other effects without interference of the APOE effect, we excluded APOE from the datasets

Table 9. MDR Analysis Results for Complete Datasets

	Number of Loci	Marker Genes (Markers)	Avg Bal Prediction Accuracy	p-value	CV Consist
Case-Control	1	APOE (rs440446)	68.32	< 0.001	5
	2	APOE (rs440446) AGT (rs5051)	67.18	< 0.001	2
	3	APOE (rs440446) PLAU (rs1916341) LRRTM3 (rs10762136)	66.18	< 0.001	2
1DSP	1	APOE (rs440446)	59.52	0.01	5
	2	APOE (rs440446) VR22 (rs7909676)	60.23	< 0.001	3
	3	APOE (rs440446) OLR1 (rs1050283) Chr.10 (rs1916341)	57.27	0.04	1
AllDSPs	1	APOE (rs440446)	62.50	< 0.001	5
	2	APOE (rs440446) AGT (rs7536290)	60.47	< 0.001	2
	3	APOE (rs440446) OLR1 (rs1050283) LRRTM3 (rs12769870)	60.27	< 0.001	3

and re-ran the analysis. In these subsequent analyses, none of the best one-, two- or three-locus models achieved average balanced prediction accuracies of greater than 53 percent or cross-validation consistency values of more than 2, and none were statistically significant ($p > 0.20$; see Table 10).

Table 10. MDR Analysis Results for Complete Datasets with APOE Excluded.

	Number of Loci	Marker Genes (Markers)	Avg Bal Prediction Accuracy	p-value	CV Consist
Case-Control	1	A2MP (rs34362)	48.13	0.93	2
	2	VR22 (rs10997591) LRRTM3 (rs10762136)	50.97	0.47	1
	3	UBQLN1 (rs2781002) VR22 (rs10997591) IDE (rs1544210)	52.37	0.21	2
1DSP	1	VR22 (rs7909676)	48.91	0.86	1
	2	OLR1 (rs1050283) Chr.10 (rs1898071)	48.47	0.90	1
	3	OLR1 (rs1050283) Chr.10 (rs1898071) LRRTM3 (rs10762122)	46.99	0.97	1
AllDSPs	1	Chr.10 (rs6482044)	50.55	0.57	2
	2	OLR1 (rs1050283) LRRTM3 (rs2147886)	48.24	0.95	1
	3	OLR1 (rs1050283) Chr.10 (rs1898071) AGT (rs5051)	48.97	0.88	1

Detection of Heterogeneity

Bayesian Classification was applied to each of the complete case-control and family-based datasets. Only affected individuals are used in the cluster analysis. The family-based dataset produced twelve clusters, and the case-control dataset produced four

clusters. To reduce the number of clusters produced by the family-based dataset and to focus on heterogeneity that might be present in both datasets, we took the top 30 markers from each dataset with the highest influence values and selected those markers present in both datasets (31 markers) (Table 11). Recall that a marker's influence value provides a rough heuristic measure of relative influence that marker had in differentiating the clusters from the overall dataset. Then, we performed the cluster analysis again using

Table 11. Top 30 Highest-Influence Markers Common to Both Datasets

Chrom	Gene	Marker	FamInfluValue	CCInfluValue
1	AGT	rs2148582	0.016	0.161
1	AGT	rs5051	0.020	0.178
9	UBQLN1	rs2781003	0.118	0.065
9	UBQLN1	rs10868038	0.116	0.024
9	UBQLN1	rs11140213	0.131	0.042
10	VR22	rs922347	0.019	0.118
10	VR22	rs4463744	0.016	0.115
10	VR22	rs2939947	0.027	0.108
10	LRRTM3	rs1001016	0.009	0.089
10	LRRTM3	rs1925617	0.417	0.359
10	LRRTM3	rs1925622	0.393	0.404
10	LRRTM3	rs1925632	0.799	0.940
10	LRRTM3	rs1952060	0.562	0.521
10	LRRTM3	rs2147886	0.840	1.000
10	LRRTM3	rs2251000	0.818	0.932
10	LRRTM3	rs2764807	0.542	0.615
10	LRRTM3	rs10762136	0.441	0.503
10	VR22	rs11593235	0.307	0.279
10	VR22	rs10997591	0.015	0.294
10	VR22	rs3096244	0.024	0.291
10	TNFRSF6	rs1800682	0.033	0.091
10	TNFRSF6	rs1324551	0.023	0.083
10	IDE	rs7076966	0.018	0.073
10	IDE	rs4646954	0.016	0.078
10	IDE	rs3758505	0.017	0.101
10	IDE	rs1544210	0.015	0.078
12	PZP	rs3213832	0.023	0.071
15	LIPC	rs6083	0.019	0.072
17	ACE	rs4291	0.369	0.007
17	ACE	rs4646994	0.597	0.015
17	ACE	rs4343	0.759	0.080

only those 31 markers. This second analysis produced 15 clusters in the family-based dataset and 6 clusters in the case-control dataset. After again ranking the markers by their influence values, it was discovered that the top 5 markers were the same in both datasets (Table 12). Therefore, in one final attempt to produce a clustering that was similar across both datasets and produced a more reasonable number of clusters, which could be subsequently investigated for main effects and interactions, we performed the cluster analysis again using only these top 5 markers. This third and final round of clustering produced 5 clusters in the family-based dataset and 3 clusters in the case-control dataset (Table 13). Upon closer inspection, two of the five clusters in the family-based dataset contained only seven and five affected subjects, respectively, making subsequent analysis of those clusters inadvisable due to almost no power to detect an effect. Thus, for all intensive purposes, there were only three resulting clusters for each of the datasets.

Permutation testing was performed to determine whether the final clustering based on the top five high-influence markers was statistically significant. In the family-based data, the clustering results produced an average class strength value of -4.34 ($p<0.002$) and an average cross-class entropy value of 4.00 ($p<0.002$). In the case-control data, the clustering results produced an average class strength value of -2.71 ($p<0.002$) and an average cross-class entropy value of 4.43 ($p<0.012$). Thus, for each of the datasets, the clustering results were significant at our predetermined alpha of ten percent (as suggested by our simulation studies in Chapters III and IV).

Table 12. Top Five Highest-Influence Markers from Second-Round of Cluster Analysis

Chrom	Gene	Marker	FamInfluValue	CCInfluValue
10	LRRTM3	rs1925632	0.938	0.792
10	LRRTM3	rs1952060	0.623	0.944
10	LRRTM3	rs2147886	1.000	1.000
10	LRRTM3	rs2251000	0.940	0.834
10	LRRTM3	rs2764807	0.673	0.890

Table 13. Distribution of Affected Individuals in Final Clustering Results

	Number of Affecteds	
Cluster	Family-Based Data	Case-Control Data
0	673	215
1	480	157
2	257	79
3	7	-
4	5	-

Since the top 5 markers were all in the same gene (LRRTM3), we investigated whether they were in linkage disequilibrium (LD) with each other and thus were encoding a single haplotype block. LD analysis using Haploview indeed showed that the five markers and four additional flanking markers were all in high LD with each other, and it showed the first four markers to be in a haplotype block (Figures 33 and 34). Furthermore, inspection of the multi-locus genotypes at the top 5 markers across the three clusters in each dataset showed that one multi-locus genotype was predominant in each of the three clusters and that these three multi-locus genotypes were the same across the case-control and family-based datasets (Table 14).

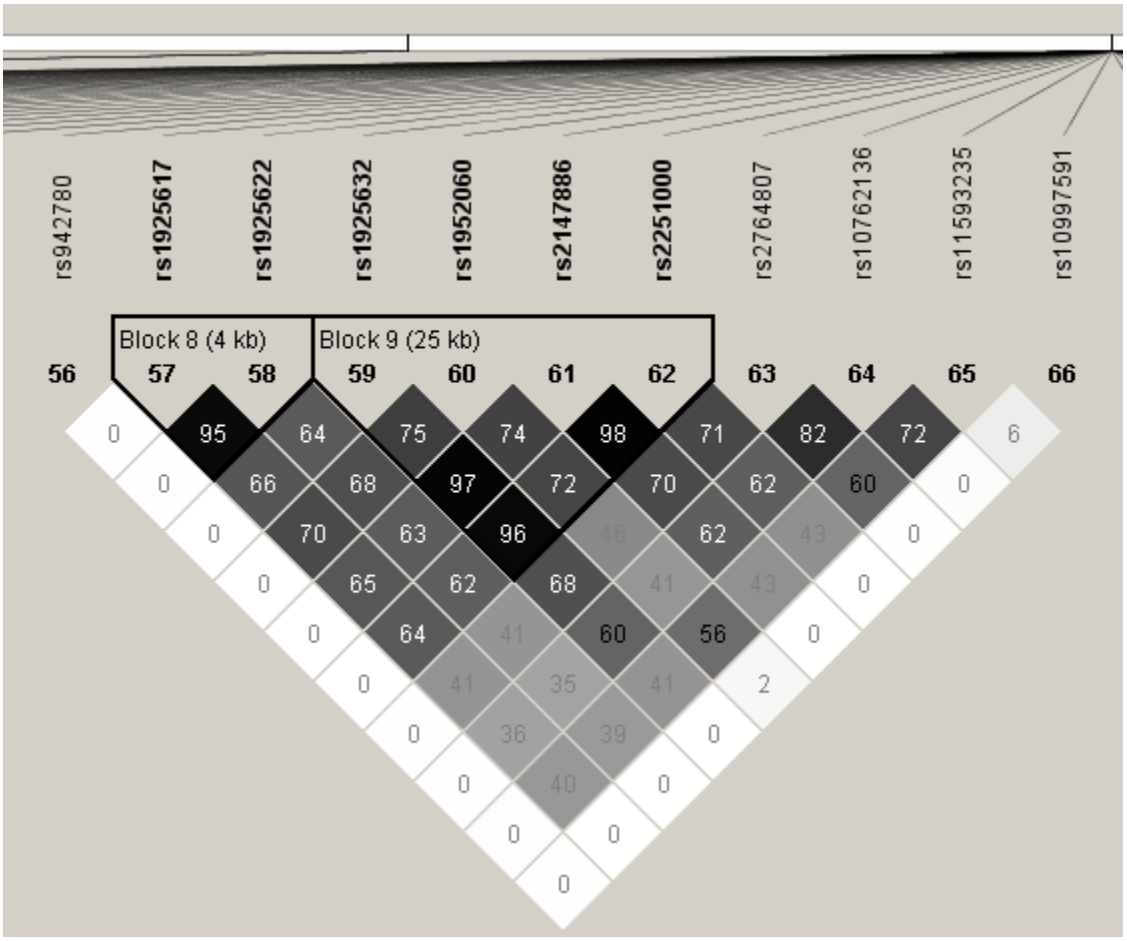


Figure 33. Linkage Disequilibrium Plot of Top 5 High-Influence Markers in Family-Based Dataset. The top five markers are: rs1925632, rs1952060, rs2147886, 2251000, and rs2764807. Numbers in each square represent pair-wise R^2 values (e.g., the number 95 in the second square from the left on the top line of the plot indicates an R^2 value of 0.95 for markers rs1925617 and rs1925622). The markers in bold are those in a haplotype block, as defined by the Haploview software program.

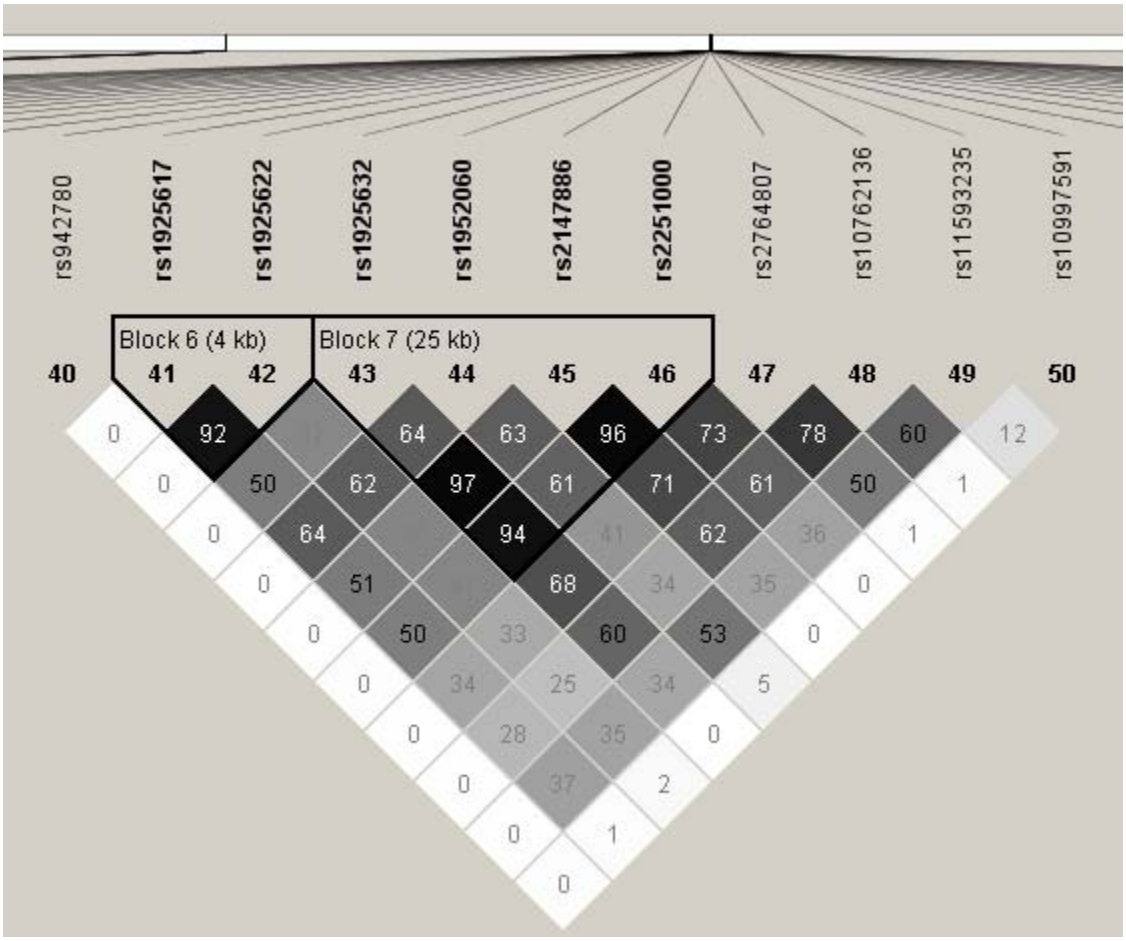


Figure 34. Linkage Disequilibrium Plot of Top 5 High-Influence Markers in Case-Control Dataset. The top five markers are: rs1925632, rs1952060, rs2147886, 2251000, and rs2764807. Numbers in each square represent pair-wise R^2 values (e.g., the number 92 in the second square from the left on the top line of the plot indicates an R^2 value of 0.92 for markers rs1925617 and rs1925622). The markers in bold are those in a haplotype block, as defined by the Haploview software program.

