Integrative Analysis Strategies for Discovering

Genetic Associations with Common Diseases

BY
JOEL FONTANAROSA
A.B., The University of Chicago, 2004

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

Yang Dai, Chair and Advisor
Hui Lu
Richard Magin
Bhaskar DasGupta, Computer Science
Larry Tobacman, Physiology and Biophysics
Charles Rhodes, Physics

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my Ph.D. advisor, Dr. Yang Dai, for her mentorship during my graduate training. Dr. Dai has been a very patient and thoughtful teacher, and her remarkable knowledge and insights regarding data analysis and research have provided me with numerous lessons that I will remember throughout my career. I am honored to have had the opportunity to train with her. I would like to thank the other faculty members of the Bioinformatics Program, Dr. Jie Liang and Dr. Hui Lu, who, with Dr. Dai, continue to provide a welcoming and stimulating environment for graduate studies and research. I am also most grateful for the generous support of Dr. Charles Rhodes. His perspective on scientific research is genuinely inspiring, and I am sincerely thankful for the enriching experiences he afforded me throughout my graduate studies. Additionally, I want to acknowledge the support of the Medical Scientist Training Program – in particular, Dr. Larry Tobacman and Roberta Bernstein, who have been very generous in providing guidance and support while helping to shape my graduate education. I would like to thank the other members of my thesis committee, Dr. Richard Magin and Dr. Bhaskar DasGupta, for taking the time out of their schedules to serve on my committee and for their dedication to graduate research in bioengineering and bioinformatics.

I would like to acknowledge the other members of the Dai Laboratory – especially Lei Huang and Damian Roqueiro. Throughout my thesis, they provided

helpful suggestions on my research projects, advice on numerous technical issues, and good company.

Lastly, I wish to extend special thanks to my family and friends, whose unwavering encouragement and support have been the most important to me throughout my entire education. Thank you.

JBF

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF ABBREVIATIONS

AUC	Area under the Receiver Operating Characteristic curve
BEAM	Bayesian Epistasis Association Mapping
CUDA	NVIDIA Compute Unified Device Architecture
eQTL	Expression Quantitative Trait Loci
GA	Genetic (or Evolutionary) Algorithm
GPU	Graphics Processing Unit
GWAS	Genome-Wide Association Studies
HWE	Hardy-Weinberg Equilibrium
KEGG	Kyoto Encycolpedia of Genes and Genomes
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MCMC	Markov Chain Monte Carlo
MDR	Multifactor Dimensionality Reduction
miRNA	Micro-ribonucleic Acid
MSE	Mean Squared Error
QTL	Quantitative Trait Loci
ROC	Receiver Operating Characteristic

LIST OF ABBREVIATIONS (continued)

SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machines
WTCCC	Wellcome Trust Case Control

ABSTRACT

Joel B. Fontanarosa
Department of Bioengineering
University of Illinois at Chicago
Chicago, Illinois (2013)

Dissertation Chairperson: Dr. Yang Dai

Genetic association studies have proven to be successful at identifying reliable associations with complex diseases. However, the majority of these results are uninformative with respect to any functional basis, and more research is necessary appreciate the mechanisms by which these associations are related to pathogenic molecular alterations. In this project, we propose a number of computational approaches to address current challenges in genome-wide association studies: detection of gene-gene interactions, utilization of high performance computing resources, development of genomic risk prediction tools, and investigation into miRNA-associated variations that may lead to problematic modulations in transcriptional activity. First, we present an adaptive evolutionary optimization algorithm that utilizes local linkage disequilibrium patterns to improve the search for gene-gene interactions associated with a phenotype of interest. Our method was applied to several simulated disease models and to a real genome-wide association study. The results indicate that our method has improved power and computational efficiency for uncovering gene-gene interactions relative to one of the most powerful competing methods. This optimization strategy was extended into a parallel algorithm that uses state of the art computing methods involving graphics

processing units to explore genome-wide association study data sets with maximal computational efficiency and minimal cost. Next, we present an improved penalized lasso regression strategy to build more accurate predictions of disease risk based on genomic and phenotypic information for case control studies. Using this approach on a simulated data set from the 1000 Genomes project, we were able to model disease risk using common and rare genetic variation in combination with quantitative trait information. Lastly, we present a framework for the determination of genomic variation associated with miRNA dysregulation. We applied our analysis method to several genome-wide association studies of common diseases to determine candidate targets for disease-associated dysfunctions in miRNA-related gene expression changes. The research in this thesis represents a set of computational tools and integrative analysis strategies that can be used to provide a detailed description of the genetic risk associated with a potentially complex inherited phenotype. Code developed in this project will be made available to the research community for further development and application to other genome-wide association studies.

SUMMARY

Technological advancements have allowed researchers to study common genetic diseases with an ever increasing level of detail. While genome-wide association studies have proven to be incredibly useful for the study of complex diseases, more research is necessary to appreciate the mechanisms by which these associations are related to pathogenic molecular alterations. In this project, we propose a number of computational approaches to address current challenges in genome-wide association studies: detection of gene-gene interactions, utilization of high performance computing resources, development of genomic risk prediction tools, and investigation into miRNA-associated variations that may lead to problematic modulations in transcriptional activity. We apply these methods to carefully designed simulated models of genetic disease as well as to case control data sets from genome-wide association studies of several diseases. The research in this thesis represents a set of computational tools and integrative analysis strategies that can be used to provide a detailed description of the genetic risk associated with a potentially complex inherited phenotype.

CHAPTER I

INTRODUCTION

Scientific understanding of human disease has evolved dramatically over the past century as a result of major advances in physiology, microbiology, biochemistry, pharmacology, and molecular biology. The discoveries in these and related areas directly led to diagnostic and therapeutic approaches that were rigorously studied and improved by clinicians and researchers throughout the twentieth century [2]. By the 1980s, molecular biology laboratory techniques were being widely used to better understand the genetic influence on cellular activity and to characterize a number of familial diseases. Since that time, an increasingly detailed and complex picture of the relationship between genetic factors and biological processes has emerged. Researchers have been able to determine the exact molecular basis for a large number of pathologic processes and phenotypes with simple Mendelian inheritance patterns. Nevertheless, many of the most important causes of morbidity and mortality in developed countries involve common diseases with complex inherited components that have defied explanation through traditional genetic analyses. Genetic research on these familial diseases (e.g. Type 2 Diabetes, Coronary Artery Disease) is undermined by the complexity of the genetic and environmental factors underlying them. These factors include both rare and common genetic variations as well as individual risks modulated by factors such as diet, age and gender. The difficulty in disentangling these risk components has limited efforts to better understand, treat, and prevent these highly prevalent conditions. The work presented

herein represents computational tools and analysis methods that aim to better understand the genetic basis for common diseases using data collected in large case control studies.

A. History of Human Gene Mapping

The concepts of intra-species variation and familial inheritance were well-recognized even in ancient civilizations. However, scientific theories and analyses regarding the origin and nature of these observations about biological variation did not occur until the nineteenth century. Modern genetics began in the early twentieth century with the rediscovery of Gregor Mendel's work in separate papers in 1900 by the German botanists Correns and de Vries [3]. The concepts of genetic inheritance introduced in this research enabled quantitative analysis of the genetic basis for biological phenotypes and contributed to the development and refinement of a detailed theoretical and statistical understanding of population genetics [3]. By 1908, the German physician Weinberg and the British mathematician Hardy [4] independently described what later became known as the Hardy-Weinberg equilibrium principle [5]. According to this principle, two alleles comprising a certain genotype will remain at the same frequency in a population in the absence of immigration, mutation, selection, non-random mating, or sampling errors. Within the next decade, researchers had described chromosomal linkage and the linear arrangement of genes [6], and the influence of inbreeding and genetic linkage on the probability of homozygosity in animal models [7]. Between 1918 and 1947, a description of the relationship between allele frequencies and the natural selection of

phenotypes in the population was developed by, most notably, Wright, Haldane, and Fisher in what became known as Modern Evolutionary Synthesis [3, 8, 9]. By the time the structure of DNA was first described in 1953 [10, 11], geneticists already had a well-described set of established principles including genetic drift, chromosomal linkage, and the relationship between allele frequencies and the natural selection of phenotypes in a population [8, 12, 13].

Concurrent with the advances in population genetics during the first half of the twentieth century, morbidity and mortality related to infectious diseases and nutritional diseases had decreased substantially in developed countries as a result of medical intervention and public health improvements [14, 15], allowing the inherited components of other common diseases in human populations to become more apparent. Using the principles of population genetics established in the first half of the century, distinctive pedigree patterns and genetic characteristics of certain inherited phenotypes (e.g. color-blindness, ABO blood groups) could be described and understood. However, direct measurement of the specific genotypic changes underlying such phenotypes would not be possible until recombinant DNA [16], hybridization [17], and sequencing [18-21] technologies were developed and refined in the 1960s and 1970s. These techniques were applied in studies conducted primarily throughout the 1980s to better understand the pathological genetic changes underlying a variety of Mendelian disease phenotypes [22-25]. By using carefully designed family studies, researchers were able to map the highly

penetrant genetic modifications responsible for these disease traits to measurable locations in the genome [26].

The promising results from these investigations helped fuel enthusiasm and support for the initiation in 1990 of the controversial Human Genome Project, a costly, decade-long effort that led to the determination of the DNA sequence and the estimation of probable genetic elements in the human genome [27-29]. In the process of working towards this landmark achievement, technology and analysis methods for automated sequencing were greatly improved [30]. While the Human Genome Project was being conducted, other researchers continued to explore molecular biology in other ways. Microarray technology was developed in the early 1990s, allowing the mRNA expression levels of numerous genes to be measured at once [31, 32]. These specialized hybridization arrays set the stage for the development high-throughput, genome-wide methods that have become common in contemporary biology laboratories. In addition to using this new chip-based hybridization approach to measure expression changes, microarray technology also could be used to measure Single Nucleotide Polymorphisms (SNPs) – a DNA sequence variation in which a single position of the genome differs among individuals in a population. While the Human Genome Project and microarray technology were in their early stages, other investigators were developing new methods to readily identify SNPs [33, 34]. By the mid-1990s, it was clear that linkage-based family studies were effective for detecting rare, highly penetrant disease variants. However, it was also clear that these studies were not appropriate for studying complex

diseases that may result from a combination of common variants with smaller effect sizes [33]. To study this type of genetic contribution to complex diseases, researchers had to design population-based studies of carefully selected individual candidate genetic markers. To avoid this limited approach, reliable markers of genetic variation would need to be discovered across the entire genome so that systematic scans could be completed in large case-control studies to obtain a global picture of variation associated with a disease of interest. To identify these markers, several groups collaborated in a large research effort parallel to the Human Genome Project to develop a map of human genome sequence variation, leading to the discovery of well over 1 million different SNPs by the time the human genome was published [35].

B. Genetic Association Studies

By 2001, a draft of the human genome sequence had been published, a large number of measurable genetic variants (SNPs) had been found, and the local correlation structure of the genome (i.e. the haplotype structure) was becoming better known [36]. To design an efficient way to study human variation in populations, another large consortium, The International HapMap Project, was initiated to study linkage and variation information in several different populations around the world (269 individuals, each genotyped at over 1 million SNPs in Phase I) [37]. Completion of the initial phase of this project allowed genetics researchers to design reasonably inexpensive, high-throughput approaches for measuring common variation across the genome by taking

advantage of the known correlation structure in the HapMap study populations [37, 38]. By 2005-2006, using the tools made possible by the HapMap Project, the first Genome-Wide Association Studies (GWAS) were conducted to explore the molecular causes of common diseases. In this type of study, between 100,000 and 1,000,000 genetic markers are measured at calculated intervals across the genome for hundreds to thousands of people with and without a disease to discover those variants and regions that are most dissimilar. Through GWAS initiatives like the Wellcome Trust Case Control Consortium (WTCCC) [39] and the National Center for Biotechnology Information Database of Genotypes and Phenotypes (NCBI dbGaP) [40], genetic association data have been made available for a wide range of common diseases for several different populations.

To date, more than a thousand genomic regions have been associated with susceptibilities to a wide variety of common diseases and with inheritance of other observable phenotypes [41]. Findings from the initial single-locus analyses of GWAS of common diseases have been shown to be meaningful and largely reproducible for numerous specific disease phenotypes [40, 42]. To further extend knowledge on human variation, the 1000 Genomes Project was initiated to provide researchers with an in-depth reference with a more detailed picture of human sequence variation than was possible in the HapMap Project [43]. The 1000 Genomes Project can be used by researchers to explore rarer, potentially causal variants in a region of interest and thereby inform further analyses, impute genotypes from existing studies, and design new studies [43, 44]. As methods to rapidly sequence individual genomes methods decline in cost, it is likely that

future GWAS will include full sequence information. In the meanwhile, analysis methods are being developed to address the computational and statistical challenges already present at the current scale to better utilize GWAS results and to generate hypotheses for future research.

C. Motivation

In the past few years, hundreds of large case-control studies of various diseases have been conducted to measure the common genetic variation for SNPs at strategic locations across the genome. These GWAS have allowed researchers to make exciting new discoveries related to the genotypes, expression patterns, and other pathophysiological changes that may be associated with common genetic diseases. GWAS have led to the discovery of more reproducible genetic associations than had been discovered using any other approach [45]. However, GWAS have been subject to several criticisms and have several limitations. First, only a minority of GWAS findings to date have led to meaningful insights about the molecular pathophysiology of disease phenotypes [42, 46, 47]. Second, the variants detected using GWAS typically have limited predictive power [42, 48]. Perhaps more importantly, the cumulative genetic risk found for most diseases studied with GWAS typically does not fully explain the strong heritability observed in corresponding epidemiologic studies [42, 45, 49, 50]. This missing heritability may be due to several causes. In some cases, the missing heritability may result from disease risk that is mediated by less common sequence variations and not

by the common, stable variations currently measured in genome-wide association studies [45, 47, 49, 50]. Another hypothesis is that the missing heritability may result from interactions between genetic loci. Many diseases are genetically complex, and it is generally suspected that multiple interacting genetic and environmental factors contribute to the pathogenesis and maintenance of the clinical features of multifactorial disease traits. Previous efforts to understand these data by looking for complex interactions within the data have yielded promising results [1, 51]. We hypothesize that improved analyses of existing data sets will uncover meaningful findings related to the elusive pathogenesis of several common genetic diseases, leading to a better understanding of individual genetic risk and to improved application of personalized treatment strategies. Herein we describe a set of tools and analysis methods developed in this work to address several of the substantial computational challenges involved with understanding the relationship between genomic variation and disease.

D. Project Overview

In Chapter III, we describe an adaptive evolutionary optimization algorithm that integrates linkage disequilibrium information while searching for multi-locus interactions in genetic association studies. This method reduces the need for exhaustive search for genotype combinations by taking advantage of inherent genomic structure, thereby improving both the power and computational efficiency of the analysis relative to conventional methods. We compare our algorithm with the most powerful competing

methods to determine its ability to manage genomic data sets efficiently while maintaining adequate power to detect simulated disease loci.

In Chapter IV, we extend the optimization strategy presented in Chapter III and demonstrate the utility of high performance computing methods for integrative GWAS analysis. We describe a parallel algorithm that implements state-of-the-art computing methods using graphics processing units (GPU) to explore GWAS data sets with maximal computational efficiency and minimal cost. The software resulting from this research will be made freely available as a user-friendly, expandable parallel genetic association analysis tool that will be increasingly necessary as genomic data sets continue to increase in size and complexity.

In Chapter V, we addresses the problem of using high density genetic association study data to develop predictive models for bivariate (case-control status) and quantitative/continuous traits**.** That is, rather than seeking to identify regions of the genome that are associated with a disease phenotype, we seek the genetic variants that best describe the likelihood of an individual's phenotype. Several studies have addressed this problem, but the ability to predict disease risk accurately using GWAS markers has been quite variable. Regression-based approaches have proven useful for this type of study, but it is necessary to avoid potentially high degrees of collinearity and overfitting that may occur when applying these methods to typical, high-dimensional genetic studies. Our analysis addresses these problems by utilizing a modification of lasso regression, a powerful and well-studied shrinkage and selection method for general linear models. The

described approach integrates information about the genes and biological pathways associated with SNPs in the data set to analyze a simulated exome sequencing data set based on the 1000 Genomes Project data.

In Chapter VI, we describe a framework for analyzing genomic variation related to microRNA (miRNA) modifications in common diseases. miRNAs comprise a large family of ~22-nucleotide-long RNAs that have been shown to perform key post-transcriptional regulation of gene expression in a wide variety of cellular processes [52, 53]. For heritable multifactorial diseases, genotypic variation is only one component of the pathogenesis [54-56]. In the past few years, a number of studies have sought to analyze the extent to which genomic variation may pathologically modulate transcriptional activity in disease related genes [57]. However, data directly relating expression and variation information within a GWAS population are currently limited. To better understand possible relationships between genomic variation and mechanisms of disease-related expression changes, we present an analysis framework for the determination of genomic variation associated with miRNA dysregulation. We applied our analysis methods to several GWAS of common diseases to determine candidate targets for disease-associated dysfunctions in miRNA-related gene expression changes.

E. Significance

The research described herein has been conducted to addresses several major challenges in the analysis of common complex genetic diseases. In this research, we

present efficient algorithms and powerful computing tools that can be used to better understand the genetic risks underlying a number of diseases. While definitively determining causality through the analysis of GWAS data is not possible due to the nature of the study design, the analytical improvements that result from this work may uncover biologically interesting associations or reveal testable research hypotheses not yet discovered in existing data sets. By releasing our code in a user-friendly package that can take advantage of state-of-the-art computing methods, we expect that other researchers in the field will be able to use our analysis tools to aid their own analyses of genome-wide association data.

CHAPTER II

PRINCIPLES OF GENETIC ANALYSIS

A. Introduction to Population Genetics

The research of Mendel, Correns, and de Vries established the principles of genetic transmission between parents and offspring. In the simplest case, if a gene has two variants (or alleles) *A* and *a*, then a cross between two heterozygous parents yields a predictable set of genotype probabilities for any offspring:

$$\left(\mathrm{Aa}_{maternal}\right) \text{ x } \left(\mathrm{Aa}_{paternal}\right) \rightarrow \left\{\mathrm{P(AA)} = \frac{1}{4}, \mathrm{P(Aa)} = \frac{1}{4}+\frac{1}{4}, \mathrm{P(aa)} = \frac{1}{4}\right\}.$$

Historically, the term *allele* refers generally to one of two or more possible forms of a gene. At the time these principles were conceived, the concepts of alleles and allele frequencies were somewhat abstract. Today, allele frequencies can be directly measured using single nucleotide polymorphisms (SNPs), which, as their name implies, are changes at a single position in genomic DNA. SNPs most commonly have two alleles, referred to as "major" and "minor" alleles based on their observed frequency in a population (major allele frequency > 0.5, minor allele frequency < 0.5). Population geneticists typically use allele frequencies as a primary measure when considering genetic variation in population. The Hardy-Weinberg Equilibrium (HWE) principle is an extension of the abovementioned Mendelian probabilities and states that allele frequencies in a population

remain constant in the absence of immigration, mutation, selection, non-random mating, or sampling errors – factors that would interfere with independent random sampling from a constant distribution. More precisely, for a genetic locus with alleles *A* and *a*, if allele A is observed with a frequency of *p* and allele *a* is observed with frequency *q*, then under HWE, homozygous AA would have frequency p^2, heterozygous Aa would have frequency *pq*, and homozygous *aa* would have frequency q^2.

Genetic material is transmitted from parents to offspring via meiosis. In this process, diploid germ cells (23 paired chromosomes) divide to produce haploid gametes (spermatozoa or ova with 23 chromosomes). In the process of separation, the

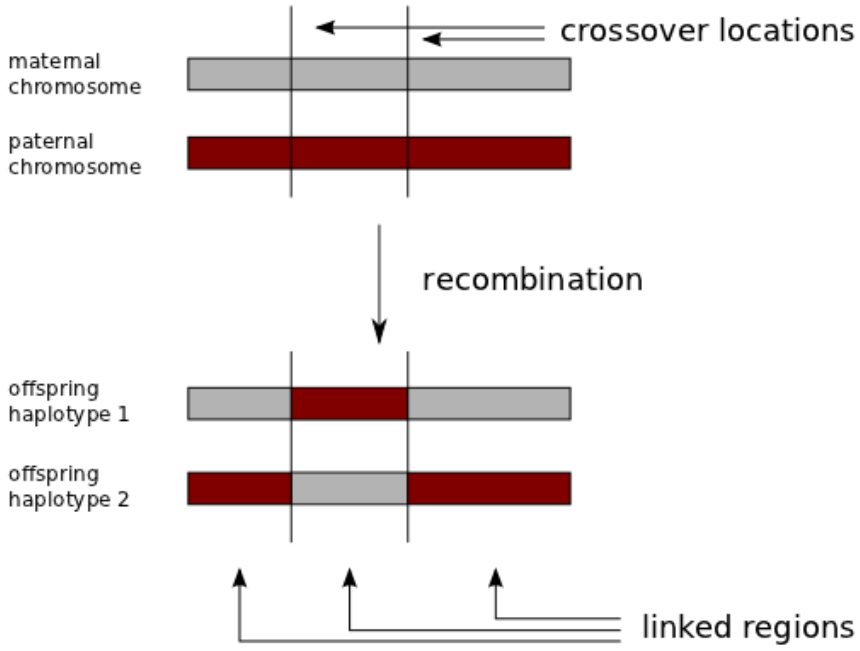


Figure 2.1. Diagram showing how recombination events during meiosis affect haplotype inheritance and linkage disequilibrium patterns. Recombination events are less likely to occur between two proximal loci, so genes that are close together have correlated genotypes.

chromosomal pairs cross over at various points, resulting in two offspring chromosomes that are each a combination of the two parental chromosomes (Figure 2.1). This process of chromosomal crossover is also referred to as recombination. The combination of alleles that are inherited together along one parental chromosome is known as a haplotype. If two genes from the same parental chromosome are typically inherited together within in a haplotype, they are said to be linked. The concept of linkage disequilibrium was described shortly after HWE, and refers generally to the non-independence (or correlation) of alleles at different sites [58]. For the purposes of this work, we are primarily concerned with linkage disequilibrium between proximal regions along a chromosome as measured in a population. The likelihood of a recombination event between two regions is less likely if those regions are close together, and linkage disequilibrium measured between two markers generally decreases as a function of the genetic distance between them [58]. Linkage disequilibrium is most commonly measured using the statistics D and r, which correspond to the covariance and correlation between loci [59]. For two loci on the same chromosome, suppose alleles *A* and *a* are at the first locus and alleles *B* and *b* are at the second locus with allele frequencies π_A, π_a, π_B, and π_b. There are four possible haplotype combinations between the two loci, with frequencies π_{AB}, π_{Ab}, π_{aB}, and π_{ab}. One of the first measures introduced was [60]:

$$D = \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}.$$

For the purpose of measuring linkage disequilibrium, D is typically normalized to the measure:

$$D' = \left| \frac{D}{D_{\max}} \right| \text{ , where } D_{\max} = \begin{cases} \min(\pi_A \pi_b, \pi_a \pi_B) \text{ if } D \geq 0 \\ \min(\pi_A \pi_B, \pi_a \pi_b) \text{ if } D < 0 \end{cases}$$

which is on the interval [0,1]. Alternatively, linkage disequilibrium can be calculated using the correlation measure:

$$r = \frac{D}{\sqrt{\pi_A \pi_a \pi_B \pi_b}}$$

which is commonly reported as r^2 (also on the interval [0,1]). If two loci are statistically independent, then D and r will both be zero. If two loci are perfectly correlated, both D and r will be 1. While these two measures are similar, they may behave differently in certain circumstances [58]. Here, we use r^2, because it has been shown to have less random variation at a given recombination distance [58].

