

COMPARISON OF THREE CLUSTERING METHODS
FOR DISSECTING TRAIT HETEROGENEITY
IN SIMULATED GENOTYPIC DATA

TRICIA A. THORNTON-WELLS

Thesis under the direction of Professor Jonathan L. Haines

Trait heterogeneity, which exists when a trait has been defined with insufficient specificity such that it is actually two or more distinct traits, has been implicated as a confounding factor in traditional statistical genetics of complex human disease. In the absence of detailed phenotypic data collected consistently in combination with genetic data, unsupervised computational methodologies offer the potential for discovering underlying trait heterogeneity. The performance of three such methods—Bayesian Classification, Hypergraph-Based Clustering, and Fuzzy k -Modes Clustering—that are appropriate for categorical data were compared. Also tested was the ability of these methods to additionally detect trait heterogeneity in the presence of locus heterogeneity and gene-gene interaction, which are two other complicating factors in discovering genetic models of complex human disease. Bayesian Classification performed well under the simplest of genetic models simulated, and it outperformed the other two methods, with the exception that the Fuzzy k -Modes Clustering performed best on the most complex genetic model. Permutation testing showed that Bayesian Classification controlled Type I error very well but produced less desirable Type II error rates. Methodological limitations and future directions are discussed.

Approved _____ Date _____

COMPARISON OF THREE CLUSTERING METHODS
FOR DISSECTING TRAIT HETEROGENEITY
IN SIMULATED GENOTYPIC DATA

By

Tricia A. Thornton-Wells

Thesis

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2005

Nashville, Tennessee

Approved:

Professor Constantin F. Aliferis

Professor Jonathan L. Haines

Professor Mike McDonald

Professor Jason H. Moore

Professor Thomas J. Palmeri

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 Department of Biomedical Informatics, the Neuroscience Graduate Program, or the National Library of Medicine Training Grant Fellowship. I would like to acknowledge the contribution of Scott Dudek, who programmed the Fuzzy k -Modes Clustering method. I would also like to thank Marylyn Ritchie, Lance Hahn and Bill White for their thoughtful input on study design.

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 Thesis 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 am especially 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 University, 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.

TABLE OF CONTENTS

	Page
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	v
LIST OF FIGURES	vi
LIST OF ABBREVIATIONS	vii
Chapter	
I. BACKGROUND	1
Complex human genetic disease	1
Statistical analysis	3
Cluster analysis	4
II. METHODS	6
Data simulation	6
Clustering methods	11
Statistical analysis	14
Comparison of clustering methods	14
Analysis of cluster-specific metrics	15
III. RESULTS	17
Descriptive Statistics	17
Multiway analysis of variance and Chi-Square Analysis	20
Correlation of ARI_{HA} and Bayesian Classification Internal Clustering Metrics	24
Permutation testing	29
IV. DISCUSSION	33
Data simulation	33
Method comparison	33
Permutation testing	35
REFERENCES	37

LIST OF TABLES

Table	Page
1. Power calculations for multiway ANOVA.....	8
2. Confidence intervals around ARI_{HA} means by method.....	18
3. Confidence intervals around ARI_{HA} means by method & model.....	19
4. Overall results of Chi-Square Test of Independence.....	22
5. Results of Chi-Square Test of Independence for Trait Heterogeneity Only datasets.....	23
6. Correlations between ARI_{HA} and Bayesian Classification internal clustering metrics	25

LIST OF FIGURES

Figure	Page
1. Factors complicating analysis of complex genetic disease: definitions, diagrams and examples	2
2. Structure of genetic models used for data simulation	7
3. Recessive genetic model.....	9
4. Novel data simulation algorithm	10
5. “Zagzig” genetic model.....	11
6. Comparison of ARI_{HA} means by method	17
7. Comparison of ARI_{HA} means by method and model.....	18
8. Distribution of ARI_{HA} means by method.....	21
9. Percentage of clustering results achieving cluster recovery levels by method.....	22
10. Percentage of clustering results achieving cluster recovery levels by method and model....	23
11. Average log of class strength versus ARI_{HA}	26
12. Average cross-class entropy versus ARI_{HA}	26
13. Average log of class strength versus ARI_{HA} paneled by number of nonfunctional loci.....	27
14. Average cross-class entropy versus ARI_{HA} paneled by number of resulting clusters	28
15. False positive rate by significance level.....	30
16. False negative rate by significance level	30
17. False negative rate by significance level paneled by number of nonfunctional loci.....	31
18. False negative rate by significance level paneled by number of affecteds.....	32

LIST OF ABBREVIATIONS

ARI_{HA}	Hubert-Arabie Adjusted Rand Index
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

BACKGROUND

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 (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. 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 Figure 1). 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.

Despite the consensus that common genetic disease is likely to be complex, statistical geneticists continue primarily using traditional methodologies to attack the 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. Among the possible reasons for this failure are false positives due to population stratification and true differences in genetic etiology between study populations (Hirschhorn NJ *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

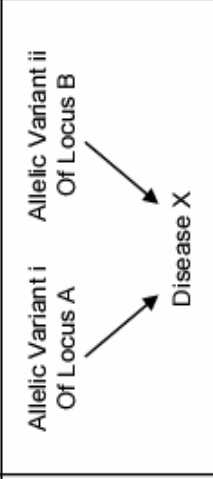

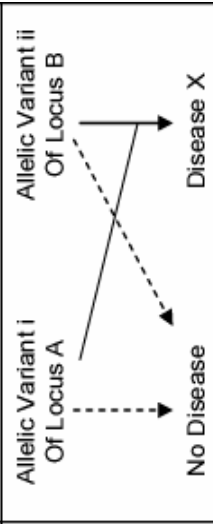
	Locus Heterogeneity	Trait Heterogeneity	Gene-Gene Interaction
Definition	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	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
Diagram			
Example One	<p>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 <i>et al.</i>, 2002; http://www.sph.uth.tmc.edu/RetNet)</p>	<p>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 <i>AE</i>, 1982; Rosenberg <i>RN</i>, 1995), and different genetic loci have been associated with the different subtypes (Devos <i>D et al.</i>, 2001)</p>	<p>Hirschsprung Disease (OMIM# 142623) - variants in the <i>RET</i> (OMIM# 164761) and <i>EDNRB</i> (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 <i>MM et al.</i>, 2002)</p>
Example Two	<p>Tuberous Sclerosis (TS, OMIM# 191100) - out of families informative for linkage analysis, half have mutations in the <i>TSC1</i> gene (located at 9q34), and the other half have mutations in the <i>TSC2</i> gene (located at 16p13) (Kulczycki <i>LL et al.</i>, 2003; Povey <i>S et al.</i>, 1994; Young <i>J and Povey S</i>, 1998)</p>	<p>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 <i>H and Joseph RM</i>, 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 <i>Y et al.</i>, 2001; Shao <i>Y et al.</i>, 2002)</p>	<p>Creutzfeldt-Jakob Disease (CJD, OMIM# 123400) and Fatal Familial Insomnia (OMIM# 176640.0010) - the Met129Val polymorphism and Asp178Asn mutation in the <i>PRNP</i> 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 <i>K et al.</i>, 1989; Owen <i>F et al.</i>, 1990; Collinge <i>J et al.</i>, 1991; Palmer <i>MS</i>, 1991; De <i>Silva et al.</i>, 1994; Doh-ura <i>K et al.</i>, 1991; Goldfarb <i>LG et al.</i>, 1992)</p>

Figure 1. Factors Complicating Analysis of Complex Genetic Disease: Definitions, Diagrams and Examples (Thornton-Wells *TA et al.*, 2004)

factor for diabetes (Shao Y *et al.*, 2002; Spielman RS *et al.*, 1993). Current statistical approaches to detecting heterogeneity, such as the admixture test (Ott J, 1992; Smith CAB, 1963), are neither sensitive nor powerful and can merely account for, not resolve, any underlying heterogeneity. 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 MD *et al.*, 2001). 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. However, for most diseases, especially neurological ones, in which I am particularly interested, little detailed phenotypic data has been collected consistently in combination with genotypic data. It is for these reasons that an unsupervised method, which does not rely on phenotypic data, is needed to mine potentially heterogeneous genotypic data as a means of data stratification and hypothesis generation.

For complicating genetic factors involving heterogeneity, 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 locus heterogeneity, multiple predictor variables (e.g., multiple loci) are present, some of which may be unmeasured or unobserved and, therefore, unavailable for inclusion in the disease model. In the case of trait heterogeneity, multiple outcome variables are present, which cannot or have not been distinguished based on the available phenotypic information. Gene-gene interactions 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 probably ubiquitous (Doh-ura K *et al.*, 1991; Goldfarb LG *et al.*, 1992; Moore JH, 2003; Tong AH *et al.*, 2004). Thus, it is critical that complex genetic data sets be properly interrogated for possible underlying interactions.