Table 14. Predominant Genotypes for the Top Five High-Influence Markers by Cluster

	Cluster		
Marker	**0**	**1**	**2**
rs1925632	A / C	C / C	A / A
rs1952060	C / T	C / C	T / T
rs2147886	C / T	C / C	T / T
rs2251000	A / G	A / G	A / G
rs2764807	C / T	C / C	T / T

Detection of Main Effects in Subsets of Data

For each of the three clusters (0,1,2) in the family-based dataset, linkage analysis and association analysis by FBAT and PDT were conducted. For each of the three clusters (0, 1, 2) in the case-control dataset, the chi-square test of independence was performed. Since the three clusters in each dataset correspond exactly, due to their definition by the same multilocus genotypes at the top 5 high-influence markers, analysis results are presented in the following subsections by cluster number.

Cluster 0 Results

Table 15 presents results for cluster 0 for all markers with significant scores on at least one of the three statistical tests performed (two-point linkage, FBAT and PDT). For cluster 0, linkage analysis found large HetLOD scores (greater than 10) for all five of the top high-influence markers plus three flanking markers in the LRRTM3 gene. Seven additional markers in the VR22 gene, which contains the LRRTM3 gene, produced HetLOD scores greater than 3. APOE produced a HetLOD score of 3.75 (reduced from 7.963 in the complete family-based dataset). For cluster 0, the FBAT found very strong associations with one marker in UBQLN1 (χ^2=6.864, df=1, p=0.009), two markers in ACE (χ^2=13.494, df=1, p<0.001 and χ^2=10.875, df=1, p<0.001) and with the APOE marker (χ^2=59.407, df=2, p<0.001). Ten other markers in LTA, VR22, LRP1, ACE and the ROI on chromosome 10 showed marginal association (χ^2>3.9, p<0.05). For cluster 0, the PDT found very strong association with APOE (χ^2=59.407, df=2, p<0.001) and marginal association with 15 other markers in VR22, LRP1, ACE and the ROI on chromosome 10 (χ^2>3.9, p<0.05).

Table 15. Main Effect Analysis Results for Cluster 0 Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	2-Pt Linkage Max HetLOD	PDT Chi-Square	PDT p-value	FBAT Chi-Square	FBAT p-value
6	LTA	rs1799724	1.0995	0.467	0.494	0.069	0.793
6	LTA	rs1800629	0.1523	0.711	0.399	4.175	0.041
9	UBQLN1	rs2781003	0	3.046	0.081	6.864	0.009
10		rs6482044	0.6086	4.898	0.027	4.781	0.029
10		rs1904018	1.7161	1.519	0.218	2.566	0.109
10		rs10763551	1.4194	0.863	0.353	1.42	0.233
10	CDC2	rs7919724	1.1955	0.534	0.465	1.603	0.206
10	CDC2	rs2448341	1.1791	0.249	0.618	2.427	0.119
10		rs7089698	0.1562	4.902	0.027	0.854	0.356
10	VR22	rs1786927	1.4381	0.337	0.561	0.017	0.895
10	VR22	rs2126750	1.9445	0.067	0.796	0.227	0.633
10	VR22	rs4745886	4.7019	0.004	0.948	0.002	0.967
10	VR22	rs7911820	4.0837	0.085	0.771	0.086	0.770
10	VR22	rs7070570	2.5327	0.339	0.953	0.08	0.777
10	VR22	rs7074454	4.9523	0.017	0.897	0.141	0.707
10	VR22	rs6480140	1.125	1.114	0.291	0.536	0.464
10	VR22	rs922347	1.619	2.028	0.154	1.049	0.306
10	VR22	rs4463744	1.7541	0.183	0.669	0.785	0.376
10	VR22	rs2441718	4.0431	2.469	0.116	5.021	0.025
10	VR22	rs2939947	3.9423	0.205	0.651	0	0.987
10	VR22	rs2456737	1.1742	3.462	0.063	5.388	0.020
10	VR22	rs4746606	1.2506	0.183	0.669	0.229	0.633
10	VR22	rs7909676	3.1293	5.431	0.020	4.46	0.035
10	LRRTM3	rs1001016	1.0904	0.502	0.479	0.003	0.959
10	LRRTM3	rs12769870	1.8426	1.24	0.266	0.333	0.564
10	LRRTM3	rs2394314	1.6074	0.454	0.501	0	0.995
10	LRRTM3	rs1925577	2.2237	0.453	0.501	0.027	0.869
10	LRRTM3	rs942780	1.8796	0.794	0.373	1.78	0.182
10	LRRTM3	rs1925617	11.1942	0.329	0.566	0.602	0.438
10	LRRTM3	rs1925622	11.2218	0.111	0.740	0.596	0.440
10	LRRTM3	rs1925632	20.2925	0.113	0.737	0.319	0.572
10	LRRTM3	rs1952060	12.638	2.367	0.124	2.107	0.147
10	LRRTM3	rs2147886	20.147	1.076	0.300	1.293	0.256
10	LRRTM3	rs2251000	20.4094	0.221	0.638	0.447	0.504
10	LRRTM3	rs2764807	13.0929	0.159	0.690	0.557	0.455
10	LRRTM3	rs10762136	10.2544	0.159	0.690	2.013	0.156
10	VR22	rs11593235	4.8028	0.677	0.411	1.978	0.160
10	VR22	rs10997591	2.5524	0.053	0.818	0.007	0.932
10	VR22	rs7903421	0.6966	4.57	0.033	1.064	0.302
10	VR22	rs3096244	2.5206	2.081	0.149	0	0.992

Table 15, continued. Main Effect Analysis Results for Cluster 0 Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

			2-Pt Linkage	PDT		FBAT	
Chrom	Gene	Marker	Max HetLOD	Chi-Square	p-value	ChiSquare	p-value
10		rs870801	1.6058	1.214	0.271	2.132	0.144
10	PLAU	rs2227564	1.07	0.134	0.714	0.199	0.655
10	PLAU	rs2227566	1.0844	0.027	0.869	0.054	0.816
10	PLAU	rs2227568	1.5063	0.6	0.439	0.127	0.721
10	PLAU	rs4065	1.0192	0.414	0.520	0.401	0.527
10		rs1439042	1.9356	0.082	0.775	0.638	0.424
10		rs1336439	1.194	1.004	0.316	0.905	0.341
10	IDE	rs7076966	1.0434	0.051	0.821	0.102	0.750
10	IDE	rs7099761	1.1965	0.509	0.475	1.019	0.313
10		rs225590	0	4.447	0.035	3.471	0.062
12	LRP1	rs1800181	0.0002	5.433	0.020	3.001	0.083
12	LRP1	rs1800154	0.0154	4.306	0.038	2.354	0.125
12	LRP1	rs1800165	0.0141	4.976	0.026	3.077	0.079
12	LRP1	rs9669595	0.0914	5.371	0.021	3.954	0.047
12	LRP1	rs7956957	0	3.918	0.048	2.328	0.127
17	ACE	rs4291	0	6.339	0.012	13.494	< 0.001
17	ACE	rs4295	0	4.831	0.028	10.875	0.001
17	ACE	rs4311	0	2.683	0.102	6.534	0.011
17	ACE	rs4646994	0	5.236	0.022	4.38	0.036
17	ACE	rs4343	0	4.67	0.031	4.766	0.029
17	ACE	rs4978	0	5.657	0.017	5.099	0.024
19	APOE	rs440446	3.7521	66.373	< 0.001	59.407	< 0.001

In the case-control dataset, very strong associations were found for the top 5 high-influence values in LRRTM3 and three flanking markers, plus one marker in IDE and the APOE marker ($\chi^2>38$, $p<0.001$). Four other markers in the PLAU, A2MP and ACE genes showed marginal association ($\chi^2>8$, $p<0.05$). Table 16 presents results chi-square results for all markers showing significant association ($p < 0.05$) for the cluster 0 case-control dataset.

Table 16. Main Effect Analysis Results for Cluster 0 Case-Control Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	Pearson's Chi-Square	p-value
10	LRRTM3	rs1925617	47.383	< 0.001
10	LRRTM3	rs1925622	45.252	< 0.001
10	LRRTM3	rs1925632	185.361	< 0.001
10	LRRTM3	rs1952060	80.66	< 0.001
10	LRRTM3	rs2147886	197.482	< 0.001
10	LRRTM3	rs2251000	171.91	< 0.001
10	LRRTM3	rs2764807	101.962	< 0.001
10	LRRTM3	rs10762136	105.74	< 0.001
10	VR22	rs11593235	38.462	< 0.001
10	PLAU	rs2227568	9.118	0.028
10	IDE	rs7099761	10.815	0.013
10	IDE	rs1544210	19.355	< 0.001
12	A2MP	rs34362	8.182	0.042
17	ACE	rs4291	8.414	0.038
19	APOE	rs440446	118.292	< 0.001

When comparing results across the family-based and case-control datasets for cluster 0, thirteen markers were found significant ($p < 0.05$) by the chi-square test in the case-control dataset and by at least one test (linkage, FBAT or PDT) in the family-based dataset. These markers include the top 5 high-influence markers in LRRTM3 and four flanking markers, plus one marker each in the PLAU, IDE, ACE and APOE genes—rs2227568, rs7099761, rs4291 and rs440446, respectively.

Cluster 1 Results

In the cluster 1 family-based dataset, linkage analysis showed very high HetLOD scores (greater than 5) for all five of the top high-influence markers plus four flanking markers in the LRRTM3 gene. Five additional markers in VR22 produced HetLOD

scores greater than 3. Marginal HetLOD scores (greater than 1) were found in another 31 markers in VR22, LRRTM3, PLAU, IDE, APOE and the ROI on chromosome 10. Both the FBAT and the PDT found very strong association (χ^2>39, $p < 0.001$) with the top 5 high-influence markers in LRRTM3 and four flanking markers, plus APOE. The PDT found 13 additional markers in CDC2, VR22, PLAU, IDE, A2M, ACE, GAPDHS and the ROI on chromosome 10 that showed marginal association (χ^2>3.8, $p<0.05$). The FBAT found marginal association (χ^2>4.3, $p<0.05$) with four of the same markers PDT found (in PLAU, IDE and the ROI on chromosome 10).

In the cluster 1 case-control dataset, the chi-square test of independence found very strong association (χ^2>15, $p<0.001$) with the top 5 high-influence markers in LRRTM3 and six flanking markers, plus one marker in GAPDH and APOE. In addition, 22 other markers in AGT, UBQLN1, VR22, CDC2, PLAU, IDE, GAPDH, A2MP, LIPC and ACE showed marginal association (χ^2>6, $p<0.05$). Table 18 presents chi-square results for all markers showing significant association ($p < 0.05$) for the cluster 1 case-control dataset.