Large-scale identification of SNPs in the 1990s allowed researchers to study linkage along a chromosome at higher resolution, and genotyping studies of dense sets of markers quickly revealed a structure in the human genome that consisted of discrete sets of haplotype blocks with a high degree of linkage disequilibrium [36]. A SNP with two alleles *A* and *a* has the three possible genotypes *AA*, *Aa*, and *aa*, and the individual parental contributions to the genotype are unknown. Thus, without further information it is not possible to determine the haplotype frequencies π_{AB}, π_{Ab}, π_{aB}, and π_{ab} for a population of individuals measured at several SNPs on a chromosome. In direct studies of haplotypes (most notably [36] and the HapMap project [37, 38]), parent-child trios were genotyped so that it would be possible to infer which alleles were on the same

chromsome in the offspring (*n.b.* this information is known as the phase of a genotype), allowing further study of the haplotypes in the observed population. Phase information is not directly observed in genetic case-control studies (e.g. GWAS), so it is necessary to estimate the haplotype frequencies to calculate linkage from these data. This can be done using several approaches. For all linkage disequilibrium calculations herein, we use a standard two-marker expectation-maximization procedure to estimate the haplotype frequencies in our linkage calculations for case-control populations [59, 61, 62]. This procedure has been previously described in detail [62]. Briefly, this method considers the observed joint genotype frequency counts for two (or more) SNPs in a population with alleles *A* and *a* and *B* and *b*, and uses these frequency counts in conjunction with the assumption of Hardy-Weinberg proportions to derive the maximum likelihood estimates for the of the molecular haplotype frequencies for *AB*, *Ab*, *aB*, and *ab* in the population.

B. Measures of Genetic Association for Single and Multi-Locus Analysis

Genetic associations in case-control studies are determined by comparing allele frequencies between the healthy and diseased populations. For example, suppose a SNP (SNP1) with alleles *A* and *a* is measured at a disease locus in a case-control study, and assume that the disease-causing allele is *a*. Designating the case counts as *u* (unhealthy) and the control counts as *h* (healthy), the resulting contingency table would be as shown in Table 2.1. Now suppose measurements are available for a second SNP (SNP2) with alleles *B* and *b* at some distant locus. To determine whether there is a gene-gene

interaction between SNP1 and SNP2, the joint genotype frequencies are counted in cases and controls for the 9 possible genotype combinations (*AABB, AaBB, aaBB, AABb, AaBb, aaBb, AAbb, Aabb, aabb*) to construct a contingency table. There are a number of ways to describe association from these two measurements. For the single locus case, descriptive epidemiologic terms are often used. For example, the odds ratio is the ratio of the odds that the cases were carriers of the *a* allele ($\frac{u_{Aa}+u_{aa}}{u_{AA}}$) to the odds that the controls were carriers of the *a* allele ($\frac{h_{Aa}+h_{aa}}{h_{AA}}$) (*n.b.* the odds ratio can also be defined using ratios of total allele counts, rather than by allele carrier counts as shown here). Genotypic relative risk can be defined similarly by using probabilities instead of odds. If it is assumed that the risk allele *a* has a stronger effect in homozygous individuals than in heterozygous individuals, then the Cochran-Armitage test trend is commonly used [63]. However, if such a trend in effect size is not observed, this test may have lower power than a more general test statistic such as the Pearson χ^2. Because the underlying genetic model in complex diseases is typically unknown, we use the Pearson χ^2 as the main

	AA	Aa	aa	Total
Cases	u_{AA}	u_{Aa}	u_{aa}	n_{cases}
Ctrls	h_{AA}	h_{Aa}	h_{aa}	$n_{controls}$
Total	n_{AA}	n_{Aa}	n_{aa}	N

Table 2.1. Contingency table displaying the frequency counts necessary to test for disease association at a single SNP.

measure of genetic association in our analyses of both single-locus and multi-locus frequency counts. For a genetic combination of k loci with a 2 x 3^k contingency table G with index i used for disease status and index j for the 3^k genotypes, this measure can be generally defined as:

$$\chi^2_{3^{k-1}df}(G) = \sum_{i=1}^{2}\sum_{j=1}^{3^k}\frac{\left(G_{ij}-\mu_{ij}\right)^2}{\mu_{ij}}$$

where:

$$\mu_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}, \text{ and } n_{.j} = \sum_{i=1}^{2} G_{ij}, \; n_{i.} = \sum_{j=1}^{3^k} G_{ij}, \text{ and } n_{..} = \sum_{i=1}^{2}\sum_{j=1}^{3^k} G_{ij}.$$

The Yates continuity correction was used to adjust for sparsity in genotype counts [63].

C. Simulation Models for GWAS

The scale and potential complexity of typical GWAS has necessitated the development of new analytical methodologies. When designing analysis methods for genetic association studies of diseases with an unknown disease model, simulation data are necessary to properly evaluate the performance of the proposed method under known conditions. There are a number of issues when considering disease simulations to evaluate a method designed to analyze gene-gene interactions in GWAS. The three most important problems regarding simulating GWAS effects are: (1) accurately modeling a disease-causing genetic interaction effect given the allele frequencies, effect size, and

penetrance; (2) including realistic genomic features such as linkage disequilibrium; and (3) considering appropriate gene-gene interaction disease models to evaluate the power of the method.

A fairly long history of simulation models have attempted to address issue 1 [64-69]. Earlier models were largely based on coalescent genetic theory, using direct

Model 1: Multiplicative 1			
	BB	**Bb**	**bb**
AA	α	ω_2	$(\omega_2)^2$
Aa	ω_1	$(\omega_1)(\omega_2)$	$(\omega_1)(\omega_2)^2$
aa	$(\omega_1)^2$	$(\omega_1)^2(\omega_2)$	$(\omega_1)^2(\omega_2)^2$

Model 2: Multiplicative 2			
	BB	**Bb**	**bb**
AA	α	α	α
Aa	α	$(\omega_1)(\omega_2)$	$(\omega_1)(\omega_2)^2$
aa	α	$(\omega_1)^2(\omega_2)$	$(\omega_1)^2(\omega_2)^2$

Model 3: Flat			
	BB	**Bb**	**bb**
AA	α	α	α
Aa	α	ω	ω
aa	α	ω	ω

Model 4: Additive			
	BB	**Bb**	**Bb**
AA	α	α	α
Aa	α	2ω	3ω
aa	α	3ω	4ω

Table 2.2. Risk models for genotype combinations. A and a are the alleles for locus 1, and B and b are the alleles for locus 2. α is the baseline effect size, and ω is the disease effect size. The baseline effect is defined as having a relative risk of 1 plus a small amount of random genetic variation. The disease effect size is modeled as a relative risk that an individual with a specific genotype also has the disease.

simulation of an evolutionary process using the basic principles of population genetics to model the disease-causing genetic effect in a population. These models are useful, but they can be quite inefficient and potentially inaccurate in large case-control studies [66, 68, 70]. Two more recently published peer-reviewed software packages, *gs* and *GWASimulator*, have been released for simulating genetic effects (both single locus effects and effects caused by interacting loci) using the population structure in HapMap populations that were used to design the SNP genotyping platforms most commonly used in GWAS [68, 70]. Thus, these two software packages address issues 1 and 2 as stated above. The remaining issue 3 can be addressed by carefully simulating different types of gene-gene interaction models for a disease-causing genetic effect. Models of gene-gene interaction effects, also known as epistatic effects, have been carefully studied [71, 72]. When evaluating a proposed method, the most common approach is to consider several of these interaction models and to determine the ability of the method to ascertain the causal effects [39, 73-76]. The independent contribution of individual SNPs to disease risk is known as the marginal effect. If one of the SNP alleles in a causal gene-gene interaction has no marginal effect, then it does not contribute to the disease in the absence of the second disease-causing SNP allele. The 4 standard types of disease models that we use for studies in this work are similar to those considered by comparable methods, and are given in Table 2.2: (1: "Multiplicative 1") a model that includes marginal effects with multiplicative interactions between all risk alleles; (2: "Multiplicative 2") a model without marginal effects that has multiplicative interactions between the risk alleles; (3: "Flat") a model without marginal effects that has the same effect size regardless of the

which the effect size increases linearly with the number of interacting disease alleles in the genotype.

D. Statistical Modeling of Genetic Disease Risk

Rather than seeking to identify regions of the genome that are associated with a disease phenotype using one of the methods described above, we now consider the problem of selecting the genetic variants that best describe the likelihood of an individual's phenotype. There are a number of studies that have considered the predictive power using sets of the most significant SNPs, genotypic risk scores, or some combination of genotypes with a phenotypic measure [42, 77-82], but these models have limited ability to predict true positives [42, 81]. There are a number of variable selection methods that can be used to build predictive models from GWAS data [83, 84]. Categorical and quantitative variables such as age, smoking status, waist size, or family history have been shown to be critical when developing predictive models for some diseases [49, 81], so the ability of a model to incorporate these features is highly desirable. Analysis of quantitative traits can readily be conducted using standard regression models, but the problem for large GWAS becomes ill-posed and the results may not be reliable.

There are a variety of methods that have been used to deal with this type of analysis, but penalized regression methods are among the most flexible and efficient [78, 85, 86]. Numerous penalized regression methods have been shown to be effective for

genome wide association studies in general [78, 87, 88], and more recent studies indicate that such methods retain their ability to detect association even for studies containing both rare and common variants [85, 89]. To understand these methods, we now introduce a standard regression framework [63, 86].

Let a continuous response vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$, for $i = 1, ..., n$, (***Y'*** is used to denote the transpose of vector ***Y***) containing the outcome variables for *n* subjects and a matrix $(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$, where $\mathbf{X}_i \in R^{m+1}$, with R^{m+1} the set of real numbers in the dimension *m*+1. Let $\mathbf{X}_{i0} = 1$ for $i = 1, ..., n$, and let $\mathbf{X}_{ij}$ for $i = 1, ..., n$, $j = 1, ..., m$ include the set of data measured at *m* predictor variables for all *n* subjects in the study. The standard linear regression model can be written:

$$Y_i = \sum_{j=0}^{m} X_{ij}\beta_j, \ i = 1, ..., n$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)'$ is the vector of regression coefficients. The ordinary least squares (OLS) solution to this regression is obtained by solving the problem:

$$\min_{\boldsymbol{\beta} \in R^{m+1}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{m} X_{ij}\beta_j \right)^2$$

However, this standard regression model is not well-suited for large studies with far more variables than samples, as it often results in inaccuracies due to model instabilities, collinearities, and overfitting. Several penalized regression methods have become popular in the analysis of large scale genetic data sets [90, 91] for their improved ability

for variable selection. In this study, we use a modification of the OLS optimization problem known as lasso ("Least Absolute Shrinkage and Selection Operator") regression, in which the L_1-norm of the non-intercept coefficients in $\boldsymbol{\beta}$ is used as a penalty to achieve a sparse solution [86, 92]. Using this approach, $\boldsymbol{\beta}$ is determined as the solution to the optimization problem:

$$\min_{\boldsymbol{\beta}\in R^{m+1}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{m} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{m}\left|\beta_j\right|,$$

where λ is a shrinkage parameter.

For case-control studies, our outcome variable $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$, for $i = 1, ..., n$, is the case-control status vector, with $Y_i \in \{0,1\}$. In this situation, logistic regression is commonly used. In this model, we use the observations at m predictor variables $(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$, where $\mathbf{X}_j \in R^{m+1}$ and $\mathbf{X}_{i0} = 1$ for $i = 1, ..., n$, to fit the following model :

$$\log\left(\frac{P(Y=1\,|\,\mathbf{X})}{P(Y=0\,|\,\mathbf{X})}\right) = \sum_{j=0}^{m} X_j\beta_j$$

where:

$$\begin{aligned} P(Y=1\,|\,\mathbf{X}) &= \frac{1}{1+e^{+\left(\sum_{j=0}^{m} X_j\beta_j\right)}} \\ P(Y=0\,|\,\mathbf{X}) &= \frac{1}{1+e^{-\left(\sum_{j=0}^{m} X_j\beta_j\right)}} \\ &= 1 - \mathrm{P}(Y_i = 1\,|\,\mathbf{X}). \end{aligned}$$

Above, we showed that the best linear model was determined by the minimization of the least squared error between the responses predicted by the fitted regression model and the observed response variable **Y**.

In logistic regression, the best model is determined by finding the model that maximizes the binomial log-likelihood:

$$\ell(\beta \mid \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \left(\sum_{j=0}^{m} X_{ij}\beta_j \right) - \log\left(1 + e^{\left(\sum_{j=0}^{m} X_{ij}\beta_j\right)} \right)$$

Thus in lasso-regularized logistic regression, the optimal penalized model is determined by solving:

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{m+1}} \ \ell(\beta \mid \mathbf{X}) - \lambda \sum_{j=1}^{m} \left| \beta_j \right|$$

Characteristics of this optimization problem cause the coefficients $\boldsymbol{\beta}_j$ for the more important predictor variables in the model to be larger, while the coefficients for the less important features are reduced towards zero. This reduction in the effective number of variables is commonly referred to as "shrinkage". The amount of shrinkage that occurs is controlled by the size of the parameter λ.

Selection of the best value for λ is a model selection problem. A common approach to finding the best λ is through v-fold cross-validation. In this method, the data set is first split into v equally sized groups. For each of the v groups, the remaining v -1 groups (the training folds) are used to fit the lasso model, and then the error in using this

model to predict the v^{th} group of observations (the testing fold) can be determined. This process is repeated v times, and the value of λ that minimizes the average prediction error across all v models is chosen.

For the lasso analyses discussed in this work, the evaluation measures to determine prediction accuracy were Area Under the Receiver Operating Curve (AUC) for logistic models and mean squared error (MSE) for continuous linear regression models [86, 92]. Mean squared error is defined as:

$$\text{MSE} = \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{m} X_{ij}\beta_j\right)^2 / df$$

where df is the degrees of freedom in the model. In the classification of a binary trait such as case-control status, we are primarily concerned with the proportion of correct predictions for cases (true positive rate) and controls (true negative rate). The Receiver Operating Characteristic (ROC) Curve is a plot of the true positive rate against the false positive rate for a binary classifier. In the case where the prediction accuracy is 100%, the AUC value will be 1. In a truly random prediction of a binary trait (with equal numbers of the two classes), the AUC will be 0.5.

The penalized lasso regression strategy described above has been used to build a predictive models using GWAS [77, 78, 93]. In another risk prediction study, the Support Vector Machines (SVM) algorithm was used to build large and accurate predictive models for Type 1 Diabetes [79]. SVM is a supervised learning method that seeks to find the hyperplane using observed data to maximally separate two (or more)

classes. In GWAS, the two classes are cases and controls, and SVM is used to determine a set of coefficients for a linear combination of SNP genotypes that maximally separate the observed classes. In the simplest case, this separating hyperplane is the solution to the optimization problem:

$$\max_{\boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} C$$

$$\text{subject to: } Y_i \sum_{j=0}^{m} X_{ij}\beta_j \geq C, i = 1, ..., n$$

where C is the size of the margin between cases and controls. Most implementations of SVM expand this model using slack variables to increase flexibility of the margin in combination with kernel functions that map the input features in *X* to a higher dimensional space in which the variables are more likely to be linearly separable. Note that regularized logistic regression may give a very similar fit as SVM. However, recall that the lasso logistic regression algorithm develops its predictions by fitting a regression line by minimizing the weighted sum of the deviation between the penalized likelihood function and the observed data. In contrast, SVM develops its predictions by mathematically transforming the input variables to a higher dimensional space to more easily compute a separating hyperplane between the classes [86, 94]. With several exceptions, the predictive accuracies reported from these methods (except for the Crohn Disease results) are in sharp contrast to the typically modest predictive accuracy found using other approaches [42, 81].

CHAPTER III

AN EVOLUTIONARY ALGORITHM TO INVESTIGATE GENE-GENE INTERACTIONS IN GENETIC ASSOCIATION STUDIES

A. Introduction

High density genetic association studies have provided researchers with a wealth of information about the genotypic features contributing to common complex diseases. These studies have led to the discovery of numerous genetic variants shown to be reliably associated with specific disease phenotypes [95]. While these high profile studies have improved the understanding of genetics, the associated genetic variants typically explain only a small proportion of the observed heritability for the multifactorial phenotypes that have been considered [40, 47, 96]. There is growing evidence in support of the importance of interaction effects in GWAS [40, 47, 49, 96]. However, reliable detection of these effects remains an unresolved problem.

A number of combinatorial and statistical methods have been developed to address the problem of detecting gene-gene interactions, including χ^2 tests [97], logistic regression [98], logic regression [99], neural networks [100-102], random forests [103], and others [39, 73-76, 103-110]. These analysis methods represent a group of diverse strategies used to approach a problem with severe computational and statistical challenges, but they can be divided into several groups: exhaustive search, heuristic/local search, and Bayesian modeling methods.

Of the data mining methods related to exhaustive search, Multifactor Dimensionality Reduction (MDR) is among the most popular and has been used to analyze a variety of real data sets [1]. MDR seeks to identify combinations of k loci that influence a disease outcome. The main strategy of the method is to avoid sparse data cells and over-parameterization by reducing the number of dimensions by collapsing the multi-way contingency table into a single-dimensional model within a 10-fold cross validation. For each combination, the joint genotype is classified according to the ratio of cases to controls. The main problem with this approach is that even moderately sized analyses can be computationally prohibitive, so a filtering or data reduction step must be taken before MDR is used to detect gene-gene interactions in a full-scale GWAS.

The most commonly cited Bayesian method is Bayesian Epistasis Association Mapping (BEAM) [39]. Bayesian model selection techniques specify a set of prior distributions for SNPs that are unassociated, associated, or interacting with respect to the disease trait, and the posterior distribution for these parameters given the observed data is optimized using the Markov Chain Monte Carlo (MCMC) technique [1, 39, 74, 76]. These methods are quite powerful, but they can be limited by lack of power if the linkage is not properly included in the calculation of the posterior likelihood.

While a large number of local and stochastic search methods have been proposed, one method shown to have superior power for a variety of two locus models while maintaining adequate efficiency to analyze full-scale GWAS is SNPHarvester [75]. SNPHarvester uses a filtering approach to limit the number of SNPs with single-locus

effects. The algorithm is initialized using random loci across the search space, and a local search procedure is used to build a "path" from each initialization point that considers combinations of SNPs until the end of the iteration or until the path reaches statistical significance. At the end of each iteration, significant paths are "harvested" and removed from analysis, increasing the efficiency in successive runs.

We propose an alternative approach that considers linkage-based partitions in an evolutionary optimization scheme to efficiently search for interactions. As discussed above, GWAS are designed such that the SNPs genotyped have a high degree of local correlation (Figure 3.1). Partitioning the genome into haplotype blocks can be

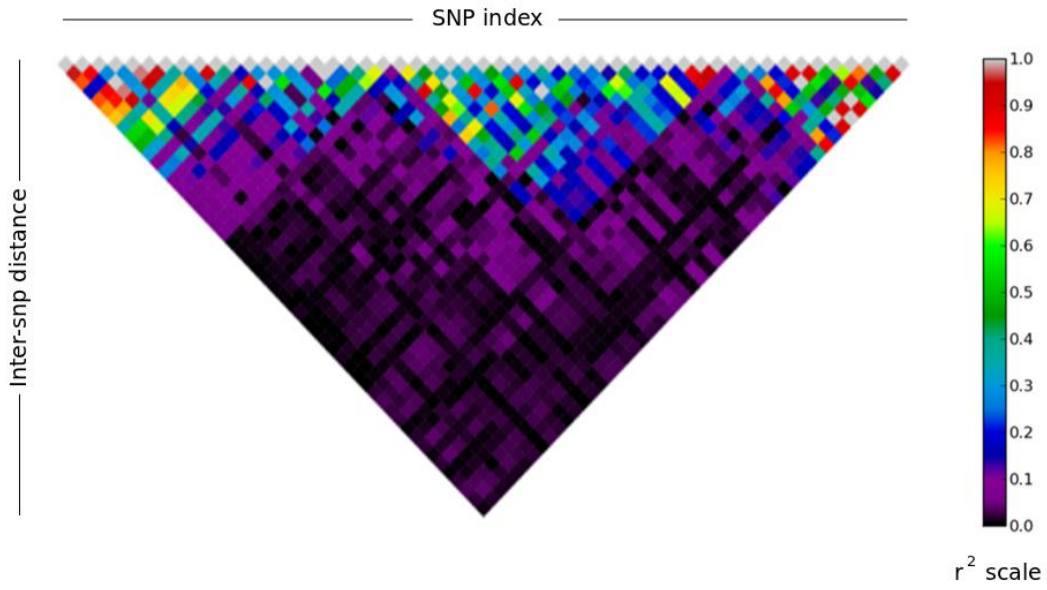


Figure 3.1. Example of linkage structure in a typical genetic association study (r^2 values). Each square in this figure corresponds to a linkage disequilibrium measured between two SNPs.

accomplished by a number of methods [61, 111-114]. Generally, these methods optimize within-block linkage above some predefined statistical threshold, and each algorithm has a means to select whether to join or split blocks given a linkage pattern. Using linkage block structure has been shown to be an effective data reduction strategy for genetic association studies of single loci [115, 116], but relatively few methods take advantage of this useful genomic feature for the purposes of identifying interactions between genetic markers [39, 103, 117, 118]. The main problem with using the standard block approaches is that they require an unacceptably large amount of computational overhead when applied to GWAS data. In this study, we present an evolutionary algorithm that takes advantage of local linkage structure to improve the reliability and computational efficiency for detecting genetic interactions. Using simulation studies modeled on realistic population and linkage structure, we show that our method is able to take advantage of this genomic feature to more efficiently and reliably detect epistatic interactions in a number of disease models.

B. Methods

B.1. Simulation Model

As discussed above, the generation of simulated data with a known genetic effect and a realistic linkage disequilibrium (LD) and population structure is essential for the proper evaluation of our method's ability to distinguish true gene-gene interactions. We considered a number of two-locus models (described above in Chapter II, Table 2.2) that

are similar to those used in the evaluation of other methods [39, 75, 108], using parameters similar to those described previously [119]. In each model, the disease prevalence was fixed at 0.01, the effect size (ω) was varied between 1.25 and 2.5, and the Minor Allele Frequency (MAF) was fixed at 0.15, 0.30, or 0.45. In this model definition, effect sizes (ω, a measure of relative risk used in [68]) include any underlying latent effects (α).

For each set of parameters, 50 replicate data sets were generated using *GWASimulator*, a peer-reviewed simulation method that retains local linkage disequilibrium patterns [68, 120]. Each simulation data set contained 2000 SNPs for 1500 individuals using a case control study design. The population and linkage disequilibrium features in our simulation data were based on HapMap CEU phased data (CEPH samples with ancestry from Northern and Western Europe) for autosomal SNPs corresponding to the Illumina HumanHap 550K SNP array [38, 68]. In order to avoid extreme LD effects, the average LD (measured by r^2 [58, 62]) was sampled across each chromosome at 25 SNP intervals for windows of 50 markers wide, and we only buried disease markers at genomic loci sampled from the middle LD quintile with a HapMap MAF within ±0.025 of the target parameter. For each simulation, Haploview [61] was used to ensure Hardy-Weinberg Equilibrium and data integrity.