Statistical Analysis

No one analytic method is superior in all respects for the range of complicating factors that might be present in a specific data set. Given the relative shortcomings of our current analyses in complex diseases, we need to greatly extend the range of available analytical tools. There is a critical 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, while 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, though, 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. While no single method adequately investigates heterogeneity and interaction issues simultaneously, we propose routinely taking a two-step approach to analysis. For example, clustering or ordered subset analysis (Hauser *et al.*, 2004) 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 (Hoh J *et al.*, 2001; Ott J and Hoh J, 2003) or the multifactor dimensionality reduction method (Ritchie MD *et al.*, 2001; Ritchie MD *et al.*, 2003) could 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.

Cluster Analysis

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 data sets. 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 the next chapter) 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.*, 1997), and Fuzzy *k*-Modes Clustering (Huang Z and Ng MK, 1999)—all of which are appropriate for categorical data.

CHAPTER II

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 2 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, (2) five percent, which is characteristic of a relatively common disease such as prostate cancer, (3) one percent, which is characteristic of a less common disease such as schizophrenia, and (4) one tenth of one percent, which is characteristic of a more uncommon disease such as multiple sclerosis. 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 when one is searching for the functional locus or loci 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, resulting in a total of 19,200 simulated data sets, were simulated. Table 1 shows the power to detect a certain effect size using multiway analysis of variance given 100 datasets per set of data simulation parameters. An effect size of 0.10 is considered small, one of

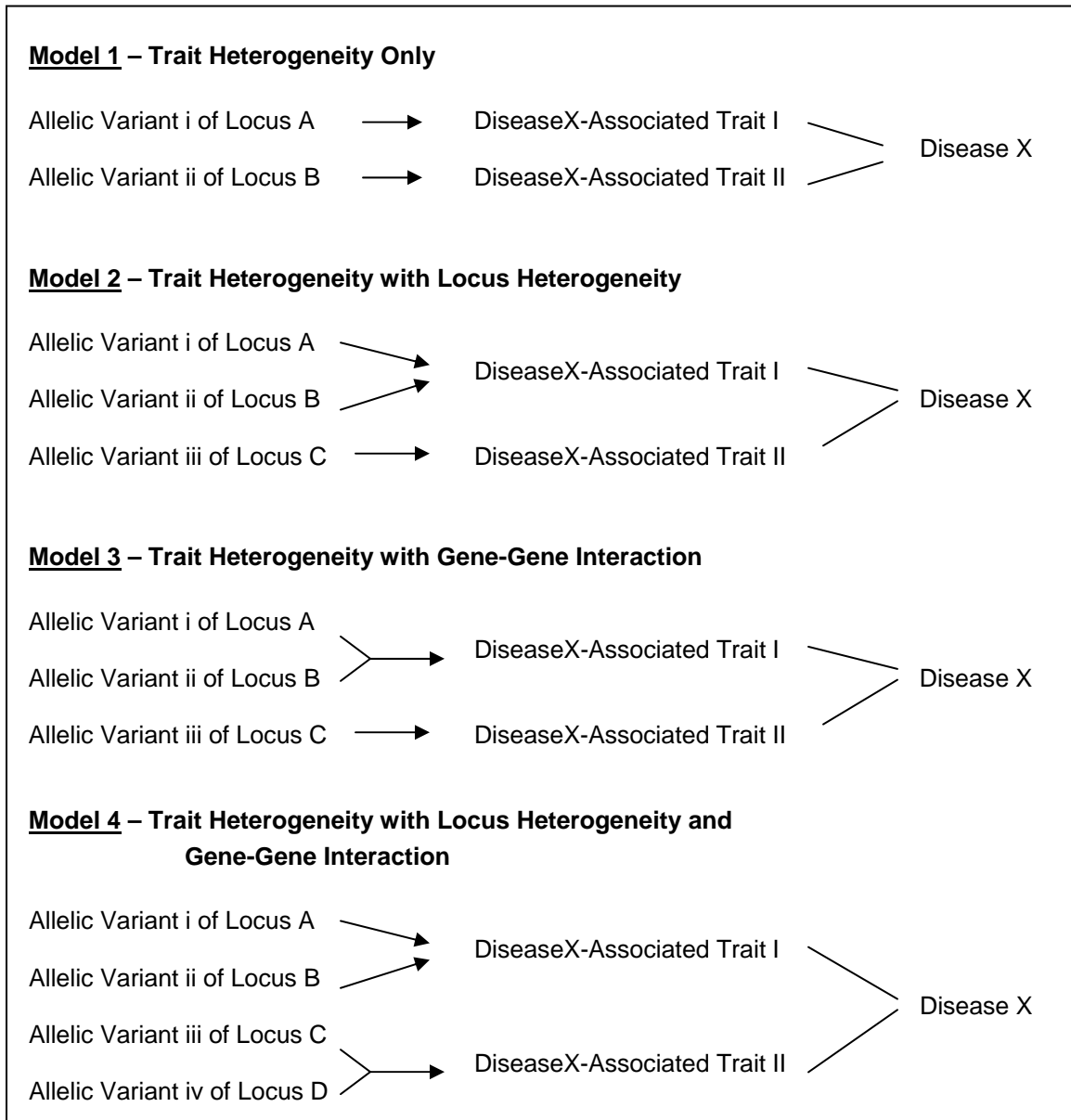


Figure 2. Structure of Genetic Models Used for Data Simulation

0.25 is medium, and one of 0.40 is large. Three of the four data simulation parameters (model type, prevalence and number of nonfunctional loci) have four levels, or groups, each. The fourth parameter (number of affecteds) has three groups. A different power calculation is provided for the factor combinations that differ by number of groups and degrees of freedom. Assuming a small effect size, reasonable power can be expected for comparing performance across methods and by number of affecteds plus one of the 4-group factors. For medium effect sizes, good power can be expected for comparing method performance by up to three of the four-group factors. Beyond that, power falls to unacceptably low levels.

Table 1. Power calculations for multiway ANOVA, given N=100 for each set of factors.

Factors	N	# Groups	df	α	Effect Size	Power
Method	19,200	3	2	0.05	0.10	1
Method * NumAffecteds	6400	9	4	0.05	0.10	1
Method * (Model or Prevalence or NumNFLoci)	4800	12	6	0.05	0.10	0.99
Method * NumAffecteds * (Model or Prevalence or NumNFLoci)	1600	36	12	0.05	0.25 0.10	1 0.76
Method * (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci)	1200	48	18	0.05	0.25 0.10	1 0.51
Method * NumAffecteds (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci)	400	144	36	0.05	0.40 0.25	0.99 0.71
Method * (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci)	300	192	54	0.05	0.40 0.25	0.88 0.36
Method * NumAffecteds (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci) * (Model or Prevalence or NumNFLoci)	100	576	162	0.05	0.40	0

For the purposes of simulating this data, a novel data simulation algorithm capable of incorporating these complex genetic factors in an epidemiologically-sound manner was designed and developed (see Figure 4). Penetrance is the probability of having a particular trait given a specific genotype (single or multilocus). Prevalence, on the other hand, is the percentage of

individuals in a population that have a particular trait. 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.

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 one or both of the disease-associated “functional” loci (see Figure 3). A naturally occurring degree of overlap results, such that some affected individuals have the high-risk genotypes from both loci.

	AA	Aa	aa
BB	0	0	0.1
Bb	0	0	0.1
bb	0.1	0.1	0.1

Figure 3. Recessive genetic model in which the disease is penetrant only when two copies of the high risk allele at one locus are present (in this case the a and b alleles are high risk). Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype.

In the second quarter of the datasets, locus heterogeneity was also simulated 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 (creating fifty percent locus heterogeneity). Again each locus operates under a recessive model for disease, such that the disease-associated genotype consists of two copies of the high-risk allele. A naturally occurring degree of overlap results, such that some affected individuals have the high-risk genotypes from two or even all three loci.

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

$$\prod_{i=1}^L \sum_{j=1}^{A_i} j \quad \text{where } L \text{ is the total number of loci and } A_i \text{ is the number of alleles for locus } i.$$

2. For each penetrance function $p(\text{Status}=\text{Affected} \mid \text{Multilocus Genotype})$
==>Distribute $1-p$ across relevant cells of Unaffecteds probability array
3. Populate cells of the Affecteds probability array with $1-(\text{cell probability})$ 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 4. 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.

In the third quarter of the datasets, a gene-gene interaction was simulated for one of the two traits. The “zagzig” gene-gene interaction model, first described by Frankel and Schork (Frankel WN and Schork NJ, 1996), which is nonlinear and nonadditive in nature, was used (see Figure 5). 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. By chance, some affected individuals are simulated to have both sets of high-risk genotypes associated with the two traits.

	AA	Aa	aa
BB	0	0	0.1
Bb	0	0.05	0
bb	0.1	0	0

Figure 5. “Zagzig” genetic model first described by Frankel & Schork (Frankel WN and Schork NJ, 1996). Two loci—A and B—are involved, each with two alleles—A and a; and B and b, respectively. Cell values indicate penetrance level, or the probability of having the trait, given the corresponding multilocus genotype.