When comparing across the family-based and case-control datasets for cluster 1, 17 markers were found significant ($p < 0.05$) by the chi-square test in the case-control dataset and by at least one test (linkage, FBAT or PDT) in the family-based dataset. These markers include the top 5 high-influence markers in LRRTM3 and four flanking markers, plus three additional markers in VR22 (rs4463744, rs10997591 and rs3096244), three markers in CDC2 (1920, rs2448341 and rs2448347), three markers in PLAU (rs1916341, rs2227566 and rs4065), and one marker in IDE (rs1832196) and APOE (rs440446). Worth noting, there are two markers in ACE (rs4353 and rs4978) that were

Table 17. Main Effect Analysis Results for Cluster 1 Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	2-Pt Linkage Max HetLOD	PDT Chi-Square	PDT p-value	FBAT Chi-Square	FBAT p-value
10		rs10826594	1.2548	0.714	0.398	0.333	0.564
10		rs1023207	1.0589	0.126	0.722	0.21	0.647
10		rs1319013	0.3697	4.306	0.038	1.345	0.246
10		rs6593491	1.5791	0.035	0.851	1.603	0.205
10		rs4998401	1.9603	0.795	0.373	0.948	0.330
10		rs4935648	2.6603	0.088	0.766	0.098	0.754
10		rs10763551	1.2115	0.124	0.725	0.964	0.326
10	CDC2	1920	2.0918	0.405	0.524	1.4	0.496
10	CDC2	rs2448341	0.4463	4.469	0.035	1.071	0.301
10	CDC2	rs2448347	1.2699	0.540	0.463	0.045	0.833
10		rs7090884	1.6566	0.045	0.831	0.024	0.878
10		rs1935	1.644	0.837	0.360	0.497	0.481
10		rs7089698	1.1446	0.168	0.682	0.053	0.818
10	VR22	rs1786927	0.3379	3.882	0.049	3.799	0.051
10	VR22	rs2126750	1.3995	2.579	0.108	0.73	0.393
10	VR22	rs4745886	1.8666	1.927	0.165	1.13	0.288
10	VR22	rs7911820	0.4904	4.311	0.038	4.311	0.038
10	VR22	rs7070570	3.1473	2.359	0.307	0.682	0.409
10	VR22	rs6480140	0.3839	4.953	0.026	4.953	0.026
10	VR22	rs922347	4.5899	0.129	0.720	0.223	0.637
10	VR22	rs4463744	2.1485	0.220	0.639	0.708	0.400
10	VR22	rs2441718	2.1851	0.841	0.359	1.68	0.195
10	VR22	rs2939947	3.6337	0.008	0.929	0.955	0.329
10	VR22	rs2456737	3.2556	1.485	0.223	1.796	0.180
10	VR22	rs4746606	1.4451	0.116	0.733	0.323	0.570
10	VR22	rs7909676	1.7719	1.735	0.188	1.689	0.194
10	LRRTM3	rs12769870	2.6682	0.333	0.564	0.788	0.375
10	LRRTM3	rs1925583	2.6807	1.189	0.276	0.482	0.488
10	LRRTM3	rs2394314	2.8638	0.919	0.338	0.684	0.408
10	LRRTM3	rs1925577	2.3376	0.113	0.737	0.005	0.944
10	LRRTM3	rs1925617	8.2048	45.075	< 0.001	55.866	< 0.001
10	LRRTM3	rs1925622	9.3456	46.140	< 0.001	57.161	< 0.001
10	LRRTM3	rs1925632	14.638	55.764	< 0.001	70.219	< 0.001
10	LRRTM3	rs1952060	7.8202	54.554	< 0.001	67.365	< 0.001
10	LRRTM3	rs2147886	16.7244	53.720	< 0.001	66.162	< 0.001
10	LRRTM3	rs2251000	15.4586	55.748	< 0.001	68.044	< 0.001
10	LRRTM3	rs2764807	10.721	50.449	< 0.001	58.393	< 0.001
10	LRRTM3	rs10762136	10.2257	46.537	< 0.001	57.414	< 0.001
10	VR22	rs11593235	5.976	39.195	< 0.001	45.584	< 0.001
10	VR22	rs10997591	2.55	1.186	0.276	0.277	0.599
10	VR22	rs3096244	1.4214	0.162	0.688	0.603	0.438

Table 17, continued. Main Effect Analysis Results for Cluster 1 Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	2-Pt Linkage Max HetLOD	PDT Chi-Square	PDT p-value	FBAT Chi-Square	FBAT p-value
10	PLAU	rs1916341	1.0987	1.801	0.180	1.598	0.206
10	PLAU	rs2227564	1.5754	0.025	0.874	0.373	0.541
10	PLAU	rs2227566	1.3058	1.848	0.174	1.982	0.159
10	PLAU	rs2227568	0.2907	6.470	0.011	5.032	0.025
10	PLAU	rs4065	1.2024	1.594	0.207	3.203	0.074
10		rs1439042	1.5462	0.601	0.438	0.424	0.515
10		rs11816558	1.073	0.000	1.000	0.065	0.798
10	IDE	rs2251101	0	7.388	0.007	4.788	0.029
10	IDE	rs1832196	0.2703	5.028	0.025	6.31	0.012
10	IDE	rs4646954	1.1098	1.817	0.178	2.64	0.104
10		rs4372378	0.4067	5.704	0.017	4.995	0.025
12	A2M	rs3832852	0.0837	6.674	0.010	1.357	0.244
17	ACE	rs4353	0	4.265	0.039	2.425	0.119
17	ACE	rs4978	0	3.991	0.046	2.254	0.133
19	GAPDS	rs4806173	0.25	4.464	0.035	3.6	0.058
19	APOE	rs440446	2.1577	40.994	< 0.001	43.475	< 0.001

significant by the FBAT and PDT in the family-based dataset but are not present in the case-control dataset. In the family-based dataset, these markers are in linkage disequilibrium with two other markers (rs4646994 and rs4343) that are were found significant by the Pearson chi-square test of independence in the case-control dataset.

Cluster 2 Results

In the cluster 2 family-based dataset, linkage analysis produced HetLOD scores greater than 3 for the top 5 high-influence markers in LRRTM3 and four flanking markers, plus one additional marker in VR22. Marginal HetLOD scores (greater than 1) were found in another 18 markers in AGT, VR22, ACE and the ROI on chromosome 10.

Table 18. Main Effect Analysis Results for Cluster 1 Case-Control Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	Pearson's Chi-Square	p-value
1	AGT	rs2148582	11.144	0.004
1	AGT	rs5051	12.809	0.002
1	AGT	rs1326886	8.652	0.013
9	UBQLN1	rs2781003	6.305	0.043
9	UBQLN1	rs2780995	6.794	0.033
9	UBQLN1	rs12344615	7.624	0.022
9	UBQLN1	rs11140213	8.023	0.018
10	CDC2	1920	11.509	0.021
10	CDC2	rs2448341	6.269	0.044
10	CDC2	rs2448347	7.161	0.028
10	VR22	rs4463744	11.652	0.003
10	LRRTM3	rs1925617	73.726	< 0.001
10	LRRTM3	rs1925622	56.225	< 0.001
10	LRRTM3	rs1925632	225.507	< 0.001
10	LRRTM3	rs1952060	92.221	< 0.001
10	LRRTM3	rs2147886	241.493	< 0.001
10	LRRTM3	rs2251000	226.643	< 0.001
10	LRRTM3	rs2764807	138.128	< 0.001
10	LRRTM3	rs10762136	119.639	< 0.001
10	VR22	rs11593235	39.889	< 0.001
10	VR22	rs10997591	26.598	< 0.001
10	VR22	rs3096244	23.893	< 0.001
10	PLAU	rs1916341	6.591	0.037
10	PLAU	rs2227566	6.866	0.032
10	PLAU	rs4065	7.26	0.027
10	IDE	rs1832196	7.803	0.020
10	IDE	rs7076966	9.976	0.007
12	GAPD	rs7307229	8.618	0.013
12	GAPD	rs1060620	15.79	< 0.001
12	GAPD	rs1060619	14.489	0.001
12	A2MP	rs16918212	6.051	0.049
12	A2MP	rs17804080	8.722	0.013
15	LIPC	rs6083	8.748	0.013
17	ACE	rs4646994	6.004	0.050
17	ACE	rs4343	8.903	0.012
19	APOE	rs440446	91.857	< 0.001

Both the FBAT and the PDT found very strong association (χ^2>11, $p < 0.001$) with the top 5 high-influence markers in LRRTM3 and four flanking markers, plus APOE. The FBAT found five additional markers in LTA, LRRTM3, PLAU and ACE that showed marginal association (χ^2>4, $p<0.05$). The PDT found one more LRRTM3-flanking marker with a very significant association (χ^2=12.255, df=2, $p < 0.001$) and three other markers in CDC2, PLAU and LRP1 that showed marginal association (χ^2>4.9, $p < 0.05$). Table 19 presents results for cluster 2 for all markers with significant scores ($p<0.05$) on at least one of the three statistical tests performed (two-pt linkage, FBAT and PDT).

In cluster 2 case-control dataset, the chi-square test of independence found very strong association (χ^2>67, $p<0.001$) with the top 5 high-influence markers in LRRTM3 and four flanking markers, plus APOE. In addition, three other markers in VR22 and A2MP showed marginal association (χ^2>7, $p<0.05$). Table 20 presents chi-square results for all markers showing significant association ($p < 0.05$) for the cluster 2 case-control dataset.

When comparing across the family-based and case-control datasets for cluster 2, 10 markers were found significant ($p < 0.05$) by the chi-square test in the case-control dataset and by at least one test (linkage, FBAT or PDT) in the family-based dataset. These markers include the top 5 high-influence markers in LRRTM3 and four flanking markers, plus APOE.

Table 19. Main Effect Analysis Results for Cluster 2 Family-Based Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	2-Pt Linkage Max HetLOD	PDT Chi-Square	PDT p-value	FBAT Chi-Square	FBAT p-value
1	AGT	rs5051	1.7146	0.155	0.694	0.819	0.365
1	AGT	rs2148582	1.7798	0.003	0.953	0.38	0.538
6	LTA	rs1800629	0	1.573	0.210	4.594	0.032
10		rs10826594	1.0955	0.268	0.605	0.096	0.757
10		rs1319013	1.3621	1.613	0.204	1.747	0.186
10		rs1148247	1.0737	0.109	0.741	1.053	0.305
10		rs6482044	1.3444	0.822	0.365	0.142	0.706
10		rs6593491	1.1512	0.297	0.586	0.985	0.321
10		rs1890739	2.627	0.261	0.610	0.064	0.800
10		rs14327	1.0027	0.43	0.512	0.852	0.356
10		rs10763551	2.3693	0.166	0.684	0.014	0.905
10	CDC2	rs7919724	0.6497	4.955	0.026	2.602	0.107
10	VR22	rs2126750	1.6167	1.152	0.283	1.716	0.190
10	VR22	rs7074454	1.1114	0.674	0.412	0.557	0.456
10	VR22	rs6480140	1.8565	0.653	0.419	0.01	0.920
10	VR22	rs2441718	3.5833	0.576	0.448	0.041	0.839
10	VR22	rs2939947	1.2757	0.272	0.602	0.573	0.449
10	VR22	rs2456737	1.4679	0.052	0.820	0.17	0.681
10	VR22	rs4746606	1.4457	0.193	0.661	0.081	0.776
10	LRRTM3	rs942780	0.1022	12.255	< 0.001	9.648	0.002
10	LRRTM3	rs1925617	3.7431	11.904	0.001	19.368	< 0.001
10	LRRTM3	rs1925622	3.9166	11.062	0.001	18.319	< 0.001
10	LRRTM3	rs1925632	8.2935	23.69	< 0.001	27.201	< 0.001
10	LRRTM3	rs1952060	6.8637	21.041	< 0.001	22.93	< 0.001
10	LRRTM3	rs2147886	9.664	28.35	< 0.001	33.476	< 0.001
10	LRRTM3	rs2251000	9.1483	26.098	< 0.001	33.166	< 0.001
10	LRRTM3	rs2764807	8.4098	21.525	< 0.001	29.282	< 0.001
10	LRRTM3	rs10762136	8.3662	21.407	< 0.001	31.142	< 0.001
10	VR22	rs11593235	6.5029	19.458	< 0.001	20.156	< 0.001
10	PLAU	rs2227568	1	7.042	0.008	7.36	0.007
10		rs4933194	1.0466	0.428	0.513	0.039	0.843
12	LRP1	rs1800154	0	5.141	0.023	2.807	0.094
17	ACE	rs4291	0.001	1.373	0.241	4.729	0.030
17	ACE	rs4343	0.0855	2.33	0.127	4.077	0.043
19	APOE	rs440446	0.5671	36.984	< 0.001	25.785	< 0.001

Table 20. Main Effect Analysis Results for Cluster 2 Case-Control Dataset. Significant results are highlighted in pale ($p<0.05$) or fluorescent ($p<0.01$) yellow.

Chrom	Gene	Marker	Pearson's Chi-Square	p-value
10	LRRTM3	rs1925617	96.408	< 0.001
10	LRRTM3	rs1925622	93.924	< 0.001
10	LRRTM3	rs1925632	182.472	< 0.001
10	LRRTM3	rs1952060	134.996	< 0.001
10	LRRTM3	rs2147886	202.584	< 0.001
10	LRRTM3	rs2251000	195.167	< 0.001
10	LRRTM3	rs2764807	146.342	< 0.001
10	LRRTM3	rs10762136	101.079	< 0.001
10	VR22	rs11593235	72.618	< 0.001
10	VR22	rs10997591	11.588	0.003
12	A2MP	rs16918212	7.025	0.030
12	A2MP	rs17804080	10.425	0.005
19	APOE	rs440446	67.132	< 0.001

Detection of Gene-Gene Interactions in Subsets of Data

For each of the three clusters in both the family-based and case-control datasets, an MDR gene-gene interaction analysis was conducted. APOE and the top 5 high-influence markers, plus the four flanking markers in linkage disequilibrium with those top markers, dominated the best MDR models (data not shown). To allow other effects to be detected over these known effects, these ten markers were excluded and the MDR analyses were repeated. Tables 21, 23 and 25 present the best MDR models for clusters 0, 1 and 2, respectively. Cross-validation (CV) consistency is provided as the number of times (out of 5) that the reported best model was the best in the fold, or split, of the data. The average (across all five cross-validation intervals) of the balanced prediction accuracy and its corresponding significance level (p-value) is also reported.

