RICE UNIVERSITY

Methods for detecting multi-locus genotype-phenotype association

by

Jeffrey R. Kilpatrick

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Science

Approved, Thesis Committee:

Luay K. Nakhleh, Chair
Assistant Professor of Computer Science

Lydia Kavraki

Lydia E. Kavraki
Professor of Computer Science

Marina Vannucci,
Professor of Statistics

Houston, Texas

November 2009

Abstract

Methods for detecting multi-locus genotype-phenotype association

by

Jeffrey R. Kilpatrick

Solutions to the genotype-phenotype problem seek to identify the set of genetic mutations and interactions between them which modify risk for and severity of a trait of interest. I propose association graph reduction (AGR), a novel algorithm to detect such genetic lesions in genome-wide data, particularly in the presence of high-order interactions. I describe several existing methods and evaluate their performance in terms of computational cost and power to detect associations. An objective comparison of the results shows that AGR successfully combines high power with computational efficiency, while providing a detailed account of interactions present in the data. No other known method combines these three properties. When applied to real data, AGR can be used to discover genetic causes of common diseases such as arthritis, hypertension, diabetes, asthma, and many others, which will facilitate the discovery of novel diagnostic tools and treatment protocols.

Acknowledgments

I thank my advisor Luay Nakhleh for his steadfast support, endless motivation, and great ideas. He is a truly wonderful advisor. His expectations are exceeded only my his encouragement.

I am also appreciative of the other members of my committee, Lydia Kavraki and Marina Vannucci. In addition to their contributions here, I am grateful for their dedication in the classroom. Each exposed me to novel computational approaches to problem solving in engaging manner.

I owe a large debt of gratitude to John B. Harley. He is a wise mentor, excellent role model, and the primary reason I am pursuing a career in science.

I thank my family for their commitment to me and constant support throughout my long academic career.

I am very grateful for the support of the Department of Energy through their Computational Science Graduate Fellowship Program, which is provided under grant number DE-FG02-97ER25308.

Computational support for this work was supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

Most importantly, I thank Emily for her patience, unwavering support, and wonderful ideas. Her contagious creativity inspires me to think differently.

Contents

Illustrations

Tables

Definitions

Allele One possible state of a polymorphic locus. For example, a SNP may have alleles G and T.

Complex trait A phenotype with multi-factorial etiology, often consisting of several genetic and environmental components.

Hardy-Weinberg Equilibrium (HWE) A situation in which the frequencies of alleles and genotypes remain constant in a population over several generations. When in HWE, the frequencies of the alleles A and B for a biallelic locus in a diploid population are expected to be related to that of its genotypes by $\pi_{AA} = \pi_A^2$, $\pi_{AB} = \pi_A \pi_B$, and $\pi_{BB} = \pi_B^2$.

Genotype The set of alleles present at a particular locus. Human genotypes have two alleles, one from each parent.

Linkage disequilibrium (LD) Association of alleles at two loci due to a phenomenon other than random chance.

Locus A heritable quantity that follows Mendel's laws of inheritance.

Marker A locus which can be reliably observed, typically through genotyping.

Penetrance The conditional probability of observing a particular phenotype given a specific factor, such as a genotype.

Phenotype An observable trait, such as eye color, blood pressure, or presence of a particular disease.

Polymorphism A locus found to be in more than one state in a population in appreciable quantities. Typically, a locus is considered polymorphic when its more frequent allele has a population frequency less than 95% [1].

Prevalence The proportion of a population with a particular phenotype.

Single nucleotide polymorphism (SNP) A locus with a single base substitution. Due to their abundance and ease of detection, SNPs are often used as markers in genome-wide association studies.

Chapter 1
Introduction

This thesis introduces association graph reduction (AGR), a novel method to identify genomic regions associated with heritable traits. Complex biological systems make discovery of such effects particularly difficult. As a result, many algorithms have already been developed to tackle this problem, of which I describe several. In contrast to others, AGR uses a greedy strategy to identify both interacting and non-interacting regions in order to discover the structure of genetic etiologies. As a result, AGR is as accurate as any known method and over an order of magnitude faster under most circumstances. When applied to any of the genome-wide data sets that proliferate recent literature, AGR may help elucidate the genetic causes of disease, resulting in better diagnostic tools and treatments.

1.1 Motivation

Much as the 20th century was heralded by many as the Century of Physics, a rapid increase in the rate of development of genomic and other basic biological research has led many to believe that this is the Century of Biology. Thanks to promised advances in our understanding of biological processes and the causes of their dysfunction, we expect medical treatment to become increasingly tailored to individual patients. Additionally, gene therapy may provide means to correct genetic aberrations leading to disease.

To fulfill these promises, scientists have embarked on several projects to provide the genetic infrastructure and technologies on which future research can be conducted. The most widely known of these efforts, the Human Genome Project, has determined the complete sequence of letters making up one subject's genome. The resulting genetic map has catalyzed the development of technologies that have reduced the effort required to search the human genome for mutations that predispose to disease and other traits of interest. In particular, with genotyping microarrays, a small lab may determine the state of over a million genetic markers in a matter of days. Many researchers have taken advantage of these technologies, which has led to the publication of several genome-wide association studies (GWAS) seeking to find the genetic causes of many diseases and other traits.

Unfortunately, these GWAS have proved less useful than one may hope. A review of 600 positive associations, 166 of which have been studied multiple times, found only six consistently replicated effects [2]. Faced with such a terrible result, it is clear that the community of biological researchers must determine and address the difficulties in discovering the mechanisms of heritable traits. Several causes of this failure have been proposed. Among these are the expected weakness of association signals due to the influence of other susceptibility factors, the effects of genomic differences among population strata, the baffling and hard to predict correlation of markers in close proximity, and the inherent complexity of cellular machinery.

1.2 Context

The complexity of cellular mechanics has two principal implications for genomic research. First, interactions between molecular species suggest that many mutations may be involved in determining the fate of an organism with respect to a given trait. Second, the redundancy of cellular processes implies that loss of function of a particular protein may be tolerated, while two or more may not. As a result, individual lesions may display little or no statistical effect when tested individually, though examining the correct combination of interacting markers shows strong association. For these reasons, I believe a methods to detect genetic associations must take into account the effect of several individual markers and interactions among them.

Of course, approaches to finding mutations responsible for phenotypes (observable traits) of interest that take into account multiple markers must face unique challenges. In particular, testing all possible sets of markers for interactions is not only impossibly difficult, but statistically unsound. Consider a modern data set with one million markers and a computer capable of executing $1,000$ tests per second. While it is reasonable to test all pairs of markers, one would have to wait over $5,000$ years to check all sets of three. Even if one could wait that long, any responsible correction to account for the possibility of positive results by random chance would likely negate any real results.

To cope with these and other challenges, novel methods frequently appear in the literature. With very few exceptions, these approaches fall into one of three categories. *Exhaustive* searches apply brute force in an attempt to uncover all interactions up to the largest computable size. In practice, such a strategy is limited to small candidate marker studies and not applicable to genome-wide data sets. *Stochastic* algorithms perform a random investigation of the search space. Some start with a model consisting of a random set of markers and attempt to improve its classification accuracy while others use expensive methods on a small, randomly selected subset of the data. While intriguing, such methods often rely on random chance to select interacting markers and as data sets grow larger, the chances of guessing correctly drop. Finally, *greedy* methods simply make the best choice at a given time based on available information. In general, such a tactic will find many effects, but is likely to miss interactions between markers showing no statistical significance on their own.

For the most part, methods presented in the literature fail in one of three ways. Several, particularly those implementing an exhaustive search, are simply too slow. Most researchers lack the massive computational resources required to deploy these on a genome-wide scale. In spite of optimistic initial publications, some approaches simply do not work. A few, notably those which fit a fixed-size model, do not identify all markers which confer risk. For example, one group of methods attempts to find the single best set of markers of size k for some user-specified value of k. These failures motivate a different approach to the problem of finding genetic causes of interesting traits.

1.3 Association Graph Reduction

To address these shortcomings, I present association graph reduction (AGR), a novel method to detect genotype-phenotype association. AGR is designed to be a hybrid greedy/exhaustive approach based on two ideas. First, the association graph is a data structure which represents everything interesting known about markers and possible interactions among them. Specifically, it is a graph with both edge and vertex weights. Nodes represent non-empty sets of markers and edges possible interactions. After initial construction, this structure is reduced through a series of elementary operations. If the weight of an edge is sufficiently significant, it may be contracted, resulting in the construction of a new vertex to replace the ends of the edge. On the other hand, if such a weight is less than that of its ends, the ends are assumed to not interact and they may therefore be removed from the graph. These steps are repeated in a specific order until the graph is empty, at which time nodes are reported along with their weight for consideration by the user.

AGR operates as a greedy or exhaustive algorithm depending on operation parameters. At each step, the weight of a potential edge or vertex must exceed a minimum significance threshold to be admitted into the graph. Use of an insignificant vertex threshold value will cause all pairwise interactions to be tested. The flexibility gained by the thresholds allows AGR to perform eitheran exhaustive search when the size of a data set allows or a partial greedy search of the set of possible interactions when appropriate.

1.4 Results

I constructed a series of benchmarks based on simulated data sets to determine the performance characteristics of AGR and provide a meaningful comparison with other methods. In addition to AGR, ten algorithms representing several search strategies were evaluated in terms of computational costs and statistical power. Anticipating long execution times by some algorithms, each method was run on small 1,000-marker data sets generated under a variety of genetic models. I then performed similar testing of the three best performing methods using 10,000 markers. In every case in which the generative model predicted some level of single-marker significance without considering interactions, AGR was the fastest and most powerful algorithm tested. When considering a model displaying no such marginal effects, AGR remained powerful.

1.5 Outline

The remainder of this thesis begins with background material in Chapter 2. I then present a review of the literature with details of several algorithms in Chapter 3, followed by a thorough description of association graph reduction in Chapter 4. The methodology used to evaluate the methods is presented in Chapters 5 and 6 on the SimGE simulation package and genetic models used for testing, respectively. Chapter 7 contains complete performance results. I conclude and suggest future work in Chapter 8.

Chapter 2

Background

In recent years, many large-scale genome-wide association studies have been conducted in an attempt to discover the genetic etiology of complex diseases. Examples include efforts to uncover the causes of Alzheimer disease [3], inflammatory bowel disease [4], age-related macular degeneration [5], systemic lupus erythematosus (SLE) [6], cardiac repolarization [7], and seven other common diseases in a set of related cohorts [8]. These impressive studies have been possible due to the collection of large numbers of samples and advances in genotyping technology. With genotyping microarrays, even small laboratories can produce millions of genotypes per day. But masses of data alone cannot solve the genotype-phenotype problem.

There are a number of issues that make the discovery of complex disease etiology a difficult task. If one wishes to consider the effect of multiple causative mutations influencing a phenotype, the problem becomes considerably more difficult. Aside from the obvious computational challenges posed by what is essentially a high-dimensional data mining problem, there are a number of biological and statistical issues that need to be taken into account. Biologically, the discovery of the genetic etiology of common complex diseases is confounded by the complicated mechanisms that underlie cellular processes. Statistically, testing a potentially large number of interactions may require a severe correction for multiple testing to maintain an experiment-wide significance level.

2.1 Biology

The basic description of biological structures and processes is encoded in deoxyribonucleic acid (DNA), which can be regarded as a blueprint for life. The information encoded in the nucleotides of DNA is transcribed into ribonucleic acid (RNA), which may then be translated into proteins which form structures and react to stimuli internal and external to the cell. Many of these proteins can react to perturbations and other events by activating or inhibiting the production of still other proteins. In this framework, life can be seen as a system of molecular processes which are mediated by the production of proteins, their interaction with themselves or other molecules, and their eventual degradation.

A simple example of how a specific sequence of nucleotides can influence a phenotype can be seen in the melanocortin 1 receptor (MC1R) [9]. This protein plays a key role in the determination of skin and hair color in humans: specific variants of MC1R are associated with red hair, fair skin, and freckles [10, 11].

Unfortunately, most phenotypes are not determined by the presence or absence of any single genetic variant. Rather, many traits have each been associated with several distinct genetic loci. For example, even before a large consortium of researchers pooled their data to identify several SNPs associated with SLE [6], over a dozen different mutations were widely accepted to play a role in pathogenesis of the disease [12]. Currently, there are over 20 candidate genes with significant evidence for association with SLE and researchers appear poised to discover still more.

The complexity of this etiology raises some fundamental questions. How do these implicated mutations modify risk for developing disease? Is risk additive in the sense that more genetic lesions confer proportionally more risk? Are there high-level interactions among these genes or their protein products? Unfortunately, there are no clear answers in the case of SLE or many other complex phenotypes, though we can draw from biological design principles to gain some hints.

Cellular mechanisms rely on complex interactions. For example, consider the partner-switching network motif depicted in Figure 2.1. In the absence of external stimuli, anti-σ (hereafter referred to as $A\sigma$) phosphorylates anti-anti-σ (denoted AAσ) and binds any free σ, rendering σ unable to activate transcription of a specific target. Upon receiving an external signal, AAσ is dephosphorylated, causing it to bind Aσ, freeing σ to perform its intended function.

The partner-switching network motif, which is found in multiple stress response pathways of *Bacillus subtilis* [13] is a complicated mechanism which may respond variously to genetic changes, depending on their nature. For example, loss of or sufficiently serious damage to σ might induce the same phenotype as a change to Aσ that rendered it unresponsive to AAσ. In this case, a phenotypically homogeneous population may harbor two distinct genetic lesions. Alternatively, it is not difficult to imagine a combination of less severe mutations in any of these genes that might result in pathway dysfunction, while any individual change may be only mildly detrimental.

Most phenotypes are the result of complex systems such as partner-switching networks and others which may consist of considerably more molecular species, physical interactions, and functional redundancy. Consequently, most common diseases are the result of several distinct genetic lesions. The relative paucity of single genetic

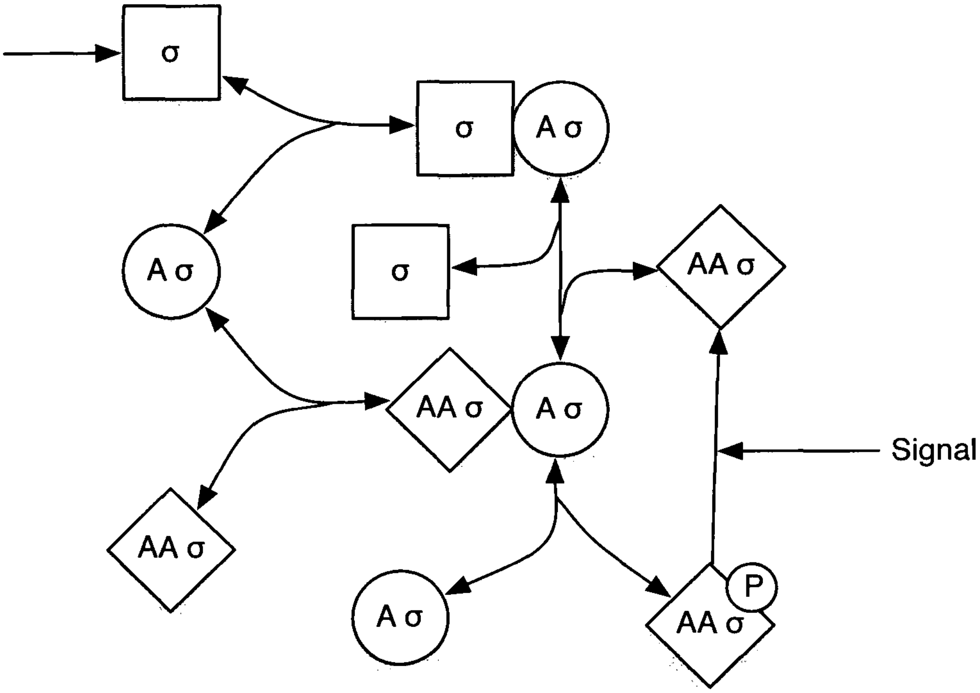

Figure 2.1 : A partner-switching network showing complex molecular interactions (adapted from [13]). σ is a transcription factor, such as the general stress factor σ^{B} or first sporulation-specific factor σ^{F} in the bacterium *Bacillus subtilis*. Depending on the presence of a specific signal, Aσ may bind to σ, rendering it inactive, or AAσ, allowing σ to function.

mutations with serious deleterious effect can also be explained in evolutionary terms. It is much simpler to de-select a single genetic lesion that is only detrimental to an organism. However, a set of many mutations, some of which may be neutral or even beneficial when considered alone, are unlikely to be evolutionarily removed from a population.

Given such reasoning and other arguments put forth by others [14, 15], it is clear that genome-wide studies for association should take interactions into account. Indeed, considerable evidence has been published demonstrating the evidence of *epistasis*, the interaction of genes. The heat shock pathway described above contains several examples of *functional epistasis*, which refers to the physical interaction of proteins [16, 17]. At the genetic level, *compositional epistasis,* in which the effect of one allele is masked by another has been shown in many species [18, 19] and is illustrated in Figure 2.2. Finally, the term *statistical epistasis* was established to describe a deviation from the multiplicative* combination of the effects of two loci on a phenotype. While few studies explicitly test for statistical epistasis, it has been observed in a few circumstances in both model species [20, 21, 22, 23, 24, 25, 26, 27] and humans [28, 29, 24, 30, 31, 27, 32], suggesting the phenomenon is real and possible to detect.

*When he proposed the root term *epistacy*, R.A. Fisher described a deviation from an additive model. A better understanding of population genetics and evolution led to the corrected meaning used today [16].

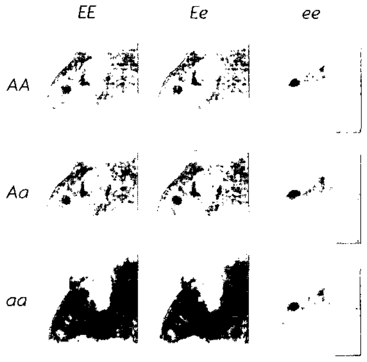

Figure 2.2 : Mouse coat color demonstrating compositional epistasis [16]. Presence of an *A* allele of the *Agouti* gene shown in rows normally results in a light coat. However, the *ee* genotype of *Mc1r* performs a downstream modification in the responsible pathway, resulting in a light coat, irrespective of the state of *Agouti*.

2.1.1 Models of Genetic Phenotypic Etiology

To explain genetic variation in phenotypic etiology, two broad classes of models have come into common use. The first, Mendelian, is the classic formulation proposed by Gregor Mendel and later generalized to support incomplete penetrance and additive risk. For a given biallelic locus, there are three possible genotypic outcomes. A Mendelian model is constructed by assigning the risk of developing a specific trait, conditional on the presence of each genotype. Three simple models are portrayed in Figure 2.3. These basic modes of inheritance—dominant, recessive, and additive—and others describe the effect a single mutation can have on risk for genetic phenotypes.

While such single-locus models remain the basis of most methods for detecting genotype-phenotype association, they do not accurately represent the complexity of the etiologies present in most traits studied today. As indicated in the previous section, many mutations may play a role in the genetic causes of a phenotype. Broadly, such effects may each individually confer a small risk or they may exhibit epistasis.

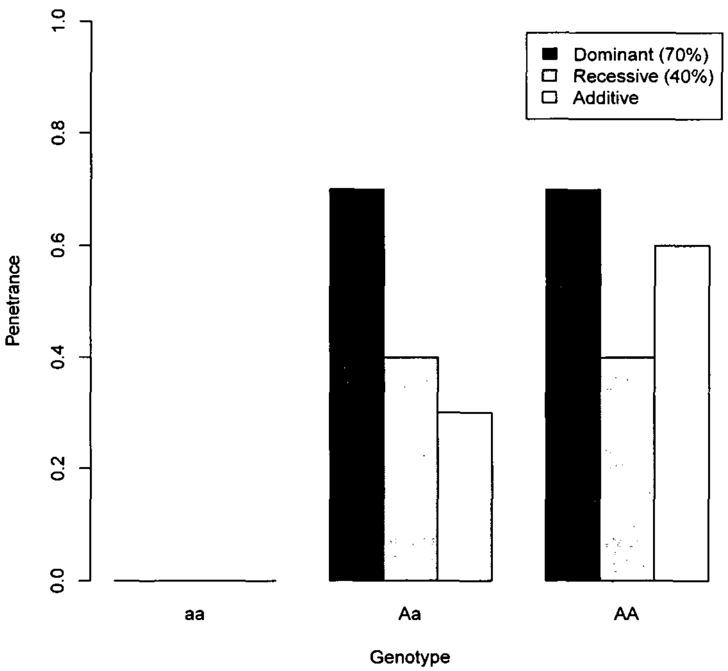

Figure 2.3 : Three examples of Mendelian inheritance. The dominant model (black) requires only one risk allele (A) to achieve maximum penetrance (70%). Recessive models, such as that shown in a dark grey, require the presence of two risk alleles to confer risk (40% in this case). A more general additive model is shown in the light grey; each additional allele increases risk.

Locus heterogeneity describes the effect of several genetic factors individually contributing risk. Such a situation may arise at an individual or population level: several genetic lesions may raise the chance of developing a phenotype in a single person or different sets of mutations might arise in different populations, giving rise to indistinguishable phenotypes. Locus heterogeneity among population strata may confound case-control studies in particular. Stratification is primarily due to ethnic background, though there are exceptions. For example, a homogeneous sub-population of Ashkenazi Jewish women harbors polymorphisms in two distinct genes that cause familial breast cancer [33].

Multi-locus epistatic inheritance can be particularly difficult to detect [34]. While few realistic interaction models predict components displaying no marginal effect [35, 36], we must consider the action of non-linear interactions [37]. Failure to acknowledge the complexity of the systems we wish to understand seems likely to impair our understanding of complex genetic etiology.

2.2 Statistical Considerations

Unfortunately, exploring the space of interactions among loci on a genome scale may prove extraordinarily difficult. The most obvious stumbling block is the computational infeasibility of examining all k-way interactions in a data set with 10^6 markers, as is common in genome-wide association studies. While such analysis is possible with modest computational resources for $k \leq 2$ [38], it is unlikely to become possible for the average researcher to test higher order interactions (Table 2.1).