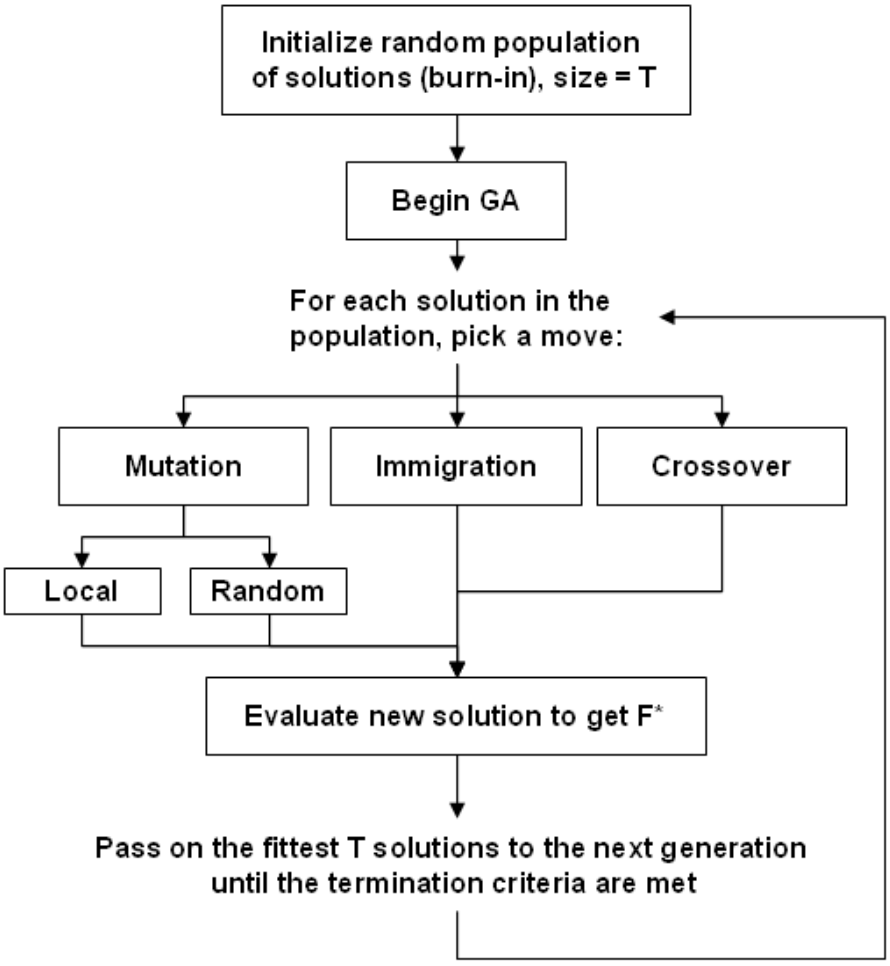


Figure 3.2. Flowchart diagram of our evolutionary algorithm structure.

B.2. Block Determination

As noted above, a number of methods exist for statistically determining block structure [61, 111-114] as it relates to either linkage or case-control information. While these methods are useful, they introduce a large degree of unnecessary computational overhead. To avoid this, we used a fixed block width that was determined on-the-fly by

sampling the data set for the width that gives an average intra-block r^2 value above 0.25. We compared our approach with the methods implemented in Haploview [61]: the method by Gabriel et al (GAB) [112], the 4 Gamete method (GAM) [111], and the Solid Spine of LD (SPL) method [61].

B.3. Proposed Algorithm

Evolutionary algorithms, also known as Genetic Algorithms (GA), are a well-established strategy in optimization and artificial intelligence applications [121]. In this approach, a set of solutions or "population" is randomly generated and iteratively improved with respect to some predefined fitness function. The solutions are modified by introducing random variety (immigrants), small modifications (mutations), or by combining solutions within the set (crossover), and the solutions with the highest fitness are selected following each iteration. Figure 3.2 describes the basic structure of the GA approach, in which F^* denotes the updated fitness measure of a solution in the population. A set of probabilities is used to randomly select the solution modifications, and these probabilities can be adaptively modified as the algorithm progresses to maximally improve the solution fitness. The pseudocode is provided in Figure 3.3.

GA Algorithm

Input:
Genotype Data
Genotype Block Information (optional)
Modifiable Functions and Parameters:
T = Population size (100)
k = order of the gene-gene interaction (2)
$|S|$ = random search depth for SNP combinations (max)
F = fitness function (χ^2)
Z = max number of iterations ($M/10$ for M = # SNPs)
tol = tolerance criteria to measure population change between iterations

Output:
List of k-block combinations with the strongest association measures

Algorithm
Initialization
Randomly select T k-block combinations
Organize these into a "population" ordered by F
Optimization:
while *Terminate* is False
define F_{least} as the least fit combination K_{least} in the population
for i=1..T **do:**
F_i = fitness for population block combination i
Randomly select a method to generate a new k-block combination K_i:
1. *Immigration*: p=0.6, randomly select a new block combination
2. *Mutation:* p=0.2
   a. Substitute one of the blocks with a random block
   b. Modify one of the two blocks by randomly choosing a nearby block
3. *Recombination*: p=0.2, substitute one of the blocks with a block from another randomly selected combination in the population

Calculate the fitness of the ith newly generated solution F^*_i
if $F^*_i > F_{least}$ **then**
K_i is put into the population and the K_{least} is removed
end if
end for
Sort the population by fitness value, re-define F_{least} and K_{least}
If termination criteria Z or *tol* is not met, *Terminate* is False
end while
Return a list of the block combinations with the best fitness values

Figure 3.3. Pseudocode for a simple version of our optimization algorithm

In our implementation, we used this optimization strategy to determine a list of k-block combinations with the best fitness. The objective of genetic association studies is to determine a set of *k* loci with the highest statistical association, and the discrete GA optimization algorithm is particularly well-suited for this purpose. We define the fitness measure for a set of blocks $B_1, B_2 \ldots B_k$ as follows. Letting $|B_m|$ represent the number of SNPs in the m^{th} block, we define S to be the set of all $\prod_{m=1}^{k} |B_m|$ possible combinations of *k*-SNP genotypes with exactly one SNP from each block. The genotypes in each B_m ($1 \leq m \leq k$) all have 3 possible genotypes, and the observed frequency counts for the 3^k possible genotypes for each genotype combination are tallied in a 2 x 3^k contingency table G_s of case and control genotype counts. For each s in S with contingency table G_s, we consider the χ^2 statistic with $3^k - 1$ degrees of freedom. The Yates continuity correction was used to adjust for sparsity [63]. The fitness of each block combination is then determined as:

$$F = \max\{\chi^2_{3^k-1}(G_s) : s \in S\}$$

The maximum search depth of S in each iteration is controlled as a parameter. When k is small (≤ 3), a deep search for the optimal F from each set S is computationally feasible and was used in this study. For larger k, F is determined by random sampling.

The fitness measure can be any statistic or function that maximally distinguishes cases from controls at the true loci. The χ^2-based fitness parameter used in this study is desirable because it is a widely-used, powerful, and efficient statistic that allows

comparison with other methods as well as a measure of statistical significance. In this study, the set of solutions was a list of unique block combinations with the number of blocks (*k*) fixed at 2. The significance p-value threshold used was 2.5×10^{-6}.

The probability of selecting a random new solution ("immigration") was assigned a value of 0.6 (adaptive range 0.4-0.8). A point modification of an existing solution ("mutation") was assigned a probability of 0.2 (adaptive range 0.1-0.3), and it involved

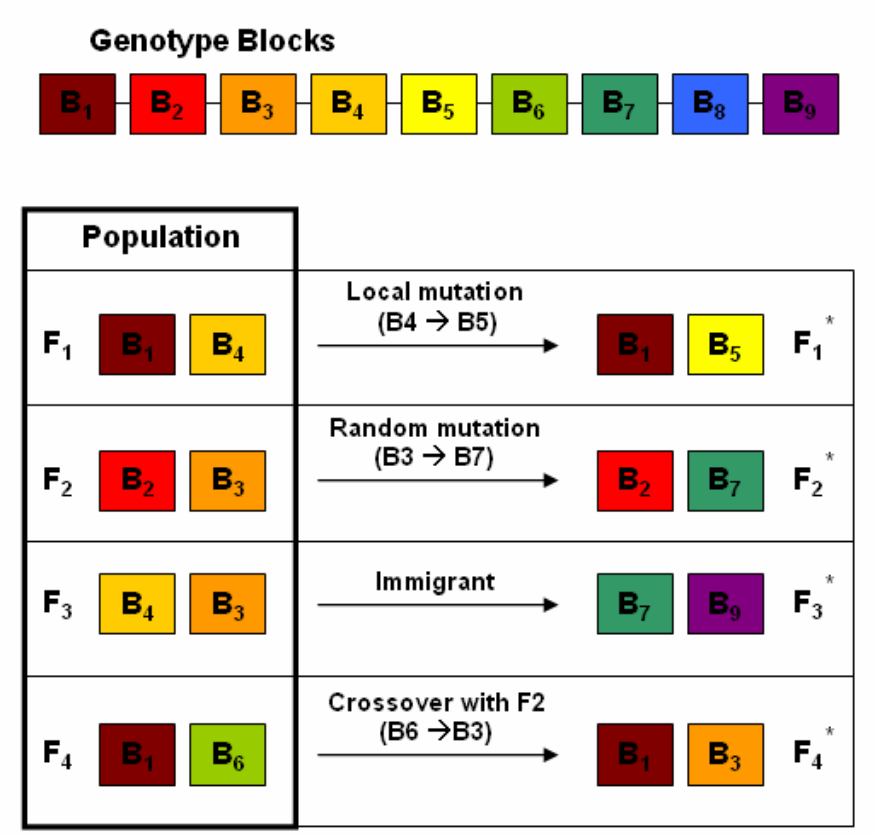


Figure 3.4. Illustrative example of a population of 2-block solutions and GA moves in one GA iteration. Blocks $B_1 \ldots B_9$ represent groups of linked genetic markers in a GWAS. F_1, F_2, F_3, and F_4 are the fitness measurements for the two-way block combinations in the population, and the corresponding F^* values are the fitness measurements for the solution proposed by the GA moves.

either (a) substituting one of the solution blocks with a random block or (b) modifying one of the two blocks by randomly choosing a nearby block. The "crossover" or "recombination" move involved substituting one of the solution blocks with a block from another randomly selected solution in the overall set. This move was assigned a probability of 0.2 (adaptive range 0.1-0.3). This random selection was biased such that fitter solutions were more likely to be selected for recombination. For a solution with fitness rank x, the probability for recombination is defined as: $\Pr(x) = 1 - e^{-(\lambda x/T)}$. We fixed $\lambda = 5$ and $T = 100$, the size of the solution set in our experiment.

A schematic of the GA moves is shown in Figure 3.4. The algorithm was allowed to run until one of the two termination criteria was reached: (1) a pre-set maximum number of iterations (default = M/10, where M is the number of SNPs) or (2) a set number of iterations pass without any change in the set of fittest solutions.

B.4. Evaluation

To assess the performance of our algorithm, we chose to SNPHarvester [75] for comparison. This algorithm provides a good comparison with our method because (1) it was shown to have impressive power relative to competing methods (e.g. BEAM) [39], (2) it has a well-written program that could be executed for comparison, and (3) it was shown to be computationally efficient enough for use with real GWAS. In SNPHarvester, we set k=2 and paths=50 as suggested by the authors in the publication of their method [75].

For both methods considered, we set the true solution to be the entire block containing the simulated causal SNP. Power was defined as the proportion of simulations in which the true solution was found using each algorithm.

C. Results

C.1. Block Method Comparison

To empirically determine that the power of our algorithm would not be affected by our proposed fixed block method, we compared our approach with 3 standard haplotype blocking algorithms using simulations for several disease models. We also computed the power of our method for the case in which there are no blocks (all blocks defined as containing a single SNP). Each of the 9 comparisons shown in Figure 3.5 consists of 50 simulation data sets using the disease simulation parameters described in Table 3.1. Blocks for GAB, GAM, and SPL were determined using Haploview using the parameters suggested by the authors [61]. For each method, the entire block containing

Label	Model	ω
A	Model 2* (lower penetrance)	2.5
B	Model 2	2.5
C	Model 2 (MAF = 0.45)	2.5
D	Model 3* (one marginal effect)	1.5
E	Model 2* (one marginal effect)	2.5
F	Model 3	2.5
G	Model 4	1.5
H	Model 4	2.5
I	Model 1	1.5

Table 3.1. Disease models used for the power comparison shown in Figure 2.5 for Models 1-4 as defined in Table 2.2 with effect size ω and modifications as indicated (*). MAF = 0.3 except in model C.

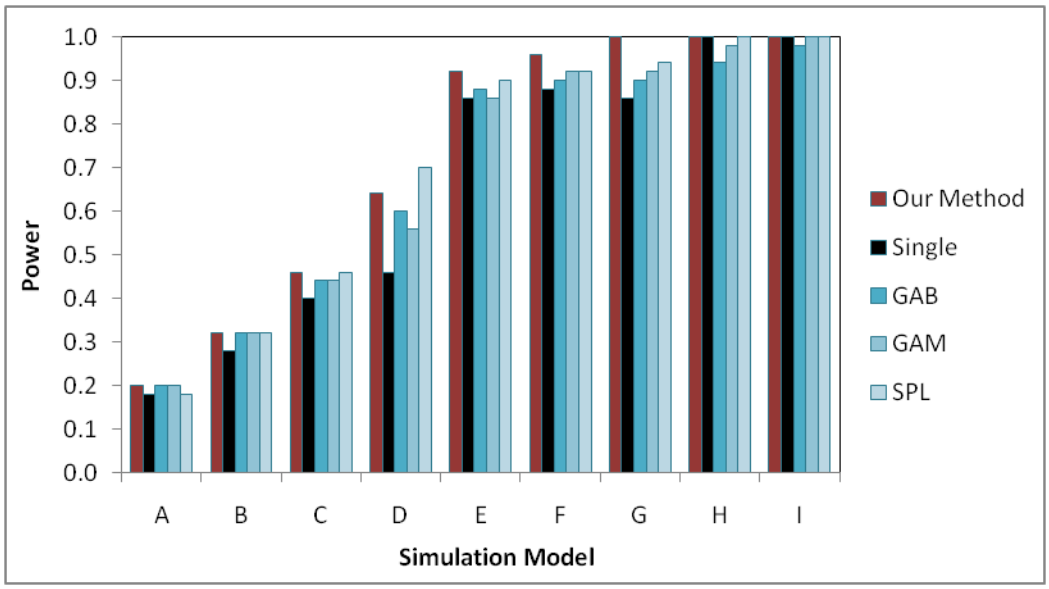


Figure 3.5. Power comparison of our GA approach run using standard block methods (teal), the block approximation approach described in the text (red), and a run with no blocks (or single-SNP blocks, shown in black) for a number of interactive models with varying strength (labeled A-I). Models are described in Table 3.2.

the causal SNP was defined as being causal for the purpose of these power calculations. As is evident in Figure 2.5, the power across these approaches is very similar (average standard deviation in power between the 4 methods across the 9 models is 0.024). This indicates that our method for determining an appropriate fixed block width was effective for determining unnecessary computations that do not lead to improved power.

C.2. Simulation Model

Results from the analysis of simulation data are shown in Figure 3.6a and Figure 3.6b. The power of SNPHarvester is comparable to that of our proposed method for the disease models described above. This approximate equivalence is expected, as these two

methods use the same statistical evaluation for the solutions considered in the search. Thus, if each algorithm samples the same set of SNP combinations in a given run, the statistical power of the two methods to detect the true solution will be equal. To assess the false positive rate for our method and for SNPHarvester, we generated 300 simulations as in Models 1-4 above, but without any simulated effects (i.e. $\omega=\alpha$ for all loci). SNPHarvester detected false positives in 11% of the null simulations with an average of 0.19 false positives per simulation (median 0.0). Our method detected false positives in 38% of the null simulations, with an average of 0.87 per simulation (median 0.0).

Performance benchmarks were computed on an Intel Core i5 750 2.67GHz Linux system with Python 2.6, gcc 4.4.3, and Java 1.6.0_20 using a set of null simulations. The average running time for one simulation with default parameters and 1500 subjects was about 45 seconds for our method and about 159 seconds for SNPHarvester. Our method achieves additional search efficiency by utilizing linkage disequilibrium information. As shown in Table 3.2, the improvements in running time are more evident when the study size is larger.

n	Our method	SNPHarvester
1500	45.4s	159.4s
3000	64.5s	415.1s
4500	84.9s	967.8s
6000	104.5s	1551.3s

Table 3.2. Running time comparison between our method and SNPHarvester (in seconds) for varying numbers of case-control subjects (n).

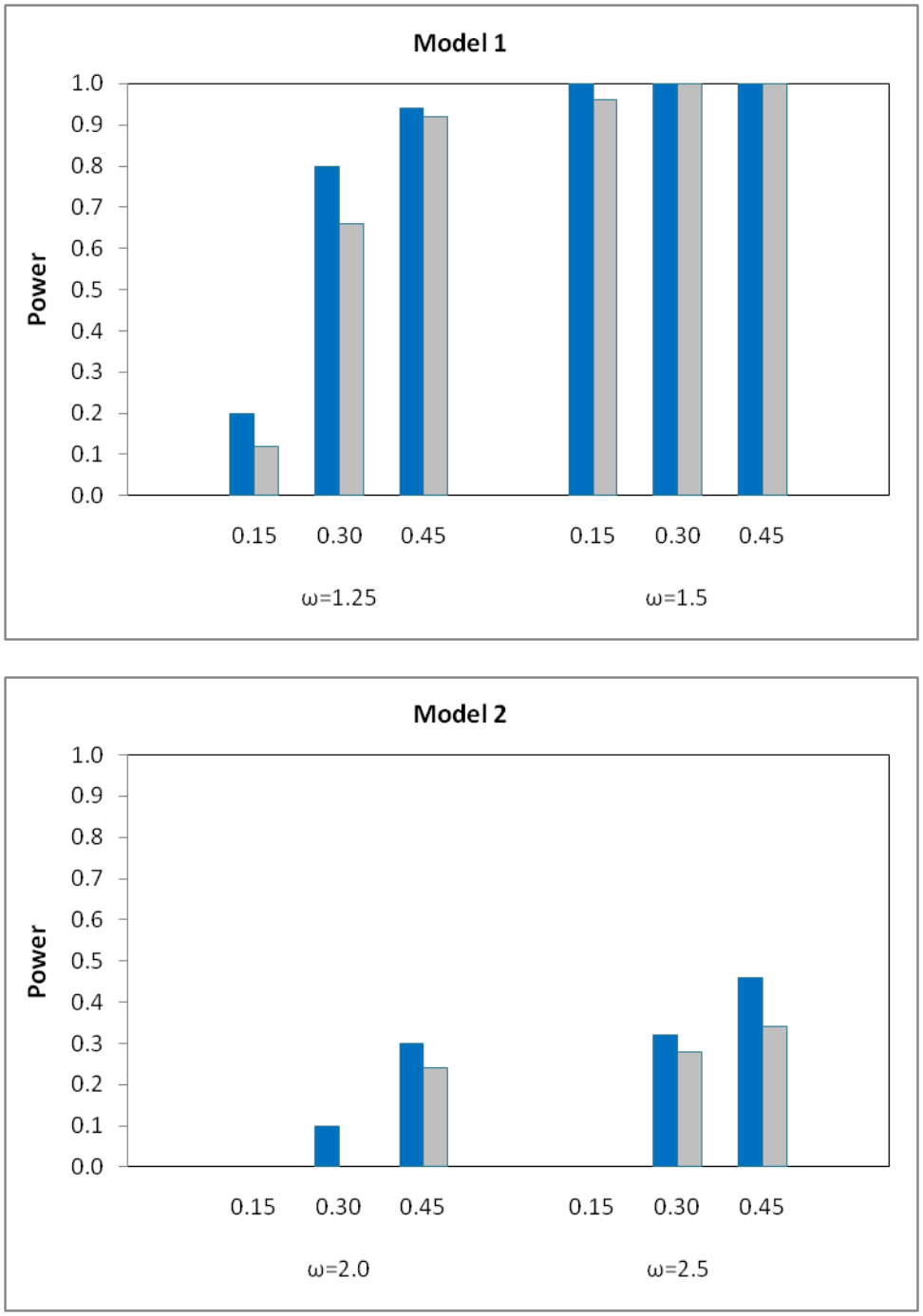
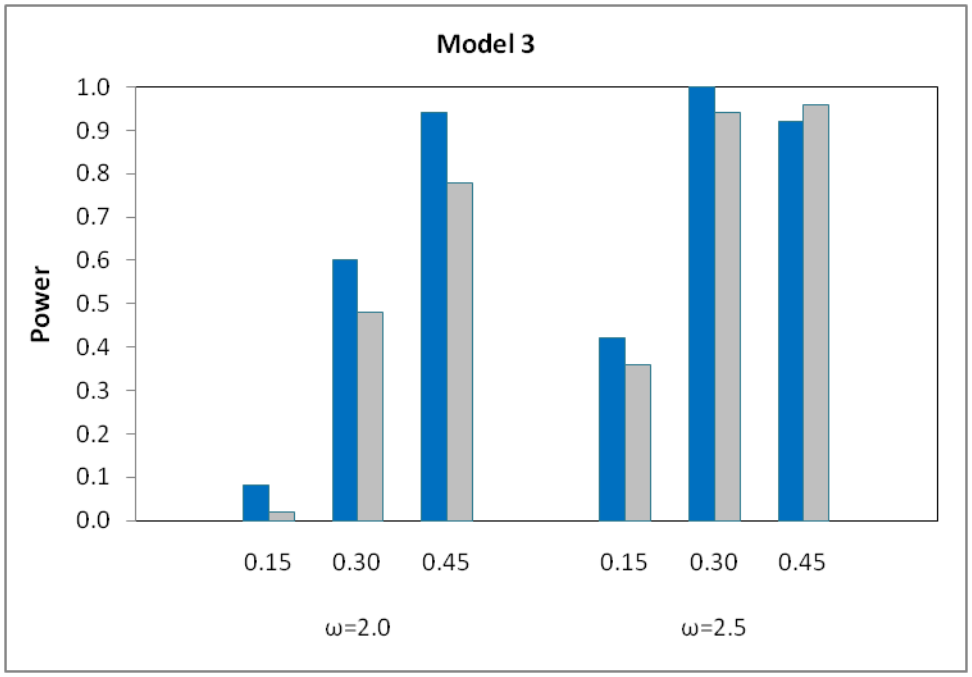


Figure 3.5a. Power of our proposed method (blue) compared with SNPHarvester (gray) for the 4 simulation models described in Table 2.2 for minor allele frequencies (0.15, 0.30, and 0.45) and effect sizes (ω) for the multiplicative disease models.



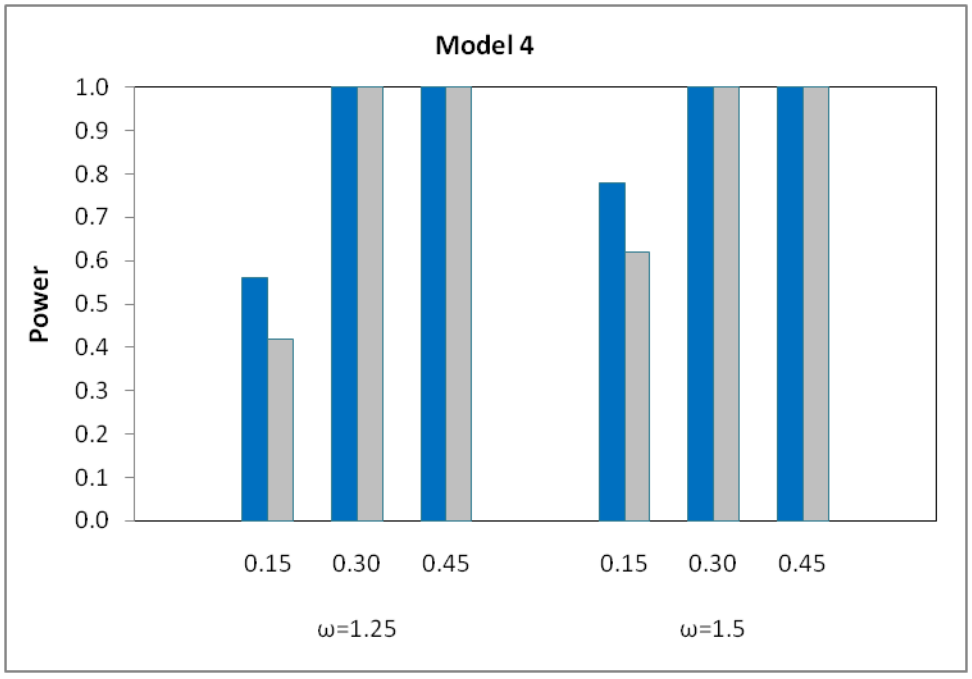


Figure 3.5b. Power of our proposed method (blue) compared with SNPHarvester (gray) for the 4 simulation models described in Table 2.2 for minor allele frequencies (0.15, 0.30, and 0.45) and effect sizes (ω) for disease models 3 and 4.