Cluster 0 Results

For cluster 0, in the family-based 1DSP dataset, the best one-locus MDR model was rs4291 in ACE ($p = 0.10$) and the best two-locus model was rs4291 in ACE and rs7909676 in VR22 ($p = 0.17$), which is not in LD ($r^2 \leq 0.01$) with any LRRTM3 marker in the dataset (Table 21). These two models were the only MDR models for cluster 0 that achieved a prediction accuracy of approximately 55 percent or greater. It is worth noting that the best one-locus MDR model in the case-control dataset, which had a lower prediction accuracy of 48.5 ($p = 0.86$), was also rs4291 in ACE. A statistically significant full factorial model was fit to the cluster 0 family-based dataset using rs4291 and rs7909676 ($\chi^2 = 19.264$, df=3, $p = 0.0002$), but the individual parameter estimates indicate that the significant effect in the model is primarily coming from marker rs 4291 (Table 22). The heterozygote and the A/A homozygote for rs4291 increased risk for disease by 2.066 ($p = 0.0106$).

Cluster 1 Results

For cluster 1, in the case-control dataset, the best one-locus model was rs3096244 in VR22 ($p = 0.11$), which is not in LD ($r^2 \leq 0.04$) with any LRRTM3 marker in the dataset, and the best two-locus MDR model involved rs3096244 in VR22 and rs4343 in ACE ($p = 0.08$). In the 1DSP family-based dataset, the best two locus model was rs2255901 in the chromosome 10 ROI and rs 922347 in VR22 ($p = 0.13$), which is not in LD with any LRRTM3 marker in the dataset,. These three models were the only MDR models for cluster 1 that achieved a prediction accuracy of greater than 55 percent (Table 23).

Table 21. MDR Analysis Results for Cluster 0

	Num Loci	Marker Genes (Markers)	Avg Bal Prediction Accuracy	p-value	CV Consist
Case-Control	1	ACE (rs4291)	48.50	0.86	2
	2	A2MP (rs34362) PLAU (rs1916341)	52.15	0.34	2
	3	VR22 (rs1786927) IDE (rs1544210) AGT (rs5051)	49.30	0.76	1
1DSP	1	ACE (rs4291)	56.48	0.10	4
	2	ACE (rs4291) VR22 (rs7909676)	54.99	0.17	3
	3	ACE (rs4646994) OLR1 (rs1050283) VR22 (rs4745886)	48.62	0.87	1
AllDSPs	1	Chr.10 (rs6482044)	54.37	0.12	3
	2	OLR1 (rs1050283) VR22 (rs7909676)	49.42	0.80	1
	3	OLR1 (rs1050283) CDC2 (rs7919724) AGT (rs2148582)	50.54	0.62	1

Table 22. Logistic Regression Results for Cluster 0 Family-Based Data Using Markers from Significant Two-Locus MDR Model

					95% Hazard Ratio Confidence Limits	
Factor	**χ^2**	**df**	**p Value**	**Hazard Ratio**	**Lower**	**Upper**
VR22(rs7909676)	3.6832	1	0.0550	1.916	0.986	3.723
ACE(rs4291)	6.5254	1	0.0106	2.066	1.184	3.606
rs7909676 * rs4291	0.3573	1	0.5500	0.881	0.582	1.335

Table 23. MDR Analysis Results for Cluster 1

	Num Loci	Marker Genes (Markers)	Avg Bal Prediction Accuracy	p-value	CV Consist
Case-Control	1	VR22 (rs3096244)	55.74	0.11	3
	2	VR22 (rs3096244) ACE (rs4343)	56.73	0.08	2
	3	VR22 (rs3096244) VR22 (rs922347) PZP (rs3213831)	53.73	0.24	2
1DSP	1	Chr.10 (rs2255901)	52.61	0.41	3
	2	Chr.10 (rs2255901) VR22 (rs922347)	56.29	0.13	3
	3	ACE (rs4646994) LRP1 (rs1800154) OLR1 (rs1050283)	48.39	0.86	1
AllDSPs	1	GAPDS (rs4806173)	52.74	0.31	4
	2	GAPDS (rs4806173) VR22 (rs7074454)	50.68	0.59	2
	3	GAPDS (rs4806173) VR22 (rs4745886) Chr.10 (rs6482044)	50.51	0.62	1

A statistically significant full factorial model was fit to the cluster 1 case-control dataset using rs3096244 in VR22 and rs4343 in ACE from the best two-locus MDR model ($\chi^2 = 20.646$, df=3, $p < 0.001$) (Table 24). Both markers displayed significant main effects, and the interaction effect, which had the opposite effect on risk, was also significant (Table 24). At marker rs3096244 in VR22, the heterozygote and T/T homozygote decreased risk by 0.464, and at marker rs4343 in ACE, the heterozygote and G/G homozygote decreased risk by 0.425. However, in reference to any genotype combination that included the A/A homozygote for rs3096244 or the A/A homozygote for rs4343, those same genotypes when considered together actually increased risk by 1.696.

Table 24. Logistic Regression Results for Cluster 1 Case-Control Data Using Markers from Significant Two-Locus MDR Model

					95% Odds Ratio Confidence Limits	
Factor	**χ^2**	**df**	**p Value**	**Odds Ratio**	**Lower**	**Upper**
VR22(rs3096244)	14.498	1	< 0.001	0.464	0.309	0.694
ACE(rs4343)	14.363	1	< 0.001	0.425	0.270	0.671
rs3096244 * rs4343	9.072	1	0.003	1.696	1.199	2.400

Using the two markers included in the best two-locus MDR model for the 1DSP family-based dataset, logistic regression was used to fit a full factorial model to the data. However, the full model was not statistically significant ($\chi^2 = 1.4917$, df=3, $p > 0.68$) ; nor were any of its factors (data not shown).

It is perhaps worth noting that in the AllDSPs dataset, rs7074454 and rs4745886 in VR22 were each in the best two- and three-locus MDR models, respectively. These markers are in linkage disequilibrium with each other in the complete family-based dataset but are not in LD with the VR22 markers found in the best case-control MDR models. Marker rs4745886 in VR22 was out of Hardy-Weinberg equilibrium in the complete family-based dataset. None of these family-based models in cluster 1 achieved prediction accuracy greater than 55 percent.

Cluster 2 Results

For cluster 2, the best one-locus MDR model in the case-control dataset was rs10997591 in VR22 ($p < 0.04$), which is not in LD ($r^2 \leq 0.12$) with any LRRTM3 marker in the dataset, and the best one-locus MDR model in the 1DSP family-based dataset was

rs11816558 in the ROI on chromosome 10 (p = 0.08). These two models were the only MDR models in cluster 2 that achieved a prediction accuracy of greater than 55 percent (Table 25).

Table 25. MDR Analysis Results for Cluster 2

	Num Loci	Marker Genes (Markers)	Avg Bal Prediction Accuracy	p-value	CV Consist
Case-Control	1	VR22 (rs10997591)	60.09	0.04	5
	2	VR22 (rs10997591) CDC2 (1920)	49.49	0.72	1
	3	VR22 (rs10997591) IDE (rs1544210) COG2 (rs3789662)	44.73	0.98	1
1DSP	1	Chr.10 (rs11816558)	59.32	0.08	5
	2	OLR1 (rs1050283) CDC2 (1920)	51.60	0.58	1
	3	OLR1 (rs1050283) ACE (rs4646994) Chr.10 (rs1916341)	50.15	0.73	1
AllDSPs	1	PZP (rs12230214)	48.03	0.89	1
	2	ACE (rs4646994) Chr.10 (rs870801)	51.98	0.46	1
	3	OLR1 (rs1050283) ACE (rs4646994) CDC2 (rs1920)	47.30	0.93	2

It is worth noting that rs10997591 in VR22 was also present in the best two- and three-locus MDR models for the case-control dataset, although their corresponding prediction accuracy was below 50 percent. The marker rs1050283 in OLR1 appeared in the best two- and three-locus MDR models for the 1DSP dataset and in the best three-locus MDR model for the AllDSPs dataset. In addition, the best two-locus MDR models

in both the case-control and the family-based 1DSP datasets and the best three-locus model in the family-based AllDSPs dataset, all included marker 1920 in CDC2. Finally, the marker rs4646994 in ACE was included in the best three-locus model for the 1DSP dataset and in the best two- and three-locus models for the AllDSPs dataset.

Discussion

Simulation studies of the Bayesian Classification method presented in Chapters III and IV were performed using simulated case-control data. The current application of the clustering method involves both family-based and case-control data. Family-based data naturally have intrafamily correlations among markers, which may not be relevant to the disease in question. Large families with particular multilocus genotype patterns may bias the choice of high influence markers more so than smaller families, leading to choices that may not generalize to a large family-based dataset or case-control dataset. No attempt was made to control for such intrafamily correlations directly. However, our decision to perform multiple rounds of clustering, choosing only those markers common to both datasets, may have averted some of this potential bias. It is encouraging that, at least in this particular application, the same five markers were selected in both the family-based and case-control datasets as being the highest influence markers.

Another issue created by family-based data involves the way in which the clustered affected individuals are recombined with the set of unaffected individuals. Since the main effect analysis methods for family-based data use pedigree information and leverage family structure and intrafamily correlation, any splitting of families threatens to reduce the informativeness of such families and to subsequently reduce the

power of the analyses. For this reason, it might have been ideal to have all affected individuals from a family always be clustered together, thereby avoiding any disruption of family structure. However, there was no way to implement such constraints within the existing (closed source) clustering software, and as it turned out, the clustering method did not choose to cluster together all individuals of the same family. Thus, the power of main effect analyses on family-based cluster subsets was likely reduced to some degree.

The power of our analyses on the cluster subsets may also have been lowered (in comparison to the complete datasets) because the number of affected subjects in each subset is only a fraction of what is present in the complete dataset. Since clustering is performed only on the affected individuals in the dataset, for the purpose of subset analysis, the resulting clusters of affected individuals are recombined with the full set of unaffected individuals. Therefore, this also means that the data in most of the subsets is substantially unbalanced. The complete case-control dataset was already somewhat unbalanced, with a ratio of cases to controls of 0.65. Thus, the ratios in the cluster subsets for the case-control data were even more unbalanced—0.31, 0.22 and 0.11—for clusters 0, 1 and 2, respectively. The complete family-based dataset was unbalanced but in the opposite direction, with a ratio of cases to controls of 1.91. Thus, the ratios in the cluster subsets for the family-based data were not as badly affected as those in the case-control dataset—0.90, 0.65, 0.35—for clusters 0, 1 and 2, respectively.

Another difference between our simulation studies and the current application is that the simulation studies used markers which had no linkage disequilibrium (LD) with each other, while the current application involved markers with considerable LD, comprising multiple haplotype blocks. The clustering method chose to focus on a set of

markers in LRRTM3 that were in high LD with each other to cluster affected subjects into more homogeneous subsets. The fact that the Bayesian Classification method essentially used (a readily discoverable) haplotype block to cluster the datasets may not be a particularly interesting result. Afterall, one could have used the results from the linkage disequilibrium analysis directly to choose haplotype blocks upon which to stratify the data, although the choice among haplotype blocks would have been arbitrary. Perhaps the fact that the clustering method could have found other multilocus genotype patterns but did not means that there were no other interesting patterns to be found. Alternatively, it is possible that there were other multi-locus genotype patterns in the datasets but that these patterns simply were not as strong or as consistent as those in the haplotype block of LRRTM3 and hence were not chosen to highly influence cluster assignment. One could try to select tag SNPs prior to clustering, with the goal of reducing the strength or dominance of such LD in the dataset, thereby allowing other weaker, perhaps more interesting, multilocus genotype patterns to be selected for use in clustering the dataset. However, initial attempts at implementing this approach on the current datasets indicate that the process of choosing the tag SNPs would be iterative, adhoc and somewhat arbitrary—in short, not at all a straight-forward solution to the situation. Additionally, eliminating markers by choosing tag SNPs could also dilute any multilocus genotype effects that are present, which the clustering method could have used to stratify the data.

Regardless of whether the clustering method's use of a haplotype block is novel or interesting, the question remains as to whether stratification or clustering based in this specific dataset using this particular LD block in LRRTM3 is meaningful. It is, indeed,

possible that there are main effects and/or interactions among other genes that are only present on certain LRRTM3 haplotype backgrounds. It is also possible that there are direct or indirect interactions between LRRTM3 and these other genes and that clustering on the LRRTM3 haplotype block allows those effects to be detected. It is also possible that the pertinent interactions involve VR22, which is the larger gene in whose intron LRRTM3 resides. Ultimately, whether these results are meaningful will be determined by whether the statistical results reported here can be replicated, and, more importantly, whether functional molecular studies can demonstrate the biological plausibility of such interactions.

VR22 or CTNNA3 (catenin, alpha 3; MIM#607667) is a binding partner of beta-catenin (Janssens et al., 2001), which interacts with presenilin 1. Presenilin 1 interacts with the gamma-secretase involved in processing the amyloid precursor protein (APP), and its mutations have been associated with increased levels of amyloid beta 42 (Citron et al., 1997; Duff et al., 1996; Qian et al., 1998), the primary component of senile plaques found in Alzheimer disease.

Leucine-rich containing proteins, like LRRTM3, are involved in protein-protein interactions, and the family of leucine-rich repeat transmembrane proteins (LRRTMs) are involved in many cellular events during nervous system development and disease (Lauren et al., 2003). Of particular relevance to Alzheimer disease pathology, LRRTM3 is highly expressed in the adult mouse hippocampus, in the granular layer of the dentate gyrus (Lauren et al., 2003). Tau-mediated neurodegeneration in this area is thought to play a role in Alzheimer disease progression (Shahani et al., 2006).