In the fourth quarter of the datasets, one trait is simulated to involve locus heterogeneity, while the other is simulated to involve the “zagzig” gene-gene interaction, as described above. Thus, there are some affected individuals who, by chance, will have one or both high-risk genotypes from the first trait as well as the high-risk genotype from the second trait.

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 is categorical, which immediately narrows the field of appropriate methods for this study to only a few. The goal of this clustering is to find a partitioning of the affected individuals based on multilocus genotypic combinations that maps onto the trait

heterogeneity simulated in the data. Three different clustering methodologies were chosen that are suitable for categorical data and are appealing due to their speed or theoretical underpinnings.

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 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. Instead, emphasis is placed on the identification of which attributes, or loci, are most important in producing the clustering.

The second method is Hypergraph Clustering (Han EH *et al.*, 1997). 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, which in this case represent simulated affected individuals, such that intracluster relatedness meets a specified threshold, while the weight of hyperedges cut by the partitioning is minimized. Hyperedge weights are determined using association rules, which are simply patterns of frequently-occurring variable instances. 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, using a series of phases—(1) Coarsening phase, (2) Initial partitioning phase, (3) Uncoarsening and refinement phase, and (4) V-cycle refinement—to create a partition of the hypergraph such that the weight of the removed hyperedges is minimized.

This results in a partitioning of the genotypes. If one were simply analyzing a single dataset, this information would be sufficient, in and of itself, since it would provide information about which multilocus genotypes were most important for partitioning out individuals. However, for the purpose of comparing the results of Hypergraph Partitioning to those of the other two methods, which produce partitionings of individuals (not genotypes), such a partitioning of individuals still needed to be created. To that end, 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. 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.

The third clustering method is Fuzzy k -Modes Clustering (Huang Z and Ng MK, 1999). k -Modes is a trivial extension of the popular k -means algorithm to categorical data. In both methods, cluster centroids are 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 not necessarily meaningful, and the k -modes algorithm instead determines the centroid as the mode vector of genotypes across individuals. 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. 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 the 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 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, one would either have to say none of those individuals were incorrectly classified, since they are all in homogenous clusters, or else one would have to consider all individuals from one of those clusters as 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 to cluster, the number of resulting clusters, and the relative size of those clusters (Steinley D, 2004).

It is, however, sensitive to the degree of class overlap, which is desirable since it will penalize more for chance good clusterings than classification error would. When interpreting ARI_{HA} values, 0.90 and greater can be considered excellent recovery, 0.80 and greater is good recovery, 0.65 and greater reflects moderate recovery, and less than 0.65 indicates poor 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. The assumptions of the multiway ANOVA were tested: (1) normality for each group, and (2) equality of error variances across groups. Since neither assumption held and data transformations were not advisable, the planned multiway ANOVA was not performed, and instead, a nonparametric method was used. Three new categorical variables were created that essentially captured the same information but could be tested using the 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. Five percent was chosen as the acceptable Type I error (false positive) rate. An evaluation was performed of whether one method significantly outperformed the others and whether that method performed satisfactorily according to the ARI_{HA} .

Analysis of Cluster-Specific Metrics

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 one does not know *a priori* to which clusters individuals belong, else 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, a correlation analysis using the Pearson correlation coefficient was performed for the ARI_{HA} and that method’s internal clustering metrics to determine how good a proxy they were for ARI_{HA} .

In addition, permutation testing was performed using the ARI_{HA} and the internal clustering metric that was the best proxy for ARI_{HA} . The ratio of 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 will create a null sample in which the frequency of multilocus genotypes is no longer associated with trait status except by chance. The superior clustering method was applied to each permuted data set and both the internal clustering metric value and the ARI_{HA} was determined. For each set of 100 permuted data sets, the distributions of the internal clustering metric values and the ARI_{HA} values were plotted. The significance of each of the simulated dataset results was determined based on whether it exceeded the values at the significance level in the 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, should one decide to use these more stringent significance criteria.

Finally, the ability of permutation testing to preserve an acceptable Type I (false positive) error rate was evaluated at the three specified significance levels. A false positive was defined as a clustering result which had a p-value according to the internal clustering metric that was significant but a p-value according to ARI_{HA} that was not significant. The Type II (false negative) error rate was evaluated at the same alpha levels to determine the sensitivity for detecting trait heterogeneity when it is present. A false negative was defined as a clustering result which had a p-value according to the internal clustering metric that was not significant but a p-value according to ARI_{HA} that was significant.

CHAPTER III

RESULTS

Descriptive Statistics

Descriptive statistics and plots for the Hubert-Arabie Adjusted Rand Index results were produced. You will recall that a score of 0.90 on the ARI_{HA} indicates excellent cluster recovery, 0.80 good recovery, and 0.65 moderate recovery. Mean ARI_{HA} values for Bayesian Classification, Hypergraph Clustering and Fuzzy k -Modes Clustering were 0.666, 0.354 and 0.556, respectively (Figure 6). Mean ARI_{HA} values differed by model type as well, with higher scores achieved on Trait Heterogeneity Only (THO) datasets for the Bayesian Classification and Hypergraph Clustering methods (see Figure 7).

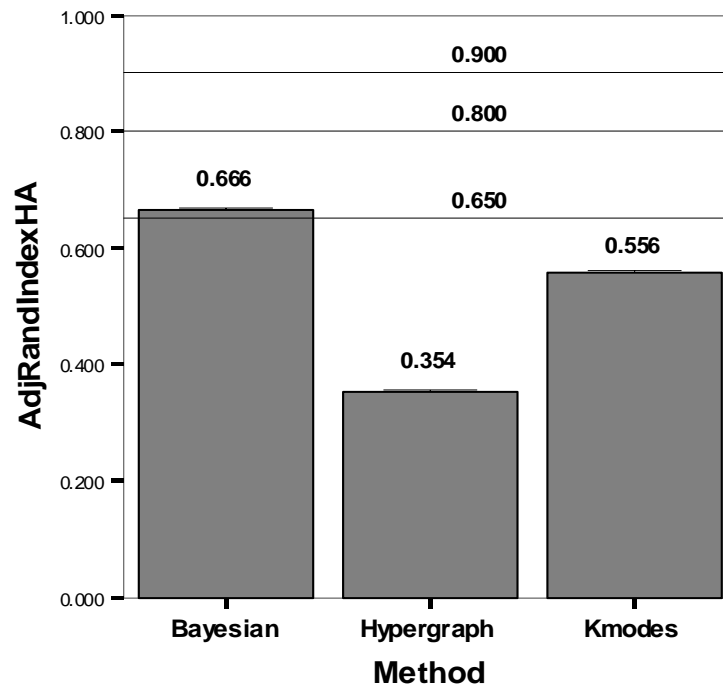


Figure 6. Comparison of Hubert-Arabie Adjusted Rand Index means by method (averaged over all parameter settings). Columns represent means. Horizontal lines represent thresholds for quality of cluster recovery: 0.90 for excellent recovery, 0.80 for good recovery, and 0.65 for moderate recovery. The barely visible error bars represent 95% confidence interval.

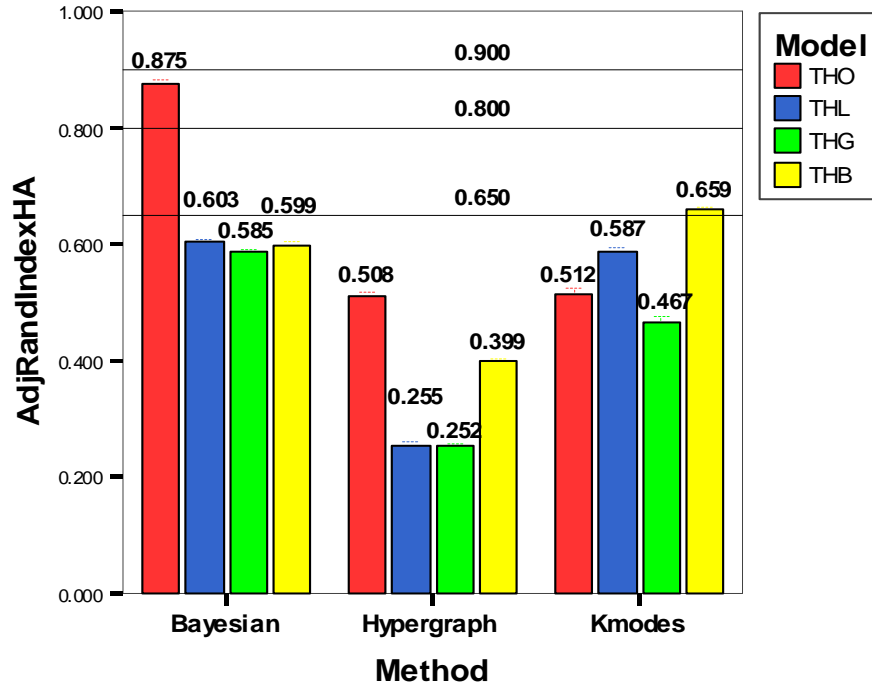


Figure 7. Comparison of Hubert-Arabie Adjusted Rand Index means by method and model. Horizontal lines represent thresholds for quality of cluster recovery: 0.90 for excellent recovery, 0.80 for good recovery and 0.65 for moderate recovery. Model abbreviations are as follows: Trait Heterogeneity Only (THO), Trait Heterogeneity with Locus Heterogeneity (THL), Trait Heterogeneity with Gene-Gene Interaction (THG), and Trait Heterogeneity with Both Locus Heterogeneity and Gene-Gene Interaction (THB)

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 2. The results for each method by model type are presented in Table 3. The intervals were very narrow, on the order of thousandths.