Even if it were possible to complete such computations, statistical concerns may

Table 2.1 : Time required to exhaustively evaluate all k-combinations of of 1,000,000 features. Figures assume one processor testing 1,000 k-combinations of features per second.

k	k-combinations	CPU Time Required
1	10^6	17 minutes
2	5×10^{11}	16 years
3	1.7×10^{17}	5,280 years
4	4.2×10^{22}	1.3×10^{12} years

prevent its use. In order to avoid discovery of false positives, a correction for executing multiple tests must be used. As the number of tests grows large, such adjustments become increasingly severe, diminishing power to detect real association effects. Even the best choice for type of correction is controversial [39]. The simplest option is Bonferroni correction, which scales the nominal significance level by the number of independent tests conducted. In the context of genomic data, this method is conservative, since nearby loci are correlated. Another popular option is to control the false discovery rate (FDR) [40], possibly by associating a q-value with each hypothesis test [41]. Rather than controlling the proportion of false positives as Bonferroni does, FDR controls the proportion of false positive rejected tests. The q-value is a measure analogous to the p-value which quantifies the significance of each test according to FDR. While the best choice of correction remains an active area of research, FDR appears to be the best option at present [42].

While avoiding false positive results is an important concern, studies must also have sufficient power to detect true associations. More observations are required to detect interactions than main effects. While there is no simple rule defining the correct number of subjects required to adequately power a given association study, it is known that 10-20 cases are required per variable when modelling with logistic regression [43]. Given the existence of genome-sized data sets with over 10^6 markers, one may wonder if it is possible to collect sufficient observations. To compound matters further, modeling interactions based on main effects may have relatively little power [44, 45]. Even standard frequentist variable selection has been ruled out as too expensive and likely to find suboptimal models [46].

With so many confounding factors and claims that various classes of methods will not work, it is no wonder there have been so many statistical and computational approaches developed to detect multi-locus genotype-phenotype associations. Not surprisingly, none of these has been found to be optimal. As we shall see in Chapter 3, many are very similar, following closely related procedures to search for locus heterogeneity and statistical epistasis.

2.3 Summary

The molecular machinery contained within cells is complex. Through evolution, pathways have developed countless physical interactions among proteins and considerable redundancy. As a result, most common phenotypes with genetic etiology are caused by multiple mutations that individually confer risk (locus heterogeneity) or interact to modify risk non-linearly (epistasis). In order to discover the genetic causes of such traits, it is therefore important to consider the effect of multiple loci.

Such consideration presents several challenges. Most importantly, it is infeasible to exhaustively test all k-sets of loci for interaction for $k > 2$ and conducting such experiments is likely to remain out of the reach of most researchers. Even given adequate computational resources, statistical challenges remain. Correction for multiple testing reduces chances of identifying true positive results. Consequently, the sample sizes required to detect interactions by simply evaluating all k-sets of loci in genome-wide data sets may never be collected.

Chapter 3
Methods

There is no shortage of methods to search for statistical epistasis in genome-wide case-control data sets. In a review of the literature, I identified 29 algorithms and frameworks in use, to which I added my own. This count excludes many specializations, tweaks, and general-purpose methods with limited success in the context of genome-wide association studies. To determine the state of the art and put my own work in context, I classified these methods and selected a representative set to explore in depth.

An overview of algorithms to detect genetic interactions is depicted in Figure 3.1. To establish an appropriate scope, algorithms which make use of expert knowledge are excluded.* The methods can be broadly classified into two groups: those that explicitly test every possible interaction up to some size and those that avoid an exhaustive enumeration of the search space. This distinction is particularly pertinent in the presence of increasingly dense microarray products which more thoroughly interrogate the genome.

As the size of available data sets exceeds one million markers, it is clear that a brute-force search for interactions will present computational and statistical difficulties as infeasibility and the need to control for multiple testing are taken into account.

*However, some of the methods discussed below can be adapted to take advantage of biological information.

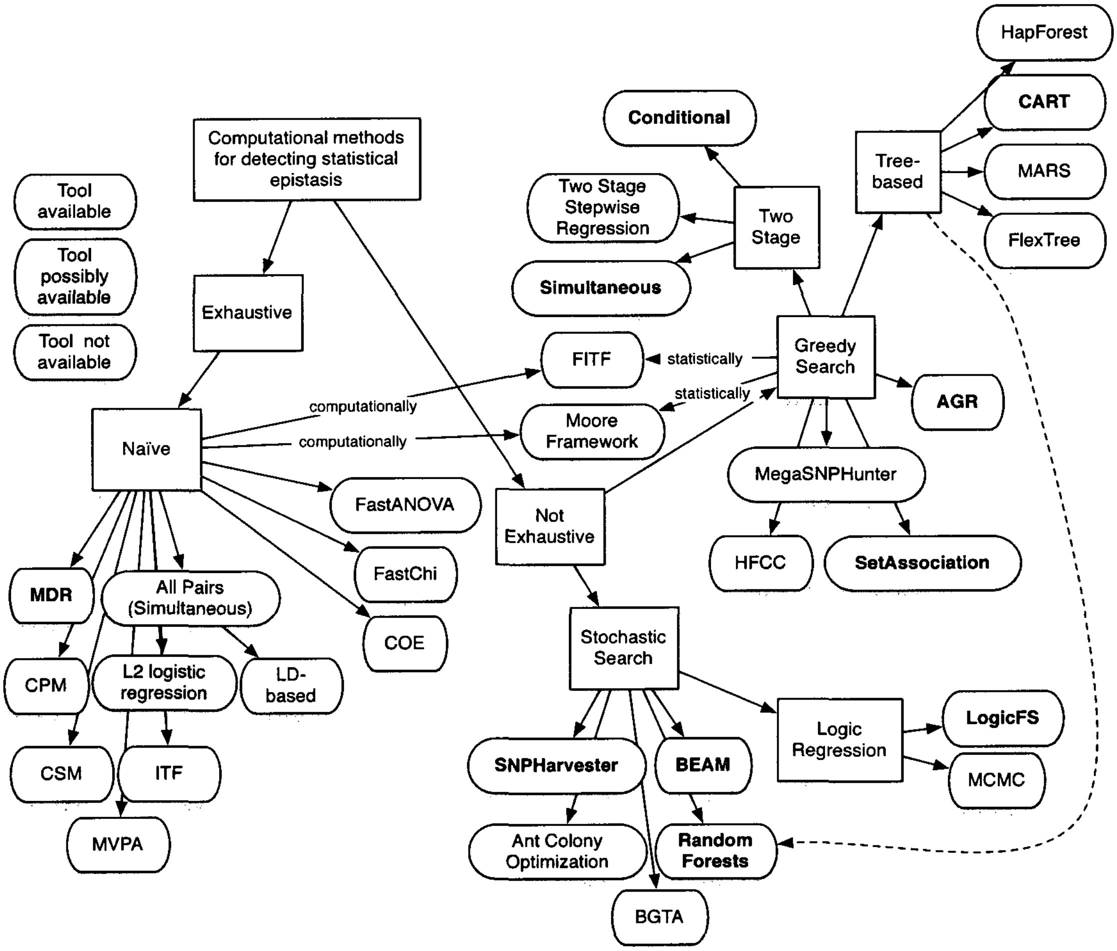

Figure 3.1 : Classification of methods to detect statistical epistasis. Methods with **bold** names are described in detail and evaluated for performance.

Still, incomplete examination of all possibilities brings with it another set of issues. How can we be certain that the interactions we choose to ignore are uninteresting? Do pure interactions really exist in genetics? How many iterations are required for a random search to reach a reasonable conclusion? It is difficult or impossible to satisfactorily answer these questions. As an approximation, I describe eleven approaches below that represent many common interaction search paradigms. In Chapter 7, I will describe the results of a head-to-head comparison of the selected algorithms.

3.1 Exhaustive Algorithms

An *exhaustive* method is one that enumerates all possible k-way interactions for some k in order to identify the effect or effects which best predict phenotypic outcomes. Some methods go even further, testing every possible partitioning of alleles [47]. This exhaustive property leads to their most important characteristic from a computational standpoint: $\Omega(f \cdot n^2)$ running time to investigate pairwise interactions, where f is the time complexity to quantify the association of a pair of markers. While this lower bound is feasible for even the largest data sets available today, generalizations to identify higher-order interactions are unreasonably expensive.

3.1.1 Multifactor Dimensionality Reduction

Introduction Multifactor dimensionality reduction (MDR) is a non-parametric data mining method [48, 49, 50]. It exhaustively searches the space of k-tuple factors (such as markers or discrete environmental influences), constructing a classifier for each combination. Each of these classifiers is tested by ten-fold cross-validation. Its authors have published several papers describing the method, comparing it to others, and applying it to several data sets (citations inline below). As a result of this effort, MDR is currently one of the most widely used methods to detect epistasis, as evidenced by the 167 papers found by PubMed when searching all fields for "multifactor dimensionality reduction."

MDR's basic premise is to reduce the multidimensional search space to a single one-dimensional predictor variable. Multi-locus genotypes are pooled into high- and low-risk categories and the resulting scheme appears to be a single dimension, though others have determined through analysis of deviance that the "effective dimensionality" of the result to be typically much larger than one.

Unlike many less popular methods, MDR has been studied extensively by the original authors and others. For example, it has been determined to be very similar to a naïve Bayes classifier when genotypes in the Bayes classifier are collapsed into aggregate multi-locus genotypes [51]. This is a particularly attractive attribute, as Bayes classifiers have been shown to be optimal with respect to minimizing classification errors [52] and particularly useful when the dimension of the feature space is high [53].

Ritchie *et al.* evaluated the power of MDR to detect gene-gene interactions in the presence of several sources of realistic error. Specifically, they simulated 100 data sets each consisting of 2 functional and 8 non-functional SNPs in 200 cases and 200 controls in six different models with either no noise or one or more of 5% genotyping error, 5% missing data[†], 50% phenocopies, or 50% genetic heterogeneity [54]. They found

[†] While they reported power for data sets with missing data, they fail to disclose how they handled

good ($> 90\%$) power to detect both functional loci in four of the six models, except in the presence of genetic heterogeneity or phenocopies and any other source of error. It is worth noting, however, that the four models for which MDR had good power had relatively high minor allele frequencies ($f_a \geq 0.25$), which implies a relative ease to detect association and/or somewhat unrealistic population prevalence ($\pi > 0.05$ for models 1, 3, and 4). The remaining models had lower causative allele frequency ($f_a = 0.1$) and more realistic population prevalence for a common trait ($\pi = 0.026$ and $\pi = 0.017$, respectively), which may account for MDR's poor performance [55].

Algorithm MDR is a four step algorithm depicted in Figure 3.2, which takes as a parameter k, the size of interaction to test:

1. Select k factors – Choose k input variables to model. In most cases, these will be SNPs, though qualitative environmental variables may be chosen. The only effective restriction is on the size of the selected set, which must be small enough for step 2 to be tractable.

2. Calculate case-control rations for each multi-locus genotype – A table is constructed with a cell for each multi-locus genotype. For example, two biallelic markers with three outcomes each will result in a 3-by-3 table. For each cell, the ratio of cases to controls with the genotypes corresponding to that cell is computed.

absent alleles. Indeed, their current software implementation has no provision for handling missing data and the best strategy for using MDR under these circumstances remains an active area of research.

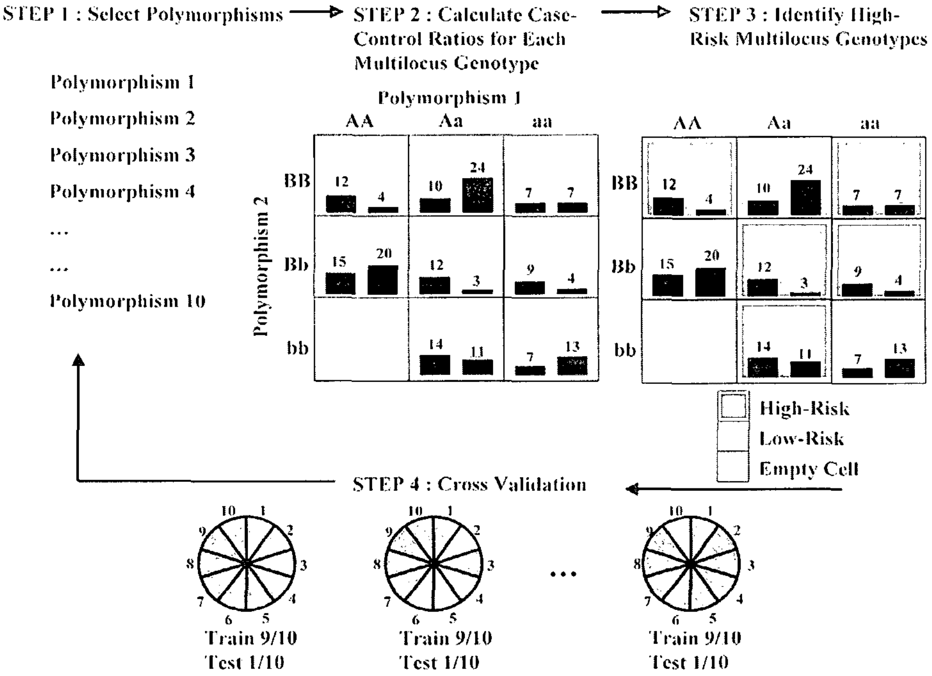

Figure 3.2 : Overview of the MDR algorithm [48]. In step 1, genetic polymorphisms or discrete environmental factors are selected. Next, case-control ratios for each multifactor outcome are computed. In the third step, cells are labeled as high- or low-risk. To conclude, ten-fold cross-validation repeats the first three steps.

3. **Identify high-risk multi-locus genotypes** – For some specified threshold T, label all cells with case-control ratio $R \geq T$ as *high-risk*, those with $R < T$ as *low-risk*, and ignore empty cells. A typical threshold value is $T = 1.0$. The authors argue this step reduces the dimensionality of the input variables by pooling high-risk cells into one group and low-risk cells into another, creating a one-dimensional predictor. This assertion has been called into question by others, who found the "effective dimension" is typically much larger than one [56].

4. **Estimate prediction error by 10-fold cross-validation** – The subjects are randomly divided into 10 equally-sized parts and steps 1-3 executed on on each possible 9/10 of the data. With each resulting model, the prediction error is determined in the remaining 1/10. To reduce the chance of poor estimates of predictor error due to random chance, the 10-fold cross-validation is repeated 10 times and the prediction errors are averaged.

Once the above procedure has been completed for all values of k to be considered (typically $k = 1 \ldots t$, where $t \approx 2$, the largest feasible value of k), the best value for this parameter is selected. The set of factors which minimizes prediction error over all tested values of k is selected as the best supported model. The significance of this result is established through Monte Carlo simulations. The average cross-validation consistency of an empirical distribution derived from 1,000 permutations generated under the null hypothesis of no association is compared to the observed cross-validation consistency. The null hypothesis is rejected when the Monte Carlo p-value is ≤ 0.05.

Summary MDR is a well known and often used method to analyze candidate-gene and other data sets of moderate size for interactions. It has been validated by multiple simulated [48, 54] and real-world data sets [48, 27]. Its strategy of building a single-variable classifier (whether or not one-dimensional) verified by ten-fold cross-validation stands apart from most well-known methods, such as χ^2 tests for trend and logistic regression analysis. A few key advantages and drawbacks are summarized below.

Advantages

- MDR has good power to detect association, assuming reasonable genetic models [54, 55].
- Built-in model validation by way of ten-fold cross validation reduces false positives [57].
- MDR is well characterized and has been evaluated and extended by different research groups [51, 58, 59, 60, 61].
- When coupled with dendrograms and other epistasis visualization tools [62] as they are in the authors' own software, results are easy to interpret.
- The authors provide an easy to use albeit slow implementation of MDR [50, 49].

Disadvantages

- While MDR is asymptotically less expensive than some other methods [47] and further gains may be made with a more efficient implementation, it is likely to remain intractable at the genome-wide scale. Its $\Theta\left(\binom{m}{k}\right)$ complexity implies that most researchers will never have access to the computational resources to evaluate modern data sets with $m \geq 1,000,000$ for $k > 2$. Given the improbability of searching higher dimensions, it is unclear whether MDR represents a substantial improvement over more traditional methods, especially those making use of cross-validation for model verification.

- The power of MDR has been shown to suffer in the presence of locus heterogeneity. Given the large number of susceptibility loci identified in common traits such as SLE by marginal testing [12, 63, 6], such difficult conditions seem likely.

- It is unclear how to cope with missing data. Three methods have been proposed: removal of subjects and/or factors with missing observations; imputing data based on observed frequencies; and imputation based on all available genomic data. The latter approach has been found to be of greatest promise [64], though it is computationally expensive and requires the use of additional third-party software.

3.1.2 All-Pairs Simultaneous Search

Introduction Perhaps the most straightforward strategy to identify interactions between all pairs of loci is to exhaustively test every such combination using a full interaction model. This method was put forth by Marchini *et al.* [38], reviewed by Ionita and Man [65], and implemented in PLINK [66, 67]. In their study of 300,000 simulated SNPs generated under three models of interaction, Marchini *et al.* found this strategy has more power to identify all markers exhibiting pairwise epistasis than a locus-by-locus search, even when Bonferroni correction was applied. Later, Ionita and Man conducted a similar power study in which they added three-locus models and concluded conditional search (Section 3.2.1.1) is likely to be more powerful.

A generalized version of simultaneous search was presented by Millstein *et al.*, which they call the Interaction Testing Framework [68]. Their algorithm considers high-order interactions constructed from significant lower-order interaction terms. In this way, it is not a pure simultaneous search, but rather a hybrid approach.

Hoh and Ott suggest a related approach that searches for loci interacting in cases but not controls or *vice versa* [69]:

1. For every pair of SNPs, construct separate 3×3 contingency tables for cases and controls to compute a χ^2_4 test for trend.
2. Form the ratio $R = c_s/c_l$, where c_s is the smaller χ^2 and c_l the larger.
3. In all pairs with $R > 7.78$ (the 90th percentile of the χ^2_4 distribution), determine significance by $5,000,000$ random R values.
4. Use FDR [40] to establish experiment-wide significance.

A further related simultaneous approach proposed by Yang *et al.* makes use of only cases [70]. They test for independence between genotype frequencies and the phenotype in the population. They conclude that the case-only model is a valid approach requiring fewer cases than the case-control design to detect gene-gene interactions.

Algorithm

1. Evaluate all pairs of loci – For each pair of loci X and Y
   (a) Evaluate the full interaction model $P \sim X + Y + X : Y$, where P is the phenotype under study.
2. Control for multiple testing – Apply Bonferroni correction or FDR [40].

Summary All-pairs simultaneous search is perhaps the most straightforward generalization of the standard locus-by-locus search. Applying familiar regression methods, one must only check all possible pairs of loci. Like many naïve approaches, the simultaneous search comes with its own set of notable problems. First, the number of tests is the square of the number of markers, resulting in a severe correction for multiple testing. While some have found the significance of interaction effects is likely to overcome any such control measures [38], others find non-exhaustive searches to be more powerful in most settings [65]. Second, interactions involving more than two loci that exhibit no main or pairwise effect will be missed. Third, there is no attempt to validate models through cross-validation or similar techniques, which may result in incorrect models. Finally, the method does not penalize complex models with non-zero interaction coefficients, which may result in over-fitted interaction models of heterogeneous effects.

Advantages

- Simple all pairs search may be the fastest exhaustive method available.

- Simultaneous search may have good power to detect association, assuming reasonable models [38].
- The method is implemented in the widely used PLINK package [66, 67].

Disadvantages

- Exhaustive testing of all pairs of loci is computationally expensive.
- Loci involved in higher-order interactions may not be identified in the absence of lower-order effects.
- Lack of validation may result in spurious models.
- Lack of control of model complexity may result in false positive interactions.
- It is unclear how to deal with missing data.

3.1.3 Restricted Search

Introduction A unique approach to effectively exhaustive search has been proposed by Xiang Zhang and adviser Wei Wang with the support of others. Their three methods exploit some property of the test statistic used to mitigate the multiple testing problem that plagues most exhaustive algorithms. Instead, they filter all pairs on a property unrelated to phenotype and evaluate only loci capable of producing significant results. The result is a set of methods with quadratic time complexity that executes only a linear number of tests, hence avoiding severe multiple testing penalties. Unfortunately, their methods were published in proceedings typically ignored by biologists and statisticians, are described in a manner difficult to understand by their potential users, and lack freely available implementations which may impair their adoption in practice.

Their first restricted search approach, FastANOVA realizes their paradigm in the context of a search for pairwise interactions predisposing to a quantitative phenotype using an ANOVA test and permutations to control the family-wise error rate [71]. The method consists of an algorithm to prune the search space by computing an upper bound on the value a particular test statistic can take. The computed value is based on a single-locus test statistic and a permutation-based phenotypic measure independent of phenotype.

Later, Zhang *et al.* developed FastChi, a search for pairwise interactions predisposing to qualitative phenotypes, based on the the Pearson χ^2 test [72] and permutations to control the false positive rate [73]. FastChi is essentially an adaptation of FastANOVA for use with quantitative phenotypes.

In an attempt to solve the problem once and for all, Zhang *et al.* developed COE, the Convex Optimization-based Epistasis detection algorithm, a generalization of FastANOVA and FastChi to all convex test statistics [74]. They demonstrate the convexity of common test statistics, including the χ^2 test, likelihood ratio goodness of fit test, entropy statistics based on mutual information [75, 76], and the Cochran-Armitage test for trend.

Summary Zhang and Wang's restricted search framework is an innovative approach to dampen the severity of multiple testing issues. While it remains computationally expensive with $\Theta(n^2)$ non-phenotypic tests required, their algorithms require only a small number of tests for association. Unfortunately, it seems unlikely these methods will be used in practice due to a lack of appropriate publicity or a freely available implementation.

3.2 Non-Exhaustive Algorithms

If an exhaustive method is one that searches all possible k-way interactions, a *non-exhaustive* algorithm performs a partial search of the possible interaction space to terminate relatively quickly. While they are typically faster than exhaustive procedures, it is impossible to know if any such method will identify or even test the correct solution for any given data set.

Non-exhaustive algorithms can be further classified according to their search space reduction strategy. *Greedy* methods perform filtering based on non-epistatic or lower-order interaction results to filter markers displaying no main or low-order effects. The success of the greedy strategy depends on the nature of interactions present in the data set: pure epistatic interactions displaying no main effects are likely to be missed. *Stochastic* algorithms iteratively select a small number of loci and perform a thorough test for epistasis. This strategy relies on luck to select interacting loci in at least one iteration.