Recent evidence is mounting in support of an alternative hypothesis for Alzheimer disease pathology, which implicates cell cycle reactivation as a key early event that precedes and possibly is causally related to tau and APP phosphorylation and apoptotic cell death (Andorfer et al., 2005; McPhie et al., 2003; Yang et al., 2006). Amyloid precursor protein has been purported to regulate activation of neuronal cell cycle proteins (McPhie et al., 2003); therefore, hypothetically, mutations in VR22 could indirectly affect cell cycle activation, through interactions with APP (by way of beta-catenin and presenilin 1). Additionally, since LRRTM3 is thought to be involved in neuronal development in some of the key areas that are later targets of neuronal cell death in Alzheimer disease, perhaps LRRTM3 is being re-activated in some way that facilitates the cell cycle re-entry of neurons. Thus, it would be interesting to learn whether VR22 and/or LRRTM3 are differentially expressed in the brains of AD patients versus controls.

For every cluster, the main effect and interaction subset analyses showed LRRTM3 markers exhibiting strong effects. This is an expected result since almost all (affected) individuals in a cluster had the same genotypes at those markers and in comparison to the unaffecteds in the datasets, it would appear that those genotypes were associated with disease status. Likewise, flanking or nearby markers in LRRTM3 and the larger gene, VR22, within which LRRTM3 resides, might demonstrate effects that could be attributed to the LRRTM3 haplotype block effect. Table 26 shows the NCBI map locations of all genotyped markers in the VR22 and LRRTM3 genes, along with their HetLOD scores in the complete family-based dataset and its three clusters. Figure 35 shows a plot of these HetLOD scores starting with the most distal markers that achieved a HetLOD of at least 2.

Looking across all the main effect and interaction analyses, there are a few genes for each cluster that deserve further investigation in relation to their LRRTM3 haplotype (Table 14). In some cases where there are two or more markers in LD with each other, in the case-control dataset, one of the markers is significant but in the family-based dataset, the other one is. This can be a simple case of sampling differences, since the two datasets are independent samples drawn from different populations and by chance the distribution of alleles or genotypes between affecteds and unaffecteds can be different between those samples at any given marker.

Complete Dataset Discussion

The preliminary analysis of the complete family-based and case-control datasets found three markers that were significant in both the case-control and family-based datasets—VR22 markers rs2441718 and rs2456737 and APOE marker rs440446. LRRTM3 marker rs1925617 was significant in the case-control dataset and was in LD with three other LRRTM3 markers—rs1925622, rs1925632 and rs2764807—which were significant in the family-based dataset by their PDT chi-square statistics. None of the MDR interaction analyses that excluded the known effect of APOE produced significant models.

Many of the markers that were found significant by at least one main effect statistical test in either the complete case-control or complete family-based datasets were also significant in the analysis of specific subsets produced by the Bayesian Classification

Table 26. Chromosomal Location and Linkage Analysis Results for Markers in VR22 and LRRTM3. Highlighted markers were the top five high-influence markers used in the final cluster analysis.

				HetLOD in Family- Based Dataset			
Chrom	Gene	Marker	Location (kb)	Complete	Cluster 0	Cluster 1	Cluster 2
10	VR22	4783	67,209	0.000	0.900	0.962	0.241
10	VR22	rs1786927	67,352	0.000	1.438	0.338	0.110
10	VR22	rs2126750	67,508	0.036	1.945	1.400	1.617
10	VR22	rs4745886	67,530	0.768	4.702	1.867	0.913
10	VR22	rs7911820	67,534	0.165	4.084	0.490	0.711
10	VR22	rs7070570	67,535	1.366	2.533	3.147	0.241
10	VR22	rs7074454	67,535	0.643	4.952	0.777	1.111
10	VR22	rs10822719	67,535	0.008	0.816	0.380	0.600
10	VR22	rs6480140	67,539	0.000	1.125	0.384	1.857
10	VR22	rs922347	67,653	0.119	1.619	4.590	0.468
10	VR22	rs4463744	67,778	0.551	1.754	2.149	0.413
10	VR22	rs2441718	67,807	1.407	4.043	2.185	3.583
10	VR22	rs2939947	67,808	0.560	3.942	3.634	1.276
10	VR22	rs2456737	67,825	1.038	1.174	3.256	1.468
10	VR22	rs4746606	68,061	0.201	1.251	1.445	1.446
10	VR22	rs7909676	68,105	1.068	3.129	1.772	0.510
10	LRRTM3	rs1001016	68,347	0.000	1.090	0.336	0.000
10	LRRTM3	rs12769870	68,347	0.000	1.843	2.668	0.933
10	LRRTM3	rs1925583	68,350	0.001	0.824	2.681	0.856
10	LRRTM3	rs2394314	68,350	0.015	1.607	2.864	0.852
10	LRRTM3	rs1925577	68,358	0.079	2.224	2.338	0.986
10	LRRTM3	rs10762122	68,386	0.001	0.864	0.820	0.249
10	LRRTM3	rs942780	68,407	0.000	1.880	0.692	0.102
10	LRRTM3	rs1925617	68,435	0.343	11.194	8.205	3.743
10	LRRTM3	rs1925622	68,440	0.302	11.222	9.346	3.917
10	LRRTM3	rs1925632	68,470	0.140	20.293	14.638	8.294
10	LRRTM3	rs1952060	68,473	0.260	12.638	7.820	6.864
10	LRRTM3	rs2147886	68,489	0.066	20.147	16.724	9.664
10	LRRTM3	rs2251000	68,495	0.073	20.409	15.459	9.148
10	LRRTM3	rs2764807	68,499	0.097	13.093	10.721	8.410
10	LRRTM3	rs10762136	68,514	0.492	10.254	10.226	8.366
10	VR22	rs11593235	68,546	0.636	4.803	5.976	6.503
10	VR22	rs10997591	68,672	0.379	2.552	2.550	0.394
10	VR22	rs7903421	68,952	0.000	0.697	0.136	0.812
10	VR22	rs3096244	69,080	0.000	2.521	1.421	0.671

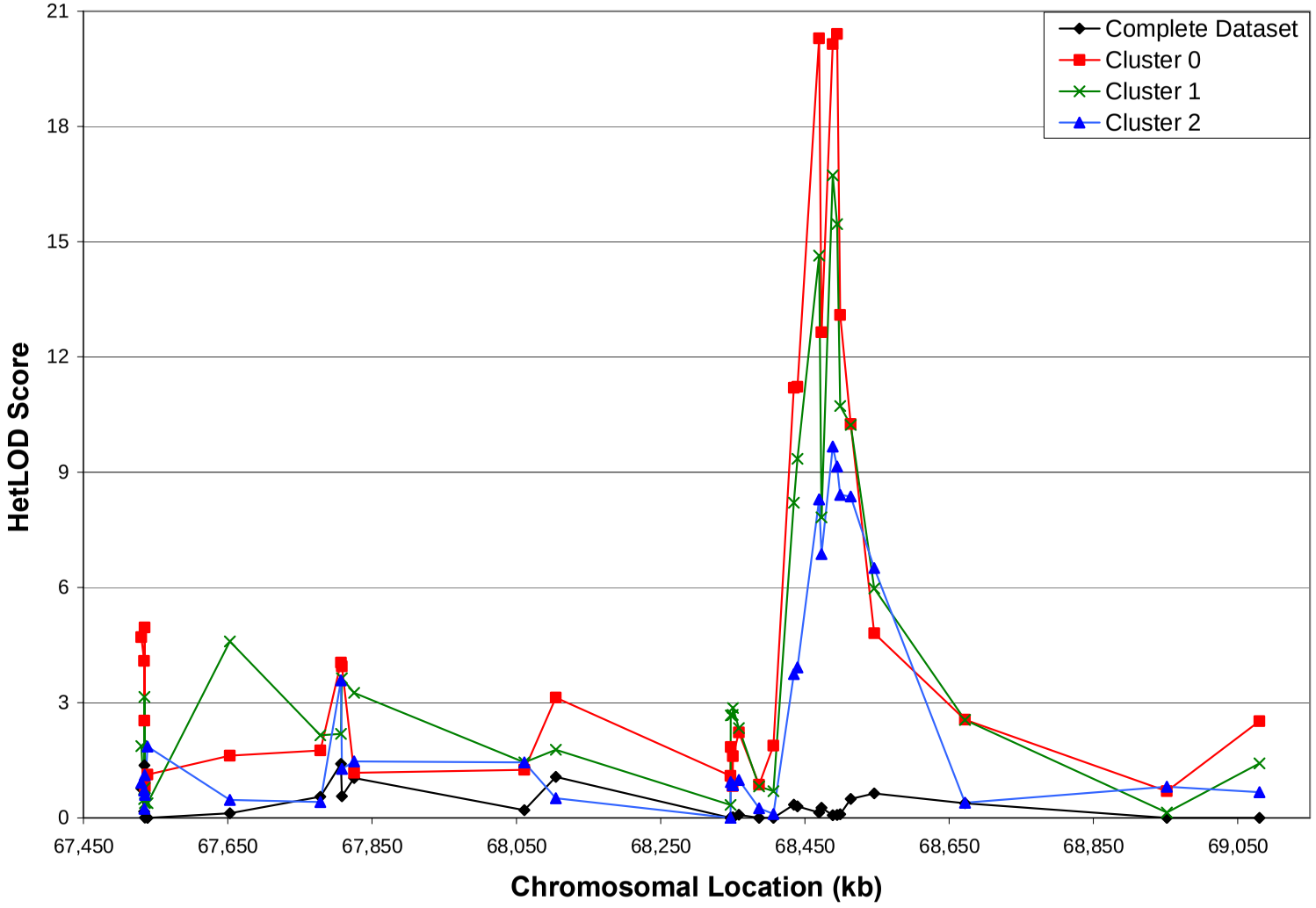


Figure 35. Linkage Analysis of Top 5 High-Influence Markers in LRRTM3 and Flanking Markers with HetLOD Scores >2

analysis. However, there was not complete consistency across the clusters, and some markers were not significant in any of the subsequent main effect analyses of the cluster subsets. Tables 27 and 28 indicate which of the markers initially found significant in the complete datasets were also found significant in one or more of the cluster subsets. Eight markers were found significant across all three clusters as well as in their respective complete dataset—APOE marker rs440446, two markers from VR22 (rs2441718 and rs2456737), four markers from LRRTM3 (rs1925617, rs1925622, rs1925632 and rs2764807) and one marker from PLAU (rs2227568). Interestingly, the effect of APOE was less in each of the cluster subsets than it was in the complete datasets, perhaps simply due to smaller sample sizes and more unbalanced data. In contrast, the VR22, LRRTM3 and PLAU marker effects were all enhanced in the cluster subsets. Since the clusters were produced basically by stratifying on an LRRTM3 haplotype block, it is not surprising that the VR22 and LRRTM3 marker effects are strengthened.

The PLAU marker rs2227568 is approximately 6.26 Mb away from the nearest genotyped VR22 marker and it exhibits no LD (r^2=0) with any of the VR22 or LRRTM3 markers. Therefore, it is unlikely that the consistency of the PLAU marker's results can be attributed to the LRRTM3 effect. PLAU (urokinase-type plasminogen activator; MIM#5328) converts plasminogen to plasmin, and plasmin is involved in the processing of the amyloid precursor protein and in the degradation of amyloid-beta (Finckh et al., 2003). The PLAU marker rs2227564 is a C/T missense polymorphism that has been associated with plasma amyloid-beta-42 levels and with LOAD in a German sample (Finckh et al., 2003) and in a United States Caucasian sample (Ertekin-Taner et al., 2005). However, at least two subsequent studies have failed to replicate these results—in

Table 27. Cluster Subset Results for Markers Found Significant in Complete Family-Based Dataset (HetLOD > 1, or FBAT or PDT $p < 0.05$). Marks in a cluster column indicate that the marker was found significant by at least one main effect family-based test (linkage, FBAT or PDT) in that cluster subset.

			Cluster		
Chrom	**Gene**	**Marker**	**0**	**1**	**2**
1	AGT	rs5051			x
10	VR22	rs7070570	x	x	
10	VR22	rs2441718	x	x	x
10	VR22	rs2456737	x	x	x
10	VR22	rs7909676	x	x	
10	LRRTM3	rs1925622	x	x	x
10	LRRTM3	rs1925632	x	x	x
10	LRRTM3	rs2764807	x	x	x
10	PLAU	rs2227568	x	x	x
10	PLAU	rs4065	x	x	
10		rs4933194			x
12	A2M	rs3832852		x	
12	A2MP	rs34362			
12	LRP1	rs1800154	x		x
12	LRP1	rs9669595	x		
12	LRP1	rs7956957	x		
17	ACE	rs4291	x		x
17	ACE	rs4295	x		
17	ACE	rs4646994	x		
17	ACE	rs4343	x		x
17	ACE	rs4353		x	
17	ACE	rs4978	x	x	
19	APOE	rs440446	x	x	x

Table 28. Cluster Subset Results for Markers Found Significant in Complete Case-Control Dataset. Marks in a cluster column indicate that the marker was found significant by at least one main effect family-based test (linkage, FBAT or PDT) in that cluster subset.

			Cluster		
Chrom	**Gene**	**Marker**	**0**	**1**	**2**
10	CDC2	rs2448347		x	
10	VR22	rs1786927			
10	VR22	rs2441718			
10	VR22	rs2456737			
10	LRRTM3	rs942780			
10	LRRTM3	rs1925617	x	x	x
12	GAPD	rs1060621			
19	APOE	rs440446	x	x	x

an Italian sample (Bagnoli et al., 2005) and in a Scottish and Swedish sample (Blomqvist et al., 2004).

Cluster 0 Discussion

For cluster 0, three genes (PLAU, IDE and ACE) showed interesting results for main effect and/or interaction analyses. In PLAU, the marker rs2227568 was significant according to both its two-point HetLOD score and its Pearson chi-square statistic. The PLAU markers rs2227564 and rs2227566, which are in LD with the former marker, were also significant by their HetLOD scores.