C.3. Age-related Macular Degeneration Genome-Wide Association Study

Having shown the utility of our method for assessing gene-gene interactions in limited simulation data sets, we now consider a real GWAS of Age-related Macular Degeneration (AMD) that consists of 103,611 autosomal SNPs genotyped for 96 cases and 50 controls (after quality control) [122]. In this landmark study, two significant SNPs, rs380390 and rs1329428, were found on chromosome 1 that lie in an intron of the

p-value	SNP1	SNP2
1.46E-08	'rs1061170'	'rs1926489'
6.69E-08	'rs10254116'	'rs3903445'
2.45E-07	'rs800292'	'rs10254116'
4.49E-07	'rs10503499'	'rs1740752'
4.53E-07	'rs931798'	'rs2828155'
4.53E-07	'rs931798'	'rs2828151'
4.77E-07	'rs1061170'	'rs9301772'
5.46E-07	'rs931798'	'rs10501439'
5.52E-07	'rs10511130'	'rs1972634'
5.61E-07	'rs10518433'	'rs3903445'
6.17E-07	'rs970476'	'rs10511467'
6.18E-07	'rs10511130'	'rs10517546'
6.26E-07	'rs6967345'	'rs3913094'
6.81E-07	'rs800292'	'rs1853882'
7.32E-07	'rs1061170'	'rs2402053'
7.51E-07	'rs1061170'	'rs284806'
7.61E-07	'rs6967345'	'rs3914244'
7.97E-07	'rs800292'	'rs1740752'
8.21E-07	'rs1978419'	'rs3913094'
8.43E-07	'rs800292'	'rs10483314'
8.97E-07	'rs1740752'	'rs7104698'

Table 3.3. Results from the AMD data set

gene for complement factor H. This discovery allowed researchers to further investigate the genetically associated molecular pathophysiology associated with AMD.

Following our quality control, there were 103,156 autosomal SNPs in our analysis. As noted elsewhere, the dominant effect is on chromosome 1 in the region near rs380390. However, our method was able to detect a number of interactions that did not include this locus but had *p*-values $< 1 \times 10^{-6}$. These loci are shown in Table 3.3.

D. Discussion

We have studied our block-based evolutionary optimization method under various conditions similar to those observed in high-density genetic association studies. We sought to assess the expected power and computational efficiency of our method by comparing it with another well-designed and efficient stochastic search method, SNPHarvester. The combination of our block strategy with the evolutionary optimization approach enables a comparatively powerful solution method with practical running time.

Other studies have shown that sets or blocks of SNPs can be used to increase the efficiency and power in association studies [103, 114-116, 120]. Evolutionary optimization is well-suited as a tool for manipulation of decision trees or logic expressions that are useful for multi-locus analysis of genetic data [123-127]. Our method can be considered an extension of these methods that utilizes LD structure in a stochastic search framework to determine the genetic loci most reliably associated with disease status. Meaningful, non-additive gene-gene interactions cannot occur between

highly correlated loci, justifying the elimination of within-block interactions from consideration in our search strategy. As with other heuristic local search strategies that have considered this problem, our method is expected to have limited ability to detect effects in the absence of either linkage disequilibrium or marginal effects.

We determined the power of our method to detect diseases caused by two interacting loci, but this model can readily be applied to uncover higher-order interactions. While our algorithm can computationally manage these higher order interactions, other analyses of genome-wide association study data for $k > 2$ have not convincingly revealed reliable results. Comprehensive power analyses of our algorithm for simulated disease models with 3 or more loci are necessary. Our evolutionary search strategy provides a flexible optimization framework that can be readily extended by using alternative fitness functions to distinguish cases from controls or by incorporating gene or pathway information to group sets of SNPs as opposed to local linkage disequilibrium structure.

CHAPTER IV

EXTENDING AN EVOLUTIONARY ALGORITHM FOR GENE-GENE INTERACTION INVESTIGATION USING HIGH PERFORMANCE COMPUTING METHODS

A. Introduction

As researchers seek to better understand genetic diseases by using analyses of gene-gene interactions, by collecting denser genetic association data, or by conducting meta-analyses of large combined data sets, the scale and complexity of the computations involved in the analysis increase dramatically [1, 128]. Concurrent with the genetic advances observed throughout the past decade, methods that use Graphics Processing Units (GPUs) for general purpose parallel computing have been engineered that make high performance computing methods more accessible and affordable [129]. There have been numerous applications of parallel GPU processing technology to scientific computing, most of which have yielded speedups of 10-100x relative to the original CPU applications [130, 131]. As discussed above, studies of epistatic interactions in GWAS sets have led to meaningful findings [45, 47, 49, 50], but the combinatorial nature of testing joint effects between genetic loci quickly leads to statistical and computational difficulties given the scale of current GWAS data sets. Among the applications that this type of high-performance computing has been successfully applied to is the detection of gene-gene interactions in GWAS [132, 133]. In this chapter, we present a parallelized

GPU implementation of the approach introduced in Chapter III to improve the efficiency of searching for gene-gene interactions. The GPU parallel platform provides researchers with a cost-effective means to obtain immense computational power to combat the ever-growing complexity of genomic studies. Several research groups have presented impressive results using GPUs to improve the computational speed of analyses for minimal cost [132-134]. These methods have significantly reduced the amount of time required to analyze interactions in large genetic association studies. Our parallel algorithm expands on the recent papers describing GPU-accelerated gene-gene interaction analysis to take advantage of the previously demonstrated improvement in algorithmic performance in a powerful evolutionary optimization framework [135]. We demonstrate the computational speed of our method using an exhaustive analyses of simulation data sets with varying sizes, and show that the power of our GPU-accelerated implementation to detect causal interactions is equivalent to that of our the method described in Chapter III [135].

To demonstrate the utility of this software tool, we consider the Wellcome Trust Case Control Consortium (WTCCC) Crohn Disease data set. The diseases studied in the WTCCC have been analyzed using a number of methods, including several gene-gene interaction approaches [75, 106, 133, 134, 136]. In this chapter, we apply our high performance tools to a real GWAS using a biochemical pathway approach to search for meaningful interactions. This integrative approach highlights the applicability of our

software to a real data set with the goal of better understanding genetic interactions underlying the molecular pathophysiology of common diseases.

B. Methods

B.1. Proposed algorithm and GPU implementation

The motivation for the development of a GPU procedure for analyzing genetic data is to extend the general optimization procedure described in Chapter III to determine the LD block combinations with the best fitness [135]. For convenience, we briefly summarize the algorithm and notation as described above. S is the set of all possible k-SNP genotype combinations with exactly one SNP from each linkage block. The observed frequency counts for each k-SNP combination are tallied in a 2 x 3^k contingency table G_s of case and control genotype counts. For each s in S with contingency table G_s, we consider the χ^2 statistic with $3^k - 1$ degrees of freedom. The fitness of each block combination is once again defined as: $F = \max\{\chi^2_{3^k-1}(G_s) : s \in S\}$. The maximum search depth of S in each iteration is set as a parameter. For studies in this chapter, we used a deep search. The evolutionary move probabilities were once again defined as follows:

1. *Immigration*: p=0.6, randomly select a new block combination
2. *Mutation:* p=0.2
    a. Substitute one of the blocks with a random block
    b. Modify one of the two blocks by randomly choosing a nearby block

3. *Recombination*: p=0.2, substitute one of the blocks with a block from another randomly selected solution in the population

The size of the population of solutions, T, is a parameter that was set at 100, and the termination criteria are the same as above.

The main computational burden in our evolutionary optimization algorithm is the calculation of genotype frequency counts for k-SNP combinations. Tallying the genotype counts for a set of k-SNP combinations in a large population of individuals can be computed very efficiently on a GPU using a modified parallel reduction algorithm (a toy example for k=2 for 4 subjects is shown in Figure 4.1) [132, 137]. We used the NVIDIA Compute Unified Device Architecture (CUDA) for our GPU implementation. The CUDA programming model is a single-instruction, multiple-data platform that executes functions using parallel *threads* within units of data grouped into *GPU blocks*. Multiple GPU blocks can then be run in parallel using the GPU multiprocessors. Our implementation uses two main memory spaces on the GPU: (1) the larger (and slower) *global memory* for genotype storage, and (2) the smaller (and faster) *shared memory* for tallying frequency counts. Global memory is the largest memory space on the GPU, can be accessed by any GPU block, and is the only data directly accessible from the CPU. By contrast, shared memory is much smaller, and is only accessible by threads within a single GPU block.

As described in previous implementations of GPU gene-gene interaction analysis, a common way to achieve optimal performance is to structure the data such that the load is balanced across the resources of the GPU, with the majority of data-operations occurring in the fastest performing shared memory cache [132, 134, 137]. Genotypes are copied onto the GPU in global memory. For a study with N patients, each k-SNP combination is then copied into a $N*3^k$ element vector in shared memory and reduced to a list of frequency counts for each of the possible 3^k genotypes (Figure 4.1). If a vector of $N*3^k$ elements does not fit within a single GPU block, the frequency counts are split into

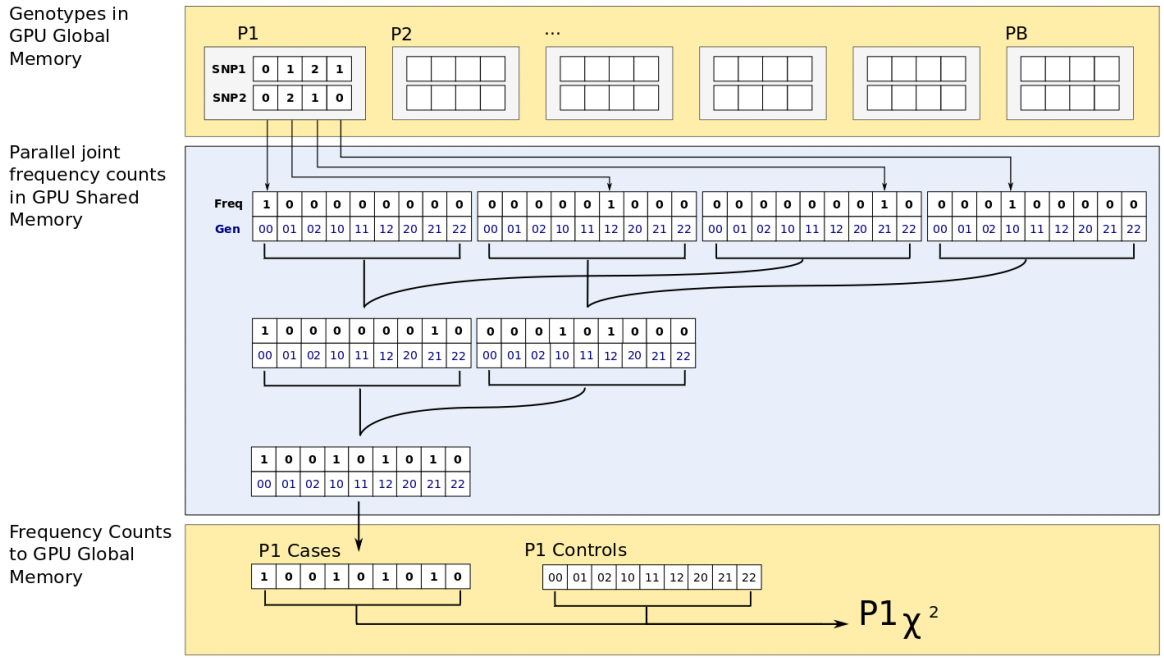


Figure 4.1. Overview of the parallel reduction approach. In this example, there are SNP combinations (k=2) *P1* through *PB* sent to *B* GPU blocks in GPU global memory. We show an example parallel reduction for *P1* with a two-SNP combination for 4 subjects in the case group. These genotype frequency counts are copied from the GPU global memory to shared memory to rapidly tally the frequency counts in parallel for all *B* GPU blocks. The frequency counts are then copied back to the GPU global memory, where they can be used to calculate the fitness measure before being copied to the CPU.

multiple GPU blocks, and the frequencies for all *N* patients are combined in global memory after the reduction. Performance of our parallel reduction on a random simulation data set as a function of shared memory usage per GPU block is shown in Figure 4.2a. Current CUDA hardware limits the number of thread processors to 32 per GPU block. Thus, designating a larger number of threads to operate in parallel in each GPU block (the "GPU Block Size", which can be up to 1024 on the NVIDIA GTX 470 and is directly proportional to shared memory usage in our algorithm) may decrease performance in smaller studies. This is because the amount of time it takes to initialize the genotype data in shared memory for our parallel reduction is not offset by the improved speed afforded by shared memory when there is less data. For larger studies, however, the performance gains obtained by increasing the usage of shared memory is worth the initialization cost. To maximally take advantage of the GPU resources and to minimize overhead due to data transfer between the GPU and the CPU, we define a parameter *B* as the number of *k*-SNP combinations calculated in parallel on the GPU during each iteration. The value of *B* depends on the specific GPU being used, and is based on the amount of GPU memory, the number of GPU multiprocessors, the maximum number of GPU blocks that may be run concurrently on the GPU, and the number of GPU blocks used per frequency count (based on *N* and the amount of GPU shared memory). Performance of our code as a function of *B* is shown in Figure 4.2b and Figure 4.3. For all calculations in this chapter, we set *B* to be the maximum value permitted for our hardware (65,535 for the NVIDIA GTX 470).

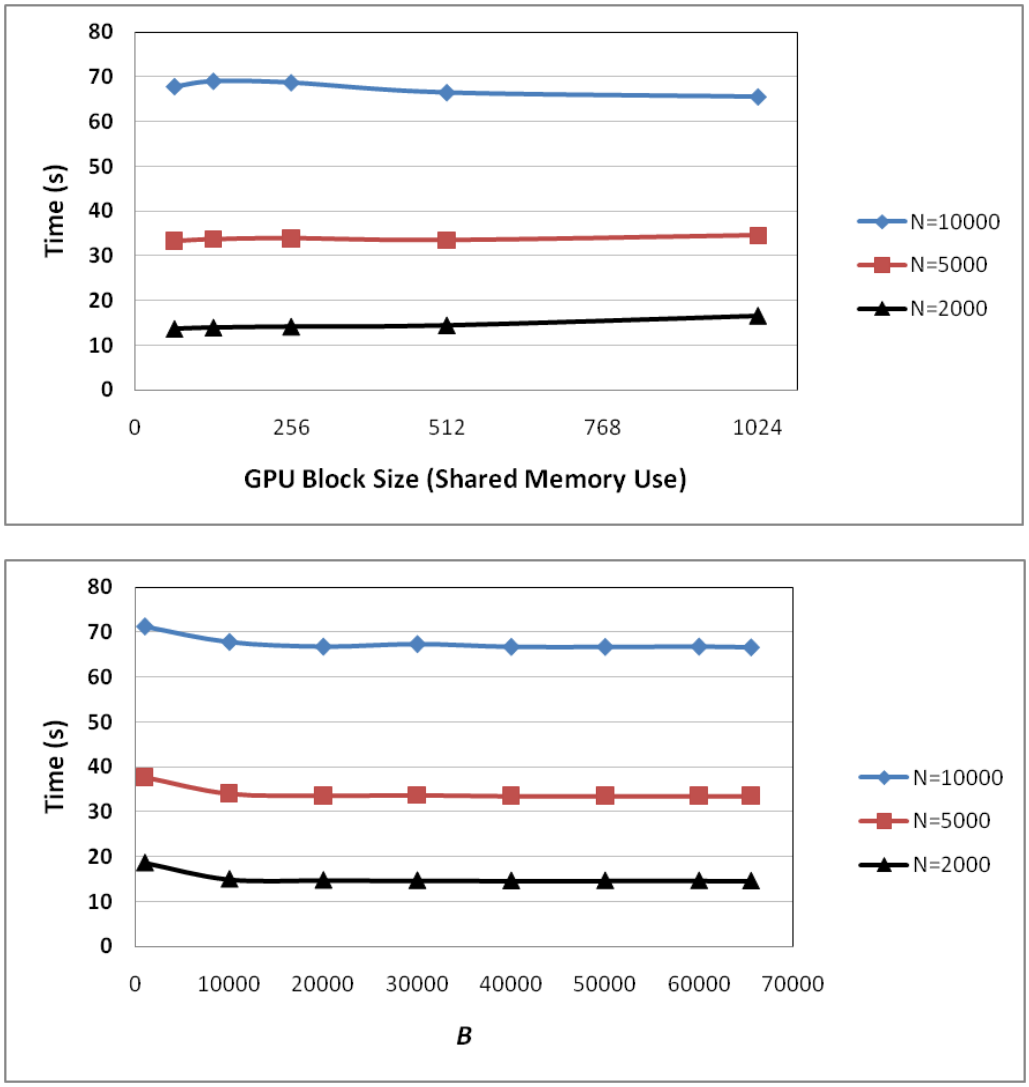


Figure 4.2. Parameter tuning for (*a*) the amount of shared memory used per joint SNP frequency calculation (GPU Block Size) and (*b*) *B*, the number of pair-wise SNP combinations calculated on the GPU during each pass. Each data point represents the average running time from 5 repeated analyses of a random case control study simulation (no causal disease loci). In each run, our exhaustive GPU procedure was used to calculate χ^2 values between all pairs of SNPs. N is the number of subjects, and the number of SNPs in each study was fixed at 2000. Genotypes were stored on the GPU in these calculations. In each pass, the indices for successive sets of *B* SNP pairs were sent to the GPU, and the frequency counts and χ^2 values were calculated on the GPU and returned to the CPU.

The evolutionary population size (*T*) that was determined to be the best for exploring genetic interactions in our previous study is much smaller than *B*. To adapt our evolutionary optimization framework to a parallel version that maximally takes advantage of the GPU resources while maintaining power to detect causal *k*-SNP interactions, we used a parallel islands approach (Figure 4.4). In this method, we initialize a set of *L* separate populations (which we refer to as *islands*), each of which has *T* solutions, such that there are a total of *B* *k*-SNP combinations across all islands. Each island population is then modified using immigration, mutation, and recombination for a

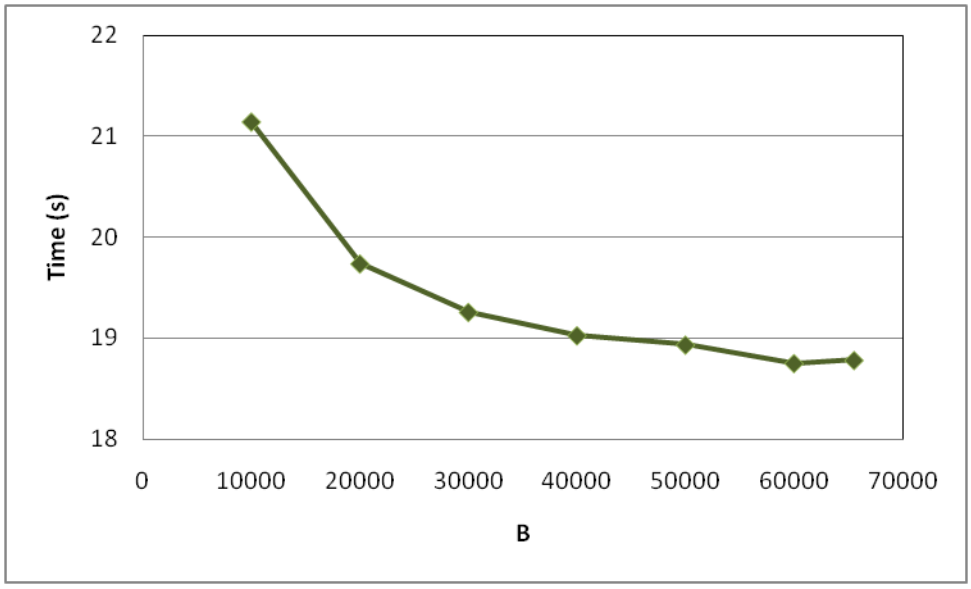


Figure 4.3. Parameter tuning for *B*, the number of pairwise SNP combinations calculated on the GPU during each pass if the genotypes are also sent to the GPU in each pass. Depending on memory constraints of a given GPU, it may not be possible to store all the genotypic data on the GPU. As in Figure 4.3, each data point represents the average running time from 5 repeated exhaustive pairwise χ^2 analyses of a random case control study simulation (no causal disease loci). In this case, the genotypes were not stored on the GPU, and instead were sent to the GPU in each pass -- making the parameter *B* more important. This situation arises for our software when the number of SNPs and/or the number of subjects becomes large. Data in this figure were collected for *N*=1000, *M*=1000.

finite number of iterations. The evolutionary moves, the move probabilities, and the definition of F are the same as in the CPU implementation. The island populations are independently optimized for a set number of iterations until the fitness of the solutions in each island is improved. The combinations across all islands are then shuffled together and regrouped into *L* populations, and the procedure is repeated until a stable set of gene-gene combinations results.

The number of independent islands (*L*) is a parameter that was defined as a function of the number of GPUs, *B,* and the number of fitness measures to calculate in each population. This approach allows our model to be adjusted to fit the resources of the specific GPU being used in order to maximize performance. We determined the performance of our code using a computer with an i5 Quad-core 2.66GHz processor, 8GB of RAM and a NVIDIA GTX 470 GPU running CUDA 3.2 on Ubuntu Linux 10.10. GPU code was implemented and tested in CUDA or C for best performance and then wrapped with Python 2.6 using the PyCUDA library (2011.1) [138].

B.2. Simulation model and performance analysis

To establish equivalence in power of our GPU implementation with that of the CPU implementation of our evolutionary optimization strategy for moderately sized studies, simulation models were built as described previously [68, 135] using a multiplicative two-locus disease model (Model 1). In each simulation data set of 5000 SNPs, the disease prevalence was fixed at 0.01, the genotypic effect size was set at a

relative risk of ω=1.25, the Minor Allele Frequency (MAF) was set at 0.15 or 0.30, and the population was set to 1000, 2000, and 5000. The disease model included two loci: each with a marginal effect and an effect size of ω*ω for any interaction between disease alleles at the two causal loci.

Because the running time of our optimization procedure is dependent on several parameters and random chance, we used exhaustive analyses over smaller sets of SNPs

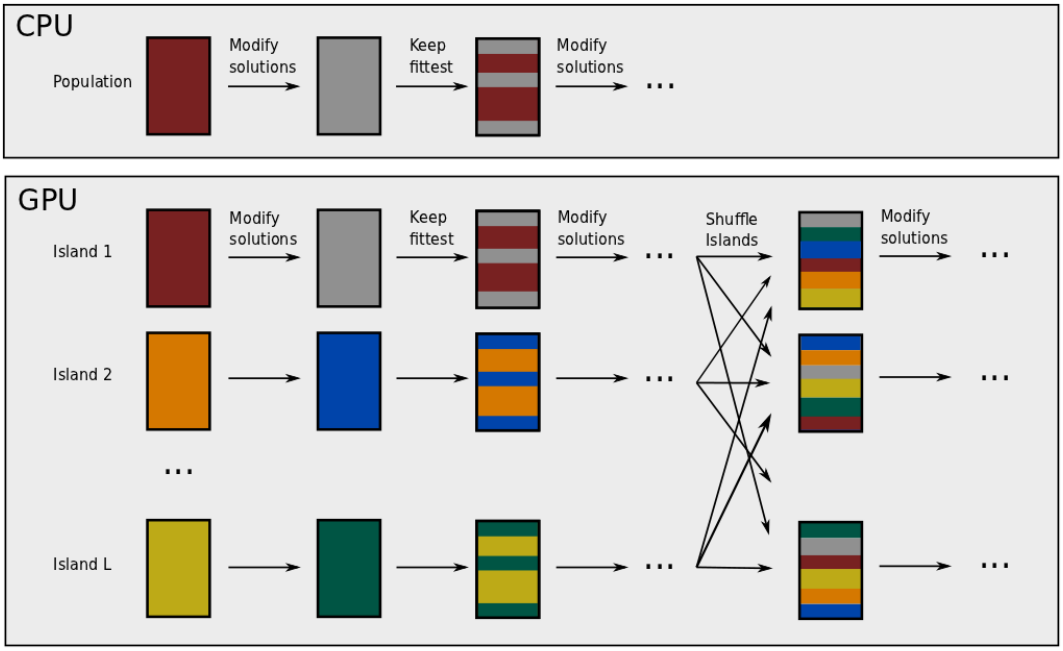


Figure. 4.4. Schematic comparing the CPU Evolutionary Algorithm with the GPU Parallel Islands implementation. In our CPU procedure, an initial "population" of combinations is iteratively modified using the immigration, mutation, and recombination moves described in the text. This process continues until a maximum number of iterations is reached or a convergence criterion is satisfied. In the GPU implementation, we initialize a set of *L* "island" populations. Each island is a population of combinations that is independently modified using the evolutionary moves. After a set number of iterations, the combinations across all islands are shuffled and the process repeats until the convergence criteria is satisfied.

on random simulation data (no simulated effect) to measure the performance of our code: N = (2000, 5000, 10000), M = (1000, 2000, 5000), where N is the number of samples and M the number of SNPs.