3.2.1 Greedy Search

3.2.1.1 Two-Stage Search

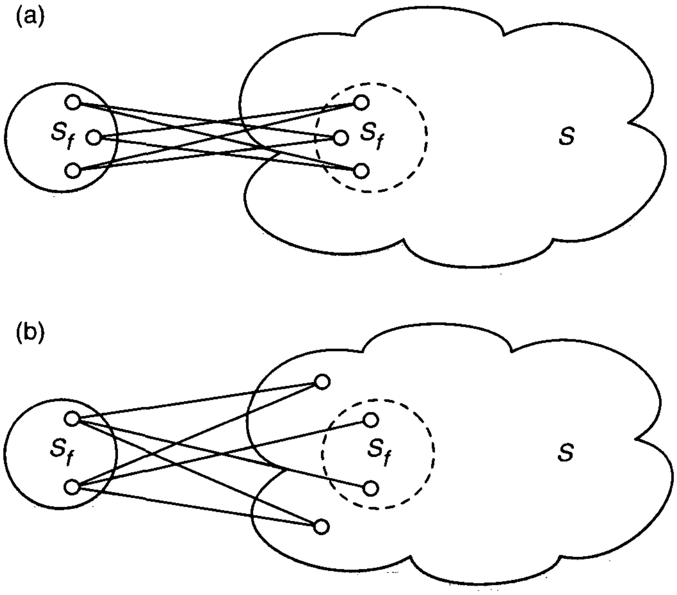

Figure 3.3 : Two-stage search methods. In both procedures, a set of marginally associated loci S_f is identified from the set of all loci under study S. Panel (a) shows a simultaneous search, which tests all possible pairs of loci in S_f. A conditional search, as shown in panel (b), tests interactions between members of S_f and S.

Introduction There are are two very similar procedures to detect statistical epistasis that involve two stages (Figure 3.3). In the first, a simultaneous search (Section 3.1.2) of all loci meeting some low level of significance is conducted. This procedure, as presented by Marchini *et al.* [38], relies on the existence of main effects: in the presence of pure epistasis, it has little power to identify epistatic interactions. The second procedure, as described by Daly and Altshuler [77], identifies the set of loci meeting a more stringent level of significance and examines all possible interactions between members of that set and all other loci. This *conditional* search was evaluated by Ionita and Man, who found that although it is not an exhaustive method, conditional search is likely to have better power to detect association than all-pairs simultaneous search [65]. They cite the penalty imposed by correction for multiple testing as a possible explanation, though Hoh and Ott remind us that the conditional strategy will fail to identify epistatic interactions in the absence of marginal effects, which suggests simultaneous search may still be required [69]. Both methods are implemented in PLINK [66, 67].

Algorithm

1. Define significance levels – Let α_f and α_e be the filtering and experiment-wide significance levels, respectively. If conducting a conditional search, α_f should be stringent. Otherwise, a minimal level such as $\alpha_f = 0.1$ may be used.

2. Stage 1 – Let S_f be the set of loci from all loci S whose significance meets α_f according to a single-marker test for association.

3. Stage 2

   (a) For simultaneous search, evaluate all pairs $\{(x, y) | (x, y) \in S_f \times S_f, x \neq y\}$ for association with the phenotype under study.

   (b) For conditional search, evaluate all pairs $\{(x, y) | (x, y) \in S_f \times S, x \neq y\}$ for association with the phenotype under study.

4. Correct for multiple testing – Apply an appropriate correction for multiple testing, such as Bonferroni or FDR [40]. Reject all pairs (x, y) with adjusted probability $p_{(x,y)} > \alpha_e$.

Summary Two-stage search methods are an intuitive and reasonably powerful method to search for epistasis in genome-wide data sets. Their simplicity and ease of implementation makes them a attractive candidates to screen data. Of course, the lack of conceptual complexity stems from the lack of features available in other methods. For example, there is typically no provision for model validation. Typical two-stage scans lack formal tests of improvement in interactive models over single-locus versions, making it difficult to clearly distinguish between epistasis and heterogeneity. Nevertheless, I suspect two stage scans will remain a dominant approach in analyzing genome-wide association data for the foreseeable future.

Advantages

- Two-stage scans are tractable, with asymptotic bounds $o(n)$ and $O(n)$, corresponding to the extreme cases when $S_f = \emptyset$ and $S_f = S$, respectively.
- The method is conceptually simple and easy to implement.
- Both procedures are implemented in PLINK [66, 67].

Disadvantages

- Two-stage scans do not test higher-order interactions.
- Identification of epistasis in the absence of marginal effects is hindered by an incomplete search of the space of possible interactions.

3.2.1.2 Classification Trees

Introduction Classification trees, such as those constructed by the Classification and Regression Trees (CART) method, are widely used in data mining applications

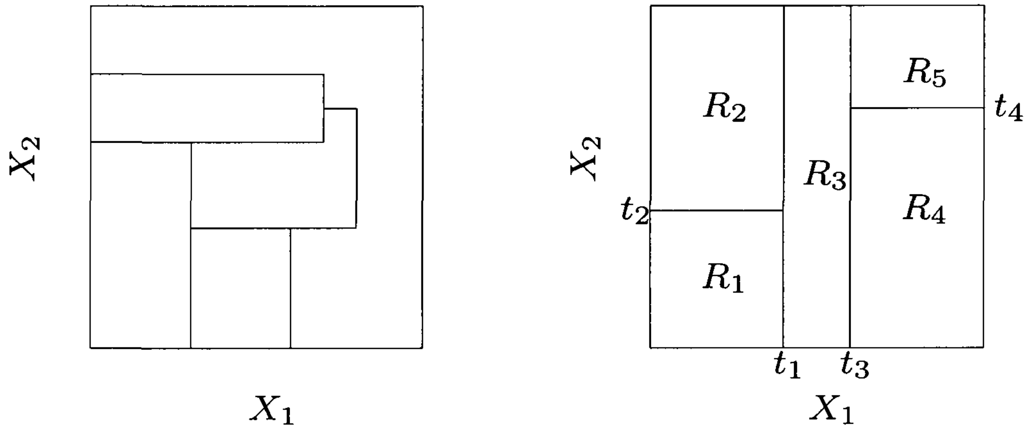

Figure 3.4 : Because classification trees recursively bifurcate their input, they cannot create arbitrary divisions such as those shown on the left. Instead, the result is a "rectangular" partitioning as shown at the right [53].

[78, 53, 79]. In this context, a *tree* is a recursive partitioning of an input on some predictor based on its classification performance. A predictor is typically defined as a single locus, though it could also be a linear combination of variables [80]. Once the initial input is split by the classifier associated with the tree's root, the resulting leaves may be further split on other classifiers. The result is a "rectangular" partitioning of the sample space: general partitioning cannot be achieved with the binary splits provided by classification trees (Figure 3.4).

Zhang and Bonney applied regression trees to genotypic data [82]. They computed splits based on the number of risk alleles at a single locus (for example "none" vs "1 or 2"), as illustrated in Figure 3.5. With Bonney and others, Zheng applied this method to the small genome-wide data sets GAW 9 [81], GAW 12 [83], and GAW 14 [84] provided by the Genetic Analysis Workshop ??. In their study of the GAW 14 data, the authors allowed for multiple deterministic trees to explain different

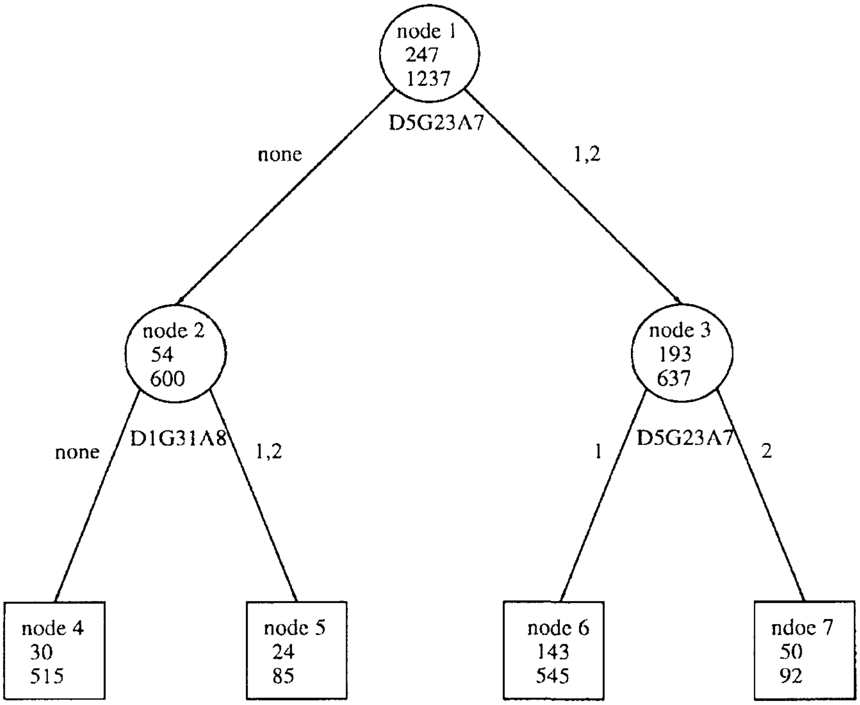

Figure 3.5 : A classification tree from [81]. At each non-leaf node of this particular tree, a decision is made based on the number of risk alleles present at the locus indicated under the node. For example, the root bifurcates the population according to the presence of any risk variants at marker D5G23A7.

pathways to the same outcome. This strategy is similar to deterministic forests [85] and related to random forests (Section 3.2.2.4), though the authors adamantly assert the reproducibility of their data, a feature lacking in random forests.

CART and the related Multivariate Adaptive Regression Splines (MARS) method [86], a generalization of CART intended to improve its regression performance, were used to study gene-gene interaction models for ischemic stroke in 92 polymorphisms in 319 cases and 56 controls [87]. Unlike previous studies, they explicitly use a variable coding to facilitate consideration of additive, dominant, and recessive effects. The authors noted the increased power of MARS to detect interactions in the absence of strong main effects, though both methods identified the same pair of SNPs, which were found to confer additive risk.

The idea of constructing forests grown from estimated haplotypes to detect epistasis was explored by Chen *et al.* [88]. They reported good power to detect association with high specificity across a range of models that included epistasis and locus heterogeneity. In addition, they replicated a known association between a marker and age-related macular degeneration and uncovered a novel result.

Algorithm The most widely used method to construct classification trees is CART [78]. In the context of classification, its goal is to choose a set of splits on predictors that minimize node impurity $Q_m(T)$. Different measures of impurity include [53]:

- Misclassification error: $Q_m(T) = \frac{1}{N_m}\sum_{i\in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$, where each of the regions (leaves) R_m contains N_m observations and $I(y_i \neq k(m))$ is 1 if observation i is in class m and 0 otherwise;

- Gini index: $Q_m(T) = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, where K is the number of classes; and

- Cross-entropy or deviance: $Q_m(T) = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

Like the exhaustive methods presented in Section 3.1, optimal selection of binary splits is computationally infeasible. Instead, CART greedily splits the root on the predictor j which minimizes the specified measure of impurity. The data are then recursively split on the remaining inputs. The algorithm terminates when some criterion such as minimum node size is met. To reduce over-fitting, the resulting tree may be pruned to achieve the desired balance between cost and complexity.

Summary Classification trees, such as those constructed by the popular and freely available [79, 89] CART method, can be constructed quickly and should identify causative loci under a variety of genetic models. They are especially useful in the presence of heterogeneity. Unfortunately, CART relies entirely on marginal signals, which effectively eliminates the possibility of detecting interactions exhibiting no main signals. However, poor performance in the presence of interaction can be improved by bumping, a method which builds CART classifiers based on bootstrap samples and keeps the model with the smallest prediction error [90, 53].

Advantages

- Construction of classification trees requires only linear time.

- Visualization is straightforward and easy to understand.

- Locus and/or population heterogeneity is handled in an elegant manner: different population strata are split at or near the root and fitted by different subtrees [91].
- CART is implemented in the freely R package *tree* [79, 89].

Disadvantages

- Node impurity criterion only considers single predictor variables: interactions without marginal effects are likely to be missed. One possible approach to alleviating this problem is bumping [90].
- Trees have no built-in model validation.
- There is no best method or criterion to control model complexity.
- Trees can be difficult to interpret, as demonstrated by the initial interpretation by Cook *et al.* of a heterogeneous effect as interactive [87].
- It is unclear how to deal with missing data. Imputation is likely required.

3.2.1.3 Set-Association

Introduction The set-association method is a fast algorithm designed to detect loci in the presence of heterogeneity [39, 92]. Unlike many other techniques that offer sophisticated modeling [78, 93, 56] or complex algorithms [94, 95, 96], set-association simply computes the sum of single-locus statistics as a measure of association. Its goal is to combine as much information about the measured loci as possible without

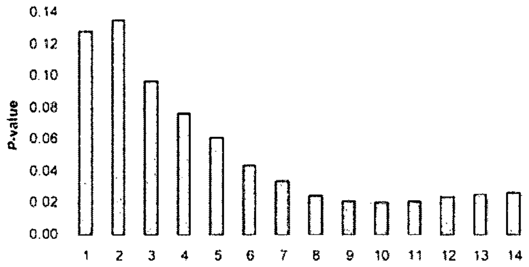

Figure 3.6 : Significance of sum statistics in a real-world sample of 779 heart disease patients in 89 SNPs in 62 candidate genes [39]. Six months after angioplasty, 342 showed restenosis and were considered cases. As the number n of statistics summed increases, maximum significance is reached at a globally minimum p-value.

resorting to interaction testing, in an effort to reduce error associated with multiple testing. While it has extremely limited capacity to detect epistatic interactions in the absence of main effects, it can quickly identify effects comprised of single loci with moderate to high significance, even when the susceptibility mutations they represent do not confer risk in all cases.

Algorithm Set-association is a simple process that attempts to combine information from several measures that have been shown to be indicative of genotype-phenotype association. In particular, it uses a χ^2 test of Hardy-Weinberg disequilibrium (HWD) in controls to filter markers with low genotyping rates and in cases or all samples as a measure of association [97]. This χ^2 is combined with a standard χ^2 test of allelic association in order to increase power over using either separately. The n most significant of these combined statistics are summed to create a summary statistic of all putative susceptibility loci. As n increases, one expects the nominal significance p_n of the sum to decrease to a global minimum before increasing as adding more noise markers are added (see Figure 3.6). The process is outlined in Figure 3.7.

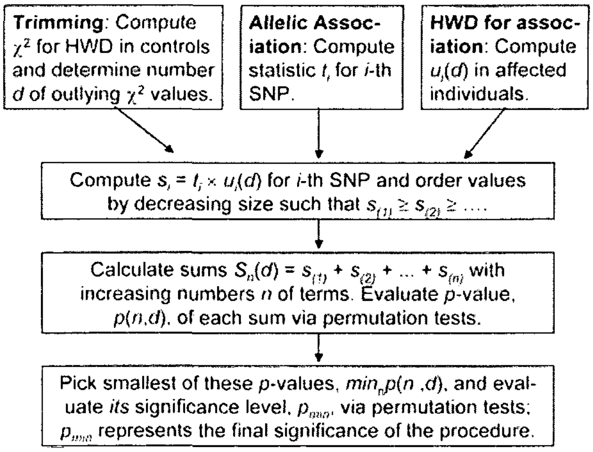
Trimming: Compute χ² for HWD in controls and determine number d of outlying χ² values.
Allelic Association: Compute statistic tᵢ for i-th SNP.
HWD for association: Compute uᵢ(d) in affected individuals.
Compute sᵢ = tᵢ × uᵢ(d) for i-th SNP and order values by decreasing size such that s(1) ≥ s(2) ≥ ....
Calculate sums Sₙ(d) = s(1) + s(2) + ... + s(n) with increasing numbers n of terms. Evaluate p-value, p(n,d), of each sum via permutation tests.
Pick smallest of these p-values, minₙp(n,d), and evaluate its significance level, pₘᵢₙ, via permutation tests; pₘᵢₙ represents the final significance of the procedure.

Figure 3.7 : Overview of set-association algorithm [39].

1. Trim loci – Compute a χ^2_{HWD} for HWD in controls only for each locus. Since high HWD can indicate the presence of genotyping error, we wish to trim outliers. For the d SNPs above the $(1-\alpha)$ percentile, set $\chi^2_{HWD} = 0$. In future χ^2 tests of HWD in cases, controls, or both, set the result to zero.

2. Weight loci – Compute a χ^2_{HWD} for HWD as u_i and χ^2_{AA} as t_i for allelic association for each locus i. If cases are considered affected and controls unaffected, use cases only in the χ^2_{HWD}, otherwise, use all subjects.

3. Compute sums – Set $s_i = t_i \times u_i$ for each locus i for $1 \leq i \leq n$ and sort such that $s_{(1)} \geq s_{(2)} \geq \ldots \geq s_{(n)}$.

4. Determine significance of sums – For each $n = 1, 2, \ldots, N$ for some parameter N, let $S(n) = \sum_{i=1}^{n} s_{(i)}$ and determine the associated p-value p_n by taking a random sample of all possible case-control relabellings and computing $S(n)$ for

each.

5. Select final sum – Select $p_{min} = \min_n p_n$, the p-value of the most significant sum of single-locus statistics.

6. Determine final significance – Permute the samples and re-compute statistics to estimate the significance of p_{min}.

Summary Set-association is a simple method that attempts to increase power through two strategies. First, it minimizes the number of tests executed in an effort to avoid the penalties associated with correcting for false positive results. Second, it extracts association information from each tested marker using multiple statistics that measure different phenomena. The result is an interesting and relatively fast algorithm that merits further exploration. I believe the ideas to combine different measures of association and filtering potentially noisy data by testing for HWD should be more widely deployed.

Advantages

- No explicit interaction testing reduces execution time.
- Combines two distinct measures of association to increase power.
- Loss of significance due to multiple testing is avoided by not testing interactions.
- Missing data handled elegantly by χ^2 statistics.
- Implemented in freely available *sumstat* software [92].

Disadvantages

- Ability to identify interaction effects without marginal is significantly impaired by lack of testing.

3.2.2 Stochastic Search

3.2.2.1 Bayesian Epistasis Association Mapping

Introduction BEAM (Bayesian Epistasis Association Mapping) is a framework developed by Zhang and Liu to perform Bayesian partitioning and compute the posterior probability that a marker or set of markers is associated with a trait under study [98, 99]. By making use of Bayesian statistics, they make possible the incorporation if expert knowledge, such as known information about gene-gene interactions. They further define a novel frequentist test statistic, the B-statistic, to evaluate the significance of a marker or set of markers.

Algorithm

1. Let $I = (I_1, \ldots, I_L)$ be the set of L markers with $I_j \in \{0, 1, 2\}$. $I_j = 0$ denotes marker j is unassociated, $I_j = 1$ means it independently influences phenotype risk, and $I_j = 2$ means it interacts with other loci. Define $P(I)$, the prior probability on I. The authors suggest an uninformative Dirichlet prior.

2. Use the Metropolis-Hastings (MH) MCMC algorithm to approximate the posterior distribution

$$P(I|D, U) \propto P(D_1|I)P(D_2|I)P(D_0|I)P(I),$$

where $D = \{D_0 \bigcup D_1 \bigcup D_2\}$ is the set of case genotypes of unassociated (D_0), marginally associated (D_1), and interacting markers (D_2), and U is the set of control genotypes. To force the modeling of high-order interactions, start with a minimum number of markers in D_2 during burn-in, gradually decreasing this bound to zero. Potential moves are accepted according to the MH ratio. Use one of the following random move types:

(a) randomly change the partition of a marker or

(b) randomly swap the partitions of two markers.

3. For a frequentist interpretation of the posterior, compute the B-statistic [98].

Summary BEAM is a straightforward application of Bayesian statistics to categorize the posterior probability that a marker or set of markers is associated with a phenotype of interest. It uses the popular Metropolis-Hastings algorithm to sample from the posterior, with certain provisions that high-order interactions are explored during burn-in. While it is somewhat slower than most non-exhaustive search methods, BEAM's power to detect associations rivals or exceeds even some exhaustive methods, such as MDR [98].

Advantages

- Bayesian frameworks, such as BEAM, allow for the incorporation of expert knowledge in the prior distribution $P(I)$.

- BEAM retains astonishingly good power with low minor allele frequencies under some models of interaction [98].
- Software implementing BEAM is freely available [99].

Disadvantages

- The Monte Carlo search may not correctly classify interacting markers displaying no main effect, diminishing power to detect pure epistasis.
- Running time is slow relative to other non-exhaustive methods.
- Implementation generates random genotypes based on allele frequencies in cases and controls, rather than more sophisticated and widely accepted multi-locus imputation techniques [99].

3.2.2.2 SNPHarvester

Introduction SNPHarvester is a stochastic search method specifically designed to detect interactions between SNPs affecting trait risk in genome-wide analyses [96]. Conceptually, the algorithm can be viewed as a cross between k-means clustering and Markov chain Monte Carlo: it attempts to find the k SNPs which best explain trait status by repeatedly replacing one of the members of a current "best set" of SNPs. I find SNPHarvester to be one of the most promising methods for detecting statistical epistasis. Unfortunately, it suffers from poor marketing on the part of its authors and an implementation [100] that is likely to be unusable by its target users.

Algorithm The SNPHarvester method consists of two algorithms. *PathSeeker* finds sets of SNPs of fixed size k whose association scores exceed some threshold. The *SNPHarvester* algorithm eliminates eliminates marginally associated loci and runs *PathSeeker* for several values of k. A final post-processing step attempts to filter spurious interactions.

PathSeeker

1. Select k SNPs at random to be in the active set A.

2. For each SNP $s \notin A$, check if the association score of A is improved by removing any member of A and replacing it with s. If so, keep the best combination as a new A.

3. Return all sets A with scores more significant than some specified threshold as the best final set.

SNPHarvester

1. Select all SNPs marginally associated according to a χ^2 test with 2 degrees of freedom following Bonferroni correction. Add these as singleton sets to S, the set of groups of associated SNPs and remove them from further consideration.

2. For each k from 2 to $log_3 N_d - 1$, where N_d is the number of cases with the trait under study,

    (a) $r \leftarrow 0$

(b) While $r < \mathit{SuccessiveRun}$:

i. Run *PathSeeker* with score function $\chi^2_{3^k-1}$.

ii. If *PathSeeker* identified any sets A exceeding the significance threshold, add them to S and set $r \leftarrow 1$. Otherwise, increment r.

iii. Remove from further consideration all SNPs in the best set identified by *PathSeeker* whether or not its score exceeded the threshold.

3. Return S as the set of groups of associated SNPs.

Post Processing

1. For each group of associated SNPs with size k in S, use 3 and 3^j dummy variables to code each SNP and j-way SNP interaction, respectively, for $2 \leq j \leq k$.