IDE (insulin degrading enzyme; MIM#146680) is a metallopeptidase that can degrade peptides such as amyloid beta and may be responsible for the removal of extracellular amyloid beta (Selkoe, 2001) and the clearance of the cytoplasmic fragment of amyloid precursor protein following liberation of the amyloid-beta protein (Edbauer et al., 2002). In IDE, the marker rs7099761 was significant by its HetLOD score and its Pearson chi-square statistic. The IDE marker rs1544210, which is in LD with the former marker, was also significant by its Pearson chi-square statistic.

Perhaps the most interesting results were for the ACE gene. ACE (angiotensin 1 converting enzyme; MIM#106180) has been shown to inhibit the aggregation of amyloid beta by degrading amyloid beta 40 into less toxic products (Hu et al., 1999; Hu et al., 2001). The marker rs4291 was significant by its PDT, FBAT and Pearson chi-square statistics and appeared in the best one- and two-locus MDR models for the 1DSP family-based dataset. This two-locus MDR model was confirmed by logistic regression to be largely a main effect of rs4291. Five other ACE markers—rs4295, rs4311, rs4646994,

rs4343 and rs4978—which were all in LD with the former marker and/or each other, were also significant by their PDT and FBAT statistics.

Cluster 1 Discussion

For cluster 1, four genes (PLAU, IDE, CDC2 and ACE) showed interesting results for main effect and/or interaction analyses. In PLAU, markers rs1916341, rs2227566 and rs4065, which are in LD with each other, were all significant by their HetLOD scores and Pearson chi-square statistics. In IDE, marker rs1832186 was significant by its FBAT, PDT and Pearson chi-square statistics. In addition, IDE markers rs2251101 and rs4646954, which are in LD with rs1832186, were also significant by their FBAT and PDT chi-square statistics.

CDC2 (cell division cycle 2; MIM#116940) is a kinase involved in the abnormal phosphorylation of tau and the aggregation of tau into paired helical filaments (Pei et al., 2006), which are present in the neurofibrillary tangles of Alzheimer disease. In CDC2, markers rs2448347 and 1920 were significant by their HetLOD scores and their Pearson chi-square statistics. Another CDC2 marker, rs2448341, which is in LD with rs2448347, was also significant by its Pearson chi-square statistics and its PDT chi-square statistic.

Finally, in ACE, markers rs4646994 and rs4343, which are in LD with each other, were significant by their Pearson chi-square statistics, and rs4343 appeared in the best two-locus MDR model in the case-control dataset, which was confirmed by logistic regression to have both a main and interactive effect involving rs4343. In addition, ACE markers rs4353 and rs4978, which are in LD with rs4646994, were found significant by their PDT chi-square statistics.

Cluster 2 Discussion

In cluster 2, there were no genes, other than the expected LRRTM3, VR22 and APOE, which showed evidence for association in both the case-control and family-based datasets. This subset was the smallest and most unbalanced from each of the case-control and family-based clusters, and it is possible that its overall size and/or the extent of the imbalance between affecteds and unaffecteds made these analyses too underpowered to detect an effect, if it were there. It is also possible that no interactions exist with the cluster 2 LRRTM3 haplotype and the other markers included in the datasets.

Other Discussion

No discussion about a large data analysis project such as this would be complete without mention of the multiple-testing problem. As one increases the number of tests performed, the likelihood of generating false positive results also increases, beyond the per-comparison significance level (alpha) established at the beginning of the study. There are a number of different strategies for correcting for this inflation of the false positive rate. However, most are quite conservative, and in light of the fact that the current study is exploratory in nature, such caution at the expense of power would be imprudent. For example, if we were to use a simple Bonferoni correction (Dunn OJ, 1961), we would have to divide our per-comparison alpha by the total number of markers being tested (93 for the case-control dataset and 138 for the family-based dataset), resulting in a family-wise alpha of about 0.0005. In the overall datasets, only APOE marker rs440446 would have been considered statistically significant, and in the cluster subsets, only a few other LRRTM3 markers would have reached significance also. A

more liberal correction strategy such as False Discovery Rate could be employed instead (Benjamini and Hochberg, 1995; Storey, 2002). However, since our predominant goal is to not miss any real effects, which we could subsequently investigate further, even a slightly more liberal correction strategy is not desired. In addition, since we know there is considerable LD among our markers, the assumption that all the tests are independent is not valid. We would, in fact, expect that two markers in LD with each other would frequently produce similar results, in excess of how often two independent markers should do so. Furthermore, since all of the markers tested were chosen because they are functional and/or positional candidates for late-onset Alzheimer disease, the likelihood that a positive result is true is higher than it would be if the markers were chosen at random, for example in the case of a genomic screen. Finally and perhaps most importantly, it should be noted that we have tested two independent datasets, which serve as one test and one replication dataset, and are focusing only on those effects that were found significant (at the per-comparison level of 0.05) in both datasets. Thus, we have further reduced the chance that such a statistically significant result is a false positive.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

Summary, Conclusions and Future Studies

Common diseases with a genetic basis are likely to have a very complex etiology, in which the mapping between genotype and phenotype is far from straightforward. A new comprehensive statistical and computational strategy for identifying the missing link between genotype and phenotype has been proposed. Numerous examples of heterogeneity and gene-gene or gene-environment interactions support the theoretical basis for such an approach, which emphasizes the need to address heterogeneity in the first stage of any analysis (Chapter II). Uncovering any heterogeneity that may exist in a dataset removes a formidable source of noise, affording main effect and interaction analysis methods the best opportunity to detect effects that may be present only on particular genetic backgrounds or in individuals with particular environmental exposure(s).

It is a reality that currently a majority of genetic studies, particularly those involving neurological diseases, do not have substantial phenotypic data available, even though the quality and volume of genotypic data may be excellent. Many factors, including cost, feasibility (invasiveness), and technical limitations (reliability and interpretation) of phenotyping technologies, make the collection of rich phenotypic data more challenging. Given the lack of methods for dissecting heterogeneity that do not rely on substantial phenotypic data, a comparison of three 'unsupervised' clustering methods

was conducted (Chapter III). Bayesian Classification was chosen as the best of these methods, which allow detection of multilocus genotype patterns that may underlie or be a proxy for genetic or trait heterogeneity. It performed very well under a simple genetic model of trait heterogeneity, and it had very good control of its false positive rate and acceptably low false negative rates under specific simulation conditions.

Since it is unknown how complex the genetic models in real data are, a further evaluation was conducted of Bayesian Classification's performance and applicability under a wider range of simulation conditions was performed (Chapter IV). False positive rates were well-controlled under all conditions simulated. However, false negative rates varied dramatically between conditions. Under the specific condition of having a relatively high number of nonfunctional loci (100) and a moderate sample size (500 affected individuals), the false negative rates were unacceptably high (at or above 60 percent). However, for all other conditions, the false negative rates were at or below 20 percent, with most below five percent (at an alpha of ten percent). The other number of nonfunctional loci tested was an order of magnitude lower (10), and the other value for sample size was double (1000). Thus, further simulation studies exploring the "breakpoint" or slope of the false negative rates between the two extremes of the current simulation conditions may further aid in interpretation of negative results. For example, there may be a critical ratio of independent variables (genotypes) to instances (individuals) above 5 affecteds per marker genotyped that must be maintained in order to keep false negative rates under control, and this would be an important to know when designing a study or interpreting results from a Bayesian Classification analysis.

The application to late-onset Alzheimer disease presented in Chapter V involves a family-based dataset with 138 markers genotyped and 1422 affected individuals (yielding a ratio of over 10 affecteds per marker genotyped) and a case-control dataset with 93 markers genotyped and 451 affected individuals (yielding a ratio of just under 5 affecteds per marker genotyped). Thus, based on the simulation studies, the case-control dataset may have been underpowered to allow detection of heterogeneity by the Bayesian Classification method. However, this concern was perhaps mitigated by taking a consensus approach, looking for commonality of high-influence markers between the two datasets.

Bayesian Classification found statistically significant clusterings for both the family-based and case-control datasets, which used the same five markers as their most influential in determining cluster assignment. These markers were all in LRRTM3 and were in high linkage disequilibrium with each other. Each of the three resulting clusters could be characterized by their haplotypes at these five markers, and the same haplotypes defined the clusters in both the family-based and case-control data. In subsequent analyses to detect main effects and gene-gene interactions, markers in four genes—PLAU, IDE, CDC2 and ACE—were found to be associated with late-onset Alzheimer disease in particular subsets of the data based on their LRRTM3 haplotype. While all of these genes are viable candidates for LOAD based on their known biological function, further studies are needed to replicate these statistical findings and to elucidate possible biological interaction mechanisms between LRRTM3 and these genes.

Future Directions for Research

Molecular biologists and geneticists alike now acknowledge that the most common human diseases with a genetic component are likely to have complex etiologies. Similarly, there has been increasing appreciation for the phenotypic complexity of disease traits and for the need to collect rich phenotypic data to facilitate the elucidation of the even more complex relationships between genotypes and phenotypes. Investigation of such complexity requires well-informed study design, meticulous data collection and innovative strategies for data analysis.

Over the past twenty years, advances in genotyping technology have far outpaced those in statistical and computational methods for analyzing genetic data. Likewise, geneticists have given much less attention to phenotyping technologies. To most effectively leverage the massive amounts of genotypic data being produced, we must have comparably rich datasets of phenotypic information available for mapping genotypes to phenotypes. Thus, going forward, genetic studies will need to increasingly focus time and resources to collecting phenotypic data that can refine definitions or subcategories of traits or diseases and can serve as endophenotypes, which are more likely to have simple etiologies and to directly map to specific genetic markers.

In the case of neurological diseases, one collection of phenotyping technologies which has matured considerably over the past five to ten years is neuroimaging. Magnetic resonance imaging (MRI) and positron emission tomography (PET) have been used successfully to detect signs of disease, sometimes in advance of clinical symptoms, in such neurological diseases as Alzheimer disease (Masters et al., 2006; Small et al., 2000) schizophrenia (Ho et al., 2003; Velakoulis et al., 2006) and Tourette syndrome

(Gerard and Peterson, 2003). The more recently developed diffusion tensor imaging (DTI) method might come even closer to measuring a biologically relevant proxy for neuronal dysfunction, and it has already been applied to such neurological diseases as Alzheimer disease (Nierenberg et al., 2005), schizophrenia (Buchsbaum et al., 2006) and Turner syndrome (Holzapfel et al., 2006). Neuroimaging methods are minimally invasive and can produce data with good spatial or temporal resolution. Voxel-based morphometry methods are being developed and applied for associating differences in activation of particular brain regions with genetic markers of disease.

In addition to these neuroimaging technologies, an emphasis on possible biological mechanisms of disease has positively influenced the design of behavioral assessment tools, increasing their utility for phenotyping and producing endophenotypes that can be mapped to genotypic data. Overall, careful planning of study designs will be essential, making best use of existing resources and keeping in mind what statistical and computational analyses will be possible based on the types of data to be collected. Future genetic studies of neurological diseases will require collaboration among geneticists, behavioral neuroscientists and neuroimaging experts, particularly in the short-term. Methodologies enabling the integration of disparate data sources (genotyping and neuroimaging or behavioral) must be investigated in order to harness the power inherit in their complexity.

REFERENCES

Anderberg MR. *Cluster Analysis for Applications*. New York: Academic Press, 1973.

Andersen K, Launer LJ, Ott A, Hoes AW, Breteler MM, Hofman A. Do nonsteroidal anti-inflammatory drugs decrease the risk for Alzheimer's disease? The Rotterdam Study. *Neurology* 45: 1441-1445, 1995.

Andorfer C, Acker CM, Kress Y, Hof PR, Duff K, Davies P. Cell-cycle reentry and cell death in transgenic mice expressing nonmutant human tau isoforms. *J Neurosci* 25: 5446-5454, 2005.

Ashley-Koch AE, Mei H, Jaworski J, Ma DQ, Ritchie MD, Menold MM, DeLong GR, Abramson RK, Wright HH, Hussman JP, Cuccaro ML, Gilbert JR, Martin ER, Pericak-Vance MA. 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: 281-292, 2006.

Bagnoli S, Tedde A, Cellini E, Rotondi M, Nacmias B, Sorbi S. The urokinase-plasminogen activator (PLAU) gene is not associated with late onset Alzheimer's disease. *Neurogenetics.* 6: 53-54, 2005.

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265, 2005.

Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B* 57: 289-300, 1995.

Bergem AL. Heredity in dementia of the Alzheimer type. *Clin Genet* 46: 144-149, 1994.

Blomqvist ME, Andreasen N, Bogdanovic N, Blennow K, Brookes AJ, Prince JA. Genetic variation in CTNNA3 encoding alpha-3 catenin and Alzheimer's disease. *Neurosci.Lett.* 358: 220-222, 2004.

Bradford Y, Haines JL, Hutcheson H, Gardiner M, Braun T, Sheffield V, Cassavant T, Huang W, Wang K, Vieland V, Folstein S, Santangelo S, Piven J. Incorporating language phenotypes strengthens evidence of linkage to autism. *Am J Med Genet* 105: 539-547, 2001.

Breitner JC, Silverman JM, Mohs RC, Davis KL. Familial aggregation in Alzheimer's disease: comparison of risk among relatives of early-and late-onset cases, and

among male and female relatives in successive generations. *Neurology* 38: 207-212, 1988.