The population and linkage disequilibrium features in our simulation data were based on HapMap CEU phased data (CEPH samples with ancestry from Northern and Western Europe) for autosomal SNPs corresponding to the Illumina HumanHap 550K SNP array [38, 68]. Linkage structure was measured and included in our models as described previously [135].

B.3. Linkage Disequilibrium Calculations

To expand the utility of our GPU software, we implemented several basic functions for conducting genetic analysis (LD, HWE calculations, simple plotting functions). Single locus analysis (e.g. HWE calculations, single-locus association tests) of GWAS is not particularly computationally expensive, and these functions are included for convenience. However, for intensive analysis of LD patterns, the computational efficiency of GPU-accelerated frequency counts offer an improvement to commonly used methods. Large-scale LD patterns can be approximated using efficient matrix methods to compute a correlation matrix that estimates patterns of non-independence between SNPs [98]. However, a more precise analysis of LD patterns requires the direct estimation of two-locus haplotype frequencies, as discussed in Chapter II. One of the most commonly cited software packages that can be used for this purpose is Haploview [61, 139]. We

adapted the standard LD procedures used in Haploview to take advantage of GPU resources. In our LD calculations, the GPU was used to compute the joint frequency counts for each pair of SNPs, and these frequency counts passed back to the CPU for the calculation of the LD statistics as described above [59, 61, 62, 139]. We compared the performance of our code [139] using the abovementioned computer (i5 Quad-core 2.66GHz processor, 8GB of RAM, NVIDIA GTX 470 GPU) with Haploview 4.2 (64-bit Java 1.6.0_20). Both methods were run from the command line, and the running times were measured using the linux time utility (reported as elapsed "real" time in seconds). The number of subjects in the study (*N*) was set to 1000, 2500, 5000, and 10000, and the number of SNPs in the study (*M*) was set to 1000 and 2500. LD measures (D', the 95% confidence interval for *D'*, and r^2) were calculated for all pairs of SNPs in each simulation.

B.4. WTCCC Data Processing and Analysis

Having demonstrated the utility of our GPU-accelerated approach for GWAS analysis, we applied our software to a subset of data from the WTCCC [39]. Of the WTCCC diseases studied, the findings for Crohn Disease were unique in that they revealed a set of biochemical pathways that were consistently enriched with significant SNPs in a number of studies [140, 141]. There have also been published analyses of gene-gene interactions within this data set [1, 142]. However, the set of gene-gene

interactions within the subset of pathway SNPs in the WTCCC Crohn Disease data set has not been determined.

We accessed the Affymetrix GeneChip Mapping 500K array set called by the CHIAMO algorithm [39] for the Crohn Disease (CD) case cohort and for the 1958 birth control cohort. Following the recommendations of the WTCCC, we excluded a number of subjects (as listed the "exclusion-list-05-02-2007.txt" file accompanying the GWAS data), and any SNPs not conforming to the criteria: (1) missing data proportion > ~0.05, (2) MAF < 0.05 and missing data proportion > 0.01 or MAF < 0.01, (3) combined control group (58C+National Blood Service) HWE Exact Test p-value < 5.7e-7, (4) Inter-control group (58C vs NBS) 1df Trend Test p-value < 5.7e-7 (as measured by WTCCC) (5). 58C vs NBS 2df General Test p-value < 5.7e-7 (as measured by WTCCC). CHIAMO SNP measurements with a score < 0.9 were considered missing data. Following these exclusions, there were 1,480 control subjects and 1,748 CD subjects genotyped at 496,410 SNPs.

Next, we updated the information obtained in the WTCCC to be consistent with current genomic info for hg19/Genome Reference Consortium Human genome build 37 (GRCh37). Data about the reference positions for SNPs in the study were downloaded from NCBI dbSNP (build 132) for GRCh37, information about genetic locations were downloaded from RefSeq using the UCSC Genome Browser [143], and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways and their mappings to our data set were downloaded from existing peer-reviewed sources [144-146]. We used this

information to remove any SNPs in the WTCCC data not associated with any KEGG pathway, resulting in 39,664 autosomal SNPs in 3,666 genes in 212 pathways. As reported previously, there are several very strong single locus effects present within the set of pathway SNPs [141]. After confirming these strong, previously published single-locus associations in our data set, we removed any SNPs with a p-value < 1×10^{-3}. We then analyzed the data set for interactions for k=2 using our GPU-accelerated optimization procedure as described above. After running our analysis, any interactions between two blocks with SNPs in a nearby chromosomal region (within 100kb) with a LD measurement of either $r^2 \geqslant 0.8$ or $D' \geqslant 0.8$ were not reported as significant interactions. For this analysis, the typical missing value cutoff of 0.95 was relaxed to 0.94 to allow exploration of a previously published, top-ranked interaction between the KEGG pathway SNPs rs7154773 and rs10130695 in the WTCCC Crohn Disease Data Set [142]. While this interaction is indeed very significant ($\chi^2_{8df} = 88.62$, $p < 1 \times 10^{-16}$), these two loci (rs7154773: Chromosome 14, GRCh37 position 60749118; rs10130695: Chromosome 14, GRCh37 position 60755984) have a D' value of 0.962 – indicating that they are strongly correlated. For comparison with our results in a separate analysis, we also tested the set of SNPs involved in the top interactions reported in [1] using our objective function (rs10027689, rs10156534, rs11649428, rs12647454, rs12751992, rs1584444, rs1601668, rs17825620, rs2201677, rs2358356, rs2478836, rs301630, rs4471699, rs4677143, rs509544, rs511435, rs524731, rs636646, rs6532916, rs668394, rs7202714, rs7217284, rs7773053, rs8006622, rs9436212, rs9540533). rs4471699 was excluded because of a high proportion of missing data in the control group (~10%), and

the other SNPs were not included in our optimization algorithm since they are not on any of the KEGG pathways. Our approach seeks to provide information about meaningful gene-gene interactions between separate genetic loci by limiting our analysis to SNPs that are not in strong LD within the set of pathway SNPs. While we consider a limited data set in this analysis, our method is capable of analyzing the full data set of 496,410 SNPs, and an exhaustive search for combinations in this data set is planned in future work.

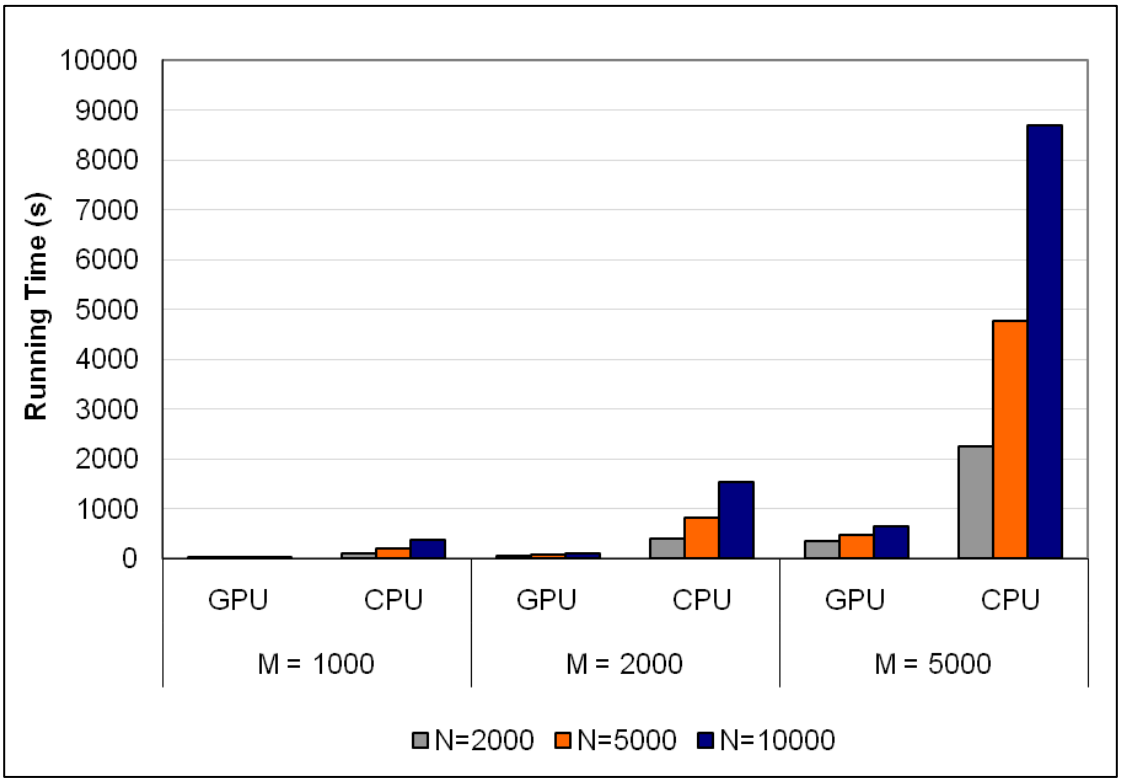


Figure 4.5. Comparison of GPU and CPU running times in seconds for exhaustive search on data sets of varying size. *N* is the number of subjects and *M* is the number of SNPs.

C. Results

C.1. Simulation Performance

Since the running time of our optimization procedure may be variable, we elected to use a limited exhaustive search to assess the comparative running time of our GPU and CPU code. The sample sizes N=(2000, 5000, and 10000) correspond roughly with a moderately sized GWAS, a large GWAS [39], and a combined analysis of more than one GWAS. As shown in Figure 4.5, the running times are comparable for small studies, but the computational benefit of the parallel implementation becomes readily evident for exhaustive analyses of even a limited set of SNPs.

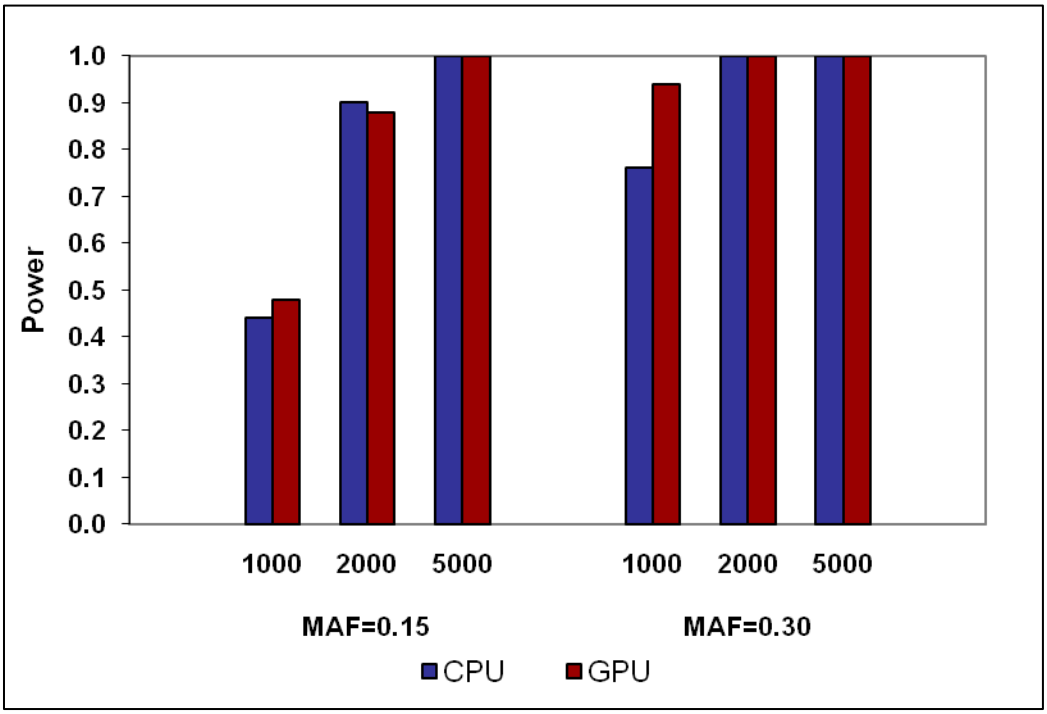


Figure 4.6. Comparison of the power of the GPU and CPU implementations to detect causal genetic markers. Power is reported as the proportion of times the causal gene-gene interaction was found in 50 simulations for *M*=5000, and *N*=(1000,2000,5000).

It was important to establish that the GPU and CPU implementations of our optimization approach had equivalent power to detect a simulated gene-gene interaction. We defined power as the proportion of times out of 50 replicate simulation data sets that the causal interaction was detected. As shown in Figure 4.6, the GPU and CPU versions of our optimization algorithm have nearly identical power.

C.2. Linkage Disequilibrium

Results for the LD calculation running time comparison between our software and Haploview for simulation data sets with varying numbers of subjects and SNPs are shown in Figure 4.7. Care was taken to ensure that the same procedures were completed in both software packages to ensure a fair comparison of computational efficiency. Our software increased the speed between 6x and 38x, with more impressive improvements in computational efficiency observed for the larger studies. Further improvements in the performance of our LD calculation may be possible in the future through the use of CPU multithreading resources to parallelize the CPU computations for improved performance relative to the single-processor results from our method shown in Figure 4.7.

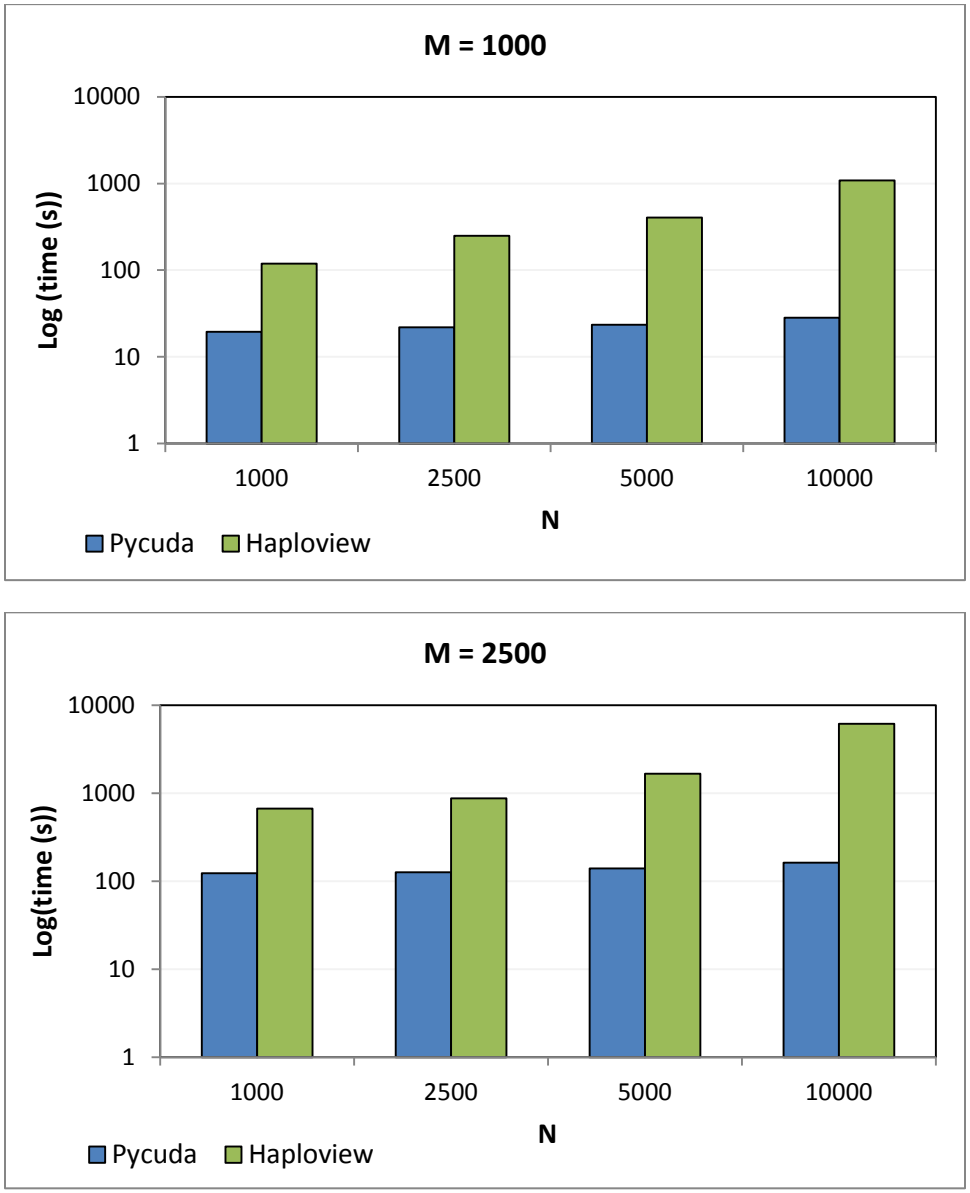


Figure 4.7. Comparison of exhaustive LD calculation running times (in log(s)) for our method (blue) and Haploview (green) for varying numbers of subjects (N) and SNPs (M).

C.3. WTCCC

Results from our analysis of the WTCCC Crohn Disease and 58C Control data set are included in Table 4.1. Our evolutionary optimization procedure was run using the same parameters and block procedures described above. Our analysis revealed a number of significant loci in different genetic regions that have not been described in previous analyses of interaction effects in this data set. The approach in [142] only yielded interactions between SNPs that are on the same chromosome, and the most significant pair of interacting SNPs reported in this paper involved SNPs from the same gene that

		SNP1					SNP2			Interaction	
ID	Chrom	Gene	χ2	p-value	ID	Chrom	Gene	χ2	p-value	χ2	p-value
rs1328199	1	VAV3	1.66	0.4354	rs6121731	20	CDH4	11.13	0.0038	51.8	1.84E-08
rs2710779	3	FGF12	13.5	0.0012	rs12709950	19	UBE2S	4.47	0.1069	49.09	6.12E-08
rs3769222	2	RAPGEF4	6.01	0.0496	rs12544197	8	TRHR	2.29	0.3189	47.85	1.06E-07
rs2686322	3	KAT2B	0.41	0.8156	rs2830075	21	APP	7.96	0.0187	47.69	1.13E-07
rs2686322	3	KAT2B	0.41	0.8156	rs2186302	21	APP	9.55	0.0084	47.61	1.17E-07
rs2686322	3	KAT2B	0.41	0.8156	rs2234988	21	APP	9.77	0.0075	47.19	1.41E-07
rs4950485	1	GJA5	5.3	0.0708	rs6592775	11	GAB2	4.2	0.1225	46.86	1.63E-07
rs17664296	3	PRICKLE2	4.07	0.1309	rs11784860	8	ZNF34	4.13	0.127	46.85	1.64E-07
rs7303842	12	CRY1	4.19	0.1233	rs7995795	13	CLDN10	9.61	0.0082	46.74	1.72E-07
rs1048156	10	PPYR1	2.54	0.2812	rs3923871	11	OR4B1	7.29	0.0261	46.02	2.35E-07
rs2246209	1	NR5A2	2.66	0.2643	rs9333117	10	ITGA8	9.56	0.0084	45.93	2.45E-07
rs8041887	15	ARNT2	12.9	0.0016	rs2837193	21	IGSF5	6.42	0.0403	45.79	2.61E-07
rs697690	1	PER3	10.9	0.0043	rs1249873	1	PLXNA2	6.22	0.0447	45.76	2.64E-07
rs3772074	2	TPO	1.01	0.6028	rs1893572	18	DCC	10.2	0.0061	45.61	2.81E-07
rs17591814	1	PLA2G4A	0.36	0.8341	rs1345423	16	GRIN2A	1.14	0.5669	45.58	2.86E-07
rs1867971	10	HK1	2.53	0.2823	rs6061892	20	CDH4	5.76	0.0562	45.3	3.23E-07
rs2566539	2	CTNNA2	4.61	0.0996	rs2538990	7	CNTNAP2	7.84	0.0198	45.2	3.37E-07
rs12613346	2	GALNT13	0.72	0.6981	rs1980846	2	ABCA12	11.88	0.0026	45.15	3.45E-07
rs3118182	1	LAMC2	6.07	0.0482	rs7192535	16	PLCG2	11.82	0.0027	45.04	3.61E-07
rs6679356	1	IL12RB2	0.31	0.8571	rs13361707	5	PRKAA1	13.8	0.001	44.9	3.84E-07

Table 4.1. Table of the most significant pairwise interactions resulting from our analysis of the WTCCC Crohn Disease and Control data sets.

had a high degree of correlation as measured by D'. While the method in [142] reported the strongest interaction occurring between SNPs rs7154773 and rs10130695 (χ^2_{8df} = 88.62, $p < 8.8 \times 10^{-16}$), our analysis found a slightly stronger interaction within the same locus between SNP rs7154773 and rs8011227 ($\chi^2_{8df} = 91.03$, $p < 2.9 \times 10^{-16}$) (Our method also returned the previously reported rs7154773 and rs10130695). These interactions were removed from our results because of the high degree of correlation between these two SNPs. The reason our algorithm was able to detect these strongly correlated interactions is that they happened to be split up in two separate linkage blocks in our analysis. Note that if either of these two SNP combinations had been within the same LD block (as is plausible, given their strong LD measures), our method would not have been able to detect these interactions. The only SNP interaction published for the WTCCC Crohn Disease data set in [147] was measured to have a χ^2_{8df} of 30.961 (p=0.00014277) in our data set, and was thus not as significant as the results reported in Table 4.1. Similarly, the interactions found using the --fast-epistasis method in PLINK for this data set [1] were not as significant using our more general χ^2_{8df} statistic, and were also not as significant the interactions found in the pathway set using our approach (Table 4.2). The one exception was the interaction between rs4677143 and rs8006622 ($\chi^2_{8df} = 51.4$, $p = 2.17 \times 10^{-08}$).

There are a several interesting features to highlight from our list of top results in Table 4.1. There are a number of very strong interactions in our list of top results that had no detectable difference between cases and controls for either of the individual SNPs.

Second, the findings using our general χ^2 statistic may be different from those calculated using the trend χ^2 statistic used by PLINK's --fast-epistasis method [1]. Lastly, the genes reported in these interactions are largely different from those reported in previous pathway analyses of Crohn Disease. In the absence of confirmatory results in separate GWAS or laboratory experiments, interpretation of these new interactions would be speculative. However, some of the results are in keeping with previously observed results from pathway analyses of Crohn Disease GWAS [140]. For example, one of the interacting SNPs is found within the IL12RB2 gene. Single-locus associations with this gene have been found in a number of GWAS, but typically at a different locus with a very strong marginal effect. The SNP found in our analysis, rs6679356, had no marginal effect, but did have a strong interaction with PRKAA1, a signaling molecule related to cellular energy stores. Further investigation into evidence supporting the disease association of these gene-gene interactions in other large-scale GWAS of Crohn Disease are necessary to confirm these findings.