2. Fit the L_2 penalized logistic regression model

$$L(\beta_0, \beta, \lambda) = -l(\beta_0,\beta) + \frac{\lambda}{2}\|\beta\|_2^2,$$

where $l(\beta_0, \beta)$ is the binomial log-likelihood and λ a regularization parameter using forward-backward variable selection with Bayesian Information Criterion as a measure of model complexity and cross-validation to select λ (see [53] for details and [56] for a straightforward application of L_2 penalized regression to genotypic data).

3. Report the epistatic interactions selected by the regression.

Summary SNPHarvester is an intriguing multi-step randomized algorithm designed to detect statistical epistasis in genome-wide SNP data sets. It reduces runtime complexity by initially filtering SNPs with marginal effects and later removing those found to interact. It further eliminates tests by randomly and incompletely searching the interaction space.

Searching the space of non-significant markers k SNPs at a time is an interesting notion. It appears the authors intended to search for up to k-way interactions at each step. However, beginning $k = 2$ and incrementally increasing model complexity eliminates the chance of finding larger interactions involving fewer than k SNPs found to be significant at an earlier stage. Since these simpler models are not discovered for use in the post-processing stage, real interactions involving several loci are likely to be missed. Given the implausibility of interaction models displaying significance only when many members are taken into account, running SNPHarvester with $k > 2$ seems unlikely to yield positive results. Further, the random nature of the algorithm that makes the algorithm applicable to large genome-scale data sets quickly becomes a liability as the number of SNPs grows: as the number of possible interactions grows, the probability that enough members are randomly selected by *PathHarvester* shrinks.

The authors of SNPHarvester acknowledge the integration of expert knowledge may improve the performance of their method. A straightforward extension would be to bias SNP selection in the first step of *PathSeeker* according to known biological interactions.

Advantages

- Complexity is effectively linear. *SNPHarvester* will continue to call the linear-time *PathSeeker* while significant results are identified and then terminate with a constant number of unproductive calls. In practice, it is reasonable to expect that most SNPs in a genome-wide study are not associated and do not interact[‡], which implies an approximately linear running time.

- Removing SNPs with significant marginal effects for the straightforward detection of epistatic interactions.

- Modeling by L_2 penalized logistic regression creates easy to interpret results.

Disadvantages

- Removing SNPs with significant marginal effects from further consideration limits the possibility of identifying spectacular results amplified by epistatic interactions.

- The method's random nature requires sufficient luck to select SNPs in epistasis the initial active set A in an invocation of *PathSeeker*. While repeated starts mitigate this problem, it is unclear whether the reduced running time facilitated by the random nature of SNPHarvester provides sufficient reasons to abandon an all-pairs search of SNPs without marginal effects.

- Its difficult to use implementation [100] may impede widespread adoption by non-programmers.

[‡] In fact, this assertion forms the basis of diverse methods, including [101] and [102].

- Algorithm does not model interactions due to linkage disequilibrium caused by locus proximity.

3.2.2.3 Logic Regression

Introduction Logic regression is a framework that builds classifiers by identifying predictive combinations of binary features, typically moving through the model space with simulated annealing [93]. Its goal is to identify additive and non-linear interactions between observed variables to predict class. As such, it appears to be ideally suited to the task of selecting SNP alleles associated with disease status.

In the logic regression framework, predictive models are represented as binary trees in which leaves correspond to predictor variables and other nodes represent AND or OR operations. For example, the Boolean expression

$$(X_1 \wedge X_2^c) \wedge [(X_3 \wedge X_4) \vee (X_5 \wedge (X_3^c \vee X_6))]$$

can be represented by the tree depicted in Figure 3.8. Once in the tree form, a may be changed by one of six atomic operations, which are illustrated in Figure 3.9.

Two studies have constructed different genetic association search strategies based on logic regression. The first, by Kooperberg and Ruczinksi, explores the interaction space by Markov chain Monte Carlo [103, 104], while the other, which its authors Schwender and Ickstadt call logicFS, performs bootstrapping the samples [105, 106]. As the latter method was designed for use with genome-wide SNP and is hence deserving of more focus in this review, I focus on logicFS.

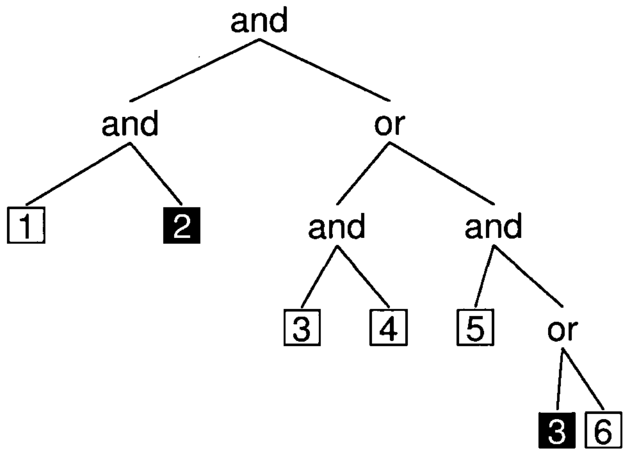

Figure 3.8 : Tree representation of the Boolean expression $(X_1 \wedge X_2^c) \wedge [(X_3 \wedge X_4) \vee (X_5 \wedge (X_3^c \vee X_6))]$ [93]. The number in each leaf indicates the index of the variable it represents. Nodes with dark background indicate the complement of the corresponding variable.

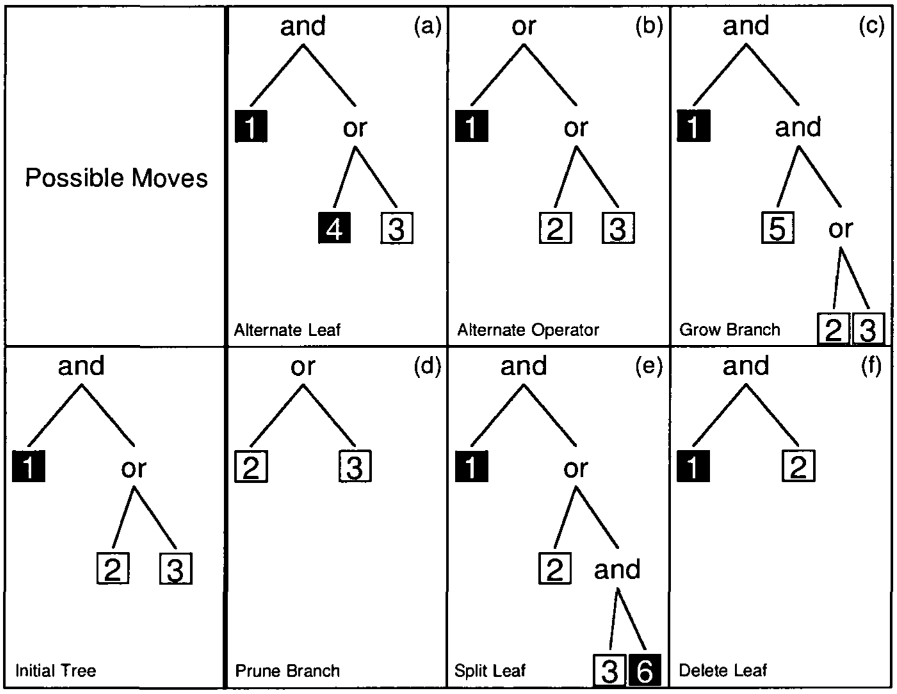

Figure 3.9 : Permissible moves on a logic regression tree [93]. The initial tree is in the lower-left panel and the results of applying the moves are shown in panels *(a)-(f)*.

Algorithm At its core, logicFS follows a familiar strategy similar to that used by classification trees (Section 3.2.1.2) and random forests (Section 3.2.2.4): construct a tree or set of trees representing Boolean expressions that explain as much variability in the outcome as possible. Apart from the interpretation of the resulting trees and a different set of moves with which the model space can be explored, the distinction between logicFS and other tree-based methods is its use of simulated annealing, which prevents convergence on local maxima.

logicFS for SNPs

1. Recode the biallelic input data such that each locus i is represented by two variables. S_{i1} indicates whether the subject is heterozygous for the less frequent variant and S_{i2} indicates homozygosity for the minor allele.

2. Identify a bootstrap sample of size n from the data set consisting of n observations.

3. Using simulated annealing, move through the search space using the tree moves shown in Figure 3.9 to identify the tree (or set of trees) that best explains the cases. When using multiple trees, logic expressions L_i $i = 1, \ldots, p$ can be constructed and combined using the generalized linear model

$$g(E(Y)) = \beta_o + \sum_{i=1}^{p} \beta_i L_i$$

   where Y indicates response, β_i are parameters, and $g(\cdot)$ is a link function. In case-control studies, $g(\cdot)$ is the logit function.

4. Convert the identified trees into Boolean expressions in disjunctive normal form. Such a transformation eliminates ambiguity among the expressions identified in each bootstrap sample.

5. Repeat steps 1-4 B times.

6. Determine the significance of the resulting tree(s) using one of the following importance measures. In both cases, the variable importance measure VIM indicates either significance of a model ($VIM > 0$), insignificance ($VIM = 0$), or the that it is obtrusive to correct classification ($VIM < 0$).

   (a) Variable importance of a single logic regression tree is defined as

   $$VIM_{single} = \frac{1}{B}\left(\sum_{b:P\in L_b}(N_b - N_b^-) + \sum_{b:P\notin L_b}(N_b^+ - N_b)\right)$$

   where L_b is the set of loci and interactions identified in the bth iteration for $b = 1 \ldots B$, N_b is the number of out of observations in the bth iteration that are classified by the regression model, and $N_b^-/N_b^+$is the number of out of bag observations correctly classified by the bth model after the locus or interaction P has been removed from/added to the model.

   (b) Variable importance of a logic regression forest is defined as

   $$VIM_{Multiple} = \frac{1}{B}\sum_{b=1}^{B}(N_b - N_b^*) = \frac{1}{B}\sum_{b:P\in L_b}(N_b - N_b^*)$$

   where N_b^* is the number of correctly classified out of bag observations.

Summary Logic regression methods, such as logicFS, attempt to identify one or more Boolean expressions that best explain case-control status. logicFS explores the search space by a combination of greedy moves and, with small probability, random mutation of the tree(s) representing the current best tree or forest. Bootstrapping is used to avoid over-fitting and important loci and interactions are identified through variable importance measures.

Advantages

- Greedy search of the interaction space is fast.
- Interpretation of resulting Boolean expressions is straightforward.
- logicFS and MCMC logic regression are implemented in the freely available R packages *logicFS* and *LogicReg,* respectively [106, 104, 89].

Disadvantages

- Identification of interactions displaying no main effects requires random selection of correct loci during a rare random step, which diminishes the method's power to detect such effects.
- Splitting of input loci into two variables may be excessive. Dominant and/or recessive coding might suffice (See Section 7.1.9).
- Interpretation of nominal importance measures is unclear.
- It is unclear how to cope with missing data.

3.2.2.4 Random Forests

Introduction A random forest, as conceived by Leo Breiman, is an ensemble of classification trees grown on a small set of input variables selected with replacement from the set of all features [107, 108, 109] (see also Section 3.2.1.2). After construction of the forest, typically using CART, the class of a testing sample is determined by majority vote. The result is an easy to implement extension to standard tree classifiers that have a theoretically better chance of detecting interactions whose components display no marginal effects.

Lunetta *et al.* assessed the ability of random forests to detect interactions displaying no significant main effects in simulated data. In the most comprehensive search of the interaction model space I am aware of, the authors generated data sets containing up to 32 loci in complex models displaying locus heterogeneity and epistasis. They determined the ability of random forests to correctly rank causative loci as highly significant when compared to "noise loci" generated under the assumption of no association. As compared to Fisher's Exact Test, random forests have significantly improved performance in nearly every observed case, especially as the number of SNPs tested grows large.

Algorithm There are many ways to adapt random forests to a problem of interest and the choices made can improve or impair the method's performance [107]. Here, I outline the algorithm as it is implemented in the RandomJungle package, a fast, multi-threaded implementation of random forests [109]. Note that other options that may change the algorithm or its results are available.

1. For a chosen value t, which defaults to 500 in RandomJungle, construct t trees:
   (a) Select n cases from the training set containing n cases with replacement.
   (b) Select m features at random on which to compute splits in the tree. By default, RandomJungle selects the square root of the number of available features.
   (c) Grow the tree as large as possible using CART.
2. Compute the importance of each variable according to its intrinsic importance (i.e., Gini index) or permutation importance.

Summary Random forests are a straightforward extension to tree classifiers in which the input data are bootstrapped and predictor variables are chosen at random. They share some properties with bagging [110] and boosting [111, 112, 113] and may, along with bumping [90], be viewed as an extension of CART that can detect non-linear interactions between features with a chance of identifying interactions among partners displaying little marginal effects.

Advantages

- Random forests are fast to construct.
- Pure epistasis might be detected.
- Random forests are implemented in the excellent RandomJungle package [109], which allows for the specification of many options, supports several file types, and makes use of threads to reduce execution time.

Disadvantages

- The ability to detect interactions between features displaying little or no marginal effect depends on their selection at random.
- The random nature of this method implies that results may not be reproducible [85].
- Consensus vote eliminates chance to meaningfully recover model, giving only a list of putative susceptibility loci without an indication of their relationships.

3.3 Summary

Algorithms to detect multi-locus genotype-phenotype association may be classified as exhaustive or non-exhaustive. Members of the former category test all possible interactions up to a user-specified size, while the others use a greedy heuristic or stochastic search strategy to quickly identify causative loci. There are several broad strategies to conduct an efficient partial search of the space of partial interactions, many of which are outlined in Figure 3.1. While it is beyond the present scope to describe and test all published methods, I have selected a set of ten which represent most search strategies.

In this chapter, I reviewed several existing methods to detect multi-locus association. Each is described in the context of other methods, presented as an algorithm, and several advantages and disadvantages enumerated. These, along with a novel method, are evaluated for statistical performance and computational efficiency in Chapter 7.

Chapter 4

Association Graph Reduction

Association graph reduction (AGR) is a tunable search framework that, depending on the parameters specified, bears resemblance to all-pairs simultaneous search (Section 3.1.2), two-stage conditional search (Section 3.2.1.1), or the Interaction Testing Framework [68]. The method explores possible locus interactions by greedily performing a series of operations on a graph structure that represents current knowledge about genotype-phenotype associations and potential epistatic interactions. It is a flexible, fast, and extremely powerful framework for the detection of both locus heterogeneity and statistical epistasis in the presence or absence of main effects.

Definitions Let $A = (V, w_v, E, w_e)$ be an association graph (AG) with vertex set V and edge set E, with vertex- and edge-weight functions $w_v : V \longrightarrow \mathbb{R}^+$ and $w_e : V \times V \longrightarrow \mathbb{R}^+$, respectively. Define G to be the set of genotype vectors and P the phenotype vector in the case-control cohort under study. An AG represents everything relevant known about an instance of the genotype-phenotype problem, including the significance of possible interactions. To accomplish this goal, each weighted vertex $v \in V$ represents a non-empty set of loci whose weight quantifies the strength of association between the loci it represents and the phenotype vector; a vertex set containing more than one locus represents an interaction between its members. Similarly, a weighted edge $v_1v_2 \in E$ represents a potential interaction between its ends: if v_1v_2 were *contracted* it would be replaced by a new combined vertex $v_{1,2}$ representing the union of the sets of loci in v_1 and v_2. Such an action would reflect the belief that v_1 and v_2 take part in an epistatic interaction.

Weights To guide the search for susceptibility loci and interactions between them, define weight measures $w_v(v_i) = \phi_v(G_i; P)$ and either

$$w_e(v_i, v_j) = \phi_e(\mathit{combine}(G_i; G_j); P) \tag{4.1}$$

or

$$w_e(v_i, v_j) = \phi_v(G_i, G_j), \tag{4.2}$$

where ϕ_v and ϕ_e are measures of association such as a χ^2 test for trend and $G_i \subseteq G$ is a vector multi-locus genotypes for the loci represented by vertex i. The helper function *combine* takes the genotypes of two loci or sets of loci and creates a new sequence of multi-locus genotype identifiers. For example, combining two biallelic makers with three genotypic outcomes each will result in a sequence comprised of nine multi-locus alleles. Similarly, combining a pair of biallelic markers with a third results in a 27-outcome sequence. Note that the association functions ϕ_v and ϕ_e need not be the same, though comparison of their results must be well-defined.

Two types of edge weights are available. The first simply defines the weight of an edge to be the strength of association between the combination of its ends and the phenotype (4.1). In this case, evaluation of an edge is equivalent to testing its ends for interaction in a manner similar to an all-pairs simultaneous search (Section 3.1.2). An alternative edge weight ignores the phenotype vector and instead quantifies similarity between the loci represented by the ends (4.2). This definition tests for locus interaction in a manner similar to one suggested by Hoh and Ott [69]. Since P is not used in the test, it can be argued that edge calculations should not be counted when applying correction for multiple testing (as seen in Section 3.1.3), resulting in a less severe penalty. Such a measure may be carried out in cases alone or in all subjects.

In addition to the requirement of weight comparability, there is one further constraint on the choice of an appropriate association functions ϕ_v and ϕ_e: they must correct appropriately for increased degrees of freedom as edges are contracted and vertices combined. For example, to test some number of loci with n possible multi-locus genotypes, a test statistic T with $T \sim \chi^2_{n-1}$ would be a reasonable selection. On the other hand, standard mutual information

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

would be inappropriate, since I will increase with the number of possible multi-locus genotypes, regardless of their strength of association. Mutual information may be adapted, however, by applying a normalization. One such corrected measure is

$$\mathrm{NMI}(X;Y) = \frac{\mathrm{I(X;Y)}}{H(X)H(Y)},$$

where $H(X) = E(I(X)) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$ is the entropy of variable X.

For any choice of weight functions, the relation "more significant than" must also be defined. Such a comparison acts as a means of model regularization, allowing the user to penalize complex models. In the case of χ^2 weights, let

$$f(e,v) = \frac{\chi^2_e/\nu_e}{\chi^2_v/\nu_v} \sim F(\nu_e,\nu_v)$$

be a test of "significant" improvement where χ^2_e and χ^2_v are the results of χ^2 tests on an edge e and vertex v with ν_e and ν_v degrees of freedom, respectively. Let $e \gg v$ indicate that edge e is more significant than v for the purposes of graph reduction if $\Pr(f) < \alpha_f$, a parameter specific to χ^2 weights. By contrast, let $x > y$ indicate that object x has more significant weight than y without regard to the magnitude of the difference.

Thresholds Beyond weight types, two further parameterizations are available. To perform a greedy search of possible interactions, we may wish to establish a minimum level of significance for inclusion of vertices and edges in the graph. Let α_v and α_e be the vertex and edge thresholds, respectively. No vertex v with $w_v(v) \not> \alpha_v$ nor any edge e with $w_e(e) \not> \alpha_e$ may be included in the graph.

4.1 Algorithm

Once suitable weight functions and other parameters have been selected, AGR analysis proceeds in two phases: construction and reduction of an AG. Construction is effectively a two-stage simultaneous search (Section 3.2.1.1) of all loci meeting some threshold α_v. Through a series of modifications of an AG, reduction explores the space of possible locus combinations to approximately identify susceptibility loci and interactions. Two possible moves are defined for reduction: vertex removal and edge contraction (Figure 4.1). At each step, the most significant object in the graph is selected and modified according to one of the moves until the graph is empty or a significant edge representing a weak interaction is selected. Note that the latter termination criterion is only possible when using the genotype similarity weight defined

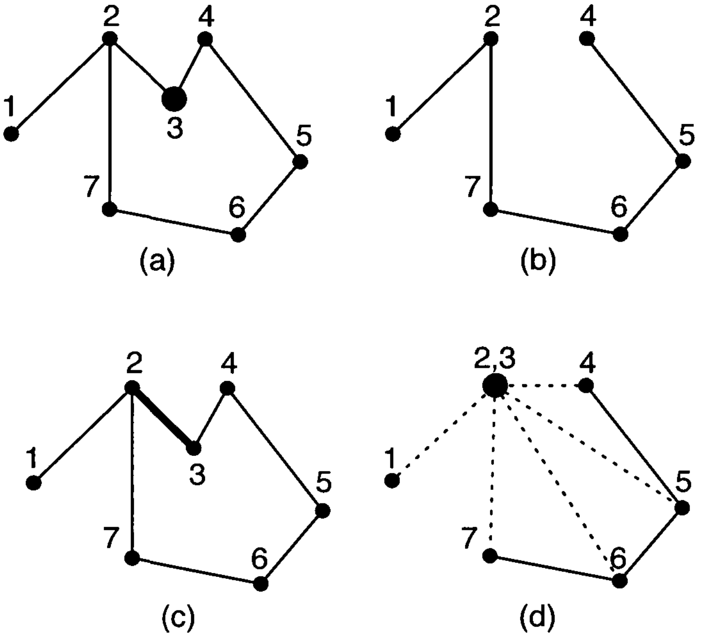

Figure 4.1 : Reduction moves on an association graph. In panel (a), vertex 3 is determined to be the most significant object in the graph and consequently removed (b). By contrast, the most significant object could be an edge, such as 2—3 in panel (c), which is contracted into vertex 2, 3 in panel (d). Dashed edges are new and may be rejected if its weight is less significant than the edge threshold α_e.

in (4.2).

Association Graph Construction

1. Initialization – $A = \emptyset$.

2. Add vertices – For each locus i, construct a vertex v_i and determine its significance $p_i = w_v(v_i)$. If $p_i > a_v$, add v_i to A.

3. Add edges – For each vertex pair $\{(v_i, v_j) | v_i, v_j \in V \wedge v_i \neq v_j\}$, construct the edge $v_i v_j$ and determine its significance $p_{i,j} = w_e(v_i, v_j)$. If $p_{i,j} > \alpha_e$, add $v_i v_j$ to A.

Association Graph Reduction

1. Initialization – Let $R = \emptyset$ be a list of vertices which have been removed from the graph, which is maintained in order with decreasing order of significance according to the relation $>$.

2. Reduction – While the association graph A is not empty:

   (a) Let x be the vertex or edge with the most significant weight w.

   (b) If $x \in V$, remove x and all edges for which x is an end from A and insert x into R.

   (c) If $x \in E$, $w \gg w_v(v_1)$, and $w \gg w_v(v_2)$ for ends v_1 and v_2, contract x, creating a new vertex v_x. For every vertex i other than v_x, compute $p_i = w_e(v_x v_i)$ and add the edge $v_x v_i$ to A if $p_i > \alpha_e$. Otherwise, add all remaining vertices to R and terminate.