Table 2. Confidence intervals around ARI_{HA} means by method

Method	Mean	Standard Error	Confidence Interval	
			Lower End	Upper End
Bayesian	0.666	0.001	0.664	0.667
Hypergraph	0.354	0.001	0.352	0.355
Fuzzy <i>k</i> -Modes	0.556	0.001	0.555	0.558

Table 3. Confidence intervals around ARI_{HA} means by method and model. Model abbreviations are as follows: Trait Heterogeneity Only (THO), Trait Heterogeneity with Locus Heterogeneity (THL), Trait Heterogeneity with Gene-Gene Interaction (THG), and Trait Heterogeneity with Both Locus Heterogeneity and Gene-Gene Interaction (THB).

Method	Model	Mean	Standard Error	Confidence Interval	
				Lower End	Upper End
Bayesian	THO	0.875	0.001	0.872	0.878
	THL	0.603	0.001	0.601	0.606
	THG	0.585	0.001	0.583	0.588
	THB	0.599	0.001	0.596	0.601
Hypergraph	THO	0.508	0.001	0.506	0.511
	THL	0.255	0.001	0.252	0.257
	THG	0.252	0.001	0.250	0.255
	THB	0.399	0.001	0.396	0.401
Fuzzy <i>k</i> -Modes	THO	0.512	0.001	0.510	0.515
	THL	0.587	0.001	0.585	0.590
	THG	0.467	0.001	0.464	0.469
	THB	0.659	0.001	0.656	0.661

Sometimes during the course of a run, the Fuzzy *k*-Modes Clustering algorithm would end up moving all individuals from a given cluster into other clusters, thereby resulting in some ‘empty’ clusters, such that the number of true ‘resulting’ clusters was smaller than the prespecified number. In fact, it often produced one cluster (effectively no partitioning) of the data, especially for datasets with a larger number of nonfunctional loci. This result was due to a convergence problem in which the algorithm initially choose as cluster modes two individuals whose multilocus genotypes were so similar to each other that the probability that any given individual in the dataset would belong to one cluster versus another was equal. In such cases, the individuals were arbitrarily assigned to the first cluster, thereby resulting in the remaining cluster(s) being empty. As one might expect, as the number of nonfunctional loci increased, this result was more common.

Multiway Analysis of Variance and Chi-Square Analysis

For the planned comparison across methods, the distribution assumptions of the planned multiway ANOVA were tested: (1) normality for each group and (2) equal variance across groups being compared. Using the Kolmogorov-Smirnov test of normality with Lilliefors significance correction, for each of the three distributions by method, the null hypothesis that they were drawn from a population with a normal distribution was rejected (Bayesian Classification $D=0.173$, $df=19200$, $p<0.001$; Hypergraph Clustering $D=0.126$, $df=19200$, $p<0.001$; Fuzzy k -Modes Clustering $D=0.221$, $df=19200$, $p<0.001$). In addition, using the Levene's test for equality of variances, the null hypothesis that the variances were equal across the three distributions was rejected ($F=225.101$, $df_1=2$, $df_2=57597$, $p<0.001$). It was readily evident from visual examination of the distributions of ARI_{HA} that not only were the distributions not normal and without equal variances, the shape of the distributions were very different across the groups. (See Figure 8.) The distribution of ARI_{HA} scores for Bayesian Classification was bimodal, that for Hypergraph Clustering was negatively skewed, and that for Fuzzy k -Modes Clustering was slightly positively skewed. In light of this, one particular data transformation was unlikely to normalize all three of them. Additionally, performing three different data transformations, even if acceptable, would make interpretation of the results extremely difficult.

Therefore, three new categorical variables were constructed that essentially captured the same information as the raw ARI_{HA} values but could be tested using the nonparametric chi-square test of independence. The three variables were calculated as the number of clustering results achieving each of the three ARI_{HA} cutoff values of 0.65 (for moderate cluster recovery), 0.80 (for good cluster recovery) and 0.90 (for excellent cluster recovery). Results are displayed in terms of percentages by clustering method (Figure 9) and by clustering method and genetic model (Figure 10). A chi-square test of independence was performed testing the null hypothesis that the number of clusterings achieving a certain ARI_{HA} cutoff value was independent of the clustering method. The three methods performed significantly differently on each of the new ARI_{HA} cutoff statistics (Table 4). The Bayesian Classification outperformed the other two methods. However, across all the dataset parameters, Bayesian Classification achieved moderate or better recovery on only 49% of the datasets—hardly a stellar performance (Figure 9).

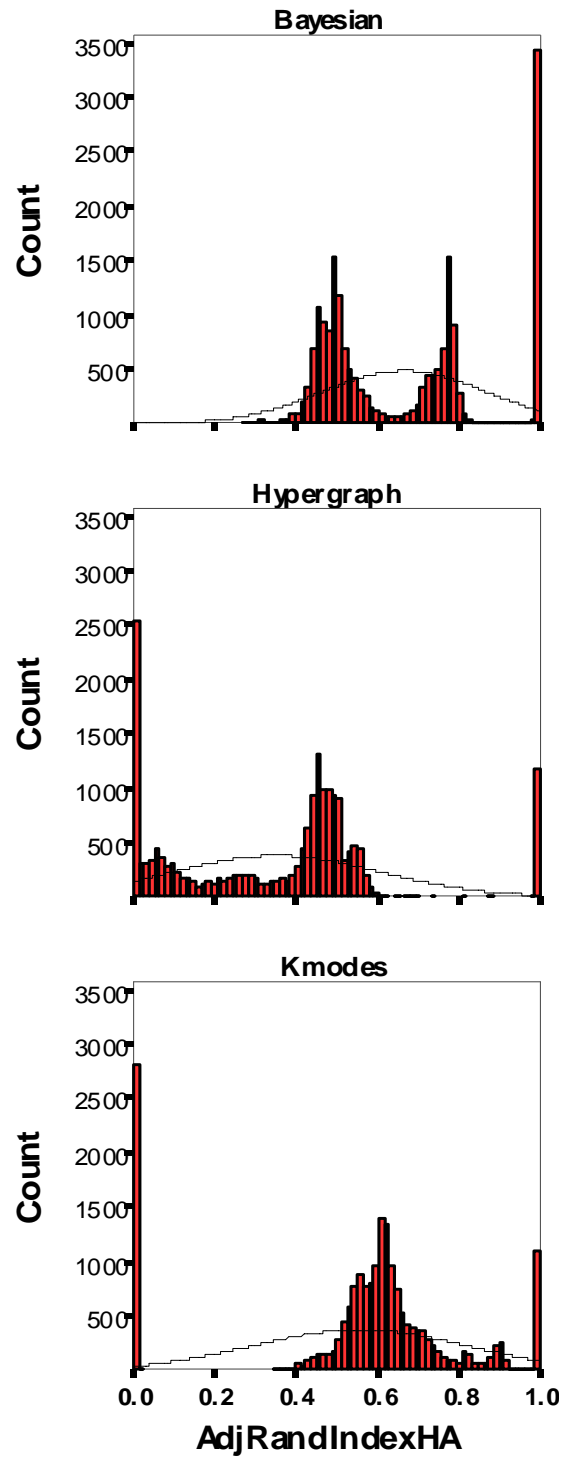


Figure 8. Distribution of ARI_{HA} means by method.

Table 4. Overall results of Chi-Square Test of Independence testing the null hypothesis that the percentage of clustering results achieving the specified cluster recovery level does not differ across clustering methods.

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

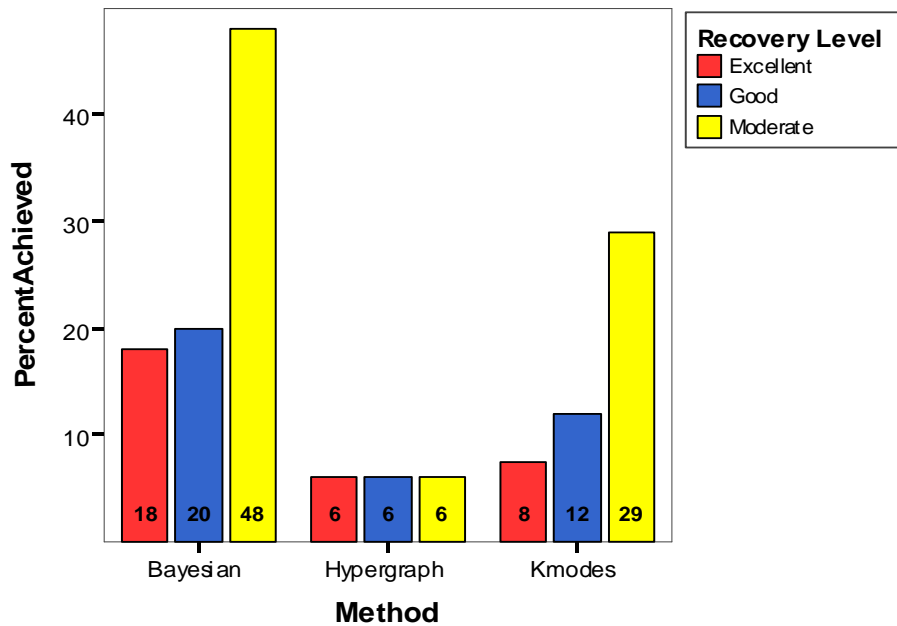


Figure 9. Percentage of clustering results achieving cluster recovery levels by method.