SNP1				SNP2				Interaction	
ID	Chrom	χ2	p-value	ID	Chrom	χ2	p-value	χ2	p-value
rs9436212	1	2.73	0.2553	rs11649428	16	3.62	0.1639	34.19	3.75E-05
rs12751992	1	2.19	0.3351	rs1601668	12	0.09	0.9560	23.48	2.80E-03
rs4677143	3	12.03	0.0024	rs8006622	14	6.33	0.0423	51.42	2.17E-08
rs1584444	4	5.40	0.0671	rs2201677	4	2.38	0.3049	19.35	1.31E-02
rs1584444	4	5.40	0.0671	rs12647454	4	1.20	0.5477	19.25	1.36E-02
rs1584444	4	5.40	0.0671	rs6532916	4	1.69	0.4298	19.73	1.14E-02
rs1584444	4	5.40	0.0671	rs10027689	4	1.88	0.3898	18.31	1.90E-02
rs668394	6	3.30	0.1919	rs10156534	9	2.33	0.3116	37.66	8.70E-06
rs511435	6	3.59	0.1661	rs10156534	9	2.33	0.3116	36.02	1.74E-05
rs509544	6	3.14	0.2081	rs10156534	9	2.33	0.3116	36.84	1.23E-05
rs524731	6	3.17	0.2045	rs10156534	9	2.33	0.3116	36.99	1.16E-05
rs7773053	6	2.99	0.2248	rs17825620	14	0.62	0.7339	37.98	7.58E-06
rs2358356	10	1.70	0.4280	rs9540533	13	0.71	0.6998	33.46	5.09E-05
rs2478836	10	2.04	0.3611	rs7217284	17	1.44	0.4867	41.70	1.54E-06
rs636646	13	4.13	0.1267	rs301630	16	7.04	0.0296	34.63	3.12E-05

Table 4.2. A list of the top interactions from [1] computed using our objective function for comparison with our method.

D. Discussion

As shown in previous studies [132-134], parallel GPU computing techniques can substantially improve the performance of genetic analysis tools. Our block-based evolutionary optimization strategy described in Chapter III was designed to improve the power and efficiency of gene-gene interaction detection by taking advantage of local LD structure. This parallel implementation of our algorithm extends this approach so that it can be more conveniently applied for exploratory analysis of genome-scale data. We demonstrated that our GPU method has equivalent power using a standard multiplicative two-locus disease model that included marginal effects. As has been shown previously,

this power is expected to be higher for our method when the sample size is large and when the causal allele is more common. Any differences in power between the two implementations (as shown in Figure 4.6) are due to random variations in our evolutionary algorithm or a slightly increased search space of the GPU implementation of the algorithm

We expanded the set of genomic association analysis tools accelerated by our parallel implementation to include basic calculations of HWE and LD, and demonstrated that our method offers substantial improvements in performance relative to a widely cited method. We demonstrated the utility of our approach by applying our analysis method to the WTCCC Crohn Disease data set. In this analysis, we report a number of gene-gene interactions that were not previously reported by other large scale analyses of epistasis in the WTCCC Crohn Disease data set [1, 142, 147]. Further study of these interactions in other GWAS to confirm our findings is necessary. A previously conducted semi-exhaustive analysis of 89,294 SNPs from this data set took 14 days on a single node of a computer cluster [1]. An analysis using BEAM on a data set of 47,727 SNPs took roughly 8 days [39, 136]. Our analysis of the set of 39,664 SNPs using our GPU evolutionary algorithm described above took slightly more than one hour. We plan to expand our analysis of the WTCCC data in exhaustive analyses of pairwise interactions. However, prior to undertaking this large-scale task, the efficiency of our analysis method can be improved even further using CPU multiprocessing techniques. The calculations reported above only take advantage of one CPU core and one GPU card at a time. While

the speed improvement we observe is quite useful for studies of limited sets of SNPs, full scale analysis (which would take about 8 days with our current GPU code and about 195 days with our CPU code, each only using one CPU core) cannot be offloaded onto a single GPU concurrently. Thus, this type of analysis has more CPU involvement and will benefit substantially from multiprocessing techniques as well as the use of multiple separate GPUs [133].

One limitation of our current software is that it will only run on devices compatible with NVIDIA CUDA – which is free for academic use, but is currently restricted to NVIDIA hardware. Forthcoming versions of our code will allow our software to be run on hardware from other vendors by translating the necessary GPU functions from CUDA to OpenCL [148]. In our future research, we plan to expand the set of genomic association analysis tools accelerated by our parallel implementation, and we will also explore new ways of further improving the performance of our code as new hardware and techniques emerge.

CHAPTER V

USING LASSO REGRESSION TO DETECT PREDICTIVE AGGREGATE EFFECTS IN GENETIC STUDIES

A. Introduction

Besides the associations with common complex diseases discussed in the previous chapters, GWAS of common variants have revealed numerous genetic loci that significantly modulate phenotypes for a wide assortment of other important clinical phenotypes ranging from the expected risk of certain malignancies [149, 150] to commonly measured clinical traits such as lipid levels [151]. These results are promising, but it is nevertheless increasingly evident that the common variants found in GWAS provide an incomplete picture of the underlying genetic risk for many of the familial diseases that have been studied [152-154]. Thanks to the increased availability of sequencing technologies and to large scale efforts such as the 1000 Genomes Project, exome scans are becoming increasingly popular in complex disease genetics. These studies represent several new challenges in genetic analysis.

Although a variety of machine learning methods have been used in GWAS [90], penalized regression methods are among the most flexible and are thus well-suited for analysis of data sets such as exome scans, which may contain both common and rare effects. Numerous penalized regression methods have been shown to be effective for both common and rare variants [87, 89, 91, 152]. Zhou et al [152] proposed a

combination of group and lasso penalties to find both rare and common variants using sets of markers grouped by pathway and gene. However, their method was evaluated using family breast cancer registry data, and its performance is unclear for larger scale data from GWAS.

To improve accuracy, some studies have imposed an arbitrary p-value cutoff to limit the number of genetic variants in the lasso model [91], whereas others have applied the model across all variants using the lasso penalty and a group penalty for the gene or pathway [152]. In this study, we propose an approach using a lasso model that first selects sets of genetic variants for each pathway and gene and then generates an optimized lasso model based on the selected marker sets. Taking advantage of information provided in the Genetic Analysis Workshop 17 (GAW17) exome data set, we can build two lasso models for each pathway or gene based on regression on disease status or on a quantitative trait. This approach is more time-consuming than optimization of a lasso model for the full set of variants. However, our strategy permits us to build individual optimal models on each of the variant sets related to the pathway and gene, allowing a more flexible and accurate model determination. In the remainder of this Chapter, we examine the performance of this new approach using the GAW17 exome scan simulation data set.

B. Methods

B.1. Data Description

The GAW17 data set contains 697 unrelated individuals from the 1000 Genomes Project genotyped at 24,487 autosomal SNPs from 3205 genes [155]. 206 pathways from the Kyoto Encycolpedia of Genes and Genomes (KEGG) [144] are represented, spanning 7,929 different SNPs and 1100 different genes. We restricted our analysis to the 13,572 non-synonymous variants in the study. Each of the 200 simulated datasets includes the following information for each individual: case-control status, 3 continuous quantitative traits (Q1, Q2, Q4), and 3 phenotypic features (Age, Smoking status, and Gender). We used a multidimensional scaling analysis (Figure 5.1) based on genomewide pairwise identity-by-state distances computed in PLINK [98] to independently verify the 3 main

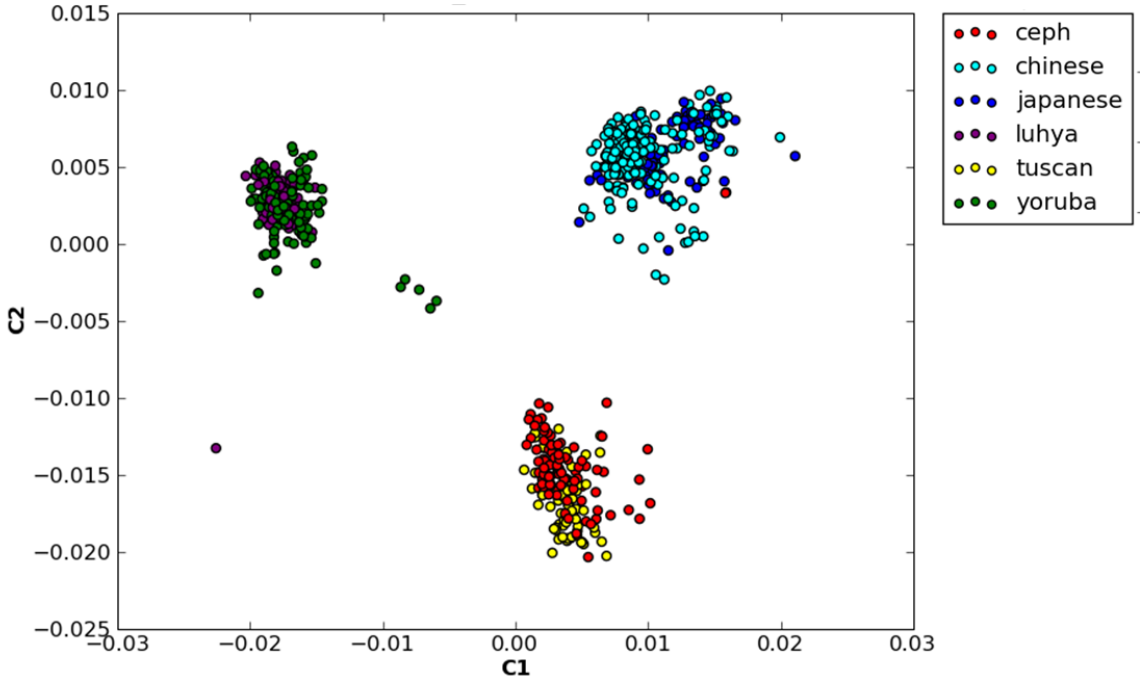


Figure 5.1. Multidimensional scaling plot showing population stratification in this simulation study. In this graph, each point is an individual, and the two axes correspond to a reduced representation of the data in two composite dimensions (arbitrarily labeled C1 and C2). We generated 3 binary features to include in our model, assigning patients to their corresponding Asian (blue and teal), European (red and yellow), and African (green and purple) strata.

continental population strata described by other groups at GAW17 (African: Luhya, Luhya - additional, Yoruba-1, Yoruba-2, Yoruba-additional; Asian: Denver Chinese, Denver Chinese-additional, Han Chinese -1, Han Chinese-2, Han Chinese-additional, Japanese-1, Japanese-2, Japanese - additional; European: CEPH - 1, CEPH - 2, Tuscan, and Tuscan – additional)) [156-158]. We then generated 3 binary features to include in our model, assigning patients to their corresponding Asian, European, and African populations. Two population outliers were removed from our analysis.

B.2. Analysis

We use the R software package *glmnet* in our analysis for lasso regression [92] and evaluate our models using a 5-fold cross-validation procedure for each simulation dataset. More specifically, we split the datasets into 5 independent folds of approximately equal size such that the case-control ratios in each population are maintained in each fold. Models are trained using 4 folds of the data and then tested using the remaining fold. This procedure is repeated for each of the 5 training and testing fold combinations. To determine an optimal value λ^* for each training set, we apply an inner loop of 10-fold cross-validation. Then λ^* is used on the entire training set to build the final model for the evaluation of the testing fold. Finally, the averaged evaluation measures over the 5 testing folds are reported as testing accuracy. In our analysis the evaluation measures are the area under the receiver operating curve (A_{ROC}) for logistic models and the mean squared error for continuous linear regression models.

We consider three basic models:

1) Lasso logistic regression with all genetic variants included

2) Lasso logistic regression for each of the (a) 3,205 genes or (b) 206 pathways, followed by a lasso regression using the combined set of selected variants from all genes or pathways; and

3) Three separate lasso linear regression models for each of the continuous quantitative traits Q1, Q2, and Q4 for each pathway, followed by a lasso logistic regression over the entire set of selected variants across all pathways.

For each of these strategies, we consider a genotype-only model, a combined model that includes phenotype information (Age, Smoking, and Gender), and a restricted model that is limited to a fixed number of variables. In this study, the restricted models are limited to have a maximum of 50 variables.

Model 1 is similar to most other applications of the lasso regression model, in which a single regularization parameter is used. This model is convenient and computationally efficient, but its ability to detect local effects within biologically meaningful subsets of genes that are of interest in an exome study may be limited. Models 2 and 3 first determine optimized models for each gene or pathway, and then run a lasso regression over the combined set of variants selected for each gene or pathway.

C. Results

C.1. Performance of the models

Results for all the models are shown in Table 5.1. Each of the 200 simulated datasets was analyzed separately. Because model 2 had a substantially longer running time, it was evaluated for only 50 (Model 2a) and 150 (Model 2b) randomly selected datasets. To determine the baseline performance for our models, we sampled several simulation datasets using 180 random variants (corresponding to the average size of the basic "Genotypes Only" Model 1 result). The expected average A_{ROC} for a randomly

	Model	Training A_{ROC}	Testing A_{ROC}	# True	Size	N
1	Genotypes Only	0.57	0.55	3.57	179.43	200
	Genotypes Restricted	0.56	0.55	0.84	22.07	200
	Combined Model	0.82	0.82	1.27	28.38	200
	Combined Model Restricted	0.82	0.82	1.06	18.70	200
2a	Genotypes Only	0.61	0.54	9.98	545.33	50
	Genotypes Restricted	0.56	0.55	0.86	21.66	50
	Combined Model	0.83	0.81	2.78	94.32	50
	Combined Model Restricted	0.83	0.82	1.14	20.57	50
2b	Genotypes Only	0.73	0.54	11.65	348.86	150
	Genotypes Restricted	0.58	0.56	2.01	29.57	150
	Combined Model	0.85	0.78	9.35	228.43	150
	Combined Model Restricted	0.83	0.82	2.48	29.26	150
3	Genotypes Only	0.62	0.54	11.32	294.68	200
	Genotypes Restricted	0.58	0.56	1.75	22.84	200
	Combined Model	0.83	0.82	3.94	64.17	200
	Combined Model Restricted	0.83	0.82	2.04	20.40	200

Table 5.1. Prediction results for various model types. Averaged results from a 5-fold evaluation procedure on N simulation datasets. "Training A_{ROC}" was obtained in the course of the R package *glmnet*'s internal 10 fold cross-validation on the training sets. "Testing A_{ROC}" was determined by applying each of the trained models to the 5 independent testing sets. "# True" is the average number of causal simulation markers included, and "Size" is the

selected set of variants was ~0.49. Similarly, we used *glmnet* to compute optimal models from the set of 160 causal simulation markers, and determined that the average A_{ROC} of this optimal set of genotypes was 0.59. This value represents the average predictive accuracy of an optimized subset of the genetic variants responsible for assigning disease status in the simulation, and is considered the target value of our models that use only genotype data. As observed in Table 5.1, the purely genetic models had A_{ROC} values closer to 0.55 for all models considered. The combined models with the phenotypic features had an A_{ROC} of 0.82, a universally higher average testing A_{ROC} value independent of any genotypic combination. Because of the high marginal effect sizes of the phenotypic variables (Age, Gender, and Smoking status), these effects frequently overpowered the effect sizes of genetic markers included in the lasso models. The unrestricted lasso models often resulted in solutions with a large number of variablD.es, limiting the practical utility of these models. The testing A_{ROC} values of the restricted models were often the same as or better than those of the unrestricted models, indicating better generalization ability for the restricted models. However, the predictive performance of the genetic component did not reach the best possible level and the models included larger numbers of non-causal variants. The use of gene and pathway information did not result in meaningful improvements in the regression models with respect to predictive capability.

	Model 1					Model 3				
	Gene	**SNP**	**Count**	**MAF**	**Causal**	**Gene**	**SNP**	**Count**	**MAF**	**Causal**
Gene Only	FLT1	C13S523	35	0.0667	Y	FLT1	C13S523	71	0.0667	Y
	ADAMTS7	C15S3360	22	0.0029	N	SRPR	C11S6885	63	0.0014	N
	TG	C8S4379	17	0.0050	N	TG	C8S4379	61	0.0050	N
	MDN1	C6S4146	15	0.0050	N	RPA3	C7S297	58	0.0007	N
	GOLGA1	C9S4013	13	0.0308	N	LAMB3	C1S10178	54	0.0007	N
	FLT1	C13S522	12	0.0280	Y	RPL27	C17S2981	52	0.0007	N
Gene Restricted	FLT1	C13S523	19	0.0667	Y	FLT1	C13S523	44	0.0667	Y
	TEX14	C17S3819	9	0.0043	N	FLT1	C13S522	24	0.0280	Y
	FLT1	C13S522	8	0.0280	Y	CYP3A43	C7S2324	21	0.0976	N
	UBA3	C3S2197	7	0.0108	N	TG	C8S4379	18	0.0050	N
	GOLGA1	C9S4013	7	0.0308	N	PRKCA	C17S4578	16	0.1664	Y
	CYP3A43	C7S2324	7	0.0976	N	PIK3C2B	C1S9189	15	0.0065	Y
Combined	age	age	200	NA	Y	age	age	200	NA	Y
	smoke	smoke	163	NA	Y	smoke	smoke	185	NA	Y
	FLT1	C13S523	49	0.0667	Y	FLT1	C13S523	81	0.0667	Y
	FLT1	C13S522	16	0.0280	Y	FLT1	C13S522	34	0.0280	Y
	PIK3C3	C18S2492	7	0.0172	Y	PIK3C3	C18S2492	18	0.0172	Y
	HFE	C6S853	3	0.0036	N	PRKCA	C17S4578	8	0.1664	Y
	ARNT	C1S6533	3	0.0115	Y	ARNT	C1S6533	8	0.0115	Y
	ACP1	C2S1	2	0.0093	N	UBA3	C3S2197	7	0.0108	N
Combined Restricted	age	age	200	NA	Y	age	age	200	NA	Y
	smoke	smoke	163	NA	Y	smoke	smoke	180	NA	Y
	FLT1	C13S523	49	0.0667	Y	FLT1	C13S523	75	0.0667	Y
	FLT1	C13S522	17	0.0280	Y	FLT1	C13S522	32	0.0280	Y
	PIK3C3	C18S2492	7	0.0172	Y	PIK3C3	C18S2492	17	0.0172	Y
	ARNT	C1S6533	3	0.0115	Y	UBA3	C3S2197	6	0.0108	N
	LARGE	C22S1540	3	0.0201	N	ARNT	C1S6533	6	0.0115	Y
	MMS19	C10S4869	3	0.0050	N	KDR	C4S1861	5	0.0022	Y

Table 5.2. Feature Selection. Table of the top most frequent variables occurred in at least 4/5 trained models for Models 1 and 3. All Models were run for the 200 simulation datasets. "Count" is the number of times a given variable was observed in 4/5 trained models. "Causal" indicates the variables were those used to determine disease risk by the GAW17 simulators. "MAF" is minor allele frequency.

C.2. Variables selected by the models

Table 5.2 shows results from each experiment for the most frequent variables that were selected in at least 4 out 5 trained models within a simulation dataset for Models 1 and 3. These results reveal that the true variants detected were predominantly common variants, but our model may also have some capacity to identify true rare variants. The gene and pathway based regression approaches did not seem to produce substantially different A_{ROC} values or find different casual variants than those found using the simpler lasso approach. However, as shown in Table 5.2, the proportion of those casual variant occurring was higher in Model 3, indicating a more robust model.

E. Discussion

In this paper, we assessed the utility of several different strategies for analyzing exome simulation data with a range of causal allele frequencies in the presence of quantitative and phenotypic information. A comparison of the three proposed approaches indicates that the simple lasso regression model may be an efficient means to determine truly associated variants, but it must be modified to reduce the number of variables to avoid unreasonably large models and overfitting. As discussed in other studies of these data at GAW17, the primary genetic effects that were expected to be observed in this study were those from common variants such as C13S523 and C13S522 in FLT1. As shown in Table 5.2, individual genetic variants were identified consistently in 4 out 5 training models in only a minority of simulation analyses. For example, FLT1-C13S523

occurred in at most 81 out of 200 simulations in the "Combined" analysis for Model 3. Some loss of power was expected in our analysis, because we developed our models using 80% of a simulation dataset to obtain an independent evaluation of our methods' predictive ability. However, if we consider the same model calculated on all 200 replicates using the entire set of patients (no training set), then FLT1-C13S523 is included in 132 of 200 datasets. In larger studies or in studies that have a pre-existing independent sample to validate the predictive model, this diminished power will not affect our method as strongly and our model may be better able to discern genetic predictors.

Some variants, for example PIK3C3, appeared much more frequently in the models that combined genotypic and phenotypic effects than in models that considered only genotypes. To further investigate this finding, we built logistic regression models for Y and PIK3C3, adjusting for either only population variables or both population and phenotypic variables. PIK3C3 was significant ($\alpha = 0.01$) in 22 out of 200 datasets for the model adjusted for population only and in 105 out of 200 datasets for the model adjusted for both population and phenotypic variables, providing an explanation for this observation. Our analysis also indicates a significant relationship in the linear regression for Q1 and PIK3C3 adjusted for population only (184 out of 200 datasets) and adjusted for both population and phenotypic variables (197 out of 200 datasets) at $\alpha = 0.01$. This may also explain more frequent occurrence of PIK3C3 in Model 3 compared with Model 1 for the combined models.

Our method was able to reliably ascertain some true variants using subsets of the data for training. In addition, the signs of the regression coefficients for the frequently selected variants were highly consistent (about 99%) over different simulation datasets. However, the ability of our model to find true variants was also accompanied by a large number of non-causal variants. Because several long-range correlations exist within the GAW17 data set, a portion of the variants classified as non-causal in our study may actually be truly associated with the disease state or phenotypic traits. The predictive ability of the lasso model using only genetic information is limited, because none of the genomic subsets examined had a predictive ability that was comparable to that of the phenotypic variables. Nevertheless, incorporating these phenotypic variables into our model increases the proportion of causal genetic variants found using our method.

Our method is able to detect some causal rare variants, but the results do not indicate that this is a promising approach for the general analysis of exome sequencing data that includes causal rare variants. Identifying optimal sets of genetic variants for every gene and pathway in a dataset may take considerably higher computation time than the standard lasso model and is expected to generate robust predictive models only when there are several adequately powered common causal variants to distinguish case subjects from control subjects. While the ability of our method to reliably detect true rare variants was limited, our results indicate that our modified approach for optimizing the lasso regression for genetic prediction is an improvement over the standard lasso approach.

Future work will involve the application and evaluation of our improved lasso strategy to develop predictive genetic models for other large GWAS.

CHAPTER VI

EXPLORATION OF miRNA GENOMIC VARIATION ASSOCIATED WITH COMMON HUMAN DISEASES

A. Introduction

In the previous chapters, we have presented analysis methods and computational tools that utilize useful genomic information like local linkage disequilibrium structure or biochemical pathway information to improve GWAS analysis. For heritable multifactorial diseases, genotypic variation is thought to be only one component of the pathogenesis [83, 159]. In the past few years, a number of studies have sought to analyze the extent to which genomic variation itself may pathologically modulate transcriptional activity in disease related genes. However, data directly relating expression and variation information within a GWAS population are currently limited. To better understand possible relationships between genomic variation and mechanisms of disease-related expression changes, we present an analysis framework to search for genetic variation at a subset of loci related to miRNA. In this chapter, we discuss a general methodological framework to incorporate domain knowledge in order to investigate areas of genomic variation as they relate to molecular regulatory targets proven to be critical in numerous cellular processes. As discussed above, GWAS are designed to give an accurate measurement of common inherited variation (MAF >0.05) as measured at SNPs at calculated intervals across the genome in case and control populations. The ultimate

goals of GWAS research are improvements in the understanding of disease pathogenesis, in the accuracy of disease prediction, and in the development of personalized approaches to therapy based on genotype. As knowledge of the genome and molecular biology has increased, there has been a growing appreciation for the numerous regulatory processes that play essential roles in the process from gene transcription to protein function. While there are many components involved in these regulatory processes, one of the more important components that can be studied with GWAS is a relatively new class of non-coding RNAs called microRNAs (miRNAs) [160].

miRNAs comprise a large family of ~20-nucleotide-long RNAs that have been shown to perform key post-transcriptional regulation of gene expression in a wide variety of cellular processes [161]. Recent estimates in mammals have indicated that miRNAs might control the activity of between 30 and 50% of all protein-coding genes, and changes in their expression or regulation are associated with at least 134 separate human diseases [160-165]. miRNAs exist in the genome as precursors known as pri-miRNAs that are either transcribed by RNA polymerase II from independent genes or that are introns of protein-coding genes[161]. With the help of miRNA machinery (RNase III enzymes Drosha and Dicer), the pri-miRNA is converted into its active ~20 nucleotide form. Numerous other cofactors and accessory proteins may be involved in their maturation and regulation. There is evidence to support a wide variety of ways that miRNAs regulate their targets and that they are regulated themselves [161]. Studies of miRNA-associated genomic variation have provided evidence that SNPs within miRNA

sequence targets as well as within miRNA related machinery can substantially modulate disease risk [160, 166-169]. Clear and extensive evidence has linked genomic variation measured at SNPs with alterations in miRNA-related sites that modulate cancer risk [160]. Saetrom et al used HapMap SNPs in conjunction with local LD patterns near miRNA-related sites to identify high-ranking SNPs in a breast cancer GWAS [170]. From this analysis, a SNP in a miR-125b target site (rs1434536) upstream of bone morphogenetic protein receptor type 1B was validated, and miR-125b was shown to differentially regulate the C and T polymorphisms at that site, providing a mechanism that explained the observed disease risk associated with the SNP.