3. Correction – Adjust the probabilities of the vertices in R for multiple testing by Bonferroni, FDR [40], or other corrections.

4. Termination – Return the elements of R whose probabilities meet the experiment-wide level of significance.

4.2 Summary

Association graph reduction is a new procedure capable of identifying individual loci and epistatic interactions that predispose to a phenotype of interest. It operates by constructing a new data structure, the association graph, to represent current knowledge about genotype-phenotype association and locus-locus interactions. This simple algorithm allows the user to conduct genome-wide association studies with a variety of approaches by simply adjusting parameters.

The present formulation of AGR is entirely computational. However, expert knowledge such as information about biological interactions could be incorporated into the edge weight definitions to presumably improve power to detect association and epistasis. One way to incorporate such knowledge is to define vertex and edge weights in a Bayesian framework. In such a scenario, the purely computational AG would make use of non-informative priors, while the expert knowledge-based version might define edge prior probabilities based on biological information about interactions. While promising, such an extension remains future work.

AGR is a simple and elegant method to approach the genotype-phenotype association problem. It is capable of extremely fast operation, especially when pure epistatic interactions are not believed to be present. When an exhaustive search of all pairs of loci is desired, AGR can also accommodate. However, unlike the one- and two-stage simultaneous searches described in Chapter 3, AGR will execute as many stages as necessary. While allowing for the exploration of very high-order interactions, appropriate controls for model complexity are provided by the use of a custom comparison relation between edges and vertices to ensure identified interactions are not merely the effect of locus heterogeneity. These and further advantages and disadvantages are summarized below.

Advantages

- AGR is extremely fast at best and tractable at worst. Depending on the parameters α_v and α_e, the expensive graph construction step has cost near its lower bound $\Omega(n)$ or upper bound $O(n^2)$ on n loci. Given appropriate model regularization through a weight comparison operator $\gg$, graph reduction requires a edge contractions proportional to the number of interactions in the data.

- AGR is capable of detecting interaction effects ranging in magnitude from heterogeneity (no interaction) to pure epistasis (no marginal effects).

- Through the parameterization of edge and vertex thresholds and custom weight definitions, AGR is extremely flexible. The extent of exploration for pure interactions and the consequential running time costs and potential power gains are chosen by the user.

- Graph construction and reduction operations are easily parallelizable.

- The contraction of edges into combined vertices facilitates interpretation of results. Single vertices in the result representing multiple loci can be interpreted as significant interaction, while the entire set of vertices returned shows the structure of locus heterogeneity.

- Missing data is handled in an elegant manner when supported by the selected weight functions.

- AGR is implemented in an easy-to-use and fast multi-threaded software package.

Disadvantages

- Users must select appropriate weights and threshold parameters α_v and α_e.
- Combining loci intensifies missing data problems. When a subject is not genotyped at any locus contained in the arguments of a call to *combine*, the result is undefined.

Chapter 5
Data Simulation

In order to evaluate the relative performance of methods to detect genotype-phenotype association, benchmark data must be used. Typically, such data may come in the form of real-world observations with known associations or synthetic data generated with known risk and error models. Real data are preferable under many circumstances, since contrived models may not accurately represent complex biological processes. Unfortunately, since few instances of statistical epistasis have been discovered and replicated, one must resort to simulations. While a few software packages to manufacture synthetic genotype-phenotype data have been described (examples include [114, 115, 116]), no currently available software supports disease models involving epistatic interactions*. To address the need to create simulated data involving interactions, I have developed SimGE, an easy to use Java program to generate random case/control data that may contain epistasis.

5.1 Algorithm

SimGE uses a simple algorithm to generate interacting and non-interacting genotypes associated with a phenotype, as well as noisy, unassociated data (Figure 5.1). In short, the program first creates a set of associated loci, some of which may actually represent

*The genomeSim package supports interactions [114], though it remains unavailable over three years after its published description.

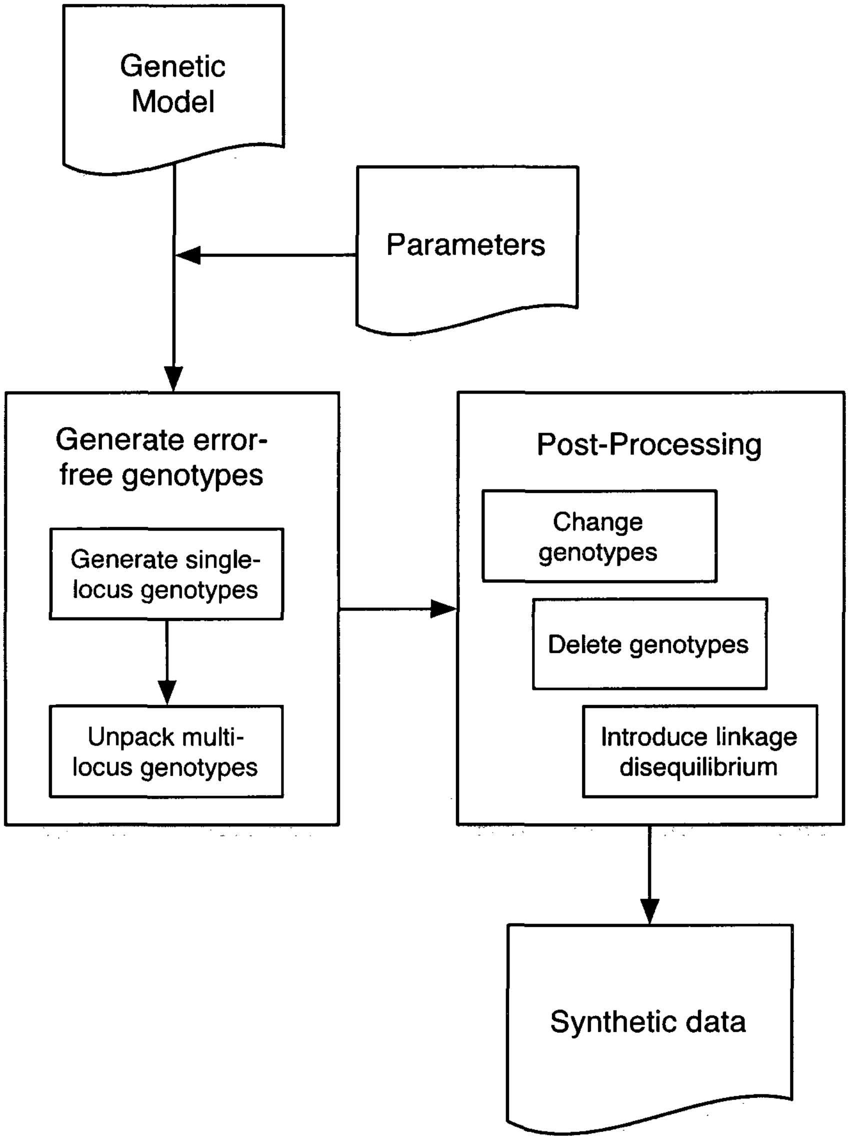
Genetic Model
Parameters
Generate error-free genotypes
Generate single-locus genotypes
Unpack multi-locus genotypes
Post-Processing
Change genotypes
Delete genotypes
Introduce linkage disequilibrium
Synthetic data

Figure 5.1 : Overview of the SimGE simulator. The user specifies a genetic model by assigning a penetrance to each multi-locus genotype and other parameters, such as the number of unassociated loci to generate. A set of single-locus genotypes are then created, which may include interaction loci. Next, interacting markers are unpacked from the interaction loci. Finally, a set of post-processing filters may be applied and the data are written to a file for further analysis.

Table 5.1 : A pure interaction model between loci A and B. In this case, 0.4 of subjects with the genotype aa at locus A and BB at locus B will develop the phenotype.

	bb	Bb	BB
aa	-	-	0.4
Aa	-	0.2	-
AA	0.4	-	-

epistatic interactions.

1. **Specify model and parameters** – The user must specify a genetic model and other parameters. Phenotype risk is specified as the penetrance associated with each possible multi-locus genotype. For example, a two-locus pure interaction may be expressed with the values in Table 5.1. Other parameters are described in Table 5.2.

2. **Generate error-free genotypes** – SimGE generates a complete set of risk and nuisance genotypes by the following steps.

   (a) From the specified associated genotype frequencies and penetrances, compute the probability p_i of each multi-locus genotype i.

   (b) Generate unassociated genotypes at risk loci for controls using the user-specified population genotype frequencies.

   (c) For each case, draw a random multi-locus genotype for an *interaction locus* representing all susceptibility loci, which is drawn from the p_is. Unpack

Table 5.2 : SimGE parameters.

Parameter	Meaning
`numCases`	The number of affected cases to generate under the specified genetic model at risk loci.
`numControls`	The number of unaffected controls to generate with completely random data.
`numVariants`	The number of variants (genotype outcomes at the risk loci).
`associatedFrequencies`	Genotype frequencies for each associated locus.
`penetrances`	Penetrances for each possible outcome in the genetic model.
`format`	The format of the output file. Supported types are `MDR`, `PLINK`, and `SimGE`.

the interaction locus into individual loci.

   (d) Generate random genotypes for unassociated loci by selecting the frequency of one allele from $U(0,1)$ and assuming Hardy-Weinberg Equilibrium.

3. Post-processing – If requested, apply post-processing to the genotypes. Possibilities include:

   (a) introduction of random errors;

   (b) random marking of some genotypes as missing; or

   (c) generation of markers in linkage disequilibrium with the susceptibility loci.

4. Output – Write the data to a file for further analysis by other means.

5.2 Example

As a concrete demonstration, I present a complete SimGE configuration file, which will be later used in Chapter 7 to evaluate the performance of the methods described in Chapters 3 and 4. With this set of parameters, SimGE will generate a total of 10,000 loci in 1,000 cases and 1,000 controls. Four of these loci, which are each biallelic, will be associated with case status. The model is mixed, containing both epistasis and heterogeneity.

```
#
# SimGE configuration file
#
# Four-locus mixed model which combines
# (a) two-locus lethal/missing genotype and
# (b) two recessive susceptibility loci conferring additive risk.
#

numCases = 1000
numControls = 1000
```

```
numLoci = 10000
numVariants = 3 3 3 3

# f(a) = 0.7; assumption of HWE
associatedFrequencies = \
0.49 0.42 0.09 \
0.49 0.42 0.09 \
0.49 0.42 0.09 \
0.49 0.42 0.09

# A four-locus model is specified as
# aabbccdd Aabbccdd AAbbccdd    aaBbccdd AaBbccdd     AABbccdd aaBBccdd     AaBBccdd AABBccdd
# aabbCcdd AabbCcdd AAbbCcdd    aaBbCcdd AaBbCcdd     AABbCcdd aaBBCcdd     AaBBCcdd AABBCcdd
# aabbCCdd AabbCCdd AAbbCCdd    aaBbCCdd AaBbCCdd     AABbCCdd aaBBCCdd     AaBBCCdd AABBCCdd
# aabbccDd AabbccDd AAbbccDd    aaBbccDd AaBbccDd     AABbccDd aaBBccDd     AaBBccDd AABBccDd
# aabbCcDd AabbCcDd AAbbCcDd    aaBbCcDd AaBbCcDd     AABbCcDd aaBBCcDd     AaBBCcDd AABBCcDd
# aabbCCDd AabbCCDd AAbbCCDd    aaBbCCDd AaBbCCDd     AABbCCDd aaBBCCDd     AaBBCCDd AABBCCDd
# aabbccDD AabbccDD AAbbccDD    aaBbccDD AaBbccDD     AABbccDD aaBBccDD     AaBBccDD AABBccDD
# aabbCcDD AabbCcDD AAbbCcDD    aaBbCcDD AaBbCcDD     AABbCcDD aaBBCcDD     AaBBCcDD AABBCcDD
# aabbCCDD AabbCCDD AAbbCCDD    aaBbCCDD AaBbCCDD     AABbCCDD aaBBCCDD     AaBBCCDD AABBCCDD
#
penetrances = \
0          0       0.0379893424 0        0.0189946712 0        0.0379893424 0        0 \
0          0       0.0379893424 0        0.0189946712 0        0.0379893424 0        0 \
0.02       0.02    0.0579893424 0.02     0.0389946712 0        0.0579893424 0        0 \
0          0       0.0379893424 0        0.0189946712 0        0.0379893424 0        0 \
0          0       0.0379893424 0        0.0189946712 0        0.0379893424 0        0 \
0.02       0.02    0.0579893424 0.02     0.0389946712 0        0.0579893424 0        0 \
0.02       0.02    0.0579893424 0.02     0.0389946712 0        0.0579893424 0        0 \
0.02       0.02    0.0579893424 0.02     0.0389946712 0        0.0579893424 0        0 \
0.04       0.04    0.0779893424 0.04     0.0589946712 0        0.0779893424 0        0

#
# Output options
#

# Output file format, which may be one of the following:
# MDR
# Plink
# SimGE
#
# The value is not case-sensitive.
# File format may be overridden from the command line.
format = MDR
```

5.3 Performance

In order to determine the cost to synthesize random genotypes using SimGE, I performed a series of benchmarks using the above parameter file. Testing took place on an 8-core MacPro running a 32-bit Java virtual machine on MacOS X 10.5. For each data set size investigated, SimGE was invoked ten times and the fastest time recorded. Results are shown in Table 5.3.

Table 5.3 : User and real time costs of of SimGE in seconds. Simulations were conducted using the above parameter file, varying only the number of loci to generate. Figures are the minimum time measured over ten runs.

Loci	User	Real
1,000	3.5	3.0
10,000	31.5	27.4
100,000	913.0	387.2

5.4 Summary

In spite of a spate of algorithms to detect genotype-phenotype association, there is a lack of tools capable of simulating statistical epistasis. To facilitate objective benchmarking and comparison of competing methods, I have presented SimGE, an easy-to-use and fast genotype-phenotype simulator. Through a simple configuration file, users may specify arbitrary generative models, including those involving epistasis and/or locus heterogeneity. With little effort on the part of its users, SimGE can quickly generate realistic genome-sized data sets.

Chapter 6

Models of Multi-Locus Risk

In order to perform a meaningful evaluation of methods to detect multi-locus association, it is important to have realistic mathematical models of the effects such genetic changes have on phenotype susceptibility. Just as there are many methods for detecting genotype-phenotype associations, several models of risk conferral have been proposed (for example, see [36, 35, 34, 38, 5, 37, 117]). Given the complexity of biological processes, it is not meaningful to select one of these as a gold standard of realism. Instead, I present a few models that have previously been used to establish the utility of several important methods outlined in Chapters 3 and 4.

While I have selected models that span much of the space defining plausible genetic disease processes, I have elected to hold constant well studied parameters such as minor allele frequency (MAF) and the amount of linkage disequilibrium between observed markers and the hypothetical disease locus. Instead, I assume risk loci are genotyped (or, equivalently, in complete linkage disequilibrium with the risk-conferring mutations), that all markers are in Hardy-Weinberg Equilibrium, and that there are no other sources of genetic risk. I set the population prevalence to 0.01 and MAF of risk loci to $f_a = 0.3$ for most models, which I select for its high power to detect association over models with lower MAF. The only exception is a purely epistatic model displaying no marginal risk, whose risk loci have MAF $f_a = 0.5$. Allele frequencies for unassociated alleles are drawn from $f_u \sim U(0.05, 0.5)$, which

Table 6.1 : Multi-locus model parameters.

Parameter	Value
MAF (associated locus)	$f_a = 0.3$
MAF (unassociated locus)	$f_u \sim U(0.05, 0.5)$
Phenotype prevalence	$p_p = 0.01$
Number of cases	$n_a = 1000$
Number of controls	$n_u = 1000$
Number of markers	$m = 1000$, $m = 10,000$

resembles the empirical distribution found in recent Affymetrix™ genome-wide SNP microarrays. These parameters, which are summarized in Table 6.1, represent very realistic conditions under which a genome-wide scan for association may take place. The only notable exception is the selection of only $m = 1000$ markers, which is insufficient for a genome-wide study. This value was chosen as way to make tractable evaluation of exhaustive methods such as MDR [48], which test all possible k-way interactions (see Section 3.1.1). In Chapter 7, three well-performing methods were selected for further analysis of $m = 10,000$ simulated SNPs under otherwise identical conditions.

6.1 Categories of Multi-Locus Association

The risk conferred by susceptibility loci can be described as either *additive* or *epistatic.* Additive effects are those that confer risk independent of other risks, while epistasis describes interactions between loci.

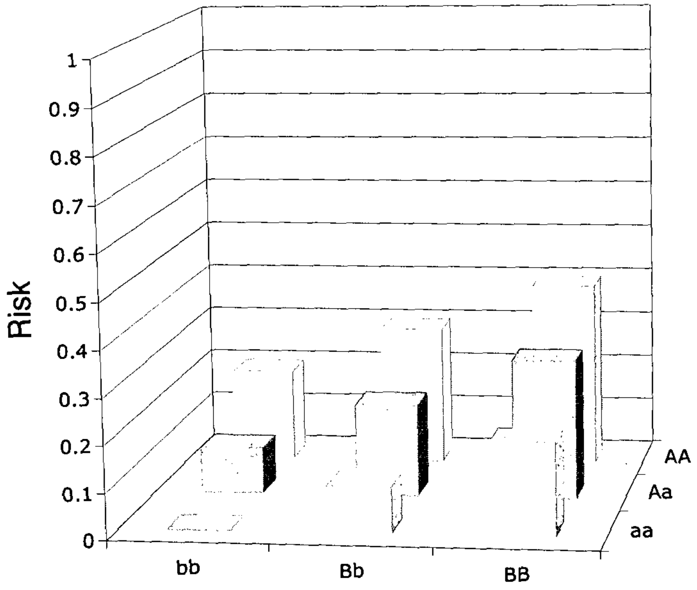

Figure 6.1 : Simple two-locus additive risk model. Loci A and B have risk alleles A and B, respectively. Each susceptibility allele confers a small amount of risk in a dominant mode of inheritance.

When multiple additive effects are present, a condition known as *locus heterogeneity*, risk may increase additively or multiplicatively. While each locus or set of interacting loci in a heterogeneous system will show a marginal statistical effect, the magnitude of risk conferred by each locus may be very small. Furthermore, loci may play different roles in different study subjects, which makes their detection more difficult than simple Mendelian genes. A simple two-locus example is depicted in Figure 6.1. A pair of susceptibility loci A and B each operate additively: each disease allele further impairs the system, resulting in increased risk or severity of phenotype.

One says that statistical epistasis is present when a deviation from additive (or multiplicative) risk is observed if the contribution of multiple susceptibility loci are considered [16]. In such a system, interacting loci may or may not show marginal statistical significance. An example of each scenario is depicted in Figure 6.2. In panel (a), the presence of two or more risk alleles from either susceptibility locus confers maximal risk. In such a *threshold* model, risk is dichotomous. The second panel depicts a *lethal genotype* model. One dysfunctional allele sufficiently impairs a process to develop disease, while more are incompatible with life.

Most methods are restricted in the type of multi-locus association effects they may discover. For example, MDR [48, 49] seeks only effects caused by an interaction between k loci for some user-specified value k. On the other hand, Set-Association [39] simply sums single-locus marginal statistics, which likely precludes the possibility of identifying purely epistatic effects. Other strategies may be capable of identifying both types, but remain restricted in other ways. For example, a typical two-stage simultaneous search will identify all loci meeting some low level of marginal association and then test all two-way interactions within that set [38]. Still other important methods might be theoretically capable of detecting all meaningful types of models, but only if entire sets of interacting loci without marginal effects are selected simultaneously in a random trial [107]. To provide a meaningful comparison and performance-based classification of methods, I have selected models consisting of marginal effects, interactions, and a combination of each.

6.1.1 Locus Heterogeneity

Methods capable of detecting locus heterogeneity (for example, Set-Association [39]) tend to be designed to detect several non-interacting susceptibility loci. Therefore, I propose two additive models consisting of two and four loci, each of which confers

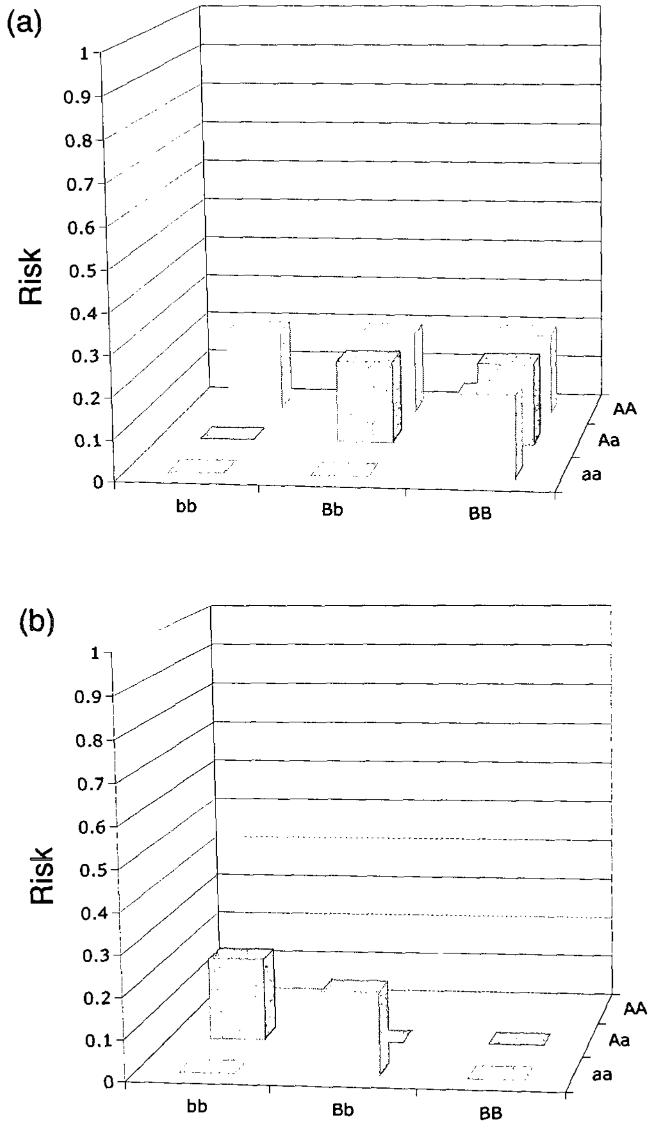

Figure 6.2 : Simple two-locus epistatic risk models. In **(a)**, two or more risk alleles are required to express the phenotype; each locus will show some marginal association. A pure epistatic interaction with no detectable marginal effect is shown in panel **(b)**, in which the presence of precisely one risk allele is required at each locus, a plausible scenario when some risk genotypes are lethal.

risk in a recessive pattern of inheritance. Marginal penetrances for the two-locus model are 0.07 and 0.0411. The four-locus model has penetrances 0.018, 0.023, 0.03, and 0.0401. These extremely low values reflect the extreme difficulty in detecting susceptibility loci when multiple risk factors are involved, even in the absence of epistatic interactions displaying no main effects.