Breitner JC, Welsh KA, Helms MJ, Gaskell PC, Gau BA, Roses AD, Pericak-Vance MA, Saunders AM. Delayed onset of Alzheimer's disease with nonsteroidal anti-inflammatory and histamine H2 blocking drugs. *Neurobiol Aging* 16: 523-530, 1995.

Brown DF, Dababo MA, Bigio EH, Risser RC, Eagan KP, Hladik CL, White CLI. Neuropathologic evidence that the Lewy body variant of Alzheimer disease represents coexistence of Alzheimer disease and idiopathic Parkinson disease. *J Neuropathol Exp Neurol* 57: 39-46, 1998.

Buchsbaum MS, Friedman J, Buchsbaum BR, Chu KW, Hazlett EA, Newmark R, Schneiderman JS, Torosjan Y, Tang C, Hof PR, Stewart D, Davis KL, Gorman J. Diffusion Tensor Imaging in Schizophrenia. *Biol Psychiatry* 2006.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22: 231-238, 1999.

Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A. Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* 32: 237-244, 2002.

Chakravarti A. Population genetics--making sense out of sequence. *Nat Genet* 21: 56-60, 1999.

Cheeseman P and Stutz J. Bayesian classification (AutoClass): theory and results. In *Advances in Knowledge Discovery and Data Mining*. Edited by: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Menlo Park: The AAAI Press, 153-180, 1996.

Chen JY, Stern Y, Sano M, Mayeux R. Cumulative risks of developing extrapyramidal signs, psychosis, or myoclonus in the course of Alzheimer's disease. *Arch Neurol* 48: 1141-1143, 1991.

Cheverud JM and Routman EJ. Epistasis and its contribution to gentic variance components. *Genetics* 139: 1455-1461, 1995.

Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47: 549-554, 2004.

Citron M, Westaway D, Xia W, Carlson G, Diehl T, Levesque G, Johnson-Wood K, Lee M, Seubert P, Davis A, Kholodenko D, Motter R, Sherrington R, Perry B, Yao H, Strome R, Lieberburg I, Rommens J, Kim S, Schenk D, Fraser P, St George HP, Selkoe DJ. Mutant presenilins of Alzheimer's disease increase production of 42-residue amyloid beta-protein in both transfected cells and transgenic mice. *Nat Med* 3: 67-72, 1997.

Collinge J, Palmer MS, Dryden AJ. Genetic predisposition to iatrogenic Creutzfeldt-Jakob disease. *Lancet* 337: 1441-1442, 1991.

Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 118: 201-210, 1993.

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

Cruts M, Backhovens H, Wang SY, Van Gassen G, Theuns J, De Jonghe CD, Wehnert A, De Voecht J, De Winter G, Cras P, . Molecular genetic analysis of familial early-onset Alzheimer's disease linked to chromosome 14q24.3. *Hum Mol Genet* 4: 2363-2371, 1995.

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

Daw EW, Heath SC, Wijsman EM. Multipoint oligogenic analysis of age-at-onset data with applications to Alzheimer disease pedigrees. *Am J Hum Genet* 64: 839-851, 1999.

Daw EW, Payami H, Nemens EJ, Nochlin D, Bird TD, Schellenberg GD, Wijsman EM. The number of trait loci in late-onset Alzheimer disease. *Am J Hum Genet* 66: 196-204, 2000.

De Silva R, Ironside JW, McCardle L, Esmonde T, Bell J, Will R, Windl O, Dempster M, Estibeiro P, Lathe R. Neuropathological phenotype and 'prion protein' genotype correlation in sporadic Creutzfeldt-Jakob disease. *Neurosci Lett* 179: 50-52, 1994.

Devos D, Schraen-Maschke S, Vuillaume I, Dujardin K, Naze P, Willoteaux C, Destee A, Sablonniere B. Clinical features and genetic analysis of a new form of spinocerebellar ataxia. *Neurology* 56: 234-238, 2001.

Ditter SM and Mirra SS. Neuropathologic and clinical features of Parkinson's disease in Alzheimer's disease patients. *Neurology* 37: 754-760, 1987.

Doh-ura K, Tateishi J, Sasaki H, Kitamoto T, Sakaki Y. Pro-to-leu change at position 102 of prion protein is the most common but not the sole mutation related to Gerstmann-Straussler syndrome. *Biochem Biophys Res Comm* 163: 974-979, 1989.

Duda RO and Hart PE. *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1973.

Duff K, Eckman C, Zehr C, Yu X, Prada CM, Perez-Tur J, Hutton M, Buee L, Harigaya Y, Yager D, Morgan D, Gordon MN, Holcomb L, Refolo L, Zenk B, Hardy J, Younkin S. Increased amyloid-beta42(43) in brains of mice expressing mutant presenilin 1. *Nature* 383: 710-713, 1996.

Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc* 56: 52-64, 1961.

Edbauer D, Willem M, Lammich S, Steiner H, Haass C. Insulin-degrading enzyme rapidly removes the beta-amyloid precursor protein intracellular domain (AICD). *J Biol Chem* 277: 13389-13393, 2002.

Ertekin-Taner N, Ronald J, Feuk L, Prince J, Tucker M, Younkin L, Hella M, Jain S, Hackett A, Scanlin L, Kelly J, Kihiko-Ehman M, Neltner M, Hersh L, Kindy M, Markesbery W, Hutton M, de Andrade M, Petersen RC, Graff-Radford N, Estus S, Brookes AJ, Younkin SG. Elevated amyloid beta protein (Abeta42) and late onset Alzheimer's disease are associated with single nucleotide polymorphisms in the urokinase-type plasminogen activator gene. *Hum Mol Genet* 14: 447-460, 2005.

Finckh U, van Hadeln K, Muller-Thomsen T, Alberici A, Binetti G, Hock C, Nitsch RM, Stoppe G, Reiss J, Gal A. Association of late-onset Alzheimer disease with a genotype of PLAU, the gene encoding urokinase-type plasminogen activator on chromosome 10q22.2. *Neurogenetics.* 4: 213-217, 2003.

Flegal KM, Carroll MD, Kuczmarski RJ. Overweight and obesity in the United States: prevalence and trends, 1960-1994. *Int J Obe Relat Metab Disord* 22: 39-47, 1998.

Fogel GB and Corne DW. *Evolutionary Computation in Bioinformatics*. San Francisco: Elsevier Science, 2002.

Fox NC and Rossor MN. Seeing what Alzheimer saw---with magnetic resonance microscopy. *Nat Med* 6: 20-21, 2000.

Frankel WN and Schork NJ. Who's afraid of epistasis? *Nat Genet* 14: 371-373, 1996.

Friedman J. Multivariate adaptive regression splines. *Ann Stat* 19: 1-141, 1991.

Gerard E and Peterson BS. Developmental processes and brain imaging studies in Tourette syndrome. *J Psychosom Res* 55: 13-22, 2003.

Glabe C. Does Alzheimer disease tilt the scales of amyloid degradation versus accumulation? *Nat Med* 6: 133-134, 2000.

Goedert M. Pinning down phosphorylated tau. *Nature* 399: 739-740, 1999.

Good IJ. A causal calculus. *British Journal of Philosophy of Science* 11: 305-318, 1961.

Good P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer, 2000.

Growden JH. Advances in the Diagnosis of Alzheimer's Disease. 139-153, 1995.

Guillozet AL, Weintraub S, Mash DC, Mesulam MM. Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Arch Neurol* 60: 729-736, 2003.

Hahn LW and Moore JH. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol* 4: 0016-2004.

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

Han EH, Karypis G, Kumar V, Mobasher B. Clustering Based on Association Rule Hypergraphs. *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery:* 9-13, 1997a.

Han EH, Karypis G, Kumar V, Mobasher B. Clustering in High Dimensional Space Using Hypergraph Models. In *Technical Report #97-063*. Computer Science and Engineering, University of Minnesota, 1997b.

Hanson R, Stutz J, Cheeseman P. Bayesian classification theory. In *Technical Report # FIA-90-12-7-01*. Artificial Intelligence Research Branch, NASA Ames Research Center, 1991.

Harding AE. Clinical features and classification of inherited ataxia. *Adv Neurol* 61: 1-14, 1993.

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M. Ordered subset analysis in genetic linkage mapping of complex traits. *Genet Epidemiol* 27: 53-63, 2004.

Hauser ER, Watanabe RM, Duren WL, Boehnke M. Stratified Linkage Analysis of Complex Genetic Traits Using Relate Covariates. *Am J Hum Genet Suppl* 63: A45-A45, 1998.

Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA. Alzheimer Disease in the US Population: Prevalence Estimates Using the 2000 Census. *Arch Neurol* 60: 1119-1122, 2003.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 4: 45-61, 2002.

Hirst C, Yee IM, Sadovnick AD. Familial risks for Alzheimer disease from a population-based series. *Genet Epidemiol* 11: 365-374, 1994.

Ho BC, Andreasen NC, Nopoulos P, Arndt S, Magnotta V, Flaum M. Progressive structural brain abnormalities and their relationship to clinical outcome: a longitudinal magnetic resonance imaging study early in schizophrenia. *Arch Gen Psychiatry* 60: 585-594, 2003.

Hoh J, Wille A, Ott J. Trimming, Weighting, and Grouping SNPs in Human Case-Control Association Studies. *Genome Res* 11: 2115-2119, 2001.

Holzapfel M, Barnea-Goraly N, Eckert MA, Kesler SR, Reiss AL. Selective alterations of white matter associated with visuospatial and sensorimotor dysfunction in turner syndrome. *J Neurosci* 26: 7007-7013, 2006.

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

Hu J, Igarashi A, Kamata M, Nakagawa H. Angiotensin-converting enzyme degrades Alzheimer amyloid beta-peptide (A beta); retards A beta aggregation, deposition, fibril formation; and inhibits cytotoxicity. *J Biol Chem* 276: 47863-47868, 2001.

Hu J, Miyatake F, Aizu Y, Nakagawa H, Nakamura S, Tamaoka A, Takahash R, Urakami K, Shoji M. Angiotensin-converting enzyme genotype is associated with Alzheimer disease in the Japanese population. *Neurosci Lett* 277: 65-67, 1999.

Huang Z and Ng MK. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans Fuzzy Syst* 7: 446-452, 1999.

Hubert L and Arabie P. Comparing partitions. *J Classif* 2: 193-218, 1985.

Janssens B, Goossens S, Staes K, Gilbert B, van Hengel J, Colpaert C, Bruyneel E, Mareel M, van Roy F. alphaT-catenin: a novel tissue-specific beta-catenin-binding protein mediating strong cell-cell adhesion. *J Cell Sci* 114: 3177-3188, 2001.

Katzman R and Fox P. The world wide impact of dementia in the next fifty years. In Mayeux R and Christen Y (Eds): Epidemiology of Alzheimer's Disease: From Genes to Prevention. Berlin: Springer, 1999.

Kaufman L and Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc., 1990.

Kulczycki LL, Kostuch M, Bellanti JA. A clinical perspective of cystic fibrosis and new genetic findings: relationship of CFTR mutations to genotype-phenotype manifestations. *Am J Hum Genet* 116A: 262-267, 2003.

Kurtzke JF. Multiple sclerosis: changing times. *Neuroepidemiology* 10: 1-8, 1991.

Lauren J, Airaksinen MS, Saarma M, Timmusk T. A novel gene family encoding leucine-rich repeat transmembrane proteins differentially expressed in the nervous system. *Genomics* 81: 411-421, 2003.

Li WT and Reich J. A complete enumeration and classification of two-locus disease models. *Hum Hered* 50: 334-349, 2000.

Liang X, Schnetz-Boutaud N, Bartlett J, Anderson BM, Gwirtsman H, Schmechel D, Carney R, Gilbert JR, Pericak-Vance MA, Haines LH. Association analysis of genetic polymorphisms in the cell division cycle 2 (CDC2) gene with late-onset Alzheimer disease. *Dement Geriatr Cogn Disord* (In Press).

Lucek P, Hanke J, Reich J, Solla SA, Ott J. Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum Hered* 48: 275-284, 1998.

Ma DQ, Whitehead PL, Menold MM, Martin ER, Ashley-Koch AE, Mei H, Ritchie MD, DeLong GR, Abramson RK, Wright HH, Cuccaro ML, Hussman JP, Gilbert JR, Pericak-Vance MA. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. *Am J Hum Genet* 77: 377-388, 2005.

Marinov M and Weeks D. The complexity of linkage analysis with neural networks. *Hum Hered* 51: 169-176, 2001.

Martin ER, Ritchie MD, Kang S, Hahn L, Moore JH. A novel method to Identify potential interactions in nuclear families: The MDR-PDT. *Genet Epidemiol* 2005.

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

Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 67: 146-154, 2000.

Masters CL, Cappai R, Barnham KJ, Villemagne VL. Molecular mechanisms for Alzheimer's disease: implications for neuroimaging and therapeutics. *J Neurochem* 97: 1700-1725, 2006.

Mayeux R, Ottman R, Maestre G, Ngai C, Tang MX, Ginsberg H, Chun M, Tycko B, Shelanski M. Synergistic effects of traumatic head injury and apolipoprotein-epsilon 4 in patients with Alzheimer's disease. *Neurology* 45: 555-557, 1995.

Mayeux R, Stern Y, Sano M. Heterogeneity and prognosis in dementia of the Alzheimer type. *Bull Clin Neurosci* 50: 7-10, 1985.

McCulloch W and Pitts W. A logical calculus of the ideas immenent in nervous activity. *Bull of Math Biophys* 5: 115-133, 1943.