Numerous other examples of genomic variation related to functional miRNA modifications can be taken from cancer biology, but as mentioned above, miRNAs themselves have been implicated in a wide variety of diseases. Furthermore, recent data from an animal model have provided direct evidence of genomic variations that modify miRNA activity [54]. Despite this growing body of evidence in support of the vital role of miRNAs in molecular pathophysiology, few studies besides those related to cancer have considered miRNA-related SNPs in GWAS populations [169, 171]. To better understand possible relationships between genomic variation and disease related expression changes, we propose an analysis of the genetic variation at miRNA loci. In this Chapter, we will use existing resources to build up a subset of miRNA-associated regions in which to analyze common SNP variations as measured in GWAS [172, 173]. We consider miRNA variation in SNPs within 100kb of a known miRNA or a gene

sequence involved in miRNA processing. We then investigate these locations using the WTCCC GWAS data to determine a set of miRNA-related variations associated with these common diseases that can be used to follow up in laboratory studies to better understand possible pathophysiological aspects of disease progression, maintenance, and/or treatment.

B. Methods

We once again investigated disease associations in the WTCCC populations using the recommended guidelines for data analysis as described in Chapter IV above [39]. We accessed the Affymetrix GeneChip Mapping 500K array set called by the CHIAMO algorithm for the 1958 birth control cohort, the National Blood Service control cohort, and the following case cohorts: Crohn Disease (CD), Bipolar Disorder (BD), Coronary Artery Disease (CAD), Rheumatoid Arthritis (RA) and Type II Diabetes (T2D) [39]. For each data set, we used the 1958 birth cohort as our control population. We excluded a number of subjects (as listed the "exclusion-list-05-02-2007.txt" file accompanying the GWAS data), and any SNPs not conforming to the criteria: (1) missing data proportion >

	Cases
BD	1868
CAD	1988
CD	1748
RA	1860
T2D	1924

Table 6.1. Number of cases used in each study. Each analysis included 1480 controls and 720 miRNA SNP sites.

0.05, (2) MAF < 0.05 and missing data proportion > 0.05 or MAF < 0.01, (3) combined control group (58C+National Blood Service) HWE Exact Test p-value < 5.7e-7, (4) Inter-control group (58C vs NBS) 1df Trend Test p-value < 5.7e-7 (as measured by WTCCC) (5). 58C vs NBS 2df General Test p-value < 5.7e-7 (as measured by WTCCC). CHIAMO SNP measurements with a score < 0.9 were considered missing data.

We updated the information obtained in the WTCCC to be consistent with current genomic info for hg19/Genome Reference Consortium Human genome build 37 (GRCh37). Data about the reference positions for SNPs in the study were downloaded from NCBI dbSNP (build 132) for GRCh37, information about genetic locations in were downloaded from RefSeq using the UCSC Genome Browser [143], miRNA information (unique ids and GRCh37 coordinates) was downloaded from Version 17 of miRbase (mirbase.org). We mapped each miRNA to the nearest SNP in the WTCCC, and then excluded any miRNAs that were not with 100kb of any SNP. We mapped the targets to

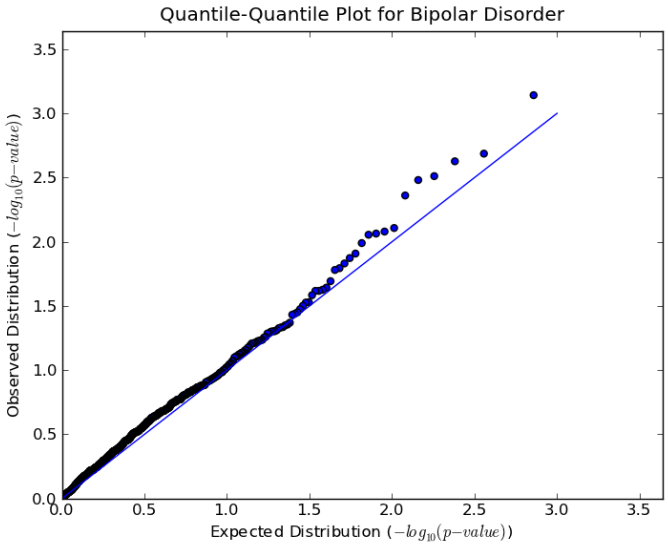
	snp	chrom	miRNA	χ^2	p-value
BD	rs2790466	10	hsa-mir-607	14.467289	7.22E-04
	rs1002095	9	hsa-mir-2861	12.370379	2.06E-03
CAD	rs1108183	1	hsa-mir-3659	16.797252	2.25E-04
	rs6792339	3	hsa-mir-3921	12.106655	2.35E-03
CD	rs8060598	16	hsa-mir-3181	22.141633	1.56E-05
	rs2304442	3	hsa-mir-4271	13.732693	1.04E-03
RA	rs845787	20	hsa-mir-663	10.901896	4.29E-03
T2D	rs10051407	5	hsa-mir-548f-3	14.681385	6.49E-04
	rs2280401	19	hsa-mir-150	13.67183	1.07E-03

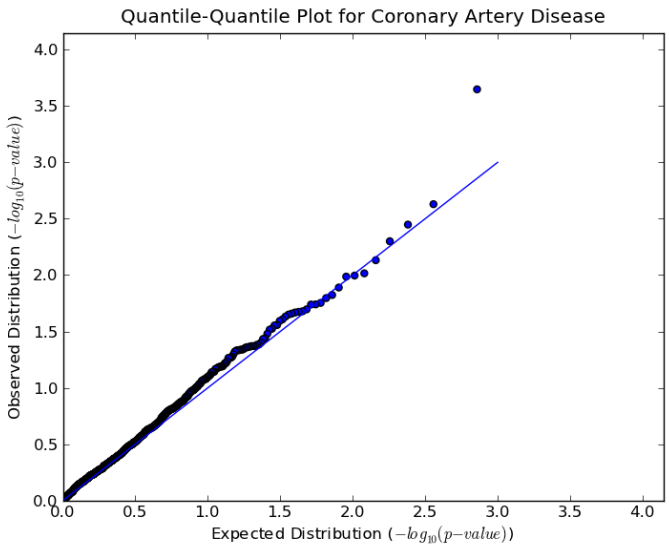
Table 6.2. Table of the most significant miRNA-related SNPs from the 5 WTCCC data sets.

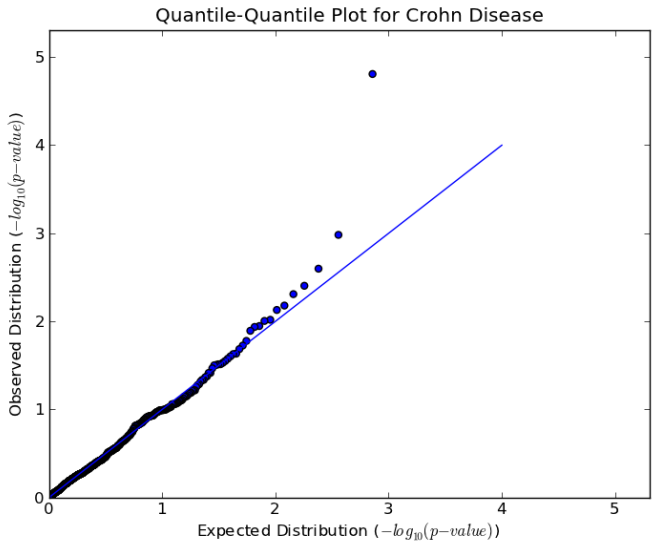
the WTCCC data in groups using the gene annotations and position information in the dbSNP annotation. We also included SNPs that were within 100kb of the miRNA processing machinery genes: DROSHA, DGCR8, XPO5, RAN, DICER1, TARBP2, GEMIN4, and TNRC6B. A summary of the data sets is included in Table 6.1.

C. Results

We calculated χ^2 values for all single SNPs in our data set and reported results for χ^2 values achieving significance at $\alpha = 0.005$. We then calculated χ^2 values for all pairwise interactions in each data set, and reported values achieving significance at $\alpha = 5 \times 10^{-5}$. We constructed quantile-quantile (Q-Q) plots by plotting the ordered list of log-transformed p-values for the single locus analysis (y-axis) against the set of expected values obtained from the theoretical null distribution (Figure 6.1). In these plots, when the distribution of p-values is equal to that expected from the null distribution, the points will all be observed on the diagonal. Our findings are summarized in Tables 6.2 and 6.3. hsa-mir-663 (RA, Table 6.2) and hsa-mir-150 (T2D, Table 6.2) have been implicated in immune [174, 175] and inflammatory and/or oxidative processes [176, 177]. Other well-studied SNPs in our results include hsa-mir-21, which we found to have an interaction with a SNP in hsa-mir-4317 in the CAD data set (Table 6.3). There is evidence linking hsa-mir-21 with regulation of inflammatory processes [178, 179], and this miRNA has also been implicated in coronary artery disease [180]. Further evaluation of these results along with validation in other large GWAS is necessary.







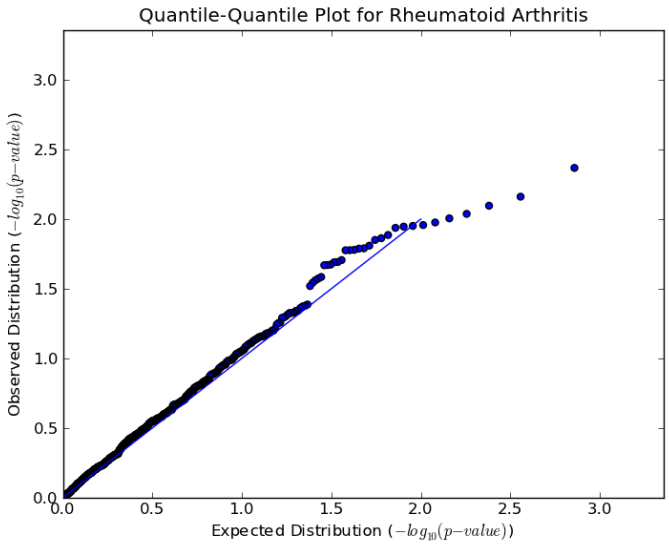
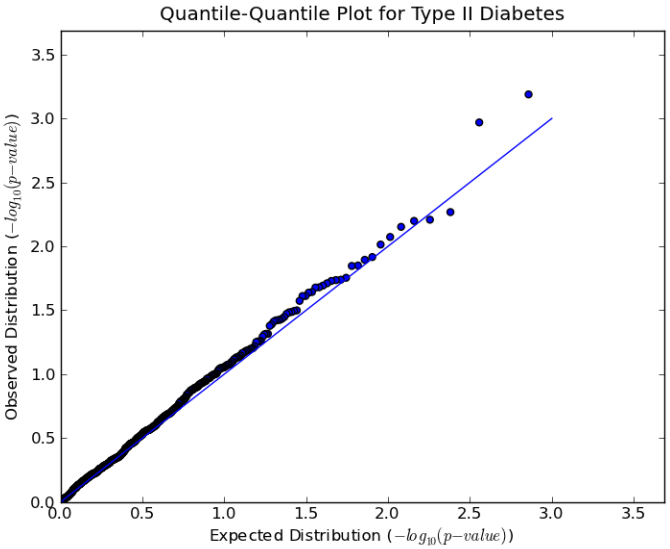


Figure 6.1. Quantile-Quantile Plots from the single locus analysis of miRNA-related SNPs for the 5 WTCCC Diseases studied: Bipolar Disorder, Coronary Artery Disease Crohn Disease, Rheumatoid Arthritis, and Type II Diabetes. These plots show the deviation of the observed -$\log_{10}$ (p-value) distribution calculated using the χ^2_{2df} test at the miRNA-related loci (y-axis) from the expected (theoretical) -$\log_{10}$ (p-value) distribution.



D. Discussion

Here we have presented an analysis that used information about the genomic locations of genes related to miRNA and its processing machinery to explore potential miRNA-related associations in several GWAS. Our use of biological knowledge to explore sets of interacting SNPs related to miRNA is intended to be complementary to other methods. By considering the variation in this subset of results, we hope to reveal meaningful statistical associations related to a subset of the genome in which polymorphisms may directly impact the regulation of downstream targets. Our results have shown a number of potentially meaningful associations, including findings at some miRNA sites for which there is direct experimental evidence of miRNA modulation of

	SNP1					SNP2					Interaction	
	snp	chrom	miRNA	χ^2	p-value	snp	chrom	miRNA	χ^2	p-value	χ^2	p-value
BD	rs4719842	7	hsa-mir-148a	6.14	4.65E-02	rs11864516	16	hsa-mir-662	12.10	2.36E-03	38.59	5.86E-06
	rs10936410	3	hsa-mir-1263	9.50	8.63E-03	rs10080387	6	hsa-mir-548b	6.00	4.97E-02	38.58	5.89E-06
	rs2790466	10	hsa-mir-607	14.47	7.22E-04	rs2682714	12	hsa-mir-548c	2.23	3.27E-01	36.29	1.55E-05
CAD	rs1292053	17	hsa-mir-21	7.60	2.24E-02	rs163750	18	hsa-mir-4317	2.18	3.36E-01	35.48	2.19E-05
CD	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	rs8099430	18	hsa-mir-3929	0.79	6.75E-01	40.64	2.43E-06
	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	rs1893321	18	hsa-mir-187	0.94	6.27E-01	40.07	3.11E-06
	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	rs192808	19	hsa-mir-935	0.31	8.56E-01	39.96	3.25E-06
	rs13377158	10	hsa-mir-3611	9.23	9.89E-03	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	38.41	6.32E-06
	rs2304442	3	hsa-mir-4271	13.73	1.04E-03	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	37.91	7.82E-06
	rs7577243	2	hsa-mir-149	8.97	1.13E-02	rs8060598	16	hsa-mir-3181	22.14	1.56E-05	37.06	1.12E-05
RA	rs2754163	14	hsa-mir-208b	9.38	9.18E-03	rs12928353	16	hsa-mir-365-1	8.68	1.30E-02	34.21	3.72E-05
T2D	rs2624183	5	hsa-mir-4280	1.37	5.03E-01	rs2280401	19	hsa-mir-150	13.67	1.07E-03	38.28	6.68E-06
	rs300917	4	hsa-mir-3139	0.94	6.25E-01	rs1332311	9	hsa-mir-491	9.27	9.69E-03	35.86	1.86E-05

Table 6.3. Table of the most significant pairwise interactions between miRNA-related SNPs from 5 WTCCC data sets.

cellular processes that may be related to the diseases considered [175, 176]. We did not find any associations with the miRNA processing machinery that met our cutoff criteria. The strongest result we found in this subset was an association in the T2D data set with SNP rs784567 in gene TARBP2 that did not meet our cutoff criteria (χ^2 = 10.168, p=0.0062). The exploratory analysis presented here has highlighted a number of findings that we hope to look into in future research. For example, we have explored the set of interactions within this group of miRNA-related SNPs, but a more complete assessment of miRNA SNP interactions with other (non-miRNA) SNPs in the WTCCC study is warranted. The miRNA analysis framework presented here is fully compatible with the GPU-accelerated tools presented in Chapter IV above, allowing us to more easily consider larger data sets in the future.

CHAPTER VII

CONCLUSION

In this thesis, we have presented computational approaches and integrative analysis strategies to overcome several current challenges in the analysis of GWAS. Methods described in the above work have addressed important issues related to gene-gene interactions, high-performance computing for genetic studies, predictive modeling of genetic disease risk, and investigations into candidates for disease-associated functional variation. We applied these methods to carefully designed simulated models of genetic disease as well as to case control data sets from real GWAS. Our gene-gene interaction analysis method used an adaptive evolutionary optimization framework that integrated local LD information to reduce the dimensionality of the search for SNP combinations. Using simulation data, we showed that our method was able to outperform one of the most powerful competing methods in terms of both power and computational efficiency. We improved this analysis approach to be even more efficient by developing a parallel optimization algorithm that takes advantage of state-of-the-art high performance computing methods for GPUs. In our analysis of GWAS data from the WTCCC, we integrated information about the genomic location of biochemical pathways to explore a set of biologically relevant gene-gene interactions.

Next, we presented an improved penalized lasso regression strategy to build more accurate predictions of disease risk based on genomic and phenotypic information for

case control studies. Using this approach on a simulated exome scan based on data from the 1000 Genomes project, we were able to model disease risk using common and rare genetic variation in combination with several simulated continuous phenotypes. While the genotypic predictive models were limited by the nature of the data, our results indicate that our modified pathway-based penalized regression procedure yields more robust results than the more commonly applied standard lasso model. Further investigation of this method in other GWAS is warranted.

Lastly, we conducted an exploratory analysis of genomic variation associated with miRNA dysregulation. We found several significant results for each of the 5 WTCCC diseases studied. More research is necessary to appreciate the mechanisms by which these associations are related to pathogenic molecular alterations, and future studies will expand on these analyses using additional data sources.

CITED LITERATURE

[1] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nat Rev Genet*, vol. 10, pp. 392-404, Jun 2009.

[2] T. E. Andreoli and R. L. Cecil, *Andreoli and Carpenter's Cecil essentials of medicine*, 8th ed. Philadelphia, PA: Saunders/Elsevier, 2010.

[3] E. Mayr, *The growth of biological thought : diversity, evolution, and inheritance*. Cambridge, Mass.: Belknap Press, 1982.

[4] G. H. Hardy, "Mendelian Proportions in a Mixed Population," *Science*, vol. 28, pp. 49-50, Jul 10 1908.

[5] J. F. Crow, "Hardy, Weinberg and language impediments," *Genetics*, vol. 152, pp. 821-5, Jul 1999.

[6] A. H. Sturtevant, "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association.," *Journal of Experimental Zoology*, vol. 14, pp. 43-59, 1913.

[7] H. S. Jennings, "The Numerical Results of Diverse Systems of Breeding, with Respect to Two Pairs of Characters, Linked or Independent, with Special Relation to the Effects of Linkage," *Genetics*, vol. 2, pp. 97-154, Mar 1917.

[8] J. F. Crow, "Population genetics history: a personal view," *Annu Rev Genet*, vol. 21, pp. 1-22, 1987.

[9] R. A. Fisher, *The genetical theory of natural selection*. Oxford,: The Clarendon press, 1930.

[10] J. D. Watson and F. H. Crick, "Genetical implications of the structure of deoxyribonucleic acid," *Nature,* vol. 171, pp. 964-7, May 30 1953.

[11] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid," *Nature,* vol. 171, pp. 737-8, Apr 25 1953.

[12] J. E. Pool, *et al.*, "Population genetic inference from genomic sequence variation," *Genome Res,* vol. 20, pp. 291-300, Mar 2010.

[13] J. F. Crow, "Was there life before 1953?," *Nat Genet,* vol. 33, pp. 449-50, Apr 2003.

[14] A. R. Omran, "The epidemiologic transition. A theory of the epidemiology of population change," *Milbank Mem Fund Q,* vol. 49, pp. 509-38, Oct 1971.

[15] G. L. Armstrong, *et al.*, "Trends in infectious disease mortality in the United States during the 20th century," *JAMA,* vol. 281, pp. 61-6, Jan 6 1999.

[16] S. B. Zimmerman, *et al.*, "Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide," *Proc Natl Acad Sci U S A,* vol. 57, pp. 1841-8, Jun 1967.

[17] E. M. Southern, "Detection of specific sequences among DNA fragments separated by gel electrophoresis," *J Mol Biol,* vol. 98, pp. 503-17, Nov 5 1975.

[18] F. Sanger, *et al.*, "DNA sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci U S A,* vol. 74, pp. 5463-7, Dec 1977.

[19] W. Gilbert and A. Maxam, "The nucleotide sequence of the lac operator," *Proc Natl Acad Sci U S A,* vol. 70, pp. 3581-4, Dec 1973.

[20] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proc Natl Acad Sci U S A,* vol. 74, pp. 560-4, Feb 1977.

[21] A. M. Maxam and W. Gilbert, "Sequencing end-labeled DNA with base-specific chemical cleavages," *Methods Enzymol,* vol. 65, pp. 499-560, 1980.

[22] D. Botstein, *et al.*, "Construction of a genetic linkage map in man using restriction fragment length polymorphisms," *Am J Hum Genet,* vol. 32, pp. 314-31, May 1980.

[23] J. F. Gusella, *et al.*, "A polymorphic DNA marker genetically linked to Huntington's disease," *Nature,* vol. 306, pp. 234-8, Nov 17-23 1983.

[24] M. Olson, *et al.*, "A common language for physical mapping of the human genome," *Science,* vol. 245, pp. 1434-5, Sep 29 1989.

[25] J. R. Riordan, *et al.*, "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA," *Science,* vol. 245, pp. 1066-73, Sep 8 1989.

[26] V. A. McKusick, "Current trends in mapping human genes," *FASEB J,* vol. 5, pp. 12-20, Jan 1991.

[27] L. Roberts, "The human genome. Controversial from the start," *Science,* vol. 291, pp. 1182-8, Feb 16 2001.

[28] E. S. Lander, *et al.*, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, Feb 15 2001.

[29] J. C. Venter, *et al.*, "The sequence of the human genome," *Science*, vol. 291, pp. 1304-51, Feb 16 2001.

[30] J. C. Venter, *et al.*, "Massive parallelism, randomness and genomic advances," *Nat Genet*, vol. 33 Suppl, pp. 219-27, Mar 2003.

[31] S. P. Fodor, *et al.*, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, pp. 767-73, Feb 15 1991.

[32] M. Schena, *et al.*, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, pp. 467-70, Oct 20 1995.

[33] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, pp. 1516-7, Sep 13 1996.

[34] U. Landegren, "Ligation-based DNA diagnostics," *Bioessays*, vol. 15, pp. 761-5, Nov 1993.

[35] R. Sachidanandam, *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, pp. 928-33, Feb 15 2001.

[36] M. J. Daly, *et al.*, "High-resolution haplotype structure in the human genome," *Nat Genet*, vol. 29, pp. 229-32, Oct 2001.

[37] "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299-320, Oct 27 2005.

[38] K. A. Frazer, *et al.*, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851-61, Oct 18 2007.

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

[40] M. D. Mailman, *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nat Genet,* vol. 39, pp. 1181-6, Oct 2007.

[41] J. H. Hindorff LA, Hall PN, Mehta JP, and Manolio TA. (2011, 2011). *A catalog of published genome-wide association studies*. Available: www.genome.gov/gwastudies

[42] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *N Engl J Med,* vol. 363, pp. 166-76, Jul 8 2010.

[43] R. M. Durbin, *et al.*, "A map of human genome variation from population-scale sequencing," *Nature,* vol. 467, pp. 1061-73, Oct 28 2010.

[44] L. Huang, *et al.*, "Genotype-imputation accuracy across worldwide human populations," *Am J Hum Genet,* vol. 84, pp. 235-50, Feb 2009.

[45] T. A. Manolio, *et al.*, "Finding the missing heritability of complex diseases," *Nature,* vol. 461, pp. 747-53, Oct 8 2009.

[46] H. R. Coleman, *et al.*, "Age-related macular degeneration," *Lancet,* vol. 372, pp. 1835-45, Nov 22 2008.

[47] M. I. McCarthy, *et al.*, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat Rev Genet,* vol. 9, pp. 356-69, May 2008.

[48] G. Gibson, "Hints of hidden heritability in GWAS," *Nat Genet,* vol. 42, pp. 558-60, Jul 2010.

[49] T. A. Pearson and T. A. Manolio, "How to interpret a genome-wide association study," *JAMA,* vol. 299, pp. 1335-44, Mar 19 2008.