6.1.2 Epistasis

Epistatic models are comprised of multiple interacting loci which display non-linear risk and often little or no marginal significance individually. There are 255 two-locus models of epistasis that confer risk [36], of which I have selected three for evaluation (Table 6.2). The first resembles panel (a) in Figure 6.2 in which multi-locus genotypes with at least two risk alleles from either or both loci elevate risk to a maximum level. It is similar to the threshold model evaluated by Marchini *et al* [38]. The second is a slightly less extreme version of Figure 6.2 panel (b), which is based on the "Missing Lethal Genotype Model" of lonita and Man [65] and is similar to the two-locus model used to demonstrate the utility of MDR [48]. One copy of a susceptibility allele from each locus raises risk to a level β while two risk alleles from one susceptibility locus confer risk 2β. Multi-locus genotypes with more than two risk alleles are fatal and hence not present in the study population. The final model displays pure epistasis: the loci appear to confer no risk when considered individually. The penetrance matrix for this third model is similar to that of the lethal genotype model, differing primarily in its MAF, which must be $f_a = 0.5$ to avoid marginal statistical significance.

Table 6.2 : Epistatic models under study: (a) a threshold effect, (b) the consequences of lethal genotypes for a study population, and (c) a model that displays pure epistasis with no marginal risk. Unlike all other models under study, (c) requires MAF $f_a = 0.5$.

(a)

	bb	Bb	BB
aa	0	0	0
aA	0	0.038	0.038
AA	0	0.038	0.038

(b)

	bb	Bb	BB
aa	0	0	0.057
aA	0	0.028	0
AA	0.057	0	0

(c)

	bb	Bb	BB
	0	0	0.04
aA	0	0.02	0
AA	0.04	0	0

6.1.3 Mixed Model

Few strategies for detecting multi-locus association are capable of efficiently detecting both interactions and locus heterogeneity. To my knowledge, no comparison of any methods has taken place under such circumstances. Therefore, I present a mixed model comprised of four risk loci, two of which (A and B) interact in a missing lethal genotype model and the others (C and D) confer risk independently. The interacting pair of loci confer risk in a manner similar to panel B of Table 6.2: risk is doubled if two susceptibility alleles are inherited from the same locus. The other loci confer risk in a recessive mode of inheritance. A complete listing of penetrances associated with four-locus genotypes is presented in Table 6.3.

Table 6.3 : Mixed model penetrances

Genotype	A/B Penetrance	C Penetrance	D Penetrance	Total Penetrance
aabbccdd	-	-	-	*0*
aabbccDd	-	-	-	*0*
aabbccDD	-	-	0.02	*0.02*
aabbCcdd	-	-	-	*0*
aabbCcDd	-	-	-	*0*
aabbCcDD	-	-	0.02	*0.02*
aabbCCdd	-	0.02	-	*0.02*
aabbCCDd	-	0.02	-	*0.02*
aabbCCDD	-	0.02	0.02	*0.04*
aaBbccdd	-	-	-	*0*
aaBbccDd	-	-	-	*0*
aaBbccDD	-	-	0.02	*0.02*
aaBbCcdd	-	-	-	*0*
aaBbCcDd	-	-	-	*0*
aaBbCcDD	-	-	0.02	*0.02*
aaBbCCdd	-	0.02	-	*0.02*
aaBbCCDd	-	0.02	-	*0.02*
aaBbCCDD	-	0.02	0.02	*0.04*

Table 6.3 : (continued)

Genotype	A/B Penetrance	C Penetrance	D Penetrance	Total Penetrance
aaBBccdd	0.038	-	-	*0.038*
aaBBccDd	0.038	-	-	*0.038*
aaBBccDD	0.038	-	0.02	*0.058*
aaBBCcdd	0.038	-	-	*0.038*
aaBBCcDd	0.038	-	-	
aaBBCcDD	0.038	-	0.02	*0.058*
aaBBCCdd	0.038	0.02	-	*0.058*
aaBBCCDd	0.038	0.02	-	*0.058*
aaBBCCDD	0.038	0.02	0.02	*0.078*
Aabbccdd	-	-	-	*0*
AabbccDd	-	-	-	*0*
AabbccDD	-	-	0.02	*0.02*
AabbCcdd	-	-	-	*0*
AabbCcDd	-	-	-	*0*
AabbCcDD	-	-	0.02	*0.02*
AabbCCdd	-	0.02	-	*0.02*
AabbCCDd	-	0.02	-	*0.02*
AabbCCDD	-	0.02	0.02	*0.04*

Table 6.3 : (continued)

Genotype	A/B Penetrance	C Penetrance	D Penetrance	Total Penetrance
AaBbccdd	0.019	-	-	*0.019*
AaBbccDd	0.019	-	-	*0.019*
AaBbccDD	0.019	-	0.02	*0.039*
AaBbCcdd	0.019	-	-	*0.019*
AaBbCcDd	0.019	-	-	*0.019*
AaBbCcDD	0.019	-	0.02	*0.039*
AaBbCCdd	0.019	0.02	-	*0.039*
AaBbCCDd	0.019	0.02	-	*0.039*
AaBbCCDD	0.019	0.02	0.02	*0.059*
AaBBccdd	-	-	-	*0*
AaBBccDd	-	-	-	*0*
AaBBccDD	-	-	-	*0*
AaBBCcdd	-	-	-	*0*
AaBBCcDd	-	-	-	*0*
AaBBCcDD	-	-	-	*0*
AaBBCCdd	-	-	-	*0*
AaBBCCDd	-	-	-	*0*
AaBBCCDD	-	-	-	*0*
AAbbccdd	0.038	-	-	*0.038*

Table 6.3 : (continued)

Genotype	A/B Penetrance	C Penetrance	D Penetrance	Total Penetrance
AAbbccDd	0.038	-	-	*0.038*
AAbbccDD	0.038	-	0.02	*0.058*
AAbbCcdd	0.038	-	-	*0.038*
AAbbCcDd	0.038	-	-	*0.038*
AAbbCcDD	0.038	-	0.02	*0.058*
AAbbCCdd	0.038	0.02	-	*0.058*
AAbbCCDd	0.038	0.02	-	*0.058*
AAbbCCDD	0.038	0.02	0.02	*0.078*
AABbccdd	-	-	-	*0*
AABbccDd	-	-	-	*0*
AABbccDD	-	-	-	*0*
AABbCcdd	-	-	-	*0*
AABbCcDd	-	-	-	*0*
AABbCcDD	-	-	-	*0*
AaBbCCdd	-	-	-	*0*
AaBbCCDd	-	-	-	*0*
AaBbCCDD	-	-	-	*0*
AaBBccdd	-	-	-	*0*
AaBBccDd	-	-	-	*0*

Table 6.3 : (continued)

Genotype	A/B Penetrance	C Penetrance	D Penetrance	Total Penetrance
AaBBccDD	-	-	-	*0*
AaBBCcdd	-	-	-	*0*
AaBBCcDd	-	-	-	*0*
AaBBCcDD	-	-	-	*0*
AaBBCCdd	-	-	-	*0*
AaBBCCDd	-	-	-	*0*
AaBBCCDD	-	-	-	*0*

6.2 Summary

In order to determine a method's ability to detect genotype-phenotype association through simulation, realistic generative models are required. In this chapter, I have presented six such models that exhibit either epistasis, locus heterogeneity, or both. When combined with the specified minor allele frequencies, the models predict population prevalence of 1%. In Chapter 7, I will use these models to evaluate the performance of our method (AGR) as well as the methods described in Chapter 3.

Chapter 7

Evaluation of Methods

In order to conduct a meaningful comparison of the methods described in Chapters 3 and 4, I evaluated their statistical and computational performance on several simulated data sets. Data were generated under the six generative models discussed in Chapter 6 using SimGE (Chapter 5). Each method was executed in turn, recording runtime expense and rates of detecting actual and spurious associations. Results for each method as well as specific parameters used can be found in the following sections. An overview and discussion of my findings are presented in Chapter 8.

Performance was evaluated according to three criteria: sensitivity, specificity, and running time. Let P_t and P_f be the number of true and false positives discovered in each data set, respectively. Similarly denote the number of true and false positive discoveries as N_t and N_f. Sensitivity $s_n = P_t/(P_t + N_f)$ is the rate at which true susceptibility loci are recovered from the data. Similarly, specificity $s_p = N_t/(N_t+P_f)$ quantifies a method's ability to avoid spurious hits. Given that million-marker data sets are likely to become increasingly common, it may be difficult to intuitively grasp the importance of a low rate of specificity. Therefore, I include as a performance measure the expected number of false positives in a million-locus experiment $s_p \times 10^6$.

Of particular interest is the feasibility of each method. To quantify an algorithm's ability to scale to genome-wide data sets, I measured average user and real (clock) time required to correctly complete analysis. User time measures non-system related computations and can be expressed in CPU-time units. For serial implementations, user and real time measurements are closely related. However, parallel implementations may decrease real time computational requirements on multi-processor systems. When evaluating software performance, it is typical to conduct several runs and select the minimum as the true computational cost. However, given the stochastic nature of some methods under study, running time statistics were aggregated by averaging to avoid biases created by fast random runs.

All experiments were conducted on the Shared University Grid at Rice (SUG@R). Though none of the tested software is capable of using multiple nodes, some are multithreaded. Parallel computations took place on eight CPU cores and serial algorithms were executed simultaneously on up to eight data sets.

7.1 1,000 Loci

I first present the results of all methods run on 1,000-locus data sets. While such a small set of markers does not accurately represent genome-wide data sets, it does provide a useful starting point, allowing expensive methods a chance to demonstrate their utility, even if larger data sets show their futility.

7.1.1 AGR

Parameters Association graph reduction allows for the specification of vertex and edge weight types and significance thresholds for accepting objects into a graph. For the present analysis, I selected χ^2 weights with $\alpha_e = \alpha_v = 0.1$ for models displaying marginal association and $\alpha_e = \alpha_v = 1$ for the model of pure epistasis. In all cases, edge weights did not take into account phenotypic status: the second edge weight formulation (Equation 4.2) was used. To control for multiple testing, I applied the conservative Bonferroni correction, multiplying the probability of each vertex removed from the graph with the total number of tests executed upon termination.

The choice to use different parameters based on *a priori* knowledge of the underlying generative model may prove controversial. In my view, the likelihood of discovering pure epistatic interactions seems minute, given their paucity of appearance in the literature.* While it is certainly possible that my inability to find a single such result is due to a bias in favor of practical methods that complete analysis in a short amount of time, I find this reasoning unsatisfying. Were there a real world example of pure epistasis, I contend that proponents of MDR and other exhaustive methods would be eager to find and publicize them. While their failure may be technical, it underscores the rarity of such circumstances, if nothing else.

Still, I wished to demonstrate that AGR is flexible enough to discover interactions displaying no effects, should one choose to conduct such a search. In so doing, I found it to be reasonably fast and quite powerful, as the results below indicate.

AGR is implemented in parallel as a multi-threaded program. To accurately quantify its performance on modern multi-core machines, eight CPU cores were allocated to computation.

*While it is not difficult to construct mathematical interaction models without marginal effects, it us unclear whether such scenarios exist in real data [118]. I am unaware of a single published instance of statistical epistasis exhibiting no marginal association.

Statistical Performance AGR's sensitivity to detect all causative loci and specificity are presented in Table 7.1. It correctly identified all susceptibility loci displaying main effects with high specificity. The model showing pure epistasis provided a particular challenge, resulting in a small decrease in sensitivity. I suspect this drop to be caused by the correction for multiple testing, which is particularly severe when all pairs are tested. The expected number of false positives in a million-marker experiment is reasonable. In all cases, the level of error would require at most a few custom microarray experiments to eliminate spurious hits.

Computational Costs The time required to conduct analysis is summarized in Table 7.2. The computational cost to identify associated loci in the 1,000-SNP data set was negligible in most cases. Not surprisingly, exhaustive enumeration of all pairs of loci required longer, though it remained quite reasonable.

7.1.2 MDR

Parameters Multifactor dimensionality reduction software [49, 50] requires the specification of one key parameter k, the size of interaction to test. For the present study, I investigated all single markers and pairs thereof ($1 \leq k \leq 2$). MDR's implementers specify a default case/control ratio threshold $T = 1$ for classification of multi-locus genotypes as high or low risk. MDR is multi-threaded and was allowed to use eight CPU cores during execution.

Statistical Performance MDR successfully identified all risk factors in models with two risk loci while maintaining a low level of false detection (Table 7.3). Unfor-

Table 7.1 : AGR performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	100.0	99.0	100.0	100.0	100.0
Specificity (%)	99.8	99.8	99.8	99.8	99.6	99.6
Expected False Positives	2,041	2,047	1,984	2,042	4,049	4,041

Table 7.2 : AGR cost in 1,000 loci averaged over 1,000 simulated data sets. In cases where edge and vertex thresholds were used ($\alpha_e < 1, \alpha_v < 1$) termination was nearly instantaneous. Examining all pairs of loci took longer, but remains very reasonable.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	0.1	0.1	66.0	0.1	0.1	0.1
Real (s)	0.2	0.2	8.6	0.2	0.2	0.2

Table 7.3 : MDR performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	100.0	100.0	100.0	50.0	62.8
Specificity (%)	99.8	99.8	99.7	99.8	99.8	99.7
Expected False Positives	2,004	2,004	3,003	2,004	2,008	4,061

tunately, it consistently failed to identify all associated loci in more complex models. In the 4-locus heterogeneous model, MDR appears to have identified one locus during its scan of single markers and then paired the same marker with another risk locus during its search for interactions. It performed somewhat better on the mixed model, often selecting a different marker in its first stage than those identified by its second.

Computational Costs MDR's exhaustive search should have consistent and predictable computational costs. This notion is supported by runtime statistics, which remained nearly constant across all tested models (Table 7.4). Unfortunately, this expense is substantial. Assuming a fixed cost of $c_m = \sqrt{\bar{c}}/1000$ per marker for observed average cost $\bar{c}$, it would take over a CPU-day to analyze 10,000 SNPs and nearly 122 CPU-days to evaluate all pairs in a 100,000-SNP data set.

7.1.3 All-Pairs Simultaneous Search

Parameters I conducted all-pair simultaneous search with PLINK [66, 67]. For each pair of markers A and B, PLINK performs logistic regression on the model

Table 7.4 : MDR cost in 1,000 loci. In all cases, MDR examines all pairs of loci, resulting in predictable runtime costs.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	1049.9	1051.6	1047.7	1048.9	1051.1	1051.8
Real (s)	136.6	137.3	136.3	136.6	136.8	136.9

$P \sim \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB + \epsilon$, where P is the phenotype under study. For the purposes of identifying interactions, β_0, β_1, and β_2 are treated as nuisance parameters. While investigators typically elect to perform a locus-by-locus scan for main effects as a first step, such analysis was not conducted here. After PLINK analysis, I applied Bonferroni correction to obtain a set of associated loci.

Statistical Performance Despite its examination of every pair of loci under study, the all-pairs simultaneous search performs poorly for most models considered. For all except the *threshold* and *pure* epistasis models, sensitivity was poor. For models displaying main effects, this shortcoming may be explained and made acceptable by the fact that a marginal scan would have had perfect sensitivity for all models except one (data not shown).

In the face of poor performance by an all-pairs search, one may be consoled by the fact that most associated loci are, in fact, likely to be identified by a scan for main effects. Unfortunately, the simultaneous search fails to find strong interactions, which may abandon valuable information about phenotype etiology and render results

Table 7.5 : All-pairs simultaneous search performance in 1,000 loci. Figures do not include discovery by a genome-scan for marginal effects as would often accompany an all-pairs simultaneous search.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	70.4	100.0	0.0	54.3	51.3
Specificity (%)	99.7	99.8	99.7	99.9	99.7	99.7
Expected False Positives	2,548	2,005	2,548	573	2,749	2,633

Table 7.6 : All-pairs simultaneous search cost in 1,000 loci. In all cases, a simultaneous search examines all pairs of loci, resulting in predictable runtime costs.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	818.3	818.0	817.5	817.2	819.2	818.7
Real (s)	824.3	824.0	823.6	823.3	825.2	824.6

impossible to reproduce.

Computational Costs Like MDR, an all-pairs simultaneous search has predictable and nearly constant cost (Table 7.6). While somewhat lower than MDR's, the fixed cost associated with exhaustive search remains prohibitive at a genome-wide scale. Assuming the same fixed cost model as above, a scan of 10,000 SNPs would require over 22 CPU-hours and 100,000 markers would consume nearly 95 days of CPU time.

7.1.4 Two-Stage Search

Parameters There are two types of two-stage searches, simultaneous and conditional, both of which were conducted with PLINK [66, 67]. In both cases, PLINK identified a set of marginally associated markers at the $\alpha = 0.1$ level after either Bonferroni or FDR correction for multiple testing. Simultaneous search considered all pairs of loci identified in the first stage, while conditional tested all marginally suggestive markers with all others, regardless of evidence for association. After the second stage, I applied Bonferroni correction to identify a final set of associated loci.

Statistical Performance A two-stage scan necessarily conducts a locus-by-locus scan for main effects as its first step. In that context, I am interested in determining what information, if any, can be gained from a second stage. Indeed, a marginal scan at an appropriate experiment-wide significance threshold of $\alpha = 0.05$ has 100% sensitivity in all except the pure epistasis model with reasonable specificity (data not shown). Table 7.7 summarizes statistical performance of two-stage search methods, ignoring results which may arise as part of a related scan for marginally associated loci. Neither method had better sensitivity for most models nor were they likely to identify associated loci displaying no main effect.

Computational Costs Runtime expense to conduct a two-stage search is summarized in Table 7.18. If α_l is the lax level of significance required for a marker to pass from a stage one scan of m markers, a simultaneous scan requires $\Theta\left((n\alpha_l)^2\right)$ steps, while conditional takes $\Theta\left(n^2\alpha_l\right)$, assuming in each case that most markers follow the

Table 7.7 : Two-stage simultaneous and conditional search performance in 1,000 loci. Figures do not include discovery by a genome-scan for marginal effects as would often accompany these methods. A one-stage marginal scan with an appropriately stringent α level has 100.0% sensitivity in all models except that which displays no main effect.

(a) **Simultaneous**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	70.4	0.2	0.0	66.8	54.2
Specificity (%)	99.8	99.8	100.0	100.0	99.7	99.8
Expected False Positives	2,056	1,731	4	60	2,729	2,228

(b) **Conditional**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	67.2	0.0	0.0	50.2	50.8
Specificity (%)	99.8	99.8	100.0	100.0	99.8	99.8
Expected False Positives	2,006	1,350	8	4	2,019	2,042

Table 7.8 : Two-stage simultaneous and conditional search cost in 1,000 loci. Figures include the cost of the first-stage marginal scan, one second in every case.

(a) **Simultaneous**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	11.0	11.2	10.6	11.0	11.5	11.5
Real (s)	11.0	11.2	10.7	11.1	11.5	11.5

(b) **Conditional**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	20.0	20.4	19.2	20.1	21.0	21.0
Real (s)	20.0	20.4	19.3	20.2	21.0	21.0

null distribution. This cost is small for 1,000 loci and scaling to genome-wide data sets seems possible with a fast enough implementation.

7.1.5 Classification Trees

Parameters There are two popular algorithms for constructing classification trees, of which I chose the more readily available CART method, as implemented by the R package *tree* [78, 79, 89]. Splits were selected by minimizing deviance.

Statistical Performance In models displaying any marginal significance, CART performed extremely well (Table 7.9). Sensitivity for each of these models was 100% and specificity was no lower than most other methods. Unfortunately, CART had no power to recover pure interactions.

Computational Costs CART was not the fastest method tested, though it was not unreasonably slow (Table 7.10) when testing $1,000$ loci. However, the method must recompute impurity at each split, which costs $\Theta(n)$ and suggests CART may not scale to genome-wide data sets with such a high cost per locus.

7.1.6 Set-Association

Parameters I evaluated the performance of the original authors' implementation of Set-Association, sumstatS [92, 39]. Relevant analysis parameters which may be specified include the maximum number of terms in a sum of statistics and the number of permutation samples to draw. I performed analysis using the authors' defaults, allowing $N = 10$ terms in a sum and $P = 2,000$ permutations.

Statistical Performance Set-Association performs a scan for marginal association signals and attempts to identify locus heterogeneity by summing test statistics. It is designed to exploit multiple independent measures of association, such as a standard χ^2 test for association, as well as a χ^2 test for Hardy-Weinberg Disequilibrium. Unfortunately, the simulated data do not include alternative hints of association, so sumstatS must rely only on allelic association.

Even though the data were not generated in a manner to most clearly demonstrate the power of the Set-Association method, the algorithm performed well under nearly all circumstances (Table 7.11). It successfully identified all effects with any marginal signal and maintained a predictable 1% false positive rate. Its only shortcoming is an apparent inability to identify interactions displaying no main effect, which is expected.

Table 7.9 : Classification tree performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	100.0	0.0	100.0	100.0	100.0
Specificity (%)	99.8	99.8	100.0	99.8	99.6	99.6
Expected False Positives	2,005	2,006	15	2,004	4,020	4,031

Table 7.10 : Classification tree cost in 1,000 loci. Cost is dominated by calculating single-locus statistics, resulting in nearly predictable runtime expense.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	22.3	22.0	21.8	22.1	22.2	22.4
Real (s)	22.6	22.4	22.1	22.4	22.5	22.7

Table 7.11 : Set-association performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	100.0	1.1	100.0	100.0	100.0
Specificity (%)	99.0	99.0	99.0	99.0	99.0	99.0
Expected False Positives	10,020	10,020	10,020	10,020	10,040	10,040

Computational Costs The cost of Set-Association is linear in the number of markers analyzed. However, this analysis hides a large constant associated with determining the significance of the N terms through P permutation samples. When n is of the same order of magnitude as P, the cost of determining significance can be a relatively large proportion of runtime complexity.

Empirical runtime measurements were unsurprising (Table 7.12). sumstatS' cost was similar to other methods employing a greedy search and permutation-based determination of significance.