The performance of the three clustering methods across different dataset parameters was evaluated in an attempt 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 10). For this subset of the datasets, Bayesian Classification

Table 5. Results of Chi-Square Test of Independence for Trait Heterogeneity Only datasets, the testing the null hypothesis that the percentage of clustering results achieving the specified cluster recovery level does not differ across clustering methods.

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

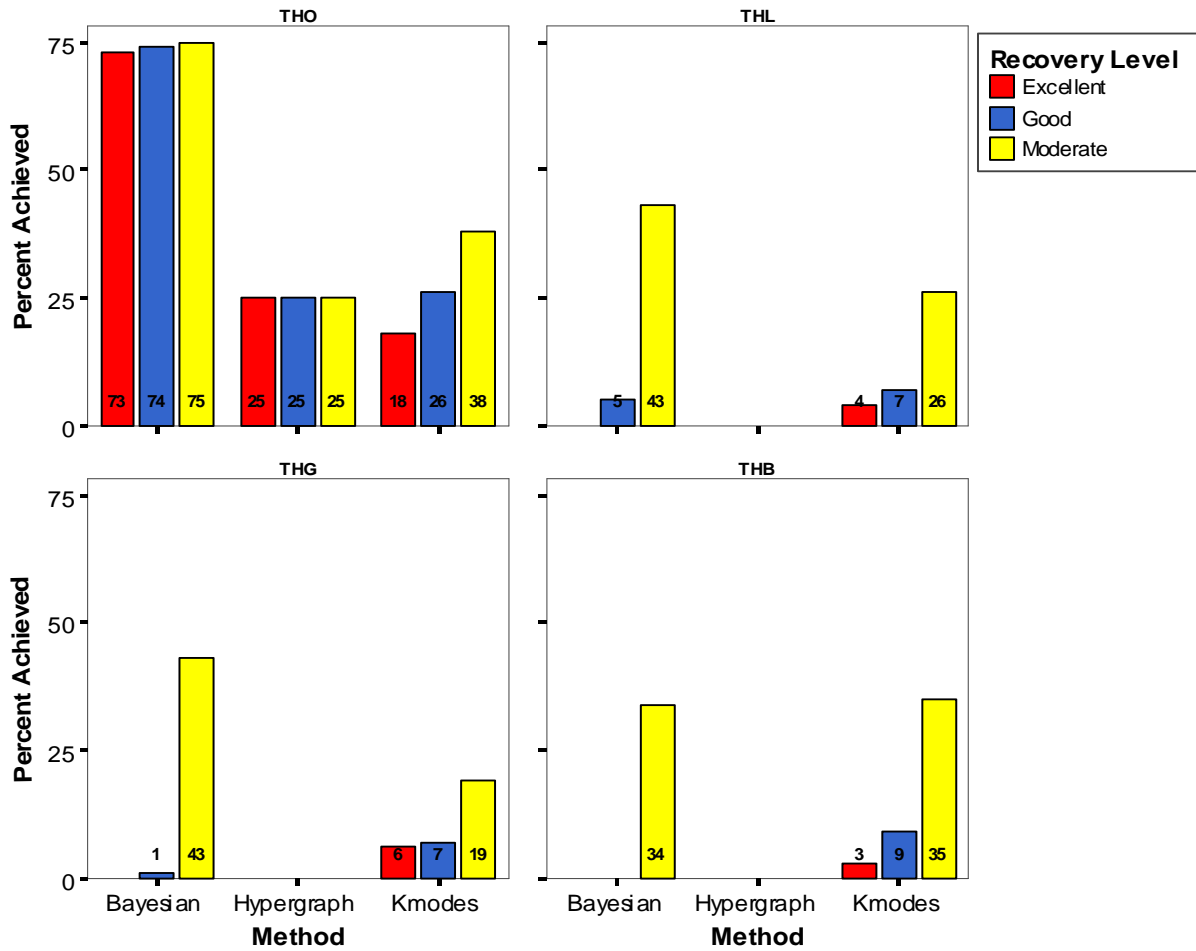


Figure 10. Percentage of clustering results achieving cluster recovery levels by method and model.

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 5). 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. 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.

Correlation of ARI_{HA} and Bayesian Classification Internal Clustering Metrics

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 calculated per resulting cluster, or class, two derivative measures were calculated for each metric: (1) the average metric value across clusters, and (2) the maximum metric value across clusters. To evaluate the validity of using one of the Bayesian Classification internal clustering metrics as a proxy for the ARI_{HA} (since ARI_{HA} is unknown for real data) a correlation analysis using the Pearson correlation coefficient was performed for each of those derivative measures with ARI_{HA} . The maximum and the average derivative measures correlated almost perfectly with each other, for each of the internal clustering metrics ($r=1.000$, $p<0.001$ for log of class strength; $r=0.963$, $p<0.001$ for cross class entropy). Although significant, due to large sample sizes, as evaluated by Pearson correlation coefficient, the correlations with ARI_{HA} were not particularly strong (Table 4). The strongest correlation was between average log of class strength and ARI_{HA} ($r=0.584$, $p<0.001$).

Figures 11 and 12 plot ARI_{HA} versus average log of class strength and average cross-class entropy, respectively. As is visually apparent, the relationship between the measures is nonlinear in both cases. For average log of class strength, there are four distinct groupings of values along the scale. After plotting average log of class strength versus ARI_{HA} separately for each number of nonfunctional loci, it became clear that the groupings were related to this noise parameter (Figure 13). The highest grouping of average log of class strength resulted from

datasets with the fewest number of nonfunctional loci (0), and likewise, the lowest average log of class strength values resulted from datasets with the greatest number of nonfunctional loci (100). Still, even after this examination of each grouping on a smaller scale, the relationship between the average log of class strength and ARI_{HA} was not straightforward.

Table 6. Correlations between ARI_{HA} and Bayesian Classification Internal Clustering Metrics. Pearson correlation coefficients (r) and significance (p-value) provided.

Internal Clustering Metric	Hubert-Arabie	Adjusted Rand Index
	r	p Value
Average Log of Class Strength	0.584	< 0.001
Maximum Log of Class Strength	0.582	0.001
Average Cross-Class Entropy	0.024	< 0.001
Maximum Cross-Class Entropy	-0.042	< 0.001

As for the relationship between average cross-class entropy and ARI_{HA} , it is also nonlinear and not particularly strong. There appears to be an effect based on the number of resulting clusters in which most of the nonlinearity is attributable to datasets resulting in 2 clusters (see Figure 14). What is most troublesome is that, other than for very low entropy values for datasets where there were 2 resulting clusters, high scores for average cross-class entropy, which should indicate good clustering, actually correspond to low (poor) ARI_{HA} values. It is this characteristic that makes using cross-class entropy as a proxy for ARI_{HA} ill-advised. Therefore, despite its complex relationship with ARI_{HA} , average log of class strength is a better option for an ARI_{HA} proxy than is average cross-class entropy.

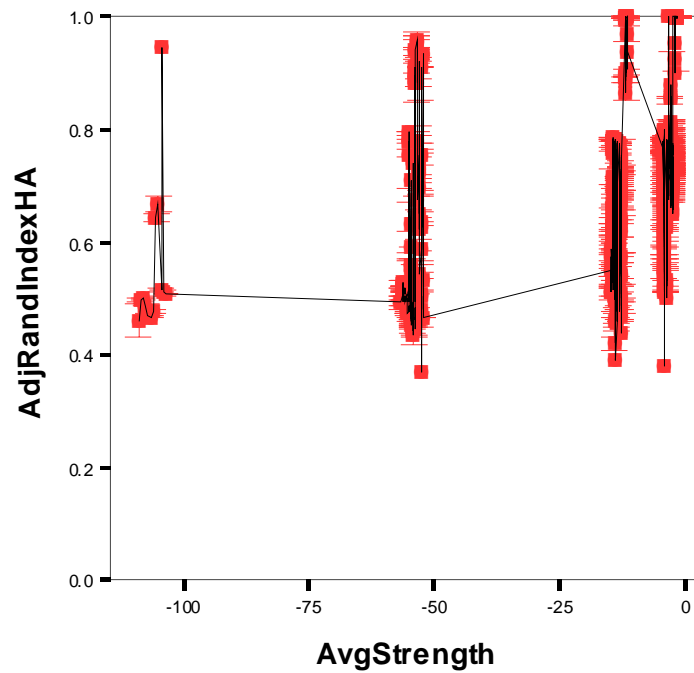


Figure 11. Average Log of Class Strength versus Hubert-Arabie Adjusted Rand Index. Points represent means over 100 datasets per set of simulation parameters. Red bars represent standard error.