[50] P. M. Visscher, *et al.*, "Heritability in the genomics era--concepts and misconceptions," *Nat Rev Genet,* vol. 9, pp. 255-66, Apr 2008.

[51] J. H. Moore and S. M. Williams, "Epistasis and its implications for personal genetics," *Am J Hum Genet,* vol. 85, pp. 309-20, Sep 2009.

[52] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nat Rev Genet,* vol. 5, pp. 522-31, Jul 2004.

[53] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and microRNAs," *Nat Rev Genet,* vol. 8, pp. 93-103, Feb 2007.

[54] W. L. Su, *et al.*, "Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques," *Mol Syst Biol,* vol. 7, p. 490, May 24 2011.

[55] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature,* vol. 461, pp. 218-23, Sep 10 2009.

[56] H. Zhong, *et al.*, "Integrating pathway analysis and genetics of gene expression for genome-wide association studies," *Am J Hum Genet,* vol. 86, pp. 581-91, Apr 9 2010.

[57] W. Cookson, *et al.*, "Mapping complex disease traits with global gene expression," *Nat Rev Genet,* vol. 10, pp. 184-94, Mar 2009.

[58] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: models and data," *Am J Hum Genet,* vol. 69, pp. 1-14, Jul 2001.

[59] A. R. Rogers and C. Huff, "Linkage disequilibrium between loci with unknown phase," *Genetics,* vol. 182, pp. 839-44, Jul 2009.

[60] R. Lewontin and K. Ken-ichi, "The evolutionary dynamics of complex polymorphisms," *Evolution,* vol. 14, pp. 458-472, 1960.

[61] J. C. Barrett, *et al.*, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics,* vol. 21, pp. 263-5, Jan 15 2005.

[62] L. Excoffier and M. Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Mol Biol Evol,* vol. 12, pp. 921-7, Sep 1995.

[63] A. Agresti, *Categorical data analysis,* 2nd ed. New York: Wiley-Interscience, 2002.

[64] D. A. Greenberg, "Simulation studies of segregation analysis: application to two-locus models," *Am J Hum Genet,* vol. 36, pp. 167-76, Jan 1984.

[65] J. F. Kingman, "Origins of the coalescent. 1974-1982," *Genetics,* vol. 156, pp. 1461-3, Dec 2000.

[66] S. M. Dudek, *et al.*, "Data simulation software for whole-genome association and other studies in human genetics," *Pac Symp Biocomput,* pp. 499-510, 2006.

[67] M. D. Ritchie and W. S. Bush, "Genome simulation approaches for synthesizing in silico datasets for human genomics," *Adv Genet,* vol. 72, pp. 1-24, 2010.

[68] C. Li and M. Li, "GWAsimulator: a rapid whole-genome simulation program," *Bioinformatics,* vol. 24, pp. 140-2, Jan 1 2008.

[69] R. R. Hudson, "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics,* vol. 18, pp. 337-8, Feb 2002.

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

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

[72] R. Culverhouse, *et al.*, "A perspective on epistasis: limits of models displaying no main effect," *Am J Hum Genet,* vol. 70, pp. 461-71, Feb 2002.

[73] M. D. Ritchie, *et al.*, "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.

[74] W. Tang, *et al.*, "Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy," *PLoS Genet,* vol. 5, p. e1000464, May 2009.

[75] C. Yang, *et al.*, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics,* vol. 25, pp. 504-11, Feb 15 2009.

[76] Y. Zhang, *et al.*, "Bayesian models for detecting epistatic interactions from genetic data," *Ann Hum Genet,* vol. 75, pp. 183-93, Jan 2011.

[77] H. Eleftherohorinou, *et al.*, "Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases," *PLoS One,* vol. 4, p. e8068, 2009.

[78] C. Kooperberg, *et al.*, "Risk prediction using genome-wide association studies," *Genet Epidemiol,* Sep 14.

[79] Z. Wei, *et al.*, "From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes," *PLoS Genet,* vol. 5, p. e1000678, Oct 2009.

[80] R. W. Davies, *et al.*, "Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies," *Circ Cardiovasc Genet,* vol. 3, pp. 468-74, Oct 1 2010.

[81] R. Plomin, *et al.*, "Common disorders are quantitative traits," *Nat Rev Genet,* vol. 10, pp. 872-8, Dec 2009.

[82] N. R. Wray, *et al.*, "The genetic interpretation of area under the ROC curve in genomic profiling," *PLoS Genet,* vol. 6, p. e1000864.

[83] J. H. Moore, *et al.*, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics,* vol. 26, pp. 445-55, Feb 15 2010.

[84] A. Bureau, *et al.*, "Identifying SNPs predictive of phenotype using random forests," *Genet Epidemiol,* vol. 28, pp. 171-82, Feb 2005.

[85] H. Zhou, *et al.*, "Association Screening of Common and Rare Genetic Variants by Penalized Regression," *Bioinformatics,* Aug 6.

[86] T. Hastie, *et al.*, *The elements of statistical learning : data mining, inference, and prediction*, 2nd ed. New York: Springer, 2009.

[87] S. Szymczak, *et al.*, "Machine learning in genome-wide association studies," *Genet Epidemiol*, vol. 33 Suppl 1, pp. S51-7, 2009.

[88] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, pp. 30-50, Jan 2008.

[89] W. Guo and S. Lin, "Generalized linear modeling with regularization for detecting common disease rare haplotype association," *Genet Epidemiol*, vol. 33, pp. 308-16, May 2009.

[90] A. Dasgupta, *et al.*, "A brief review of machine learning methods in genetic epidemiology: the GAW17 experience," *BMC Proceedings*, vol. *in press*, 2011.

[91] C. Kooperberg, *et al.*, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, pp. 643-652, 2010.

[92] J. Friedman, *et al.*, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *J Stat Softw*, vol. 33, pp. 1-22, 2010.

[93] C. J. Hoggart, *et al.*, "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies," *PLoS Genet*, vol. 4, p. e1000130, 2008.

[94] T. Verplancke, *et al.*, "Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies," *BMC Med Inform Decis Mak*, vol. 8, p. 56, 2008.

[95] D. C. Thomas, *Statistical methods in genetic epidemiology*. Oxford ; New York: Oxford University Press, 2004.

[96] L. Kruglyak, "The road to genome-wide association studies," *Nat Rev Genet,* vol. 9, pp. 314-8, Apr 2008.

[97] Y. Yang, *et al.*, "RET Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec," *Hum Mutat,* vol. 29, pp. 695-702, May 2008.

[98] S. Purcell, *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet,* vol. 81, pp. 559-75, Sep 2007.

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

[100] A. A. Motsinger-Reif and M. D. Ritchie, "Neural networks for genetic epidemiology: past, present, and future," *BioData Min,* vol. 1, p. 3, 2008.

[101] A. A. Motsinger-Reif, *et al.*, "Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error," *BMC Res Notes,* vol. 1, p. 65, 2008.

[102] A. A. Motsinger-Reif, *et al.*, "Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology," *Genet Epidemiol,* vol. 32, pp. 325-40, May 2008.

[103] X. Chen, *et al.*, "A forest-based approach to identifying gene and gene gene interactions," *Proc Natl Acad Sci U S A,* vol. 104, pp. 19199-203, Dec 4 2007.

[104] M. R. Nelson, *et al.*, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Res,* vol. 11, pp. 458-70, Mar 2001.

[105] J. Gayan, *et al.*, "A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis," *BMC Genomics,* vol. 9, p. 360, 2008.

[106] X. Wan, *et al.*, "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study," *BMC Bioinformatics,* vol. 10, p. 13, 2009.

[107] C. Kooperberg and M. Leblanc, "Increasing the power of identifying gene x gene interactions in genome-wide association studies," *Genet Epidemiol,* vol. 32, pp. 255-63, Apr 2008.

[108] R. Jiang, *et al.*, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics,* vol. 10 Suppl 1, p. S65, 2009.

[109] X. Zhang, *et al.*, "TEAM: efficient two-locus epistasis tests in human genome-wide association study," *Bioinformatics,* vol. 26, pp. i217-27, Jun 15.

[110] Y. Wang, *et al.*, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Res Notes,* vol. 3, p. 117, 2010.

[111] N. Wang, *et al.*, "Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation," *Am J Hum Genet,* vol. 71, pp. 1227-34, Nov 2002.

[112] S. B. Gabriel, *et al.*, "The structure of haplotype blocks in the human genome," *Science,* vol. 296, pp. 2225-9, Jun 21 2002.

[113] A. Katanforoush, *et al.*, "Global haplotype partitioning for maximal associated SNP pairs," *BMC Bioinformatics,* vol. 10, p. 269, 2009.

[114] C. Pattaro, *et al.*, "Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies," *BMC Genomics,* vol. 9, p. 405, 2008.

[115] Y. Guo, *et al.*, "Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies," *Eur J Hum Genet,* Dec 17 2008.

[116] J. Akey, *et al.*, "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?," *Eur J Hum Genet,* vol. 9, pp. 291-300, Apr 2001.

[117] W. Makarasara, *et al.*, "pHCR: a parallel haplotype configuration reduction algorithm for haplotype interaction analysis," *J Hum Genet,* vol. 54, pp. 634-41, Nov 2009.

[118] R. M. Cantor, *et al.*, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application," *Am J Hum Genet,* vol. 86, pp. 6-22, Jan 2010.

[119] J. Marchini, *et al.*, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nat Genet,* vol. 37, pp. 413-7, Apr 2005.

[120] C. Li, *et al.*, "Prioritized subset analysis: improving power in genome-wide association studies," *Hum Hered,* vol. 65, pp. 129-41, 2008.

[121] E. K. P. Chong and S. H. *Zak, *An introduction to optimization,* 3rd ed. Hoboken, N.J.: Wiley-Interscience, 2008.

[122] R. J. Klein, *et al.*, "Complement factor H polymorphism in age-related macular degeneration," *Science,* vol. 308, pp. 385-9, Apr 15 2005.

[123] J. K. Estrada-Gil, *et al.*, "GPDTI: a Genetic Programming Decision Tree induction method to find epistatic effects in common complex diseases," *Bioinformatics,* vol. 23, pp. i167-74, Jul 1 2007.

[124] J. H. Moore, *et al.*, "Symbolic modeling of epistasis," *Hum Hered,* vol. 63, pp. 120-33, 2007.

[125] R. Nunkesser, *et al.*, "Detecting high-order interactions of single nucleotide polymorphisms using genetic programming," *Bioinformatics,* vol. 23, pp. 3280-8, Dec 15 2007.

[126] J. H. M. a. B. C. White, "Genome-wide Genetic Analysis Using Genetic Programing: The Critical Need for Expert Knowledge," in *Genetic Programming Theory and Practice IV,* T. S. Rick Riolo, Bill Worzel, Ed., Genetic Programming Theory and Practice IV ed: Springer, 2007.

[127] W. P. Worzel, *et al.*, "Applications of genetic programming in cancer research," *Int J Biochem Cell Biol,* vol. 41, pp. 405-13, Feb 2009.

[128] E. E. Eichler, *et al.*, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nat Rev Genet,* vol. 11, pp. 446-50, Jun 2010.

[129] NVIDIA. (2011). *NVIDIA CUDA Website, SDK, and White Papers*. Available: http://www.nvidia.com/object/cuda_home_new.html

[130] U. o. Illinois. *Institute for Advanced Computing*. Available: http://iac.uiuc.edu/resources/cluster/

[131] NVIDIA. (2011). *NVIDIA CUDA Developer Zone*. Available: http://www.nvidia.com/object/cuda_home.html

[132] N. A. Sinnott-Armstrong, *et al.*, "Accelerating epistasis analysis in human genetics with consumer graphics hardware," *BMC Res Notes*, vol. 2, p. 149, 2009.

[133] X. Hu, *et al.*, "SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder," *Cell Res*, vol. 20, pp. 854-7, Jul 2010.

[134] L. S. Yung, *et al.*, "GBOOST : A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies," *Bioinformatics*, Mar 3 2011.

[135] J. Fontanarosa and Y. Dai, "A block-based evolutionary optimization strategy to investigate gene-gene interactions in genetic association studies," *Proceeding of 2010 IEEE International conference on Bioinformatics and Biomedicine Workshop*, pp. 330-335, 2010.

[136] X. Wan, *et al.*, "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *Am J Hum Genet*, vol. 87, pp. 325-40, Sep 10 2010.

[137] M. Harris. (2009). *Optimizing Parallel Reduction in CUDA*. Available: developer.download.nvidia.com

[138] A. Klöckner, *et al.*, "PyCUDA and PyOpenCL: A Scripting-Based Approach to GPU Run-Time Code Generation," *arXiv*, 2011.

[139] J. Barrett, *et al.* (2010). *Haploview 4.2*. Available: www.broadinstitue.org/haploview/haploview

[140] K. Wang, *et al.*, "Analysing biological pathways in genome-wide association studies," *Nat Rev Genet*, vol. 11, pp. 843-54, Dec 2010.

[141] K. Wang, *et al.*, "Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease," *Am J Hum Genet*, vol. 84, pp. 399-405, Mar 2009.

[142] X. Wan, *et al.*, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics*, vol. 26, pp. 30-7, Jan 1 2010.

[143] P. A. Fujita, *et al.*, "The UCSC Genome Browser database: update 2011," *Nucleic Acids Res*, vol. 39, pp. D876-82, Jan 2011.

[144] M. Kanehisa, *et al.*, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res*, vol. 38, pp. D355-60, Jan.

[145] C. Herold, *et al.*, "INTERSNP: genome-wide interaction analysis guided by a priori information," *Bioinformatics*, vol. 25, pp. 3275-81, Dec 15 2009.

[146] C. O'Dushlaine, *et al.*, "The SNP ratio test: pathway analysis of genome-wide association datasets," *Bioinformatics*, vol. 25, pp. 2762-3, Oct 15 2009.

[147] M. Emily, *et al.*, "Using biological networks to search for interacting loci in genome-wide association studies," *Eur J Hum Genet*, vol. 17, pp. 1231-40, Oct 2009.

[148] Khronos. (2011). *OpenCL - The open standard for parallel programming of heterogeneous systems*. Available: http://www.khronos.org/opencl/

[149] K. B. Meyer, *et al.*, "Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer," *PLoS Biol*, vol. 6, p. e108, May 6 2008.

[150] B. L. Chang, *et al.*, "Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk," *Hum Mol Genet*, vol. 18, pp. 1368-75, Apr 1 2009.

[151] T. M. Teslovich, *et al.*, "Biological, clinical and population relevance of 95 loci for blood lipids," *Nature*, vol. 466, pp. 707-13, Aug 5 2010.

[152] H. Zhou, *et al.*, "Association screening of common and rare genetic variants by penalized regression," *Bioinformatics*, vol. 26, pp. 2375-82, Oct 1 2010.

[153] B. Maher, "Personal genomes: The case of the missing heritability," *Nature*, vol. 456, pp. 18-21, Nov 6 2008.

[154] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nat Rev Genet*, vol. 11, pp. 415-25, Jun.

[155] J. Blangero and e. al, "Genetic Analysis Workshop 17 mini-exome simulation," *BMC Proceedings*, vol. *in press*, 2011.

[156] Y. He, *et al.*, "A new lasso and K-means based framework for rare variants analysis in genetic association studies," *BMC Proceedings (ms. in prep.),* 2011.

[157] J. Jung, *et al.*, "Identification of multiple Rare Variants associated with a disease," *BMC Proceedings,* vol. in prep, 2011.

[158] N. Pankratz, "Application of the Weighted Sum Statistic in the GAW17 dataset," *BMC Proceedings,* vol. in prep, 2011.

[159] R. D. Hawkins, *et al.*, "Next-generation genomics: an integrative approach," *Nat Rev Genet,* vol. 11, pp. 476-486, Jun 8 2010.

[160] B. M. Ryan, *et al.*, "Genetic variation in microRNA networks: the implications for cancer research," *Nat Rev Cancer,* vol. 10, pp. 389-402, Jun.

[161] J. Krol, *et al.*, "The widespread regulation of microRNA biogenesis, function and decay," *Nat Rev Genet,* vol. 11, pp. 597-610, Sep.

[162] Q. Jiang, *et al.*, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res,* vol. 37, pp. D98-104, Jan 2009.

[163] C. L. Bartels and G. J. Tsongalis, "MicroRNAs: novel biomarkers for human cancer," *Clin Chem,* vol. 55, pp. 623-31, Apr 2009.

[164] C. Xiao and K. Rajewsky, "MicroRNA control in the immune system: basic principles," *Cell,* vol. 136, pp. 26-36, Jan 9 2009.

[165] Y. S. Lee and A. Dutta, "MicroRNAs in cancer," *Annu Rev Pathol,* vol. 4, pp. 199-227, 2009.

[166] G. Sun, *et al.*, "SNPs in human miRNA genes affect biogenesis and function," *RNA*, vol. 15, pp. 1640-51, Sep 2009.

[167] S. Bandiera, *et al.*, "microRNAs in diseases: from candidate to modifier genes," *Clin Genet*, vol. 77, pp. 306-13, Apr 2010.

[168] C. Borel and S. E. Antonarakis, "Functional genetic variation of human miRNAs and phenotypic consequences," *Mamm Genome*, vol. 19, pp. 503-9, Aug 2008.

[169] G. V. Glinsky, "An SNP-guided microRNA map of fifteen common human disorders identifies a consensus disease phenocode aiming at principal components of the nuclear import pathway," *Cell Cycle*, vol. 7, pp. 2570-83, Aug 15 2008.

[170] P. Saetrom, *et al.*, "A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis," *Cancer Res*, vol. 69, pp. 7459-65, Sep 15 2009.

[171] M. Muinos-Gimeno, *et al.*, "Design and evaluation of a panel of single-nucleotide polymorphisms in microRNA genomic regions for association studies in human disease," *Eur J Hum Genet*, vol. 18, pp. 218-26, Feb 2010.

[172] S. Griffiths-Jones, *et al.*, "miRBase: tools for microRNA genomics," *Nucleic Acids Res*, vol. 36, pp. D154-8, Jan 2008.

[173] Z. Xu and J. A. Taylor, "SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies," *Nucleic Acids Res*, vol. 37, pp. W600-5, Jul 1 2009.

[174] S. R. Dalal and J. H. Kwon, "The Role of MicroRNA in Inflammatory Bowel Disease," *Gastroenterol Hepatol (N Y),* vol. 6, pp. 714-22, Nov 2010.

[175] E. Tili, *et al.*, "Resveratrol decreases the levels of miR-155 by upregulating miR-663, a microRNA targeting JunB and JunD," *Carcinogenesis,* vol. 31, pp. 1561-6, Sep 2010.

[176] C. W. Ni, *et al.*, "MicroRNA-663 upregulated by oscillatory shear stress plays a role in inflammatory response of endothelial cells," *Am J Physiol Heart Circ Physiol,* vol. 300, pp. H1762-9, May 2011.

[177] M. Estep, *et al.*, "Differential expression of miRNAs in the visceral adipose tissue of patients with non-alcoholic fatty liver disease," *Aliment Pharmacol Ther,* vol. 32, pp. 487-97, Aug 2010.

[178] A. J. Schetter, *et al.*, "Association of inflammation-related and microRNA gene expression with cancer-specific mortality of colon adenocarcinoma," *Clin Cancer Res,* vol. 15, pp. 5878-87, Sep 15 2009.

[179] D. Iliopoulos, *et al.*, "STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer," *Mol Cell,* vol. 39, pp. 493-506, Aug 27 2010.

[180] F. Fleissner, *et al.*, "Short communication: asymmetric dimethylarginine impairs angiogenic progenitor cell function in patients with coronary artery disease through a microRNA-21-dependent mechanism," *Circ Res,* vol. 107, pp. 138-43, Jul 9 2010.

VITA

NAME: Joel Bernard Fontanarosa

EDUCATION: M.D., Ph.D., Medical Scientist Training Program, University of Illinois College of Medicine. Chicago, IL. 2013

Bioinformatics Ph.D. Program (8/2007 – 6/2011). University of Illinois at Chicago, *Department of Bioengineering*. Advisor: Yang Dai, Ph.D.

A.B. in Biological Sciences with a Specialization in Neuroscience, The University of Chicago. Chicago, IL. 2004

TEACHING: Lecturer in Renal Physiology, *Summer Prematriculation Program* (PHYB 310) for incoming first year medical students at University of Illinois at Chicago, Summer 2010

Lecturer in Renal Physiology, *Summer Prematriculation Program* (PHYB 310) for incoming first year medical students at University of Illinois at Chicago, Summer 2009

Teaching Assistant, *Biostatistics* (BIOE 339) at University of Illinois at Chicago, Spring 2008

Teaching Assistant, *Biostatistics* (BIOE 339) at University of Illinois at Chicago, Fall 2007

Teaching Assistant, *Systems Neuroscience* (Bio 24205) at The University of Chicago, Spring 2005

Teaching Assistant, *Cellular Neurobiology* (Bio 24204) at The University of Chicago, Fall 2004

HONORS: UIC College of Medicine Research Day, Gold Medal Prize, 2010.

Genetic Analysis Workshop 17 Travel Award, 2010

Provost's Award and W.C. and May Preble Deiss Award (*GPU Computing Grant*), 2009

American Gastroenterological Association Academic Skills Workshop: MD-PhD Student Program, 2008

University of Chicago Dean's List, awarded in recognition of academic achievement, 2003-2004

Howard Hughes Medical Institute Summer Research Fellow in Neural Computation and Engineering, The University of Chicago, 2003

PUBLICATIONS:

Carr J, Kiefer M, Park HJ, Li J, Wang Z, Fontanarosa J, DeWaal D, Kopanga D, Benevolenskya E, Guzman G,Raychaudhuri P. FoxM1 Regulates Mammary Luminal Cell Fate. *Cell Rep*. 2012 Jun; 28;1 (6):715-29. PMID: 22813746.

Borisov AB, Khan SF, Racz E, Poopalasingam S, McCorkindale JC, Zhao J, Fontanarosa JB, Dai Y, Longworth JW, and Rhodes CK. Observation of a Curve Crossing Mechanism in the Field Ionization of Inner-Shell Excited Single Xe33+(2p) and Double Xe34+(2s2p) Vacancy States. *IEEE Journal of Quantum Electronics.* 2012 Jun; vol 48, no 6. 806-813.

Fontanarosa J and Dai Y. Using Lasso Regression to Detect Predictive Aggregate Effects in Genetic Studies. *BMC Proc*. 2011 Nov; 5 Suppl 9:S69. PMID: 22373537.

Fontanarosa J and Dai Y. An Evolutionary Optimization Strategy Using Graphics Processing Units to Efficiently Investigate Gene-Gene Interactions in Genetic Association Studies. *Conf Proc IEEE Eng Med Biol Soc*. 2011 Aug; 5547-50. PMID: 22255595.

Fontanarosa J and Dai Y. A Block-Based Evolutionary Optimization Strategy to Investigate Gene-Gene Interactions in Genetic Association Studies. *Conf Proc IEEE Bioinformatics and Biomedicine Workshop*. 2010 Dec; 330-335.

Borisov AB, Racz E, Khan SF, Poopalasingam S, McCorkindale JC, Zhao J, Fontanarosa J, Dai Y, Boguta J, Longworth JW, and Rhodes CK. "Spatially Resolved Observation of the Spectral Hole Burning in the Xe(L) Amplifier on Single (2p) and Double (2s2p) Vacancy Transitions in the 2.62å <2.94å Range". *Journal of Physics B: At. Mol. Opt. Phys*. 2010 Feb; Volume 43, Issue 4, 045402 (10pp).

Dutt S, Dai Y, Ren H, and Fontanarosa J. Selection of Multiple SNPs in Case-Control Association Study Based on a Discretized Network Flow Approach. *Proc of BICoB 2009, Springer Lecture Notes in Computer Science*. 2009 Jan; p211-223.

Fontanarosa JB, Lasky RE, Lee H, and van Drongelen W. Localization of Brainstem Auditory Evoked Potentials in Primates: A Comparison of

Techniques for Deep-Brain Sources. *Brain Topography*, 2004 Dec; 17 (2):99-108. PMID: 15754875