7.1.7 BEAM

Parameters BEAM has numerous parameters whose optimization could be the subject of an ambitious study. I pursued recommendations by the authors, which proved difficult due to contradictions between configuration file documentation and program defaults. In the present study, configuration options specified by the author in her software distribution [99] were used, which resulted in automatic detection of reasonable parameters. For a data set of L markers, BEAM runs $10 \times L$ burn-in

Table 7.12 : Set-association cost in 1,000 loci. In all cases, the method tests each locus once, resulting in predictable runtime costs.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	818.3	818.0	817.5	817.2	819.2	818.7
Real (s)	824.3	824.0	823.6	823.3	825.2	824.6

updates, $L \times \max(L, 100)$ MCMC updates, and executes L updates between posterior draws.

Statistical Performance The performance of BEAM depended highly on the generative model (Table 7.13). While it correctly identified heterogeneous effects in every data set, epistatic models proved to be a challenge, particularly when the population contained lethal genotypes. The method had reasonable specificity.

Computational Costs BEAM was by far the slowest non-exhaustive method evaluated (Table 7.14). Even a competently implemented all-pairs search required over seven times less CPU time and over 55 times less clock time (see AGR with pure epistatic model in Table 7.2).

7.1.8 SNPHarvester

Parameters SNPHarvester is implemented by the authors in a very difficult to use Java program [100, 96]. To facilitate the present study, I made minor modifications to allow for the changes in parameters without recompilation and added support for an additional input format. In the SNPHarvester analysis, interactions up to size two were sought using a χ^2 measure of interaction. Interactions with Bonferroni-corrected probability $p_B < 0.01$ were retained by *PathSeeker*. The default value of *SuccessiveRun* = 20 was used for the number of consecutive calls to *PathSeeker* which return no significant SNP sets.

Table 7.13 : BEAM performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	14.0	99.6	0.0	100.0	100.0	90.7
Specificity (%)	100.0	99.0	100.0	99.8	99.6	99.6
Expected False Positives	438	9.895	56	2,004	4,020	3,928

Table 7.14 : BEAM cost in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	471.2	533.4	477.1	506.3	519.8	533.6
Real (s)	483.9	534.0	477.5	509.8	520.4	534.4

Statistical Performance SNPHarvester performed very well under all tested circumstances (Table 7.15). The method had reasonable power to detect all causative loci under all models without excessively many spurious results.

Computational Costs The majority of SNPHarvester's expense is in a few calls to *PathSeeker*, which whose cost is directly proportional to the number of SNPs n under study. Under the reasonable assumption that a small number of SNPs are associated with the phenotype under study relative to the size of the available marker set, repeated invocations to *PathSeeker* by *SNPHarvester* can be regarded as a constant factor, resulting in an effectively linear overall runtime complexity. This analysis is supported by the empirical measurements outlined in Table 7.16. In each case, cost was a factor of *SuccessiveRun* more expensive than a locus-by-locus search as implemented by PLINK [66].

7.1.9 Logic Regression

Parameters There are two types of logic regression that have been previously applied to genomic data. Here, I used the logicFS version made available as a Bioconductor package [106, 119]. I applied simulated annealing with upper and lower annealing chain temperatures $start = 2$ and $end = -2$, respectively, and $10,000$ annealing iterations. In 20 logicFS iterations, one tree was constructed with up to 10 leaves.

logicFS software provides a dummy coding that splits a single genotype variable into two allele values. While flexible, such a transformation may lead to excessively

Table 7.15 : SNPHarvester performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	99.2	100.0	99.2	100.0	100.0	100.0
Specificity (%)	99.8	99.8	99.8	99.8	99.6	99.6
Expected False Positives	2,003	2,016	2,003	2,015	4,020	4,026

Table 7.16 : SNPHarvester cost in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	21.7	22.0	23.0	21.8	21.7	21.7
Real (s)	22.0	22.3	23.3	22.1	22.0	21.9

large trees or other problems. Therefore, I tested two further codings which explicitly model dominant and recessive effects.

Statistical Performance logicFS had fair power to detect true associations in most models (Table 7.17). The sensitivity in each model depended largely on the coding used, varying as much as 49.2% between the best and worst (4-locus heterogeneous model). In all cases, logicFS had poor specificity. The study predicts over $10,000$ false positive associations in a million-marker genome scan.

Computational Costs Logic regression's performance was not buoyed by reduced runtime complexity (Table 7.18). To analyze $1,000$ markers required about 100 seconds, which places it behind only BEAM as the second slowest non-exhaustive method evaluated.

7.1.10 Random Forests

Parameters Random forests [107] analysis was carried out using the RandomJungle package [109]. For each data set, $t = 500$ trees were constructed based on $\sqrt{n}$ markers, where n is the total number of available loci. Every tree is grown as large as possible using the CART algorithm [78] and the importance of each variable is quantified by its Gini index.

RandomJungle is multi-threaded software. To obtain realistic real time measurements, eight CPU cores were allocated for analysis.

Table 7.17 : Logic regression performance in 1,000 loci using (a) dummy, (b) dominant, and (c) recessive coding.

(a) **Dummy coding**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	99.4	100.0	8.8	98.6	79.2	94.8
Specificity (%)	98.6	98.3	97.2	98.7	98.8	98.7
Expected False Positives	13,909	16,643	28,321	12,809	11,697	12,572

(b) **Dominant coding**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	91.0	100.0	4.8	50.4	50.5	58.0
Specificity (%)	97.9	98.4	97.2	98.1	97.9	97.9
Expected False Positives	21,326	16,239	27,558	19,169	20,567	21,224

(c) **Recessive coding**	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	76.8	4.2	99.6	79.2	93.8
Specificity (%)	98.9	97.7	97.2	98.8	98.8	98.9
Expected False Positives	11,223	22,674	27,658	12,119	11,626	10,983

Statistical Performance Random forests were able to correctly determine all susceptibility loci with few spurious results for all models exhibiting marginal association (Table 7.15). Like many other algorithms, random forests were unable to detect pure interactions, though they seldom implicated unassociated markers.

Computational Costs Random forests fit the largest possible classification tree to a data set consisting of $\sqrt{n}$ markers. Since binary split computation requires determining node impurity for each marker not used in an earlier split, complexity is linear in the number of loci under study. One may expect empirical running times to be more expensive than that of a linear search as realized by PLINK [66] by a factor of t, but in reality, RandomJungle is much faster (Table 7.20). I attribute this to a particularly thoughtful and efficient implementation of random forests, which is further sped up by parallelization.

Table 7.18 : Logic regression cost in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	103.9	102.2	104.6	103.2	102.7	103.5
Real (s)	106.4	104.6	107.5	105.7	105.2	106.1

Table 7.19 : Random forest performance in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
Sensitivity (%)	100.0	100.0	4.8	100.0	100.0	100.0
Specificity (%)	99.8	99.8	100.0	99.8	99.6	99.6
Expected False Positives	2,214	2,098	465	2,095	4,113	4,224

Table 7.20 : Random forest cost in 1,000 loci.

	Epistasis			Heterogeneity		
	Lethal	Threshold	Pure	2-locus	4-locus	Mixed
User (s)	15.2	14.4	15.6	14.7	15.1	15.4
Real (s)	3.1	2.9	3.2	3.0	3.1	3.1

7.2 10,000 Loci

The human genome consists of approximately 3×10^9 base pairs. Even given the correlation of nearby sequences, a scan of 1,000 loci would be woefully inadequate to detect association without prior information. Indeed, data sets with more than 10^6 SNPs and copy number variations are currently generated with increased frequency as genotyping costs fall.

While it is not feasible in the present context to conduct a power study of the above methods with such large data sets, it may be useful to evaluate the performance of a few methods with more than 1,000 loci. To this end, I selected three algorithms based on their statistical performance and computational tractability. In particular, I chose AGR, random forests, and SNPHarvester for evaluation with 10,000 loci under the models presented in Chapter 6.

Statistical Performance The performance of the three selected methods is presented in Table 7.21. The algorithms exhibited no appreciable difference when analyzing the 2-locus heterogeneous or the threshold or lethal genotype epistatic models. Random forests, however, suffered from slightly reduced sensitivity under the 4-locus heterogeneous model (93.1%) and poor sensitivity to detect associated loci in the mixed and pure epistasis models.

The inability of random forests to consistently detect susceptibility loci generated under the mixed model may come as a surprise, as the method had perfect sensitivity when considering the smaller 1,000-locus data sets. The negative impact of a larger data set with comparatively more noise serves to emphasize the shortcomings of stochastic methods, which may only consider a small number of features at once: the set of associated loci must be selected simultaneously at random. As the number of unassociated loci and susceptibility loci increases, the probability of such an event is diminished.

Computational Performance Computationally, the selected methods performed predictably (Table 7.22). AGR was extremely fast in every case in which exhaustive pairwise enumeration was not requested, requiring only half a second to input, process, and report on 10,000 SNPs. The complexity of random forests does not depend on the data under study. Consequently, its linear runtime was practically invariant over all models. SNPHarvester was also consistent in its cost, though it took noticeably more time to analyze the data set containing a pure interaction. One possible explanation for this phenomenon is that *PathSeeker* identifies few interacting SNPs at each invocation, requiring evaluation of more markers during further calls.

7.3 Summary

In this chapter, I described an evaluation of the methods presented in Chapters 3 and 4 using data generated by SimGE (Chapter 5) under the models outlined in Chapter 6. Initially, only 1,000 loci were generated in order to identify which methods might be capable of feasibly detecting multi-locus association at a genome-wide scale.

Based on their performance in the initial small data sets, I selected three algorithms for further study using 10,000-locus simulated data sets. Of these, association graph reduction and SNPHarvester showed markedly superior performance over ran-

Table 7.21 : Performance of association graph reduction (AGR), random forests (RH), and SNPHarvester (SH) in 10,000 SNP data sets. Here, *SNS* is the sensitivity of the method, *SPC* the specificity, and *FP* the expected number of false positives in a study comprised of 1,000,000 markers with a similar generative process.

		AGR	RF	SH
2-locus	*SNS*	100.0	100.0	100.0
	SPC	99.8	99.8	99.8
	FP	2,024	2,004	2,014
4-locus	*SNS*	100.0	93.1	100.0
	SPC	99.6	99.6	99.6
	FP	4,037	3,738	4,024
Threshold	*SNS*	100.0	100.0	100.0
	SPC	99.8	99.8	99.8
	FP	2,029	2,004	2,017
Lethal	*SNS*	100.0	100.0	100.0
	SPC	99.8	99.8	99.8
	FP	2,033	2,004	2,013
No Marginal	*SNS*	97.4	0.0	100.0
	SPC	99.8	100.0	99.8
	FP	1,952	0	2,013
Mixed	*SNS*	100.0	50.3	99.8
	SPC	99.6	99.8	99.6
	FP	4,036	2,002	4,020

Table 7.22 : Performance of association graph reduction, random forests, and SNPHarvester in 10,000 SNP data sets.

		Epistasis			Heterogeneity		
		Lethal	Threshold	Pure	2-locus	4-locus	Mixed
AGR	User (s)	1.3	1.2	6275.5	1.2	1.2	1.3
	Real (s)	0.5	0.5	801.2	0.5	0.5	0.5
RF	User (s)	53.6	52.6	53.8	53.0	50.9	51.2
	Real (s)	12.2	11.9	12.2	12.1	11.2	11.4
SH	User (s)	233.2	233.7	250.1	234.2	235.1	234.8
	Real (s)	234.1	234.5	250.9	235.1	236.0	235.8

dom forests, which was incapable of identifying pure interactions. In Chapter 8, I will discuss these methods further and propose possible improvements to each as future work.

Chapter 8

Summary

Many common phenotypes are caused in part or entirely by genetic mutations. The identification of these changes is the goal of the genotype-phenotype problem. In most cases, this task is made difficult by the complexity and redundancy of cellular processes. Due to the action of evolutionary selection, few easy to identify single-gene deleterious phenotypes exist. Rather, many genetic lesions collude to increase risk either linearly (locus heterogeneity) or non-linearly through interactions (epistasis).

Identifying genotype-phenotype associations can be computationally intractable, depending on the methods used to investigate them. For example, any attempt to discover k-way interactions must somehow cope with the enormity of the space of all such combinations. Indeed, exhaustive enumeration of all combinations of markers in a genome-wide association study is likely to remain infeasible for $k > 2$ (Table 2.1). Instead, practical algorithms must resort to partial searches of this space by making use of greediness or randomness.

Even these approaches may be problematic. While greedy strategies can bring the running time of an algorithm arbitrarily close to linear in the number of markers, they necessarily rely on marginal and low-order interaction effects to identify participants of an interaction. However unlikely in real data, some generative models predict little or no marginal effect, which suggests that approximate searches may fail to find pure interactions. Stochastic algorithms face similar challenges. Rather than climbing a hill of significance, most rely on randomly selecting the correct group of interacting markers. As the number of observable loci increases, the chance of such an event decreases. No matter the method, exhaustive, greedy, or stochastic, the genotype-phenotype association problem is fraught with computational difficulty.

8.1 Contributions

In this thesis, I have described four contributions representing progress in addressing the genotype-phenotype association problem. First, I performed an in-depth literature review and classified 30 methods designed to detect multi-locus associations according to their search strategy (Figure 3.1). Detailed descriptions of a subset of these methods chosen to represent many classes of algorithms appear in Chapter 3. Having familiarized myself with the state of the art, I designed a novel approach called association graph reduction (AGR). To determine its utility, I implemented a multi-locus genotype simulator and conducted performance testing of AGR and ten other methods.

Association Graph Reduction AGR is a novel approach to the genotype-phenotype association problem. The purpose of an association graph is represent relevant knowledge about associations between individual loci or combinations thereof. After construction, the graph is reduced by contracting edges between vertices representing loci with evidence of interaction.

Through the use of α_v and α_e vertex and edge inclusion thresholds, the proportion of the interaction space examined is finely controlled. In the case of χ^2 tests where the resulting p-value is the quantity of interest, setting $\alpha_v = \alpha_e = 1$ results in an exhaustive pairwise search of all markers followed by complete testing of combined vertices with all others. Alternatively, setting $\alpha_v = 0.05$ and $\alpha_e = 1$ would cause AGR to test only 0.05^2 of all possible pairwise interactions. This parameterization provides flexibility that sets my method apart from most others. I have created a very efficient multi-threaded implementation of AGR. Testing $10,000$ markers required only half a second when $\alpha_v = \alpha_e = 0.1$, which are thresholds with high power to recover all but pure interactions.

SimGE To address the lack of genotype simulation software capable of producing epistatic effects, I have introduced SimGE. This Java-based program can quickly synthesize genome-sized data sets with markers generated under any association model that can be expressed as a set of phenotype probabilities conditional on a multi-locus genotype. To introduce realistic errors, SimGE can apply noise to already produced genotypes. The system is already capable of introducing genotyping errors and missing data and can be easily extended to provide markers in linkage disequilibrium with susceptibility loci.

Evaluation of Methods I have evaluated eleven algorithms designed to detect multi-locus association. These were tested for their ability to quickly detect associated loci while minimizing false positives. Initially, each method was applied to $1,000$ unrealistically small $1,000$-locus data sets generated under six genetic models with the aim of eliminating from further consideration algorithms which were too inaccurate or slow. Surprisingly, eight of the eleven methods evaluated had less than 90% sensitivity for at least one genetic model. Others test only pairwise interactions, which can severely reduce the chance of identifying epistasis. While a few were particularly slow, none of the remaining three was removed consideration for poor performance. Based on this analysis, I selected the AGR, random forest, and SNPHarvester methods for further consideration.

Having identified three methods fast and accurate enough to merit further consideration, I evaluated their performance on $1,000$ random $10,000$-locus data sets generated under the same genetic models. With relatively fewer susceptibility loci present when compared to the size of the entire data set, the sensitivity of random forests suffered. The method builds classification trees based on a small random set of the input variables without regard to prior information or marginal association. As a result, the models with more loci (*4-locus* heterogeneous and *mixed*) as well as that displaying no marginal statistical significance were more difficult. Given the large number of susceptibility loci already associated with many common diseases with complex etiology, the random forests algorithm appears ill-suited to detect multi-locus genotype-phenotype associations in genome-wide data.

In contrast, AGR and SNPHarvester performed well under all models studied. Both attained perfect sensitivity in five of six models. In the remaining cases, SNPHarvester had 99.8% sensitivity in the *mixed* model and AGR 97.4% in the presence of pure epistasis. SNPHarvester's loss of power may be attributed to the large size of the *mixed* model, which incorporates epistasis. It is possible that, in a few trials, *PathSeeker* paired an independently significant locus with one enhanced by interaction. As a result, the remaining interaction locus may not be identified. While shy of perfection, AGR's 97.4% sensitivity in detecting susceptibility loci displaying no main effect is impressive. Fewer than 3% of true positive results were missed, in spite of the application of conservative Bonferroni correction.

After careful evaluation, I conclude that AGR and SNPHarvester are likely to provide the best performance in genome-wide SNP data sets. Both are reasonably fast, with SNPHarvester requiring about 4 CPU-minutes and AGR about 1.2 CPU-seconds when edge and vertex thresholds $\alpha_v = \alpha_e = 0.1$ are used. Statistical performance is comparable in nearly all circumstances, though SNPHarvester may suffer when many risk loci are present and AGR requires exhaustive evaluation in the absence of main effects. However, given a notable lack of published instances of pure epistasis, AGR appears to be the best and fastest all-around method to detect multi-locus association.

In addition to providing good statistical power and speed, AGR provides insight into the roles loci play in interactions. The result of an association graph reduction is a weight-sorted list of vertices, whose top entries may exceed the experiment-wide significance level. When taken together, these may be interpreted as a linear model. For example, if the weights of the three vertices containing markers (w, x), y, and z exceed the experiment-wide significance level, we may view this as evidence of epistasis between w and x as well as heterogeneity involving that pair and the remaining loci, suggesting the model $P \sim wx + y + z$ for phenotype P. Such extra information is unavailable in every other method tested.

8.2 Future Work

The efforts of the current study provide several possibilities for enhancement and further research. Specifically, this future work may come in three areas. First and most easily, SimGE can be improved and made into a truly remarkable software package. Next, investigation into AGR parameters and improvements to other methods may be explored. Finally, the most promising methods should be applied to real-world data.

SimGE is already a fast, advanced, and easy to use genotype simulator. However, like most software products, it may be improved. While it is already efficient enough to generate genome-sized data sets in a short amount of time, it could be made much faster. Improvements to internal data structures, more refined data types, and parallelization may vastly decrease the time required to generate data, making it easier to create large numbers of large synthetic data sets. Such improvements in speed may prove particularly important as real data sets grow larger. To make the data more realistic, additional data corruption post-processing steps should be implemented. Most importantly, markers in linkage disequilibrium (LD) should replace directly observed susceptibility loci. In actual studies, researchers are seldom lucky enough to test causal variants by chance. Rather, they must rely on strong LD to identify markers near causative loci. Finally, haplotypes may be created by incorporating HapMap [120] structure according to an existing algorithm [121]. Incorporation of the HapMap data will allow for the accurate assessment of tests with realistic and varying levels of correlation in the observed data.

Though I have investigated several AGR parameterizations and choices for those parameters, the consequences of such choices deserve further research. In Chapter 4, I presented two types of edge weights, those that take into account phenotypic status and those relying only on genotypes. Additionally, a genotypes-only measure may look for strong dependence in cases only or a difference in dependence between cases and controls. While each combination of these edge weight parameters has been implemented and informally tested, no rigorous comparison or determination of their strength and weaknesses exists. Also, the effects of α_v and α_e inclusion thresholds should be quantified conditional on the model of genetic etiology. While an F-test provides effective control for model complexity, other options should be explored. Finally, with so many genotype-phenotype association test statistics available (notably entropy-based statistics such as [75] and [76]), weight measures other than χ^2 should be investigated.

One relatively unknown method showed particular promise deserving of further development. While it has been used little by parties other than its authors, SNPHarvester performed very well in the simulated tests presented in Chapter 7. The method is a stochastic search which chooses as its starting points k randomly selected SNPs for a user-specified value of k. Since genes interact in biologically meaningful ways, it is reasonable to assume that prior biological knowledge can be profitably used to bias this random selection. Specifically, I would like to investigate the effect of integrating expert knowledge from interaction databases such as GeneNetwork [122]. By comparing any changes induced by this bias to those from random assortments of this gene-gene interaction network, I might quantify possible improvements in the quality of results identified by the method.

As explained in Chapter 5, practical circumstances led me to make use of only synthetic data to compare existing methods and establish the utility of AGR. In particular, there are few instances of published statistical epistasis, most likely due to a lack of effort to discover the phenomenon. Worse, I am unaware of any publicly available genome-wide genotype-phenotype data sets containing widely-accepted interaction effects. Without real-world benchmark data, little choice remains other than to synthesize data with known generative models.

Having identified a small set of methods showing promise to identify locus heterogeneity and epistasis, it may prove useful to apply these algorithms to real-world data. Recently, genome-wide data have been made available through large scale multi-phenotype studies such as the Wellcome Trust Case Control Consortium [8] and the Database of Genotypes and Phenotypes [123]. Evaluating several such data sets with the set of best methods may provide interesting insights into the phenotypes under study and the methods, themselves. Effects found by most or all methods may represent robust associations worthy of further scrutiny, while discrepancies may point to shortcomings in one or more method. Disagreements among methods should be given intense examination.

8.3 Conclusion

Many familiar phenotypes are caused in part by genetic anomalies. Among these are hypertension, arthritis, and cancer. While biomedical research has made progress in diagnosis and treatment of many such conditions, the underlying causes often remain a mystery. Deciphering the complex etiology of such diseases through genetic studies with tools such as association graph reduction may provide valuable insights into their causes. With such information, researchers may develop vastly improved diagnostic tools and treatments, extending and improving the lives of the affected.

Bibliography

[1] J. Ott, *Analysis of Human Genetic Linkage.* Baltimiore: Johns Hopkins University Press, third ed., 1999.

[2] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," *Genet Med*, vol. 4, pp. 45–61, Jan 2002.