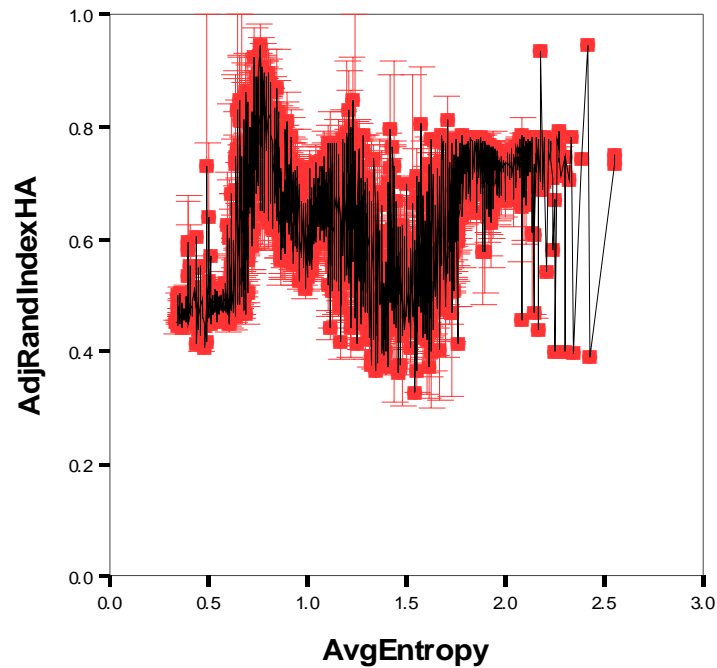


Figure 12. Average Cross-Class Entropy versus Hubert-Arabie Adjusted Rand Index. Points represent means over 100 datasets per set of simulation parameters. Red bars represent standard error.

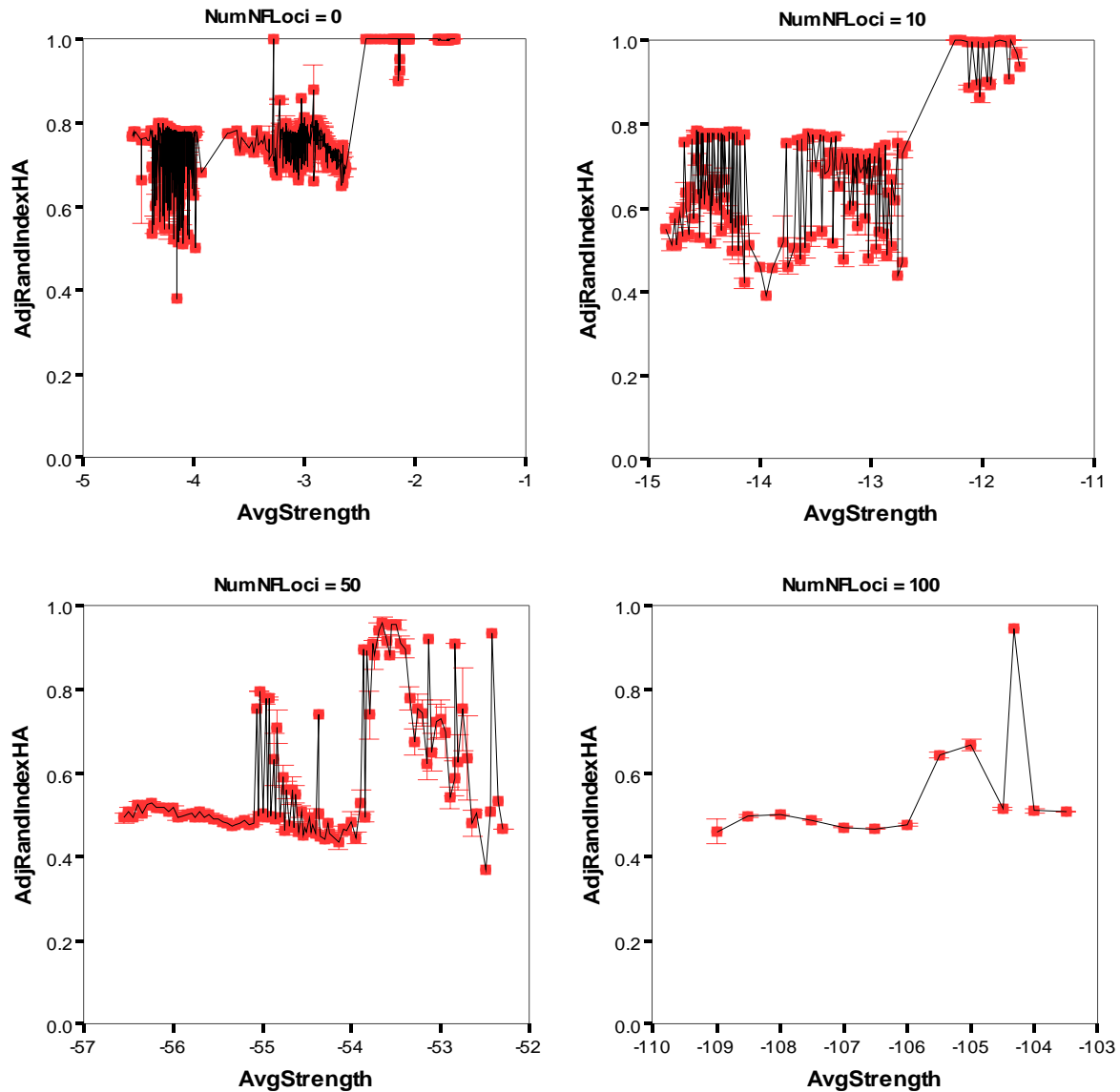


Figure 13. Average Log of Class Strength versus Hubert-Arabie Adjusted Rand Index paneled by Number of Nonfunctional Loci in simulated datasets. Notice different scales in each panel for Average Strength. Points represent means, indicating that the fewer the number of nonfunctional loci, the larger (better) the log of average class strength scores were achieved. Red bars represent standard error.