McKeith I, Mintzer J, Aarsland D, Burn D, Chiu H, Cohen-Mansfield J, Dickson D, Dubois B, Duda JE, Feldman H, Gauthier S, Halliday G, Lawlor B, Lippa C,

Lopez OL, Carlos MJ, O'Brien J, Playfer J, Reid W. Dementia with Lewy bodies. *Lancet Neurol* 3: 19-28, 2004.

McKeith IG, Galasko D, Kosaka K, Perry EK, Dickson DW, Hansen LA, Salmon DP, Lowe J, Mirra SS, Byrne EJ, Lennox G, Quinn NP, Edwardson JA, Ince PG, Bergeron C, Burns A, Miller BL, Lovestone S, Collerton D, Jansen EN, Ballard C, de Vos RA, Wilcock GK, Jellinger KA, Perry RH. Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the consortium on DLB international workshop. *Neurology* 47: 1113-1124, 1996.

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34: 939-944, 1984.

McPhie DL, Coopersmith R, Hines-Peralta A, Chen Y, Ivins KJ, Manly SP, Kozlowski MR, Neve KA, Neve RL. DNA synthesis and neuronal apoptosis caused by familial Alzheimer disease mutants of the amyloid precursor protein are mediated by the p21 activated kinase PAK3. *J Neurosci* 23: 6914-6927, 2003.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 78: 15-27, 2006.

Molsa PK, Marttila RJ, Rinne UK. Extrapyramidal signs in Alzheimer's disease. *Neurology* 34: 1114-1116, 1984.

Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56: 73-82, 2003.

Moore JH, Lamb JM, Brown NJ, Vaughan DE. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. *Clin Genet* 62: 74-79, 2002.

Moore JH and Ritchie MD. The challenges of whole-genome approaches to common diseases. *J Am Med Assoc* 291: 1642-1643, 2004.

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

Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 241: 252-261, 2006.

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

Morgan JN and Sonquist JA. Problems in the analysis of survey data and a proposal. *J Am Stat Assoc* 58: 415-434, 1963.

Morton N. Sequential tests for the detection of linkage. *Am J Hum Genet* 7: 277-318, 1955.

Mountain JL and Cavalli-Sforza LL. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am J Hum Genet* 61: 705-718, 1997.

Mufson EJ, Chen EY, Cochran EJ, Beckett LA, Bennett DA, Kordower JH. Entorhinal cortex beta-amyloid load in individuals with mild cognitive impairment. *Exp Neurol* 158: 469-490, 1999.

Narod SA, Dupont A, Cusan L, Diamond P, Gomez J-L, Suburu R, Labrie F. The impact of family history on early detection of prostate cancer. *Nat Med* 1: 99-101, 1995.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11: 458-470, 2001.

Neuman RJ, Richard T, Andrew H, Reich W, Hudziak JJ, Bucholz KK, Madden PAF, Begleiter H, Porjesz B, Juperman S, Hesselbrock V, Reich T. Evaluation of ADHD Typology in Three Contrasting Samples: A Latent Class Approach. *J Am Acad Child Adolesc Psychiatry* 38: 25-33, 1999.

Neuman RJ, Saccone NL, Holmans P, Rice JP, Sun L. Clustering Methods Applied to Allele Sharing Data. *Genet Epidemiol* 19: S57-S63, 2000.

Nierenberg J, Pomara N, Hoptman MJ, Sidtis JJ, Ardekani BA, Lim KO. Abnormal white matter integrity in healthy apolipoprotein E epsilon4 carriers. *Neuroreport* 16: 1369-1372, 2005.

Ott J and Hoh J. Set association analysis of SNP case-control and microarray data. *J Comput Biol* 10: 569-574, 2003.

Ott J. *Analysis of Human Genetic Linkage.* Baltimore: Johns Hopkins University Press, 1999.

Ott J. Strategies for characterizing highly polymorphic markers in human gene mapping. *Am J Hum Genet* 51: 283-290, 1992.

Owen F, Poulter M, Collinge J, Crow TJ. A codon 129 polymorphism in the PRIP gene. *Nucleic Acids Res* 18: 3103-1990.

Palmer MS, Dryden AJ, Hughes JT, Collinge J. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. *Nature* 352: 340-342, 1991.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 49: 1373-1379, 1996.

Pei JJ, An WL, Zhou XW, Nishimura T, Norberg J, Benedikz E, Gotz J, Winblad B. P70 S6 kinase mediates tau phosphorylation and synthesis. *FEBS Lett* 580: 107-114, 2006.

Pericak-Vance MA and Haines JL. The genetics of Alzheimer Disease. In Kind RA, Rotter JI, Motulsky AG (Eds): Genetic Basis of Common Diseases. Oxford: Oxford University Press, 2002.

Perry R, McKeith I, Perry E. Lewy body dementia--clinical, pathological and neurochemical interconnections. *J Neural Transm* 51: 95-109, 1997.

Perry RJ and Hodges JR. Fate of patients with questionable (very mild) Alzheimer's disease: longitudinal profiles of individual subjects' decline. *Dement Geriatr Cogn Disord* 11: 342-349, 2000.

Pickles A, Bolton P, Macdonald H, Bailey A, Le Couteur A, Sim C-H, Rutter M. Latent class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Am J Hum Genet* 57: 717-726, 1995.

Povey S, Burley MW, Attwood J, Benham F, Hunt D, Jeremiah SJ, Franklin D, Gillett G, Malas S, Robson EB, Tippett P, Edwards JH, Kwiatkowski DJ, Super M, Mueller R, Fryer A, Clarke A, Webb D, Osborne J. Two loci for tuberous sclerosis: one on 9q34 and one on 16p13. *Ann Hum Genet* 58: 107-127, 1994.

Province MA, Shannon WD, Rao DC. Classification methods for confronting heterogeneity. *Adv Genet* 42: 273-286, 2001.

Qian S, Jiang P, Guan XM, Singh G, Trumbauer ME, Yu H, Chen HY, Van de Ploeg LH, Zheng H. Mutant human presenilin 1 protects presenilin 1 null mouse against embryonic lethality and elevates Abeta1-42/43 expression. *Neuron* 20: 611-617, 1998.

Qin S, Zhao X, Pan Y, Liu J, Feng G, Fu J, Bao J, Zhang Z, He L. An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray. *Eur.J Hum Genet* 13: 807-814, 2005.

Rao VS, van Duijn CM, Connor-Lacke L, Cupples LA, Growdon JH, Farrer LA. Multiple etiologies for Alzheimer disease are revealed by segregation analysis. *Am J Hum Genet* 55: 991-1000, 1994.

Reich DE and Lander ES. On the allelic spectrum of human disease. *Trends Genet* 17: 502-510, 2001.

Risch N. A new statistical test for linkage heterogeneity. *Am J Hum Genet* 42: 353-364, 1988.

Risch N and Merikangas K. The future of genetic studies of complex human disorders. *Science* 273: 1516-1517, 1996.

Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, phenocopy and genetic heterogeneity. *Genet Epidemiol* 24: 150-157, 2003.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69: 138-147, 2001.

Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 24: 150-157, 2003a.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69: 138-147, 2001.

Ritchie M, White BPJHL, Moore J. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *Bioinformatics* 4: 28, 2003b.

Rivolta C, Sharon D, DeAngelis MM, Dryja TP. Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. *Hum Mol Genet* 11: 1219-1227, 2002.

Roberts GW, Gentleman SM, Lynch A. Beta-amyloid protein deposition in the brain after severe head injury. *J Neurol Neurosurg Psychiatry* 57: 419-425, 1994.

Rosenberg RN. Autosomal dominant cerebellar phenotypes: the genotype has settled the issue. *Neurology* 45: 1-5, 1995.

Roses AD, Devlin B, Conneally PM, Small GW, Saunders AM, Pritchard PA, Lock JL, Haines JL, Pericak-Vance MA, Risch N. Measuring the genetic contribution of APOE in late-onset Alzheimer Disease (AD). *Am J Hum Genet* 57: A202-1995.

Roses AD. Apolipoprotein E, a gene with complex biological interactions in the aging brain. *Neurobiol Dis* 4: 170-185, 1997.

Sadovnick AD, Irwin ME, Baird PA, Beattle BL. Genetic studies on an Alzheimer clinic population. *Genet Epidemiol* 6: 663-643, 1989.

Saunders AM, Strittmatter WJ, Breitner JC, Schmechel D, St George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Growden J, Alberts MJ, Hulette C, Crain B, Goldgaber D, Roses AD. Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43: 1467-1472, 1993.

Schultz S and Andreasen N. Schizophrenia. *Lancet* 353: 1425-1430, 1999.

Selkoe DJ. Clearing the brain's amyloid cobwebs. *Neuron* 32: 177-180, 2001.

Seno M and Karypis G. LPMiner: An Algorithm for Finding Frequent Itemsets Using Length-Decreasing Support Constraint. *Proceedings of the IEEE Conference on Data Mining:* 505-512, 2001.

Seshadri S, Beiser A, Selhub J, Jacques PF, Rosenberg IH, D'Agostino RB, Wilson PW, Wolf PA. Plasma homocysteine as a risk factor for dementia and Alzheimer's disease. *N Engl J Med* 346: 476-483, 2002.

Shahani N, Subramaniam S, Wolf T, Tackenberg C, Brandt R. Tau aggregation and progressive neuronal degeneration in the absence of changes in spine density and morphology after targeted expression of Alzheimer's disease-relevant tau constructs in organotypic hippocampal slices. *J Neurosci* 26: 6103-6114, 2006.

Sham PC, Wessely S, Castle DJ, Farmer AE, Murray RM. Further exploration of a latent class typology of schizophrenia. *Schizophr Res* 20: 105-115, 1996.

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

Sherriff A and Ott J. Applications of neural networks for gene finding. *Adv Genet* 42: 287-298, 2001.

Siegmund KD, Langholz B, Kraft P, Thomas DC. Testing linkage disequilibrium in sibships. *Am J Hum Genet* 67: 244-248, 2000.

Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet Suppl* 32: 502-508, 2002.

Slooter AJC, Cruts M, Kalmijn S, Hofman A, Breteler MM, Van Broeckhoven C, van Duijn CM. Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: the Rotterdam Study. *Arch Neurol* 55: 964-968, 1998.

Small GW, Ercoli LM, Silverman DH, Huang SC, Komo S, Bookheimer SY, Lavretsky H, Miller K, Siddarth P, Rasgon NL, Mazziotta JC, Saxena S, Wu HM, Mega MS, Cummings JL, Saunders AM, Pericak-Vance MA, Roses AD, Barrio JR, Phelps ME. Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proc Natl Acad Sci USA* 97: 6037-6042, 2000.

Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet* 27: 175-182, 1963.

Smith CAB. Testing for heterogeneity of recombination fractions values in human genetics. *Ann Hum Genet* 27: 175-182, 1963.

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

Steinley D. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychol Methods* 9: 386-396, 2004.

Storey JD. A direct approach to false discovery rates. *J R Statist Soc B* 64: 479-498, 2002.

Strittmatter WJ, Weisgraber KH, Huang DY, Dong L-M, Salvesen GS, Pericak-Vance MA, Schmechel D, Saunders AM, Goldgaber D, Roses AD. Binding of human apolipoprotein E to synthetic amyloid b peptide: isoform-specific effects and implications for late-onset Alzheimer disease. *Proc Natl Acad Sci USA* 90: 8098-8102, 1993.

Tager-Flusberg H and Joseph RM. Identifying neurocognitive phenotypes in autism. *Philos Trans R Soc Lond B Biol Sci* 358: 303-314, 2003.

Taylor W, Potts J, Cook D, Cheeseman P, Stutz J. Autoclass C Documentation (search-c.text). 2002.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Change M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikan R, Roberts T, Sdicu A, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Gurd CG, Numro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. *Science* 303: 808-813, 2004.

Tsai CT, Lai LP, Lin JL, Chiang FT, Hwang JJ, Ritchie MD Moore JH, Hsu KL, Tseng CD, Liau CS, Tseng YZ. Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 109: 1640-1646, 2004.

van Duijn CM, Farrer LA, Cupples LA, Hofman A. Genetic transmission of Alzheimer's disease among families in a Dutch population based study. *J Med Genet* 30: 640-646, 1993.

Velakoulis D, Wood SJ, Wong MT, McGorry PD, Yung A, Phillips L, Smith D, Brewer W, Proffitt T, Desmond P, Pantelis C. Hippocampal and amygdala volumes according to psychosis stage and diagnosis: a magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. *Arch Gen Psychiatry* 63: 139-149, 2006.

Wille A, Hoh J, Ott J. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genet Epidemiol* 25: 350-359, 2003.

Williams SM, Haines JL, Moore JH. The use of animal models in the study of complex disease: all else is never equal, or why do so many human studies fail to replicate animal findings? *Bioessays* 26: 170-179, 2004.

Wisniewski T, Golabek A, Matsubara E, Ghiso J, Frangione B. Apolipoprotein E: binding to soluble Alzheimer beta-amyloid. *Biochem Biophys Res Comm* 192: 359-365, 1993.

Yang Y, Varvel NH, Lamb BT, Herrup K. Ectopic cell cycle events link human Alzheimer's disease and amyloid precursor protein transgenic mouse models. *J Neurosci* 26: 775-784, 2006.

Young J and Povey S. The genetic basis of tuberous sclerosis. *Mol Med Today* 4: 313-319, 1998.

Zee RY, Solomon SD, Ajani UA, Pfeffer MA, Lindpaintner K, Heart investigators. A prospective evaluation of the angiotensin-converting enzyme D/I polymorphism and left ventricular remodeling in the 'Healing and Early Afterload Reducing Therapy' study. *Clin Genet* 61: 21-25, 2002.