[3] E. Rogaeva, Y. Meng, J. H. Lee, Y. Gu, T. Kawarai, F. Zou, T. Katayama, C. T. Baldwin, R. Cheng, H. Hasegawa, F. Chen, N. Shibata, K. L. Lunetta, R. Pardossi-Piquard, C. Bohm, Y. Wakutani, L. A. Cupples, K. T. Cuenco, R. C. Green, L. Pinessi, I. Rainero, S. Sorbi, A. Bruni, R. Duara, R. P. Friedland, R. Inzelberg, W. Hampe, H. Bujo, Y.-Q. Song, O. M. Andersen, T. E. Willnow, N. Graff-Radford, R. C. Petersen, D. Dickson, S. D. Der, P. E. Fraser, G. Schmitt-Ulms, S. Younkin, R. Mayeux, L. A. Farrer, and P. S. George-Hyslop, "The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease," *Nat Genet*, vol. 39, pp. 168–77, Feb 2007.

[4] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Bitton, H. Yang, S. Targan, L. W. Datta, E. O. Kistner, L. P. Schumm, A. T. Lee, P. K. Gregersen, M. M. Barmada, J. I. Rotter, D. L. Nicolae, and J. H. Cho, "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene," *Science*, vol. 314, pp. 1461–3, Dec 2006.

[5] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, “Complement factor H polymorphism in age-related macular degeneration,” *Science*, vol. 308, pp. 385–9, Apr 2005.

[6] The International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN), J. B. Harley, M. E. Alarcón-Riquelme, L. A. Criswell, C. O. Jacob, R. P. Kimberly, K. L. Moser, B. P. Tsao, T. J. Vyse, C. D. Langefeld, S. K. Nath, J. M. Guthridge, B. L. Cobb, D. B. Mirel, M. C. Marion, A. H. Williams, J. Divers, W. Wang, S. G. Frank, B. Namjou, S. B. Gabriel, A. T. Lee, P. K. Gregersen, T. W. Behrens, K. E. Taylor, M. Fernando, R. Zidovetzki, P. M. Gaffney, J. C. Edberg, J. D. Rioux, J. O. Ojwang, J. A. James, J. T. Merrill, G. S. Gilkeson, M. F. Seldin, H. Yin, E. C. Baechler, Q.-Z. Li, E. K. Wakeland, G. R. Bruner, K. M. Kaufman, and J. A. Kelly, “Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci,” *Nat Genet*, vol. 40, pp. 204–10, Feb 2008.

[7] D. E. Arking, A. Pfeufer, W. Post, W. H. L. Kao, C. Newton-Cheh, M. Ikeda, K. West, C. Kashuk, M. Akyol, S. Perz, S. Jalilzadeh, T. Illig, C. Gieger, C.-Y. Guo, M. G. Larson, H. E. Wichmann, E. Marbán, C. J. O’Donnell, J. N. Hirschhorn, S. Kääb, P. M. Spooner, T. Meitinger, and A. Chakravarti, “A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization,” *Nat Genet*, vol. 38, pp. 644–51, Jun 2006.

[8] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–78, Jun 2007.

[9] *Online Mendelian Inheritance in Man, OMIM ™.* Baltimore, MD: Johns Hopkins University. MIM Number: #155555: 21. October, 2009. World Wide Web URL: `http://www.ncbi.nlm.nih.gov/omim/`.

[10] P. Valverde, E. Healy, I. Jackson, J. L. Rees, and A. J. Thody, "Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans," *Nat Genet*, vol. 11, pp. 328–30, Nov 1995.

[11] P. Sulem, D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, K. P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, M. Jakobsdottir, S. Steinberg, S. Pálsson, F. Jonasson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K. R. Benediktsdottir, K. K. Aben, L. A. Kiemeney, J. H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson, "Genetic determinants of hair, eye and skin pigmentation in europeans," *Nat Genet*, vol. 39, pp. 1443–52, Dec 2007.

[12] S. K. Nath, J. Kilpatrick, and J. B. Harley, "Genetics of human systemic lupus erythematosus: the emerging picture," *Curr Opin Immunol*, vol. 16, pp. 794–800, Dec 2004.

[13] O. A. Igoshin, M. S. Brody, C. W. Price, and M. A. Savageau, "Distinctive topologies of partner-switching signaling networks correlate with their physio-

logical roles," *J Mol Biol*, vol. 369, pp. 1333–52, Jun 2007.

[14] O. Carlborg and C. S. Haley, "Epistasis: too often neglected in complex trait studies?," *Nat Rev Genet*, vol. 5, pp. 618–25, Aug 2004.

[15] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum Hered*, vol. 56, pp. 73–82, Jan 2003.

[16] P. C. Phillips, "Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems," *Nat Rev Genet*, vol. 9, pp. 855–67, Nov 2008.

[17] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Hum Mol Genet*, vol. 11, pp. 2463–8, Oct 2002.

[18] A. Balmain and C. C. Harris, "Carcinogenesis in mouse and human cells: parallels and paradoxes," *Carcinogenesis*, vol. 21, pp. 371–7, Mar 2000.

[19] J. A. Staessen, J. G. Wang, E. Brand, C. Barlassina, W. H. Birkenhäger, S. M. Herrmann, R. Fagard, L. Tizzoni, and G. Bianchi, "Effects of three candidate genes on prevalence and incidence of hypertension in a caucasian population," *J Hypertens*, vol. 19, pp. 1349–58, Aug 2001.

[20] T. F. Mackay, "Quantitative trait loci in drosophila," *Nat Rev Genet*, vol. 2, pp. 11–20, Jan 2001.

[21] S. M. Williams, J. L. Haines, and J. H. Moore, "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*, vol. 26, pp. 170–9, Feb 2004.

[22] E. Routman and J. Cheverud, "Gene effects on a quantitative trait: two-locus epistatic effects measured at microsatellite markers and at estimated QTL," *Evolution*, vol. 51, pp. 1654–1662, Jan 1997.

[23] D. Segrè, A. Deluna, G. M. Church, and R. Kishony, "Modular epistasis in yeast metabolism," *Nat Genet*, vol. 37, pp. 77–83, Jan 2005.

[24] R. Y. L. Zee, J. Hoh, S. Cheng, R. Reynolds, M. A. Grow, A. Silbergleit, K. Walker, L. Steiner, G. Zangenberg, A. Fernandez-Ortiz, C. Macaya, E. Pintor, A. Fernandez-Cruz, J. Ott, and K. Lindpainter, "Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease," *Pharmacogenomics J*, vol. 2, pp. 197–201, Jan 2002.

[25] S. M. Williams, M. D. Ritchie, J. A. Phillips, E. Dawson, M. Prince, E. Dzhura, A. Willis, A. Semenya, M. Summar, B. C. White, J. H. Addy, J. Kpodonu, L.-J. Wong, R. A. Felder, P. A. Jose, and J. H. Moore, "Multilocus analysis of hypertension: a hierarchical approach," *Hum Hered*, vol. 57, pp. 28–38, Jan 2004.

[26] C. Tsai, L. Lai, J. Lin, F. Chiang, and J. Hwang, "Renin-angiotensin system gene polymorphisms and atrial fibrillation," *Circulation*, Jan 2004.

[27] Y. M. Cho, M. D. Ritchie, J. H. Moore, J. Y. Park, K.-U. Lee, H. D. Shin, H. K. Lee, and K. S. Park, "Multifactor-dimensionality reduction shows a two-

locus interaction associated with type 2 diabetes mellitus," *Diabetologia*, vol. 47, pp. 549–54, Mar 2004.

[28] C. F. Sing and J. Davignon, "Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation," *Am J Hum Genet*, vol. 37, pp. 268–85, Mar 1985.

[29] K. E. Zerba, R. E. Ferrell, and C. F. Sing, "Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein e (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits," *Hum Genet*, vol. 107, pp. 466–75, Nov 2000.

[30] E. L. Heinzen, W. Yoon, S. K. Tate, A. Sen, N. W. Wood, S. M. Sisodiya, and D. B. Goldstein, "Nova2 interacts with a cis-acting polymorphism to influence the proportions of drug-responsive splice variants of SCN1A," *Am J Hum Genet*, vol. 80, pp. 876–83, May 2007.

[31] H. Kallberg, L. Padyukov, R. M. Plenge, J. Ronnelid, P. K. Gregersen, A. H. M. van der Helm-van Mil, R. E. M. Toes, T. W. Huizinga, L. Klareskog, L. Alfredsson, and E. I. of Rheumatoid Arthritis study group, "Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis," *Am J Hum Genet*, vol. 80, pp. 867–75, May 2007.

[32] C. C. Chen, R. B. Lu, Y. C. Chen, M. F. Wang, Y. C. Chang, T. K. Li, and S. J. Yin, "Interaction between the functional polymorphisms of the alcohol-

metabolism genes in protection against alcoholism," *Am J Hum Genet*, vol. 65, pp. 795–807, Sep 1999.

[33] E. Levy-Lahad, R. Catane, S. Eisenberg, B. Kaufman, G. Hornreich, E. Lishinsky, M. Shohat, B. L. Weber, U. Beller, A. Lahad, and D. Halle, "Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families," *Am J Hum Genet*, vol. 60, pp. 1059–67, May 1997.

[34] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: limits of models displaying no main effect," *Am J Hum Genet*, vol. 70, pp. 461–71, Feb 2002.

[35] R. J. Neuman and J. P. Rice, "Two-locus models of disease," *Genet Epidemiol*, vol. 9, pp. 347–65, Jan 1992.

[36] W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models," *Hum Hered*, vol. 50, pp. 334–49, Jan 2000.

[37] W. N. Frankel and N. J. Schork, "Who's afraid of epistasis?," *Nat Genet*, vol. 14, pp. 371–3, Dec 1996.

[38] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nat Genet*, vol. 37, pp. 413–7, Apr 2005.

[39] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in

human case-control association studies," *Genome Res*, vol. 11, pp. 2115–9, Dec 2001.

[40] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, pp. 289–300, Jan 1995.

[41] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc Natl Acad Sci USA*, vol. 100, pp. 9440–5, Aug 2003.

[42] C. Sabatti, S. Service, and N. Freimer, "False discovery rate in linkage and association genome screens for complex disorders," *Genetics*, vol. 164, pp. 829–33, Jun 2003.

[43] D. Altman and P. Royston, "What do we mean by validating a prognostic model?," *Statistics in medicine*, Jan 2000.

[44] H. K. Tiwari and R. C. Elston, "Deriving components of genetic variance for multilocus models," *Genet Epidemiol*, vol. 14, pp. 1131–6, Jan 1997.

[45] H. K. Tiwari and R. C. Elston, "Restrictions on components of variance for epistatic models," *Theoretical population biology*, vol. 54, pp. 161–74, Oct 1998.

[46] L. Breiman, "Heuristics of instability and stabilization in model selection," *The annals of statistics*, vol. 24, pp. 2350–2383, Dec 1996.

[47] M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict

quantitative trait variation," *Genome Res*, vol. 11, pp. 458–70, Mar 2001.

[48] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Am J Hum Genet*, vol. 69, pp. 138–47, Jul 2001.

[49] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, pp. 376–82, Feb 2003.

[50] J. H. Moore, "Multifactor Dimensionality Reduction." `http://www.multifactordimensionalityreduction.org/`, 2003–2009. Accessed 21. May, 2009.

[51] L. W. Hahn and J. H. Moore, "Ideal discrimination of discrete clinical endpoints using multilocus genotypes," *In Silico Biol (Gedrukt)*, vol. 4, pp. 183–94, Jan 2004.

[52] S. Theodoridis and K. Koutroumbas, *Pattern recognition.* Academic Press, Jan 2006.

[53] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, Jan 2007.

[54] M. D. Ritchie, L. W. Hahn, and J. H. Moore, "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping

error, missing data, phenocopy, and genetic heterogeneity," *Genet Epidemiol*, vol. 24, pp. 150–7, Feb 2003.

[55] W. Y. S. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd, "Genome-wide association studies: theoretical and practical concerns," *Nat Rev Genet*, vol. 6, pp. 109–18, Feb 2005.

[56] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics (Oxford, England)*, vol. 9, pp. 30–50, Jan 2008.

[57] C. S. Coffey, P. R. Hebert, M. D. Ritchie, H. M. Krumholz, J. M. Gaziano, P. M. Ridker, N. J. Brown, D. E. Vaughan, and J. H. Moore, "An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation," *BMC Bioinformatics*, vol. 5, p. 49, Apr 2004.

[58] K. Pattin, B. White, N. Barney, J. Gui, H. Nelson, K. Kelsey, A. Andrew, M. Karagas, and J. Moore, "A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction," *Genet Epidemiol*, vol. 33, pp. 87–94, Jul 2008.

[59] W. S. Bush, T. L. Edwards, S. M. Dudek, B. A. McKinney, and M. D. Ritchie, "Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction," *BMC Bioinformatics*, vol. 9, p. 238, Jan 2008.

[60] H. Mei, M. L. Cuccaro, and E. R. Martin, "Multifactor dimensionality

reduction-phenomics: a novel method to capture genetic heterogeneity with use of phenotypic variables," *Am J Hum Genet*, vol. 81, pp. 1251–61, Dec 2007.

[61] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genet Epidemiol*, vol. 31, pp. 306–15, May 2007.

[62] A. Jakulin and I. Bratko, "Analyzing attribute dependencies," *Lecture notes in computer science*, vol. 2838, pp. 229–240, Jan 2003.

[63] J. A. Kelly, K. L. Moser, and J. B. Harley, "The genetics of systemic lupus erythematosus: putting the pieces together," *Genes Immun*, vol. 3 Suppl 1, pp. S71–85, Oct 2002.

[64] J. Namkung, R. Elston, J. Yang, and T. Park, "Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method," *Genet Epidemiol*, Feb 2009.

[65] I. Ionita and M. Man, "Optimal two-stage strategy for detecting interacting genes in complex diseases," *BMC Genet*, vol. 7, p. 39, Jan 2006.

[66] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559–75, Sep 2007.

[67] S. Purcell, "PLINK." `http://pngu.mgh.harvard.edu/~purcell/plink/`, 2007–2009. Accessed 21. May, 2009.

[68] J. Millstein, D. V. Conti, F. D. Gilliland, and W. J. Gauderman, "A testing framework for identifying susceptibility genes in the presence of epistasis," *Am J Hum Genet*, vol. 78, pp. 15–27, Jan 2006.

[69] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nat Rev Genet*, vol. 4, pp. 701–9, Sep 2003.

[70] Q. Yang, M. J. Khoury, F. Sun, and W. D. Flanders, "Case-only design to measure gene-gene interaction," *Epidemiology (Cambridge, Mass)*, vol. 10, pp. 167–70, Mar 1999.

[71] X. Zhang, F. Zou, and W. Wang, "FastANOVA: an efficient algorithm for genome-wide association study," in *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York), pp. 821–829, ACM, 2008.

[72] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Phil Mag Ser 5*, vol. 50, pp. 157–175, Jan 1900.

[73] X. Zhang, F. Zou, and W. Wang, "FastChi: an efficient algorithm for analyzing gene-gene interactions," in *Pacific Symposium on Biocomputing*, vol. 14, pp. 528–539, 2009.

[74] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang, "COE: A general approach for efficient genome-wide two-locus epistasis test in disease association study," in *Research in Molecular Biology: 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009: proceedings* (S. Batzoglou, ed.), (Berlin), pp. 253–269, Springer, 2009.

[75] J. Zhao, E. Boerwinkle, and M. Xiong, "An entropy-based statistic for genomewide association studies," *Am J Hum Genet*, vol. 77, pp. 27–40, Jul 2005.

[76] C. Dong, X. Chu, Y. Wang, Y. Wang, L. Jin, T. Shi, W. Huang, and Y. Li, "Exploration of gene-gene interaction effects using entropy-based methods," *Eur J Hum Genet*, vol. 16, pp. 229–35, Feb 2008.

[77] M. J. Daly and D. Altshuler, "Partners in crime," *Nat Genet*, vol. 37, pp. 337–8, Apr 2005.

[78] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Chapman & Hall, Jan 1984.

[79] B. Ripley, "tree: Classification and regression trees." `http://cran.r-project.org/web/packages/tree/index.html`, 1998-2007, 2009. Accessed 4. Sep, 2009.

[80] J. Huang, A. Lin, B. Narasimhan, T. Quertermous, C. A. Hsiung, L.-T. Ho, J. S. Grove, M. Olivier, K. Ranade, N. J. Risch, and R. A. Olshen, "Tree-structured supervised learning and the genetics of hypertension," *Proc Natl Acad Sci USA*, vol. 101, pp. 10529–34, Jul 2004.

[81] H. Zhang and G. Bonney, "Use of classification trees for association studies," *Genet Epidemiol*, vol. 19, pp. 323–32, Dec 2000.

[82] H. Zhang and B. Singer, *Recursive partitioning in the health sciences.* Springer, Jan 1999.

[83] H. Zhang, C. P. Tsai, C. Y. Yu, and G. Bonney, "Tree-based linkage and association analyses of asthma," *Genet Epidemiol*, vol. 21 Suppl 1, pp. S317–22, Jan 2001.

[84] Ye, Zhong, and Zhang, "A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking," *BMC Genet*, vol. 6 Suppl 1, p. S135, Dec 2005.

[85] H. Zhang, C.-Y. Yu, and B. Singer, "Cell and tumor classification using gene expression data: construction of forests," *Proc Natl Acad Sci USA*, vol. 100, pp. 4168–72, Apr 2003.

[86] J. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, pp. 1–67, Mar 1991.

[87] N. R. Cook, R. Y. L. Zee, and P. M. Ridker, "Tree and spline based association analysis of gene-gene interaction models for ischemic stroke," *Statistics in medicine*, vol. 23, pp. 1439–53, May 2004.

[88] X. Chen, C.-T. Liu, M. Zhang, and H. Zhang, "A forest-based approach to identifying gene and gene gene interactions," *Proc Natl Acad Sci USA*, vol. 104, pp. 19199–203, Dec 2007.

[89] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[90] R. Tibshirani and K. Knight, "Model search by bootstrap "bumping"," *Journal of Computational and Graphical Statistics*, vol. 8, pp. 671–686, Dec 1999.

[91] K. L. Lunetta, L. B. Hayward, J. Segal, and P. V. Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests," *BMC Genet*, vol. 5, p. 32, Jan 2004.

[92] J. Hoh and J. Ott, "Sumstat." `http://linkage.rockefeller.edu/ott/sumstat.html`, 2001. Accessed 26. May, 2009.

[93] I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Logic regression," *Journal of Computational and Graphical Statistics*, Jan 2003.

[94] J. Gayán, A. González-Pérez, F. Bermudo, M. E. Sáez, J. L. Royo, A. Quintas, J. J. Galan, F. J. Morón, R. Ramirez-Lorca, L. M. Real, and A. Ruiz, "A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis," *BMC Genomics*, vol. 9, p. 360, Jan 2008.

[95] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang, and W. Yu, "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study," *BMC Bioinformatics*, vol. 10, p. 13, Jan 2009.

[96] C. Yang, Z. He, Q. Yang, H. Xue, and W. Yu, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics*, Dec 2008.

[97] D. M. Nielsen, M. G. Ehm, and B. S. Weir, "Detecting marker-disease association by testing for hardy-weinberg disequilibrium at a marker locus," *Am J Hum Genet*, vol. 63, pp. 1531–40, Nov 1998.

[98] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nat Genet*, vol. 39, pp. 1167–73, Sep 2007.

[99] Y. Zhang, "BEAM." `http://www.people.fas.harvard.edu/~junliu/BEAM/`, 2007. Accessed 20. May, 2009.

[100] C. Yang, Z. He, Q. Yang, H. Xue, and W. Yu, "SNPHarvester." `http://bioinformatics.ust.hk/SNPHarvester.html`, 2008. Accessed 20. May, 2009.

[101] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, pp. 997–1004, Dec 1999.

[102] S. Macgregor, P. M. Visscher, and G. Montgomery, "Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates," *Nucleic Acids Res*, vol. 34, p. e55, Jan 2006.

[103] C. Kooperberg and I. Ruczinski, "Identifying interacting SNPs using Monte Carlo logic regression," *Genet Epidemiol*, vol. 28, pp. 157–70, Feb 2005.

[104] C. Kooperberg and I. Ruczinski, "LogicReg: Logic Regression." `http://cran.`

r-project.org/web/packages/LogicReg/index.html, 2005–2008. Accessed 22. May, 2009.

[105] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics (Oxford, England)*, vol. 9, pp. 187–98, Jan 2008.

[106] H. Schwender, "logicFS." http://bioconductor.org/packages/2.4/bioc/html/logicFS.html, 2008. Accessed 21. May, 2009.

[107] L. Breiman, "Random forests," *Machine learning*, Jan 2001.

[108] A. Liaw, M. Wiener, L. Breiman, and A. Cutler, "randomForest." http://cran.r-project.org/web/packages/randomForest/index.html, 2001-2009. Accessed 27. May, 2009.

[109] D. F. Schwarz, "Randon Jungle version 0.7.6." http://www.randomjungle.com/, 2006-2009. Accessed 19. June, 2009.

[110] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, Jan 1996.

[111] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, pp. 197–227, Jan 1990.

[112] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and computation*, vol. 121, pp. 256–285, Jan 1995.

[113] Y. Freund and R. E. Shapire, "Experiments with a new boosting algorithm," in *Maching Learning: Proceedings of the Thirteenth International Conference*,

(San Francisco), pp. 148–156, Morgan Kauffman, 1996.

[114] S. M. Dudek, A. A. Motsinger, D. R. Velez, S. M. Williams, and M. D. Ritchie, "Data simulation software for whole-genome association and other studies in human genetics," *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 499–510, Jan 2006.

[115] B. Peng, C. I. Amos, and M. Kimmel, "Forward-time simulations of human populations with complex diseases," *PLoS Genet*, vol. 3, p. e47, Mar 2007.

[116] A. Wilson, J. E. Bailey-Wilson, E. Pugh, and A. Sorant, "The genometric analysis simulation program (G.A.S.P.): a software tool for testing and investigating methods in statistical genetics," *Am J Hum Genet*, no. Suppl 59, p. A193, 1996.

[117] H. D. Daetwyler, B. Villanueva, and J. A. Woolliams, "Accuracy of predicting the genetic risk of disease using a genome-wide approach," *PLoS ONE*, vol. 3, p. e3395, Jan 2008.

[118] H. Cordell, "Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases," *Nat Rev Genet*, vol. 10, pp. 392–404, Jun 2009.

[119] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004.

[120] The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299–320, Oct 2005.

[121] J. Li and Y. Chen, "Generating samples for association studies based on HapMap data," *BMC Bioinformatics*, vol. 9, p. 44, Jan 2008.

[122] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *Am J Hum Genet*, vol. 78, pp. 1011–25, Jun 2006.

[123] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S. T. Sherry, "The NCBI dbGaP database of genotypes and phenotypes," *Nat Genet*, vol. 39, pp. 1181–6, Oct 2007.