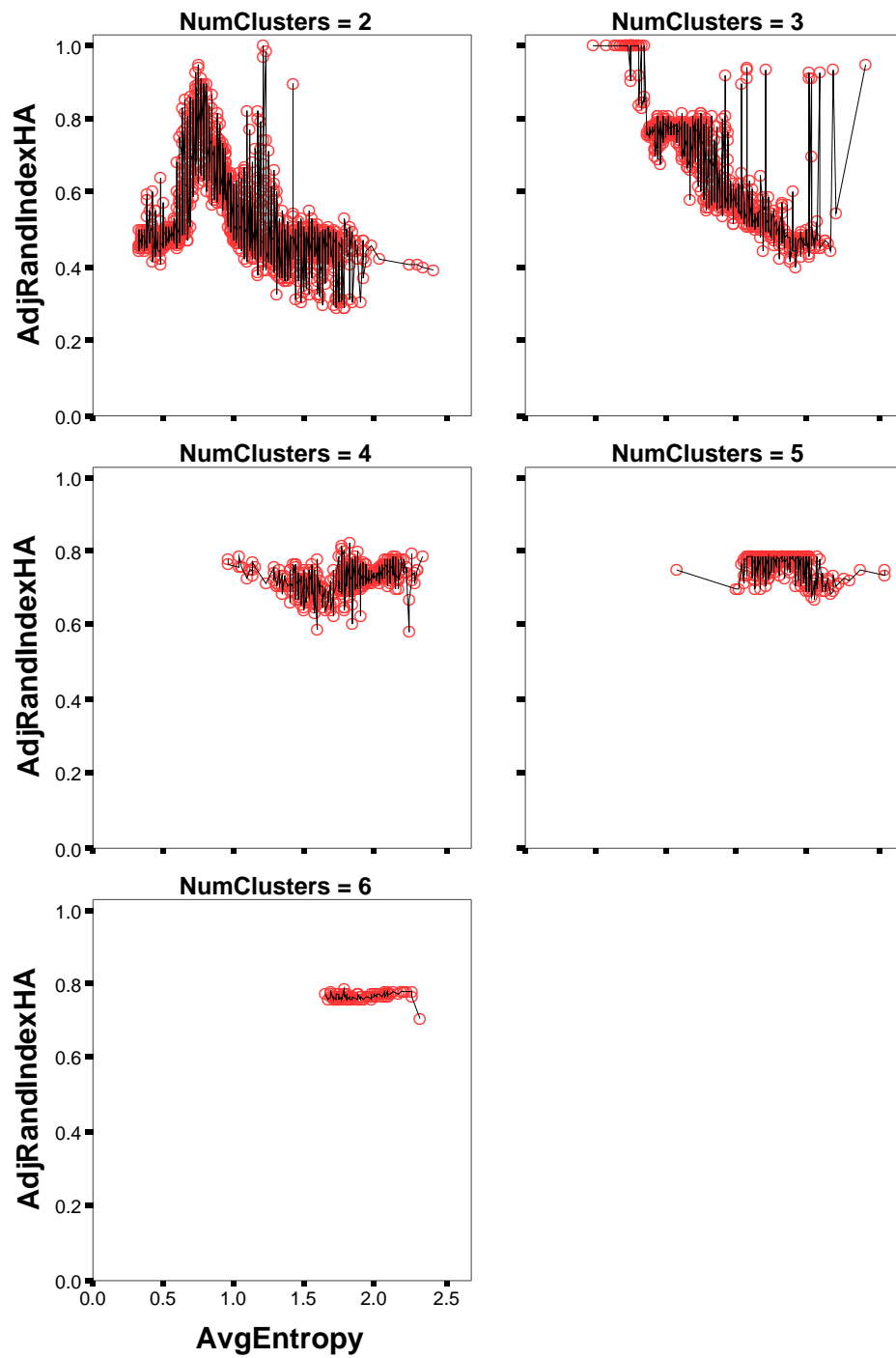


Figure 14. Average cross-class entropy versus Hubert-Arabie Adjusted Rand Index paneled by number of resulting clusters.

Permutation Testing

After choosing average log of class strength as the preferred proxy for ARI_{HA} , permutation testing was performed on the 19,200 simulated datasets, using 100 permuted datasets per simulated data set. The values corresponding to the alpha levels of 0.01, 0.05 and 0.10 for the ARI_{HA} and average log of class strength were used to calculate false positive and false negative rates. A clustering result was considered a false positive if it was ‘called’ significant according to average log of class strength but was not significant according to our gold standard, ARI_{HA} . A clustering result was considered a false negative if it was called not-significant according to average log of class strength but was significant according to ARI_{HA} . Figures 15 and 16 show the false positive and false negative rates, respectively, by alpha level.

The false positive, or Type I, error rate was controlled very well at the one percent level for all three significance levels. The false negative, or Type II, error rate was not well controlled, however. For the most stringent significance level of 0.01, almost half of the clustering results that were significant according to the ARI_{HA} were not called significant according to the average log of class strength. At the least stringent significance level ($\alpha = 0.10$), however, the Type II error rate was 21 percent, which is more acceptable. Other simulation parameters were examined for their impact on the false negative rate, and Figures 17 and 18 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).

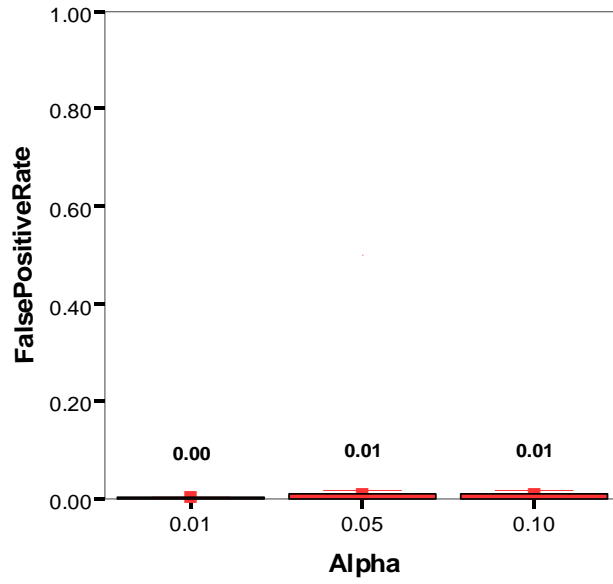


Figure 15. False positive rate by significance level (alpha).

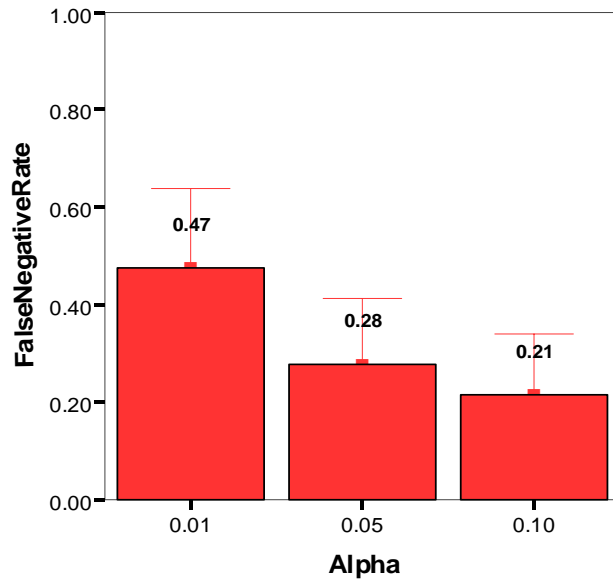


Figure 16. False negative rate by significance level (alpha).

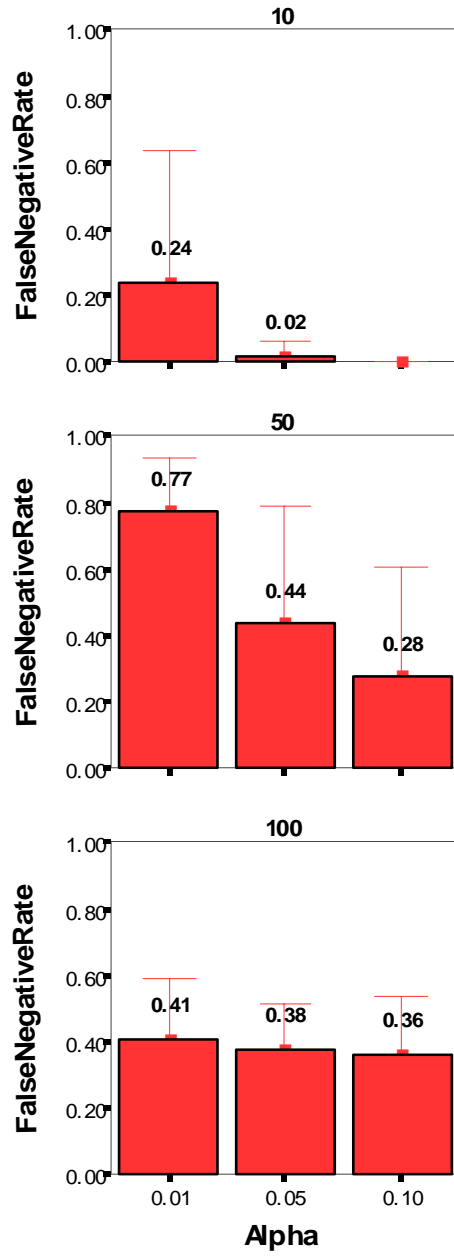


Figure 17. False negative rate by significance (alpha) level, paneled by number of nonfunctional loci.

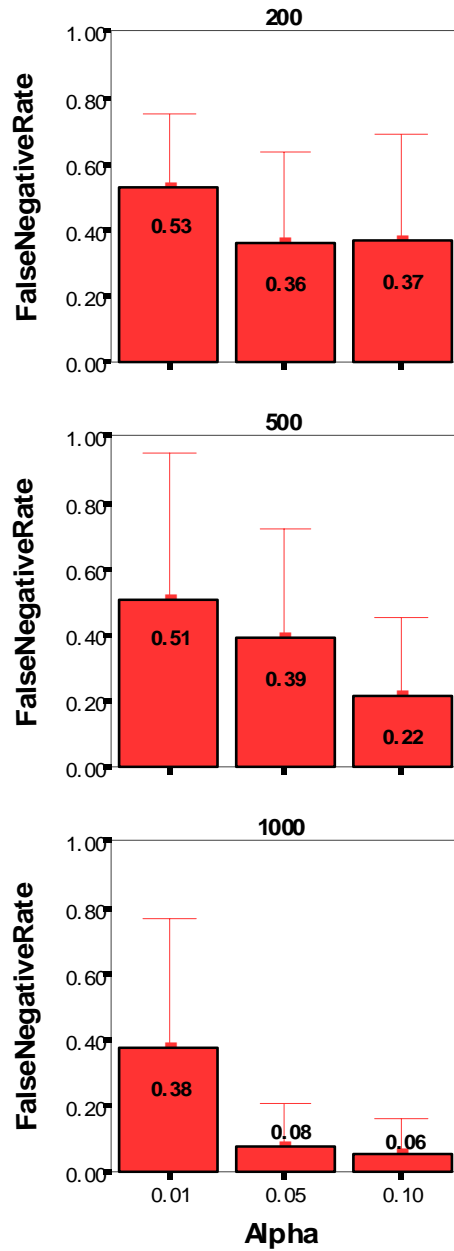


Figure 18. False negative rate by significance (alpha) level, paneled by number of affecteds (sample size).

CHAPTER IV

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 attempts to simulate heterogeneity does 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.

As expected, the simpler the model, the better the performance by the three clustering algorithms, with the exception that the Hypergraph Clustering and Fuzzy k -Modes Clustering methods performed somewhat better (although still achieved poor cluster recovery) 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 was. This novel data simulation algorithm should prove very useful for future studies of other proposed genetic analysis methods for complex diseases.

Method Comparison

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 was found to have achieved excellent recovery for 75% of the datasets with the THO model and achieved moderate recovery for 56% of datasets with 500 or more affecteds and for 86% of datasets with

10 or fewer nonfunctional loci. Neither Hypergraph Clustering nor Fuzzy k -Modes Clustering achieved good or excellent cluster recovery even under a restricted set of conditions.

Bayesian Classification was obtained as closed-source software, for which there were numerous parameters, which could have been tweaked. 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 have yielded 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. The results of this simulation study are perhaps encouraging enough to warrant further investigation of this matter.

It was certainly 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. As it turns out, the Hypergraph Clustering may not have been a good fit for this type of data. The Hypergraph Clustering method was reported to work well with very large variable sets (on the order of thousands), which have complex patterns for which large numbers of clusters (10-20+) were relevant (Han EH *et al.*, 2002). There has been no discussion in the literature about 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, the process devised to translate the resulting partitioning of genotypes into a clustering of individuals was awkward and did not seem to yield the desired result, although it was 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.

Like the Bayesian Classification method, the Hypergraph Clustering method was obtained as closed-source software (LPminer and hMETIS), and it had a number of different parameter settings also. It is possible that different settings would have yielded better results.

For example, the maximum size of an association rule was limited only by the number of variables. Perhaps it would have performed better if we had limited it to five loci, which is the largest number of functional loci in any of the datasets and in real data is perhaps the largest number of interactions one would want to find (and try to interpret). Also, it is possible that the performance could have been improved with a different level of the minimum “support” or frequency required for a pattern to be considered an association rule. The default of five percent was used, but perhaps in combination with lowering the maximum size of an association rule, the minimum support should have been raised to ten or fifteen percent. This may have reduced the number of nonfunctional loci included in the partitioning of genotypes such that the subsequent partitioning of individuals was more satisfactory.

The Fuzzy *k*-Modes Clustering method performed comparably 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. 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 dataset 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 left empty.

Permutation Testing

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. It was discovered that when using the average log of class strength as the internal clustering metric, the false negative rate was alarmingly high for the most stringent significance level of 0.01. However, it was perhaps acceptably low (21 percent) for the less stringent significance levels of 0.05 and 0.10. In addition, the average log of class strength metric controlled the false positive rate very well, at one percent or less for all three significance levels. Thus, if a clustering of data were called significant according to permutation testing using the average log of class strength,

one could be quite confident that the result were real. On the other hand, if the clustering were called statistically insignificant by the same process, using an alpha level of 0.10, there would be a twenty-one percent chance that you were mistakenly rejecting a good clustering of the data. Because these methods are meant as a preprocessing step before applying other statistical or supervised machine learning methods, one might have preferred that the method err on the side of having more false positives than false negatives. One could then further test several of the new hypotheses, feeling relatively confident that few if any potentially true hypotheses had been missed.

However, others might prefer it the way it turned out here. There is indeed a trade-off between the two types of error, and many researchers would be pleased to have a method with a one percent false positive rate, regardless of the twenty-one percent false negative rate. That is because 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. By controlling the false positive rate so well, Bayesian Classification offers a comfortable degree of certainty with regard to the hypotheses that it generates. At least, this is true when the underlying data structure is similar to that simulated under the THO model. It is not known whether data with a substantially different underlying model would lead to different behavior and different false positive and false negative rates. This needs to be further investigated to determine whether Bayesian Classification is robust to a variety of data structure conditions. There are methodological parameters concerning convergence that might be further optimized to better suit a range of data structures. These two areas should be the focus of future studies.

In conclusion, the efficacy of three clustering methodologies at uncovering trait heterogeneity in genotypic data was investigated. One method, Bayesian Classification, was found to perform very well under some conditions (THO model) and to outperform the other methods. Permutation testing confirmed that the method could be used on real data with excellent Type I error control and acceptable Type II control. Further investigation of how different parameter settings may improve the performance of Bayesian Classification is recommended.

REFERENCES

- Anderberg MR. *Cluster Analysis for Applications*. New York: Academic Press, 1973.
- 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. *American Journal of Medical Genetics* 105: 539-547, 2001.
- 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. *Nature Genetics* 32: 237-244, 2002.
- Cheeseman P and Stutz J. Bayesian Classification (AutoClass): Theory and Results. In Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (Eds). *Advances in Knowledge Discovery and Data Mining*, Menlo Park: The AAAI Press, 1996.
- Collinge J, Palmer MS, Dryden AJ. Genetic predisposition to iatrogenic Creutzfeldt-Jakob disease. *Lancet* 337: 1441-1442, 1991.
- 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. *Neuroscience Letters* 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.
- Doh-ura K, Kitamoto T, Sakaki Y, Tateishi J. CJD discrepancy. *Nature* 353: 801-802, 1991.
- 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. *Biochemical and Biophysical Research Communications* 163: 974-979, 1989.
- Duda RO and Hart PE. *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1973.
- Frankel WN and Schork NJ. Who's afraid of epistasis? *Nature Genetics* 14: 371-373, 1996.

- Goldfarb LG, Brown P, Haltia M, Cathala F, McCombie WR, Kovanen J, Cervenakova L, Goldin L, Nieto A, Godec MS, Asher DM, Gajdusek DC. Creutzfeldt-Jakob disease cosegregates with the codon 178Asn PRNP mutation in families of European origin. *Annals of Neurology* 31: 274-281, 1992.
- Good P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer, 2000.
- Han EH, Karypis G, Kumar V, Mobasher B. Clustering Based on Association Rule Hypergraphs. *Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.
- Han EH, Karypis G, Kumar V, Mobasher B. Clustering in High Dimensional Space Using Hypergraph Models. *In* Technical Report, University of Minnesota, Computer Science 97-063, 1997.
- Hanson R, Stutz J, Cheeseman P. Bayesian classification theory. *In* Technical Report, NASA Ames Research Center, Artificial Intelligence Branch FIA-90-12-7-01, 1991.
- Harding AE. The clinical features and classification of the late onset autosomal dominant cerebellar ataxias: a study of 11 families, including descendants of 'the Drew family of Walworth.'. *Brain* 105: 1-28, 1982.
- Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M. Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic Epidemiology* 27: 53-63, 2004.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genetics in Medicine* 4: 45-61, 2002.
- Hoh J, Wille A, Ott J. Trimming, Weighting, and Grouping SNPs in Human Case-Control Association Studies. *Genome Research* 11: 2115-2119, 2001.
- Huang Z and Ng MK. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems* 7: 446-452, 1999.
- Hubert L and Arabie P. Comparing partitions. *Journal of Classification* 2: 193-218, 1985.
- 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. *American Journal of Human Genetics* 116A: 262-267, 2003.
- Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity* 56: 73-82, 2003.
- Ott J. Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* 51: 283-290, 1992.
- Ott J and Hoh J. Set association analysis of SNP case-control and microarray data. *Journal of Computational Biology* 10: 569-574, 2003.
- Owen F, Poulter M, Collinge J, Crow TJ. A codon 129 polymorphism in the PRIP gene. *Nucleic Acids Research* 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.
- 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. *Annals of Human Genetics* 58: 107-127, 1994.
- 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. *Genetic Epidemiology* 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. *American Journal of Human Genetics* 69: 138-147, 2001.
- Rivolta C, Sharon D, DeAngelis MM, Dryja TP. Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. *Human Molecular Genetics* 11: 1219-1227, 2002.
- Rosenberg RN. Autosomal dominant cerebellar phenotypes: the genotype has settled the issue. *Neurology* 45: 1-5, 1995.
- Seno M and Karypis G. LPMIner: An Algorithm for Finding Frequent Itemsets Using Length-Decreasing Support Constraint. *Proceedings of IEEE Conference on Data Mining*, pp. 505-512, 2001.

- Shao Y, Raiford KL, Wolpert CM, Cope HA, Ravan SA, Ashley-Koch AA, Abramson RK, Wright HH, DeLong RG, Gilbert J.R., Cuccaro ML, Pericak-Vance MA. Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *American Journal of Human Genetics* 70: 1058-1061, 2002.
- Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics Supplement* 32: 502-508, 2002.
- Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. *Annals of Human Genetics* 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). *American Journal of Human Genetics* 52: 506-516, 1993.
- Steinley D. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological Methods* 9: 386-396, 2004.
- Tager-Flusberg H and Joseph RM. Identifying neurocognitive phenotypes in autism. *Philosophical Transactions: Biological Sciences* 358: 303-314, 2003.
- Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. *TRENDS in Genetics* 20: 640-647, 2004.
- 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.
- Young J and Povey S. The genetic basis of tuberous sclerosis. *Molecular Medicine Today* 4: 313-319, 1998.